EXTREME RESPONSE STYLE: WHICH MODEL IS BEST?

by

Brian Leventhal

Bachelor of Science, Clarkson University, 2011

Master of Arts, University of Pittsburgh, 2013

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

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

Brian Leventhal

It was defended on

March 13, 2017

and approved by

Suzanne Lane, Professor, Psychology in Education

Feifei Ye, Assistant Professor, Psychology in Education

Lan Yu, Associate Professor, Medicine

Dissertation Advisor: Clement A. Stone, Professor, Psychology in Education

EXTREME RESPONSE STYLE: WHICH MODEL IS BEST?

Brian Leventhal, PhD

University of Pittsburgh, 2017

More robust and rigorous psychometric models, such as multidimensional Item Response Theory models, have been advocated for survey applications. However, item responses may be influenced by construct-irrelevant variance factors such as preferences for extreme response options. Through empirical and simulation methods, this study evaluates the use of the IRTree Model, the multidimensional nominal response model, and the modified generalized partial credit model designed to account for extreme response tendencies. The modified generalized partial credit model was found to have the best overall fit in terms of test-level, item-level, and person-level posterior predictive model checks performed. Estimation of this model also resulted in the lowest mean squared error between observed total score and expected total score. The multidimensional nominal response model had the lowest deviance information criterion among the three models. The empirical study, data validation from the simulation study, and the simulation results provided evidence that the IRTree Model was measuring a unique construct-irrelevant variance factor compared to the two other methods. For all simulation conditions of sample size (500, 1000), survey length (10, 20), and number of response options (4, 6), the modified generalized partial credit model had the most adequate model fit with respect to mean item mean squared error. The multidimensional nominal response model was found equally suitable for surveys measuring one substantive trait when responses to 10 4-option forced-choice Likert-type items were explored.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS TABLE

2-PL	Two parameter logistic model
AIC	Akaike information criterion
AIC3	Akaike information criterion 3
ARS	Acquiescence response style
BIC	Bayesian information criterion
CAIC	Consistent Akaike information criterion
DARS	Disacquiescence response style
Dbar	Posterior mean of the deviance
DIC	Deviance information criterion
Dmean	Deviance of mean of posterior
ENJ	Enjoyment of Science
ERS	Extreme response style
GPCM	Generalized partial credit model
GR	Graded response model
IMSE	Item mean squared error
IRT	Item response theory
MAD	Mean absolute difference

MCMC	Markov chain Monte Carlo
MIRT	Multidimensional item response theory
MNRM	Multidimensional nominal response model
MPCM	Modified generalized partial credit model
MPCM-c	Modified generalized partial credit model with covariates
MPR	Midpoint responding
MSE	Mean squared error
NARS	Net acquiescence response style
NCR	Noncontingent responding
OR	Odds ratio
PCM	Partial Credit Model
PCM-ERS	Partial credit model with extreme response style
PCM-ERS-c	Partial credit model with extreme response style and covariates
pD	Effective number of parameters
PISA	Program for International Student Assessment
PPMC	Posterior predictive model-checking
PPP-values	Posterior predictive probability values
PTM	Proportional thresholds model
RR	Response range
SAS	Statistical Analysis System
SES	Socioeconomic status
VAL	Value of Science

ACKNOWLEDGEMENTS

This work is a product of my professional and academic goals. However, it would not have been possible without the generous help of others. To my academic advisor, Dr. Clement A. Stone, thank you for your support, encouragement, and guidance. I am grateful for your patience and motivation as I progressed through my graduate career. Your mentorship style, and dedication to helping others succeed, is an envious quality. To Dr. Suzanne Lane, thank you for introducing me to measurement, being an inviting mentor, and working with me every step of the way in my graduate career. To Dr. Feifei Ye, thank you for your positive energy and sharing your expertise with me. To Dr. Lan Yu, thank you for inviting me to work on a project that initially sparked my interest in response scales, and for your time and effort as a member of my committee.

To my mother and father, Barbara and Steven Leventhal, thank you for your encouragement, unwavering belief in me, and motivating ethos throughout the years. To my sister and brother-in-law, Stephanie and Andrew Handel, thank you for your tirelessly optimistic spirit and reassurance throughout my doctoral journey. This endeavor would not have been achieved without the endless uplifting enthusiasm of my grandparents, Sanford and Cynthia Leventhal, and my aunts and uncles, David and Julie Leventhal, Kim Leventhal, and Kevin Nosbisch.

Finally, to my fiancée Maria, thank you for your unconditional comfort throughout this long and arduous process. You were always there with a smile and an encouraging word throughout the good and bad moments. Thank you for being my muse and sounding board for my research ideas. Above all, thank you for being my best friend.

1.0 INTRODUCTION

Responses to items on surveys may be used as a collective instrument to measure one or more latent traits. Responses may not only be due to the trait of interest, but, may be due to a combination of the trait of interest and a response style. A response style is a systematic or stylistic tendency in how a respondent uses a rating scale when responding to self-report items (Bolt & Johnson, 2009; Cronbach, 1946, 1950; Cronbach, Snow, & Wiley, 1991; Paulhus, 1991).

Several response styles have been identified. These include acquiescence response style (ARS), disacquiescence response style (DARS), net acquiescence response style (NARS), extreme response style (ERS), response range (RR), midpoint responding (MPR) and noncontingent responding (NCR) (Baumgartner, 2001). ARS is the tendency to agree with items regardless of content (Martin, 1964; Ray, 1983) while DARS is the tendency to disagree with items regardless of content (Couch & Keniston, 1960). NARS is a lower degree of ARS where a respondent has a tendency to show greater acquiescence than disacquiescence (Greenleaf, 1992a; Hui & Triandis, 1985). ERS occurs when a respondent tends to endorse the most extreme response categories regardless of content (Greenleaf, 1992b) while MPR is the tendency to endorse middle scale categories regardless of content (Messick, 1966; Schuman & Presser, 1981). RR is the tendency to use a narrow or wide range of response categories around the mean response (Greenleaf, 1992a; Hui & Triandis, 1985; Wyer, 1969) and NCR occurs when a respondent answers items carelessly, randomly, or nonpurposefully (Watkins & Cheung, 1995).

When measuring a latent trait using a self-report questionnaire, it is important to consider the presence of response styles. An individual's response style is considered a content independent trait that may interfere with the psychometric properties of the self-report instrument. Research has shown response style confounds the interpretation of scores (Bolt & Johnson, 2009; Jin & Wang, 2014; Rorer, 1965). A respondent's stylistic use of the rating scale may introduce bias in his/her total score with respect to the intended-to-be measurement trait. For example, if the endorsement of items implies a higher level of a trait, acquiescent response style will generally lead to an overestimation of the trait (Bolt & Johnson, 2009). This problem has been found to be intensified for scales that do not balance favorable and unfavorable items (Baumgartner, 2001; Javaras & Ripley, 2007).

1.1 SIGNIFICANCE OF THE STUDY

Extreme response tendency has garnered significant research attention (Bolt & Newton, 2011; Greenleaf, 1992b; Hamilton, 1968; Thissen-Roe & Thissen, 2013; Weijters, Geuens, & Schillewaert, 2010b). It is considered one of the two (along with acquiescence) most problematic response styles in attitude and survey research (Schuman & Presser, 1981; van Herk, Poortinga, & Verhallen, 2004). Extreme response tendency (ERS) occurs when individuals systematically endorse response options at the ends of the rating scale. Consider an individual who methodically selects *1-Strongly Disagree* and *5-Strongly Agree* from a 5-point agreement scale. This individual is likely exhibiting an extreme response tendency.

Tendency to select extreme responses is associated with respondent characteristics such as trait anxiety, income, education level, age, and ethnicity (Berg & Collier, 1953; Greenleaf, 1992a, 1992b; Lewis & Taylor, 1955; Marin, Gamba, & Marin, 1992). As a result, the association between the intended substantive trait and other constructs are confounded (Bolt & Newton, 2011). Bolt and Newton (2011) found biased results studying the relationship between socioeconomic status (SES) and a self-report measure of depression. They determined this biased relationship was due to the negative association between income level and ERS.

Hamilton (1968) performed an extensive literature review concluding that ERS is an indicator of several different personality attributes. The tendency for extreme response interferes with direct cross-cultural comparison of scores due to its variation across countries (Hui & Triandis, 1989). In a study of response style between six different countries, van Herk et al. (2004) found significant differences in extreme response rates. More importantly, they concluded, that ignoring national differences in ERS may lead to invalid inferences in cross-cultural research. Weijters et al. (2010b) found that ERS is an individual trait that is considered stable across time.

1.2 CONCEPTUAL UNDERPINNINGS FOR THE STUDY

Jackson and Messick (1958) emphasized the importance of understanding response style. They stressed development and evaluation of measures accounting for response style. Additionally, they suggested that ERS may contribute systematically to the scores when measuring personality. Studies separating the extreme response trait from the substantive trait have only recently started to gain traction. This is due to the advancement of using Bayesian estimation techniques and Multidimensional Item Response Theory (MIRT).

Researchers developing MIRT models to account for extreme response have conceptualized the response process distinctly. Bolt and Johnson (2009) have theorized the trait as a unique and independent trait from all others being measured. They used the multidimensional nominal response model (MNRM) to separate out the extreme tendency trait from substantive traits.

Thissen-Roe and Thissen (2013) have suggested the response process involves a two-stage sequential decision making process. The first decision is due to the trait of interest. The second decision is due to extreme response tendency with compensatory effects from the trait of interest. To model this, they adopt an IRTree Model. The initial decision of the tree is whether or not an individual agrees with the statement. The second decision focuses on the intensity. For example, consider an individual who agrees with a 4-category Likert item. Once the individual decides they agree, they then choose the intensity. They select whether they just *Agree* or *Strongly Agree*.

Jin and Wang (2014) believe that individuals treat the threshold between response categories differently. To account for this, they suggest treating the distance between response options as individual random effects using a modified generalized partial credit model (MPCM). Although in the same MIRT paradigm, these methods approach modeling extreme response tendency very differently. These conceptual differences cause difficulty for researchers when deciding which model to use in order to account for extreme response tendencies.

1.3 SUMMARY AND STATEMENT OF THE PROBLEM

The goal of the research is to determine which method of accounting for extreme response tendencies should be used under varying conditions. Provided is an in-depth investigation into the methods designed to account for extreme response tendencies. This study was designed to determine whether the MPCM, the MNRM, or the IRTree Model is most appropriate. In addition, a determination is made about which method is more appropriate under circumstances related to sample size, survey length, and number of response options. The research is broken into two parts; an empirical study was first conducted, followed by a simulation study.

The following research questions were addressed using an empirical study:

RQ (1) How do the trait of interest estimates found using the IRTree model, the MNRM, and the MPCM correlate?

RQ (2) How do the extreme response trait estimates found using the IRTree Model, the MNRM, and the MPCM methods correlate?

RQ (3) How do the IRTree Model, the MNRM, and the MPCM methods compare in model fit?

After analysis of the empirical results, a simulation study was conducted to answer the following research questions:

RQ (4) Is the pattern of differences in the item mean square error between the expected total score and true total score among the levels of sample size significantly different among the models of estimation?

RQ (5) Is the pattern of differences in the item mean square error between the expected total score and true total score among the levels of survey length significantly different among the models of estimation?

RQ (6) Is the pattern of differences in the item mean square error between the expected total score and true total score among the levels of number of category response options significantly different among the models of estimation?

2.0 LITERATURE REVIEW

The influence of response style on item responses is a concern when performing psychological assessment using self-report questionnaires. An example response style is extreme response style (ERS). ERS is the systematic tendency to select response options on the extreme ends of the response scale regardless of the content of the item. Each item response is influenced by the substantive trait along with the extreme response trait. Several models that incorporate an extreme response style trait have been developed. Each model treats response style uniquely. For example, the multidimensional nominal response model treats ERS as a trait independent of the trait of interest. The IRTree Model treats ERS as a compensatory trait to the substantive trait. The modified generalized partial credit model treats ERS as an individual random effect. Each of these is a type of multidimensional item response theory (MIRT) model.

The development of MIRT models accounting for extreme response tendencies is still in its infancy. As a result, limited literature analyzing, comparing, or applying these models exists. The following review of literature consists of an in-depth reflection of the few available studies. Each model will be presented and described. This is followed by an in-depth investigation of empirical and simulations studies applying these methods.

2.1 THE RELATIONSHIP BETEWEEN ERS AND OTHER FACTORS

Although the development of MIRT models accounting for extreme response tendencies is still in it’s infancy, traditional descriptive statistics measuring extreme response tendencies have been studied extensively. For example, Peterson, Rhi-Perez, and Albaum (2014) studied 6-category Likert-type items. To calculate ERS tendencies they used several measures. The authors called the first measure a direct measure that calculated the number of items an individual selected extreme response options. In turn, the higher the individual’s “score”, the greater their ERS tendency. To calculate a second measure of extreme response tendencies, the authors coded response options 3, 2, 1, 1, 2, 3 instead of 1, 2, 3, 4, 5, 6. Therefore a higher score represented greater extreme response tendencies. Instead of coded 0/1 for non-extreme versus extreme, this modified score scale helped determine degrees of ERS tendencies. A third measure used calculated the standard deviation of individual responses. Although this only represents broadness or narrowness of an individual’s response pattern, the authors reasoned that higher values of standard deviation represented larger degrees of extreme response tendencies.

Peterson et al. (2014) examined 6,146 undergraduate business students representing 120 universities from 36 countries including 2,949 individuals from the United States. The students were questioned on a 27 Likert-type item survey related to ethical reasoning. All 27 items had a response scale with response options ranging from 1-*Strongly Agree* to 6-*Strongly Disagree*. There were high correlations (.82 to .93) among ERS scores between all the three methods described previously. When only US students were considered, the correlations among the ERS scores from the three methods were slightly lower (.81 to .92). Additionally, the extreme response measures were found to be significantly correlated with gender (Peterson et al., 2014). Specifically, males tended to have higher extreme response tendencies than females. Although this correlation was significant ($p < .05$), the authors note the correlation was low and significance was most likely due to the large sample size.

Of the 6,146 individuals studied, 137 were used to measure test-retest reliability (two-week gap). The reliabilities were .71 (individual standard deviation), .72 (modified version of direct measure of ERS), and .74 (direct measure of ERS). The moderate test-retest reliability led researchers to conclude that extreme response tendencies are relatively stable across time. They noted that this result was consistent with Merrens (1970) and Berg (1953). Weijters, Geuens, and Schillewaert (2010a) found that response style can be modeled using a tau-equivalent factor with a time-invariant autoregressive effect. In other words, they found that ERS tendencies are largely consistent over the course of a questionnaire. This, combined with the results from Peterson et al. (2014), suggest that ERS tendency is consistent within a survey and across time.

Wetzel, Carstensen, and Böhnke (2013) investigated extreme response style using latent class analysis. They were able to determine two groups – those who exhibited extreme response tendencies and those who did not. The two latent classes showed systematic differences in their endorsement probability for extreme response categories. They investigated two instruments with several scales and found that between 65% and 80% of the participants applied the same response style on all the scales over the two instruments. The remaining 20% - 35% of the individuals could not unambiguously be classified as one latent class or the other. The researchers found that there were differing proportions of extreme responders across the two instruments studied. Wetzel et al. (2013) found that this difference was mostly on the level of ERS. This result indicated that ERS may be more adequately measured as a continuous trait rather than a have/have-not trait. This is evidence that the MIRT models assuming ERS as a continuous latent trait used in this study are appropriate.

Liu, Conrad, and Lee (2016) investigated whether the method by which an instrument is delivered (web based surveys vs. face-to-face surveys) has a relationship with extreme response style. The authors used a latent class analysis approach and found that extreme response style is prevalent in both web-based surveys as well as face-to-face surveys. Specifically, respondents who take face-to-face surveys exhibit greater degrees of extreme response style than web-based respondents. Furthermore, they found that there is no difference in the relationship between the mode of distribution and ERS tendencies for White, Black, and Hispanic respondents.

In addition to the method of distribution of the survey, Naemi, Beal, and Payne (2009) found that those who respond to survey items quickly tend to exhibit higher degrees of extreme response tendencies. They indicated that individuals who tend to have more simplistic thinking tendencies tend to respond to items quicker and exhibit higher instances of extreme responses. The authors also investigated whether there was a relationship between intolerance of ambiguity and extreme response style. Intolerance of ambiguity is the tendency to perceive ambiguous situations as sources of threat (Stanley Budner & Budner, 1962). An ambiguous situation is one which cannot be adequately structured or categorized by the individual because of the lack of sufficient cues (Stanley Budner & Budner, 1962). Naemi et al. (2009) found that individuals with higher incidences of intolerance of ambiguity also tended to have higher degrees of ERS. They reasoned that this was due to respondents with high intolerance of ambiguity avoiding ambiguous middle categories and disproportionally selecting unambiguous endpoints.

2.2 MULTIDIMENSIONAL NOMINAL RESPONSE MODEL

Bolt and Johnson (2009) sought to model a self-reported questionnaire using the nominal response model (Bock, 1972) while accounting for extreme response style. Bolt and Newton (2011) specify a multidimensional nominal response model (MNRM) with one substantive trait, θ_1, and one trait corresponding to extreme response style, θ_{ERS}. The probability of a respondent selecting category k, k = 1,…, K, on item j, given θ_1 and θ_{ERS}, is

$$P(U_j = k | \theta_1, \theta_{ERS}) = \frac{\exp(a_{jk1}\theta_1 + a_{jk2}\theta_{ERS} + c_{jk})}{\Sigma_{h=1}^{K} \exp(a_{jh1}\theta_1 + a_{jh2}\theta_{ERS} + c_{jh})} \qquad (1)$$

where a_{jk1} is the category k slope parameter influenced by the substantive trait, a_{jk2} is the category k slope parameter influenced by the extreme response style trait, and c_{jk} is the intercept parameter for category k. Whereas the a_{jk1} and a_{jk2} identify how the propensity to select each category varies as a function of the traits, the c_{jk} parameters reflect the propensities towards category k independent of the trait. For identification purposes, $\Sigma_{h=1}^{K} a_{jh1} = \Sigma_{h=1}^{K} a_{jh2} = \Sigma_{h=1}^{K} c_{jh} = 0$. To account for extreme response tendencies, a_{jk1}, for $k = 1, \dots, K$, are fixed interval-spaced values centered at 0. The a_{jk2}, $k = 1$ and K (extreme categories) are fixed positive values and the a_{jk2}, $k = 2, \dots, K - 1$ (intermediate categories) are fixed negative values. The fixed category slope values are equivalent across items.

To see how the model accounts for ERS, consider a four-category Likert-type item. The substantive trait category slopes,a_{jk1}, are fixed to -3, -1, 1, and 3 for $k = 1, \dots ,4$, respectively. The extreme trait category slopes, a_{jk2}, are set to 1, -1, -1, and 1 for $k = 1, \dots ,4$, respectively. For the extreme response trait, the intermediate category slopes ($\mathrm{k} = 2, 3$) have equivalent negative values and the extreme response category slopes ($\mathrm{k} = 1,4$) have equivalent positive values. The $\Sigma_{\mathrm{k}}\mathrm{a}_{\mathrm{jk1}} = 0$ and $\Sigma_{\mathrm{k}}\mathrm{a}_{\mathrm{jk2}} = 0$ in order to identify the model.

In order to illustrate the influence of the extreme response category slope parameters, first consider an individual with a high tendency of selecting extreme response options ($\theta_{\mathrm{ERS}} > 0$). For this individual, the $\mathrm{a}_{\mathrm{jk2}}\theta_{\mathrm{ERS}}$ term is positive for the first and fourth categories and the $\mathrm{a}_{\mathrm{jk2}}\theta_{\mathrm{ERS}}$ term is negative for the intermediate categories. This creates a higher probability of selecting categories 1 and 4 and a lower probability of selecting categories 2 and 3. Now consider an individual with a low tendency for selecting the extreme response options ($\theta_{\mathrm{ERS}} < 0$). The $\mathrm{a}_{\mathrm{jk2}}\theta_{\mathrm{ERS}}$ term is negative for the first and fourth categories and the $\mathrm{a}_{\mathrm{jk2}}\theta_{\mathrm{ERS}}$ term is positive for the intermediate categories. This lowers the probability of selecting the extreme categories and increases the probability of selecting the intermediate categories.

To examine the use of the MNRM model, Bolt and Newton (2011) investigated a self-report instrument designed to measure two substantive traits. U.S. respondents ($\mathrm{n} = 5{,}330$) were questioned on two science subscales from the Program for International Student Assessment (PISA, 2006). The two subscales measured a respondent's "Enjoyment of Science" (ENJ) and "Value of Science" (VAL). The ENJ subscale consisted of five items and the VAL subscale consisted of 10 items. Each item was a forced-choice four point Likert-style item. The response options were 1-*Strongly Disagree,* 2-*Disagree*, 3-*Agree*, and 4-*Strongly Agree*. Responses on the two subscales were analyzed using three models. Model 1 was a two-dimensional MNRM. Model 2 was a three-dimensional MNRM with an unspecified third trait. The third model was the three-dimensional MNRM with the third trait specified to account for extreme response tendencies.

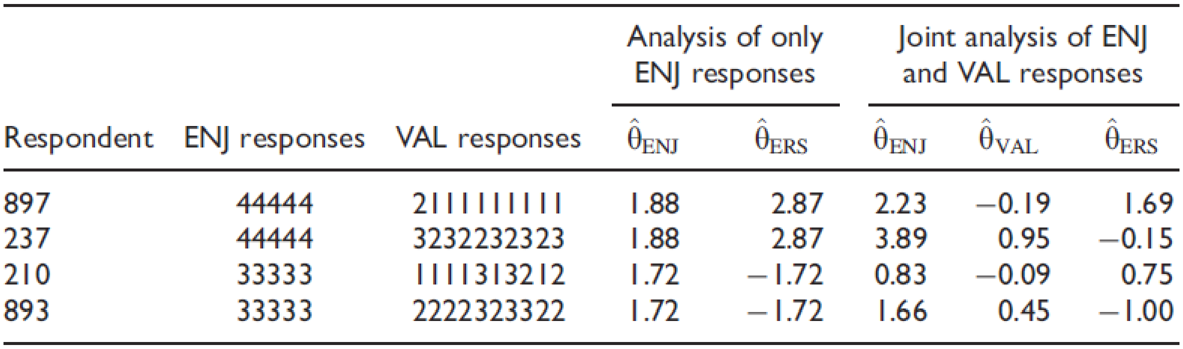
Model 1 examined the ENJ trait and the VAL trait. Each item was set to load on the dimension corresponding to its subscale. Category slopes were constrained to fixed values (-3, -1, 1, 3) for each trait. By constraining the category slopes to fixed values, Model 1 was specified as the multidimensional partial credit model. Model 2 examined the ENJ trait, the VAL trait, and an unspecified third trait. The unspecified third trait allowed for the possibility of an additional dimension influencing the individual responses. To allow the third dimension to relate to response style, the category slopes were free to vary across items. Similar to Model 1, the category slopes for the substantive traits in Model 2 were fixed to -3, -1, 1, and 3. The unspecified trait was included to confirm that a third factor was influencing the item responses. In order to determine that the third factor was the influence of extreme response tendencies, the trait needed to be defined. Model 3 specified the third dimension by fixing the category slopes to 1, -1, -1, and 1. This essentially defined the third trait as an extreme response tendencies trait. Consistent with Model 1 and Model 2, the category slopes for the two substantive traits were fixed at -3, -1, 1, and 3.

Bayesian information criterion (BIC), Akaike information criterion (AIC), Akaike information criterion 3 (AIC3), consistent Akaike information criterion (CAIC) were computed for each model. The BIC, AIC, AIC3, and CAIC were larger for Model 1 compared to Model 2. This was evidence that Model 2 had better fit. In other words, there was evidence of a third trait influencing item responses on the VAL and ENJ subscales. Model 3 was compared to Model 2 to determine if the third trait was the influence of extreme response tendencies. Model 3 had smaller BIC, AIC, AIC3, and CAIC values compared to Model 2. This provided evidence that the constraints of Model 3 seemed reasonable. In other words, Model 2 provided evidence a third dimension was present and Model 3 provided evidence that the third trait was due to extreme response tendencies (Bolt & Newton, 2011).

Bolt and Newton (2011) investigated the residual effects of the ERS trait estimate on the ENJ trait estimate when the VAL subscale items were included in the estimation process. They reasoned that the additional items from the VAL subscale would affect the estimate for the ERS trait. This effect on the ERS trait would then lead to residual effects on the ENJ trait estimates. The additional items from the VAL subscale would only load on the VAL trait and the ERS trait. They would then not contribute information to ENJ directly, but would contribute information to the extreme response trait. This is because they defined extreme response style as the tendency to select the extreme response items irrespective of the content of the item. They hypothesized that by having more items contribute to the ERS trait with the addition of VAL subscale items, the estimation of the ENJ trait would change.

To investigate the residual effects of the ERS trait estimate on the ENJ trait estimate when the VAL subscale items were included, two models were used: (1) a two-dimensional MNRM with the ENJ trait and the ERS trait and (2) a three-dimensional MNRM with the ENJ trait, the VAL trait, and the ERS trait.

Table 1. Real data examples, "Enjoy Science" (ENJ) and "Value Science" (VAL) subscales from PISA data (Bolt & Newton, 2011)



			Analysis of only ENJ responses		Joint analysis of ENJ and VAL responses		
Respondent	ENJ responses	VAL responses	$\hat{\theta}_{ENJ}$	$\hat{\theta}_{ERS}$	$\hat{\theta}_{ENJ}$	$\hat{\theta}_{VAL}$	$\hat{\theta}_{ERS}$
897	44444	2111111111	1.88	2.87	2.23	−0.19	1.69
237	44444	3232232323	1.88	2.87	3.89	0.95	−0.15
210	33333	1111313212	1.72	−1.72	0.83	−0.09	0.75
893	33333	2222323322	1.72	−1.72	1.66	0.45	−1.00

Table 1 displays item responses for the ENJ subscale, item responses for the VAL subscale, and trait estimates from the two models analyzed for four respondents. Respondents 897 and 237 have the same extreme response set on the ENJ subscale but different response sets on the VAL subscale. Respondents 210 and 893 have the same intermediate response set on the ENJ subscale but different response sets on the VAL subscale.

When only the ENJ trait and the ERS trait are considered, individuals within each pair (pair 1: 897 and 237; pair 2: 210 and 893) have equivalent ENJ trait and ERS trait estimates. The within-pair equivalent trait estimates were expected as the within-pair response sets are equivalent on the ENJ subscale. Pair 1 ($\hat{\theta}_{\text{ERS}} = 2.87$) has a higher tendency of extreme responses compared to pair 2 ($\hat{\theta}_{\text{ERS}} = -1.72$). This difference was also expected as individuals in pair 1 selected the extreme response category (4) for all items while individuals in pair 2 selected intermediate response category (3) for all items.

When responses on the VAL subscale were considered, respondent 897 had a higher tendency to select extreme response options (*1* or *4*) compared to respondent 237. On the VAL subscale, respondent 897 selected an extreme response option on 9 items. Respondent 237 did not select an extreme response option on any of the ten items on the VAL subscale. This difference in response set patterns was reflected in the ERS trait estimates from the joint analysis of ENJ and VAL responses. The ERS trait estimate from the joint analysis was $\hat{\theta}_{\mathrm{ERS}} = 1.69$ for respondent 897 and the ERS trait estimate from the joint analysis for respondent 237 was $\hat{\theta}_{\mathrm{ERS}} = -0.15$.

When the analysis was completed only considering the ENJ trait and the ERS trait, the ENJ trait estimate for respondent 897 was $\hat{\theta}_{ENJ} = 1.88$. When the analysis was completed using the ENJ trait, the VAL trait, and the ERS trait, the ENJ trait estimate for respondent 897 was $\hat{\theta}_{ENJ} = 2.23$. When accounting for both subscales, it became evident that respondent 897 tended to select extreme response options regardless of the content. Therefore, the reason respondent 897 selected response option *4* on the ENJ subscale was most likely due to their tendency to select extreme response options. It was less likely that these extreme response option selections were due to a high presence of the ENJ trait. Although the ENJ trait estimate for respondent 897 did rise when considering the VAL subscale, it did not rise with the same amount that the ENJ trait estimate did for respondent 237. When the analysis was completed only considering the ENJ trait and the ERS trait, the ENJ trait estimate for respondent 237 was $\hat{\theta} = 1.88$. When the analysis was completed using the ENJ trait, the VAL trait, and the ERS trait, the ENJ trait estimate for respondent 237 was $\hat{\theta}_{ENJ} = 3.89$. When all 15 items from the two subscales were considered, it became evident that this individual did not select extreme response options regardless of the content. Therefore, the selection of extreme response option *4* on the items from the ENJ subscale was most likely due to a high presence of the ENJ trait. For this reason, the ENJ trait estimate for respondent 237 rose dramatically when both subscales were used for analysis.

Results for respondent 893 and respondent 210 exhibited similar patterns to the results of respondent 897 and respondent 237. Both respondent 893 and respondent 210 did not select an extreme response option on the ENJ subscale. When only the ENJ trait and the ERS trait were considered for analysis, respondent 893's ERS trait estimate and respondent 210's ERS trait estimates were $\hat{\theta}_{ERS} = -1.72$. Respondent 210 selected an extreme response option on six out of the ten items on the VAL subscale. This resulted in respondent 210's ERS trait estimate to increase to $\hat{\theta}_{ERS} = 0.75$ when the ENJ trait, the VAL trait, and the ERS trait were all considered. Respondent 893 selected zero extreme response options on the VAL subscale. This resulted in respondent 893's ERS trait estimate to be lower than respondent 210's ERS trait estimate when the ENJ trait, the VAL trait, and the ERS trait were all considered for analysis. Inclusion of the additional items from the VAL subscale provided more information about the respondents' extreme response tendencies. With a lack of extreme response option selections on the ENJ subscale and the VAL subscale, it became evident that respondent 893 had a high tendency not to select the extreme response options.

In order to draw conclusions about the accuracy of estimates from the MNRM, Bolt and Newton (2011) performed a simulation study. The focus of the simulation study was to determine the recovery of a substantive trait and an ERS trait. The main objective was to provide evidence that the expected improvements in the trait of interest estimate when considering the secondary trait and the ERS trait were due to the improved ERS trait estimates. Additionally, Bolt and Newton (2011) wanted to show that this improvement was not due to collateral information from the secondary trait.

Data was simulated under a three dimensional MNRM with a substantive trait of interest, a second substantive trait, and an ERS trait. Developed to mirror the previous empirical example, 5 items were simulated for the subscale of interest and 10 items were simulated for the second subscale. The category slopes and the intercepts used for data generation were identical to the estimates from the analysis of the PISA responses when the ENJ trait, the VAL trait, and the ERS trait were considered. The correlation between the two substantive traits was simulated under five conditions (0, 0.3, 0.5, 0.58, 0.8). A correlation of 0.58 was considered to mimic the correlation found between the ENJ trait and the VAL trait from the empirical example described previously. The variances of all three traits $(\theta_1, \theta_2, \theta_{ERS})$ were held constant to match the estimated variances from the previous study. A sample size of 1000 individuals were simulated from a multivariate normal trait distribution with mean vector 0 and specified variance-covariance matrix. For each condition 100 replications were analyzed.

After data generation, three models were used to analyze the data:

Model 1: MNRM with one substantive trait (θ_1) and one ERS trait (θ_{ERS})

Model 2: MNRM with two substantive traits (θ_1, θ_2) and no ERS trait

Model 3: MNRM with two substantive traits (θ_1, θ_2) and one ERS trait (θ_{ERS})

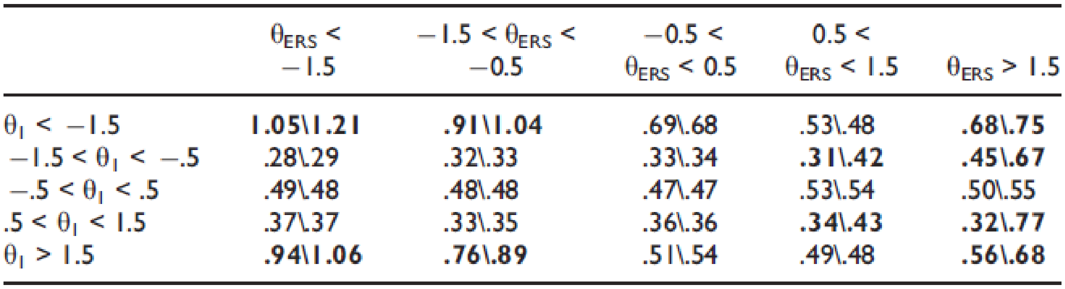
For all three models, recovery of θ_1 was relatively unaffected by the correlation between the traits. Model 3 showed a consistent one-third reduction in mean absolute difference (MAD) between true and estimated ERS when compared to Model 1. Model 3 was found to have a lower mean absolute difference between the true and the estimated trait of interest parameter values compared to the MAD between the true and the estimated trait of interest parameter values from Model 1. This result provided evidence that the second subscale items did improve estimation of the primary trait of interest. This occurred even though the items for the secondary trait did not also load on the primary trait.

It was still necessary to show that this improved estimation was due to a better estimation of the ERS trait alone and not due to collateral information from the items on the second subscale. To do this, the difference in MAD recovery of the primary trait between Model 3 and Model 2 was compared to the difference in MAD recovery of the primary trait between Model 3 and Model 1. Not accounting for extreme response tendencies was found to have the most detrimental effect on recovery of the primary trait. In other words, using both subscales but disregarding extreme response tendencies (Model 2) resulted in the poorest recovery of the primary trait. Furthermore, the difference in MAD was less between Model 3 and Model 1 than the difference in MAD between Model 3 and Model 2. There was less error in the recovery of the primary trait when estimating only the primary trait with the ERS trait compared to estimating the two traits without an ERS trait (Bolt & Newton, 2011). This shows the importance of incorporating an ERS trait when extreme response tendencies exist. This result also provided evidence that the substantive trait of interest estimate $\left(\hat{\theta}_1\right)$ benefited from the improved extreme response trait estimate in Model 3 and not collateral information from the second trait $\left(\hat{\theta}_2\right)$.

Even though Model 3 showed better recovery of the trait of interest compared to Model 1, the difference was trivial. In other words, results showed that when ERS was considered, improvement in recovery was minor. This result can be explained by the use of an overall measure. For example, there is low bias in parameter estimates for individuals with substantive trait values around $\theta_1 = 0$ when an ERS trait is not considered. Bolt and Newton (2011) stress that extreme response style is not consistent across all levels of θ_1.

Table 2 displays the MAD between the true and estimated trait of interest parameter values, conditional on different ranges of θ_1 and θ_{ERS}, for Model 3 compared to Model 1. When θ_1 is close to 0 there exists little difference between MAD results from Model 3 (including the second trait) and Model 1 (excluding the second trait). Consider when $-.5 < \theta_1 < .5$ and $-.5 < \theta_{ERS} < .5$. The mean absolute difference between the true and estimated trait levels for Model 3 and Model 1 were equivalent to .47. As θ_1 gets further from 0, the differences in recovery of the two models became more evident. For example, consider when $\theta_1 > 1.5$ and $\theta_{ERS} < -1.5$. The mean absolute difference for Model 1 is 1.06 and the mean absolute difference for Model 3 is .94. This provides evidence that Model 3 had better recovery of the true trait of interest parameters. These large differences are apparent in the values that are bolded. The bolded differences are more frequent when both θ_1 and θ_{ERS} are conditioned on values far from 0.

Table 2. Mean Absolute Difference (MAD) Results for $\boldsymbol{\theta_1}$ (Model 3\Model 1), Simulation Study (Bolt & Newton, 2011)



	$\theta_{ERS} <$ -1.5	$-1.5 < \theta_{ERS} <$ -0.5	$-0.5 <$ $\theta_{ERS} < 0.5$	$0.5 <$ $\theta_{ERS} < 1.5$	$\theta_{ERS} > 1.5$
$\theta_1 < -1.5$	**1.05\1.21**	**.91\1.04**	.69\.68	.53\.48	**.68\.75**
$-1.5 < \theta_1 < -.5$	.28\.29	.32\.33	.33\.34	**.31\.42**	**.45\.67**
$-.5 < \theta_1 < .5$	.49\.48	.48\.48	.47\.47	.53\.54	.50\.55
$.5 < \theta_1 < 1.5$	.37\.37	.33\.35	.36\.36	**.34\.43**	**.32\.77**
$\theta_1 > 1.5$	**.94\1.06**	**.76\.89**	.51\.54	.49\.48	**.56\.68**

2.3 IRTREE MODEL

The IRTree Model is an alternative approach to account for extreme response tendencies in the multidimensional item response theory framework. For this approach to work, an individual is assumed to engage in a two stage decision-making process when responding to ordinally scaled items (Böckenholt, 2012). For example, consider an item with response options 1-*Strongly Disagree*, 2-*Disagree*, 3-*Agree*, and 4-*Strongly Agree.* An individual may respond based on two distinct processes: first by deciding the direction of the response (positive versus negative); and second by deciding the intensity of the response (Thissen-Roe & Thissen, 2013).

The idea of a sequential decision-based approach is not new but is used infrequently in psychometric measures (Thissen-Roe & Thissen, 2013). Hart (1923) developed a questionnaire in order to assess this decision based approach. Initially, a respondent answered each item with assent, dissent, or neutrality. The individual then responded to each item again by underlining and double underlining responses for relative emphasis (Thissen-Roe & Thissen, 2013). Thissen-Roe and Thissen (2013) describe an example where sequential decisions are present in decision making and judgement making. If a witness identifies a suspect in a lineup, he then follows it up indicating confidence in his identification.

An IRTree Model is used to measure the sequential decision-making response process. Figure 1 shows an IRTree Model for a four-category forced-choice Likert-scaled item. The probability of *Agree*, $\mathrm{P}(\mathrm{D}_1 = 1)$, is a function of a latent trait, θ_1, which represents the substantive trait of interest. The probability of a "strong" intensity response, $\mathrm{P}(\mathrm{D}_2 = 1)$, is a function of a latent variable, θ_{ERS}, which represents individual differences in the tendency to select extreme responses (Thissen-Roe & Thissen, 2013). Here the probability of an intense response is treated independently of the first decision. In other words, the probability of a strong response is the same whether the first decision was *Agree* or *Disagree*.

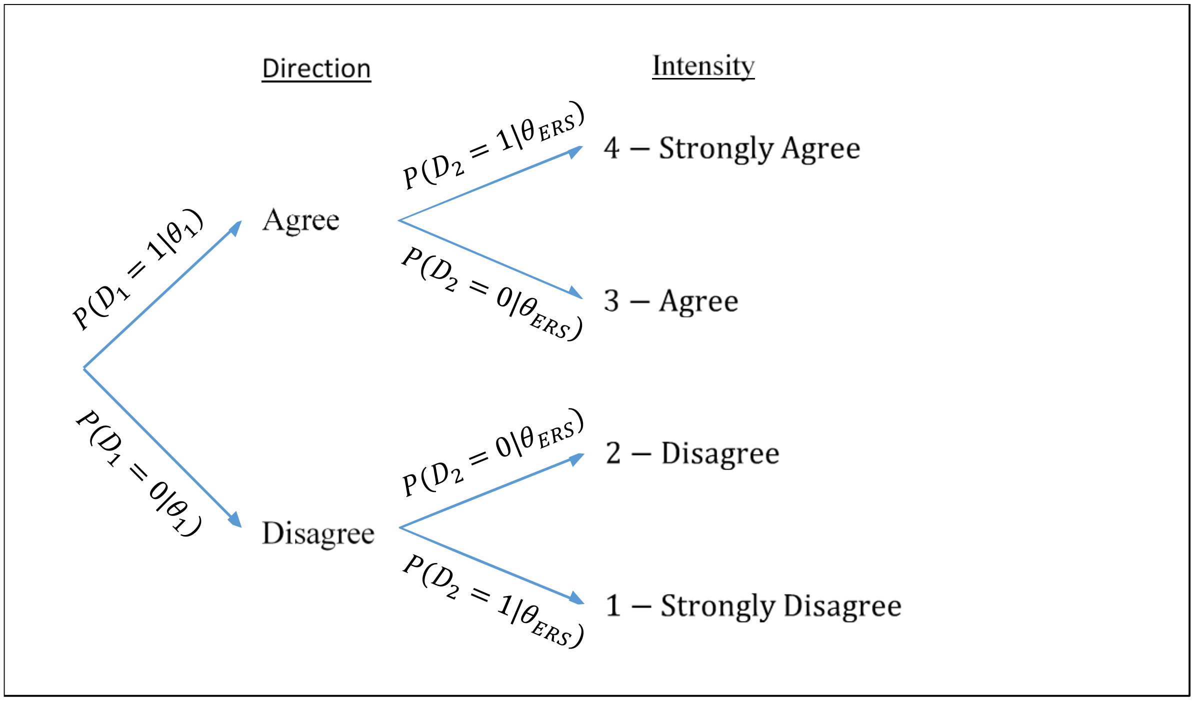


Figure 1. IRTree Model for a four-category Likert-scaled item.

The first decision is modeled using a 2-parameter logistic (2PL) model:

$$P(D_1 = 1|\theta_1) = \frac{1}{1 + e^{-(b_1 + a_1\theta_1)}} \tag{2}$$

and

$$P(D_1 = 0|\theta_1) = 1 - P(D_1 = 1|\theta_1) \tag{3}$$

where b_1 is the intercept parameter and a_1 represents the discrimination parameter that summarizes the regression of the response process on the latent variable (Thissen-Roe & Thissen, 2013). Note that the parameters are item specific but subscripts have been suppressed. The second decision is modeled using a modified 2PL model given as:

$$P(D_2 = 1|\theta_{ERS}, \theta_1) = \frac{1}{1 + e^{-(b_2 + a_2\theta_{ERS} \pm \nu a_2(b_1 + a_1\theta_1))}} \tag{4}$$

and

$$P(D_2 = 0|\theta_{ERS}, \theta_1) = 1 - \frac{1}{1 + e^{-(b_2 + a_2\theta_{ERS} \pm \nu a_2(b_1 + a_1\theta_1))}} \tag{5}$$

where b_2 is the intercept parameter and a_2 is the slope parameter describing the item-specific probability of extreme responding with the latent variable describing an individual's tendency toward extreme responses (Thissen-Roe & Thissen, 2013). The ν parameter introduces a compensatory nature of the two traits. This shift term, $\nu a_2(b_1 + a_1\theta_1)$, is additive when $k = 3$ and $k = 4$ and subtractive when $k = 1$ and $k = 2$. When ν is positive, respondents with moderate tendencies for extreme responses will tend to use the Likert-scale as intended. In other words, they will only provide an extreme response when their position has a strong intensity. For a given item, the probability of a response is the product of the probability of decision 1 and the probability of decision 2. For a given item, the probability of selecting response option k given θ_1 and θ_{ERS} is

$$P(U = k) = P(D_1) * P(D_2) \tag{6}$$

for $k = 1,2,3, and\ 4$.

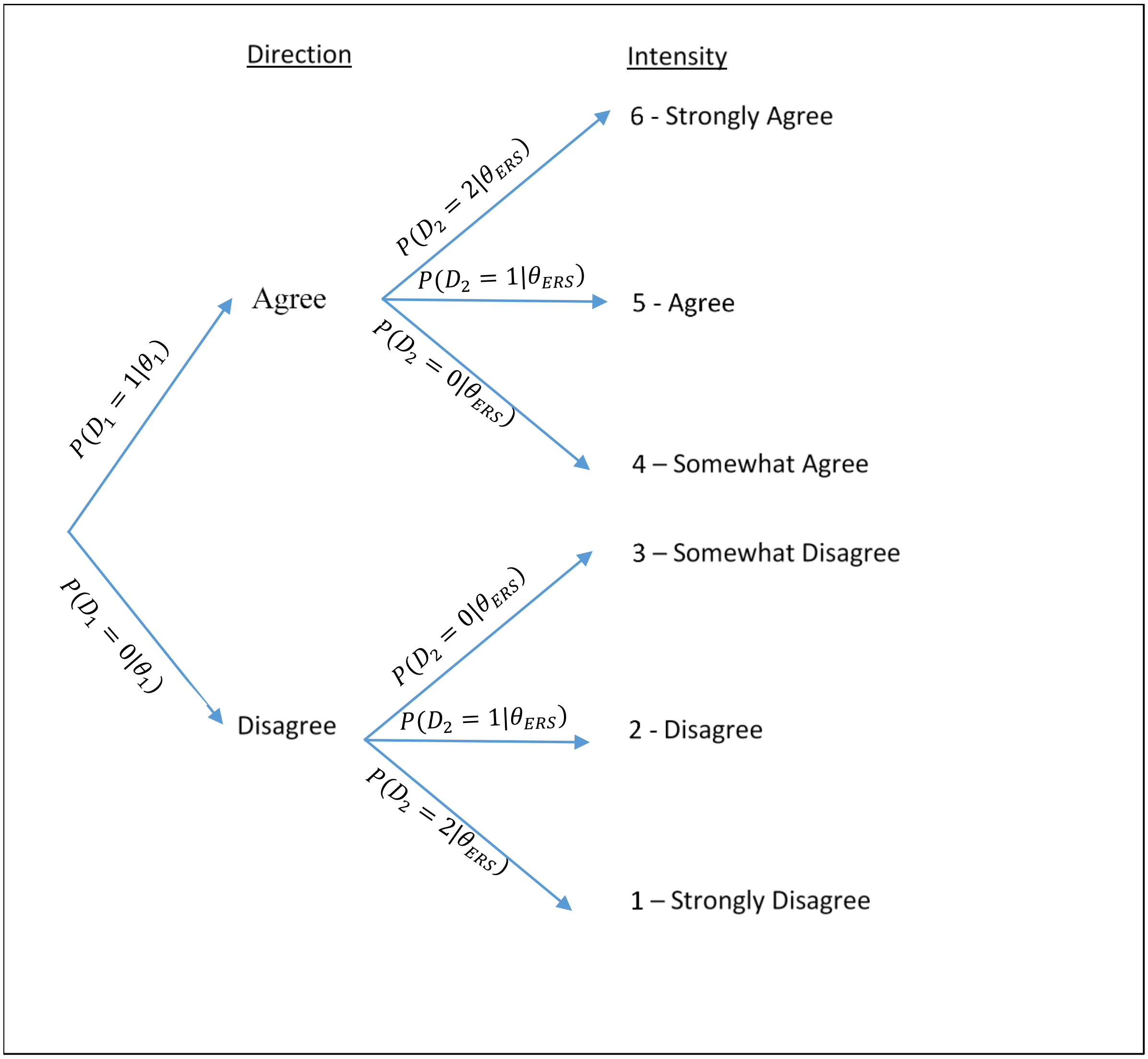


Figure 2. IRTree Model for a six-category Likert-scaled item.

An IRTree Model can be expanded for items with any number of response options. For example, Figure 2 displays a 6-category IRTree Model. The directional decision is modeled using a 2-PL. The intensity decision is modeled using a modified Graded Response Model (Samejima, 1969). The probability model for the intensity decision is given as:

$$P(D_2 = 0|\theta_{ERS}, \theta_1) = \frac{1}{1 + e^{-(c_2 + a_2\theta_{ERS} \pm \nu a_2(b_1 + a_1\theta_1))}} \tag{7}$$

$$P(D_2 = 1|\theta_{ERS}, \theta_1) = \frac{1}{1 + e^{-(c_1 + a_2\theta_{ERS} \pm \nu a_2(b_1 + a_1\theta_1))}} - \frac{1}{1 + e^{-(c_2 + a_2\theta_{ERS} \pm \nu a_2(b_1 + a_1\theta_1))}} \tag{8}$$

$$P(D_2 = 2|\theta_{ERS}, \theta_1) = 1 - \frac{1}{1 + e^{-(c_1 + a_2\theta_{ERS} \pm \nu a_2(b_1 + a_1\theta_1))}} \tag{9}$$

The c_1 and c_2 parameters are response category intercepts. The slope parameter, a_2, has the same interpretation as that of the 4-category item. The shift term, $\nu a_2(b_1 + a_1\theta_1)$, is subtractive for $k = 1{,}2,$ and 3 and additive when $k = 4, 5,$ and 6. The probability of selecting category $k, k = 1, \ldots, 6$, given θ_1 and θ_2, is

$$P(U = k) = P(D_1) * P(D_2) \tag{10}$$

The correlation between trait θ_1 and trait θ_{ERS} is treated as an unknown parameter during the estimation process.

Thissen-Roe and Thissen (2013) investigated the IRTree Model using an empirical approach. They sampled 86 items from an electronic job application survey used by 17 large and mid-size organizations in the long-term care industry. They analyzed 56 5-category Likert-type items in combination with 30 adjective dyads and forced-choice work preference items. The survey focused on work style constructs and personality constructs. They explored three disjoint representative subsamples that consisted of only first time applicants. Using the subsamples allowed replication of the study. Each sample consisted of approximately 64,000 applicants.

The generalized partial credit model (GPCM), the nominal response model (NRM), the graded response model (GR), the proportional thresholds model (PTM), and the multidimensional nominal response model (MNRM) were compared to the IRTree Model. The PTM (Rossi, Gilula, & Allenby, 2001) assumes that category threshold parameters are proportional across individuals. The item response category thresholds can be decomposed into two exclusive pieces. The first piece is a vector of values associated with the item responses. The second piece is a single value. This single value is an individual difference scale parameter. Thissen-Roe and Thissen (2013) reparametrized the PTM in order for the scale parameter to be a combination of an extreme response parameter and a common scale variance.

The MNRM, the PTM, and the IRTree Model are inherently multidimensional. Thissen-Roe and Thissen (2013) considered these models unidimensional when analysis was performed investigating one construct and an ERS trait. The MNRM, the PTM, and the IRTree Model were considered multidimensional when more than one construct was measured along with the ERS trait. For the empirical study, a multidimensional version and a unidimensional version of each model was analyzed. In each case, the 2PL model was used to analyze the adjective dyads and work preference items.

Overall, the NRM, the MNRM, the PTM, and the IRTree Model outperformed the GPCM and the GRM. The IRTree Model estimated had the lowest BIC (Bayesian Information Criterion) when the unidimensional version and the multidimensional versions were considered. This was an indication that the unidimensional version and the multidimensional version of the IRTree Model had the best fit. Small BIC differences were exhibited between the three models that accounted for extreme response tendencies (the MNRM, the IRTree Model, and the PTM). Thissen-Roe and Thissen (2013) considered these models more similar than dissimilar. The model results, however, were not completely identical. This was evident when the correlations between the estimated construct scores and the ERS scores were considered. The average correlation between the estimated construct score and the estimated ERS trait was -0.052 under the multidimensional MNRM, 0.091 under the multidimensional PTM, and 0.331 under the IRTree Model. This indicated some differences in the meaning of the traits being estimated (Thissen-Roe & Thissen, 2013).

The MNRM and the IRTree Model assume a different interaction between the item and the individual. This makes it difficult to compare estimated parameters across models. Thissen-Roe and Thissen (2013) plotted the joint consequences for the item response surfaces in order to compare models. Figure 3 displays the distribution of modal item responses across the general trait and the extreme response trait for an example item under the MNRM and the IRTree Model. The item prompt read, “I dislike having to change my plans because of other people’s mistakes” with response options *Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree*. Individuals with a high presence of the substantive trait are likely to *Disagree* or *Strongly Disagree* with this item.

Under the MNRM (left surface), individuals with high extreme response tendencies selected either *Strongly Agree* or *Strongly Disagree*. Under the IRTree Model (right surface), individuals with high ERS tendencies selected *Strongly Agree, Neutral,* or¸ *Strongly Disagree*. Under the IRTree Model, individuals exhibiting very low extreme response tendencies selected *Neutral* more often than any other response option. This was regardless of their estimated trait value. This result is strikingly different compared to those same individuals under the MNRM. When the MNRM was considered, those exhibiting high tendencies of the trait selected *Disagree* even with low tendencies of ERS. Individuals with low substantive trait estimates and low ERS tendencies selected *Agree* more often than any other response option. Individuals with low ERS tendencies and trait estimates around -2 tended to select *Neutral*.

An important difference between the IRTree Model and the MNRM is the treatment of the item parameters associated with the ERS trait. The MNRM holds these item parameters constant across all items. In other words, the extreme response trait is a function of the response options only. The IRTree Model estimates a slope parameter and an intercept parameter for the extreme response trait for each item. The IRTree Model treats ERS tendencies as a function of both response options and item stem.

Thissen-Roe and Thissen (2013) concluded that there is support for the idea that a partial-compensatory two-stage process is a plausible model for responses to Likert-type items. They defended this statement by showing the IRTree Model provided the best fit to the data compared to the several other models considered. They cautioned that this example was only in the context of employment testing and further study needs to be done to confirm the IRTree Model is useful in other psychological assessment measure.

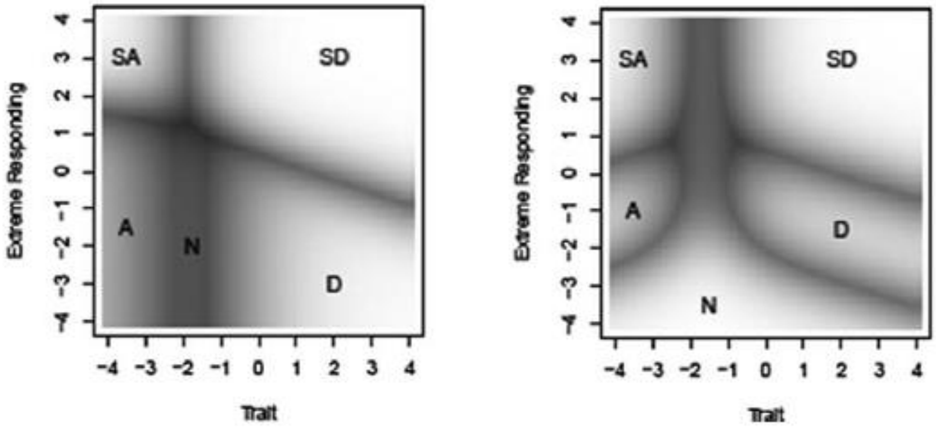


Figure 3. Dominant regions for responses to a five-category item under MNRM (left) and IRTree Model(Thissen-Roe & Thissen, 2013)

2.4 MODIFIED GENERALIZED PARTIAL CREDIT MODEL

In a third approach to modeling extreme response style, Jin and Wang (2014) introduced a modified version of the generalized partial credit model (MPCM). They sought to model a respondent's perspective of between option threshold size. To account for individual differences of between option threshold size, the threshold size is treated as a random effect. For example, one individual may consider the discrepancy between *Strongly Agree* and *Agree* to be small while others may consider this threshold large. A modified generalized partial credit model (MPCM) was developed to account for this random effect.

Jin and Wang (2014) presented the MPCM model as the log-odds of selecting option *k* over option $k - 1$ on item *j* by

$$\log\left(\frac{\mathrm{P_{njk}}}{\mathrm{P_{nj(k-1)}}}\right) = \mathrm{a_j}\left[\theta_\mathrm{n} - \left(\delta_\mathrm{j} + \omega_\mathrm{n}\tau_\mathrm{jk}\right)\right] \tag{11}$$

where a_j and δ_j are the discrimination and location parameters of the item with respect to θ, τ_{jk} is random category threshold parameter, and ω_n is a random-effect weight parameter on thresholds designed to control the propensity for an extreme response.

The random-effect weight parameter is independent of the substantive trait, θ. The weight parameter follows a log-normal distribution with a mean of zero and a variance of σ_ω^2. When ω is large, the distance between thresholds is large. When the distance between thresholds is large, the tendency to endorse the extreme response options is low. When $\omega = 1$ for all persons, the MPCM simplifies to the unidimensional generalized partial credit model. For a four-point item, if $\omega = 0.5$, the shape of the item category characteristics curves become narrower. When the item characteristic curves become narrower, the probability of selecting response option *1* and the probability of selecting response option *4* increase (Jin & Wang, 2014). As ω increases from 1, the probability to endorse an extreme response options decreases.

Jin and Wang (2014) investigated a twenty item survey designed by Lo (2001). The survey was intended to measure interpersonal conflicts. The twenty 7-point Likert type items had response options,1-*Strongly Unconfident*, 4-*Neutral*, and 7-*Strongly Confident*. Six models were used to analyze 982 responses on the survey: the partial credit model (PCM), the partial credit model with an ERS trait (PCM-ERS), the partial credit model with an ERS trait and covariates (PCM-ERS-c), the generalized partial credit model (GPCM), the modified generalized partial credit model (MPCM), and the modified generalized partial credit model with covariates (MPCM-c). The covariates used for the PCM-ERS-c and the MPCM-c were gender and the number of siblings.

Table 3. Students' Response Patterns and the Estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ (SE in Parentheses) (Jin & Wang, 2014)

					GPCM	MPCM with Covariates	
ID	Raw Responses	Mean	SD	#ER	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\theta}}$	$\hat{\omega}$
19	45333412211244442214	2.80	1.28	4	-1.47(.37)	-1.66(.50)	.95(.41)
790	41141771111177711111	2.80	2.65	18	-1.70(.38)	-1.32(.36)	.14(.09)

The deviance information criterion (DIC) suggested that the Modified Partial Credit Model with covariates had the best fit. Table 3 displays selected students' raw responses with their trait estimates from the generalized partial credit model (GPCM) and the modified generalized partial credit model (MPCM) with covariates. When the generalized partial credit model was used for analysis, student 19 had a trait estimate equal to -1.47 and student 790 had a trait estimate equal to -1.70. When analysis was conducted using the MPCM with covariates, extreme response tendencies were considered. When extreme response tendencies were considered, the trait estimate for student 19 was -1.66 and the trait estimate for student 790 was -1.32.

When analysis was performed using the MPCM with covariates, the extreme response weight parameter, ω, was estimated. The extreme response weight parameter was estimated to be 0.95 for student 19 and the extreme response weight parameter was estimated to be 0.14 for student 790. Student 19 selected an extreme response option on four items (#ER). This low tendency to select extreme response options was reflected by the estimated weight parameter value near 1. Due to student 19's tendency not to select extreme options, their true θ value may not be as high as their GPCM estimated theta suggested. In other words, their below average item response (2.80) is more likely due to a low presence of the trait and less likely due to a high tendency to select the extreme response option, *Strongly Unconfident*, regardless of the content of the item. When the analysis was done using the MPCM with covariates, their extreme response tendencies were taken into consideration, and thus, $\hat{\theta}$, decreased to -1.66.

Student 790 selected an extreme response option on 18 items. Student 790's high tendency to select extreme response options was reflected by their $\hat{\omega}$ estimate (.14) being close to 0. With such a high tendency to select extreme response options, their true value of θ is likely not as low as the GPCM indicated. In other words, Student 790's below average item response (2.80) was likely more indicative of their extreme tendencies than it was of the low presence of the trait of interest. When the MPCM was used to account for extreme tendencies, student 790's estimated trait increased to -1.32.

Figure 4 displays the relationship between the trait of interest estimates obtained from the MPCM with covariates and the trait of interest estimates obtained from GPCM. The mean difference between estimates was -0.0004 with a standard deviation of 0.22. The small average difference in trait of interest estimates between the MPCM and the GPCM was due to only 30% of respondents having a statistically significant ω. The raw differences in θ estimates ranged from -.95 to .93.

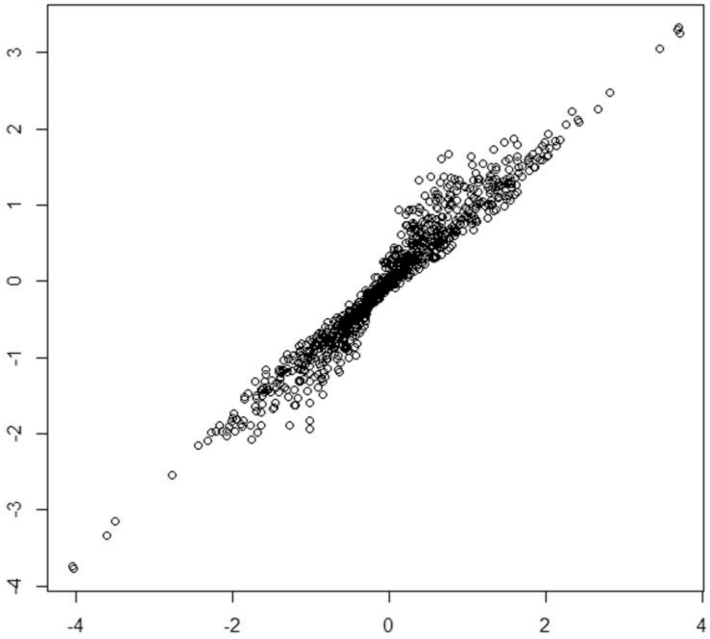


Figure 4. Trait estimates from the MPCM with covariates (y-axis) vs. GPCM(Jin & Wang, 2014)

In order to understand the accuracy of parameter recovery of the MPCM, Jin and Wang (2014) performed two separate simulation studies. The first study assessed parameter recovery when the sample size (250, 500, 1,000, 2,000), test length (20, 40), and categories per item (4, 6) varied. The ability parameters, θ, were generated from a $\mathcal{N}(0,1)$. The random effect parameters, ω, were drawn from a log-normal $(0, 0.3^2)$. Item difficulty parameters were generated from a uniform $(-2, 2)$. The threshold parameters were fixed (4-point items: $-0.6, 0, 0.6$; 6-point items: $-0.8, -0.4, 0, 0.4, 0.8$).

The second study focused on the consequences of model misspecification. The study explored using an over-parameterized model (with ERS when no ERS existed) and an under-parameterized model (without ERS when ERS was present). For this study, 1,000 individuals were simulated with $\sigma_{\omega}^{2} = 0, .3^{2}, .6^{2}$, and $.9^{2}$.

Due to lengthy computation time, only 20 replications per condition in each study were analyzed. As sample size and test length increased, recovery of item difficulties and thresholds improved. In small sample sizes (n=250), difficulties and thresholds were biased. The variance of the ω was underestimated for small samples but improved with large sample sizes and longer tests. Additionally, ERS trait estimation improved with 6-point items compared to 4-point items.

The researchers found that over-parameterized models did little harm when there was no extreme response trait exhibited. With extreme response tendencies present, the GPCM (generalized partial credit model - without an extreme response trait) yielded poor estimates compared to the MPCM with ERS. The larger the variation on the extreme trait (σ_{ω}^{2}), the worse the GPCM performed. For example, the RMSE for the slope parameters of the MPCM with ERS were $0.102, 0.098,$ and 0.104 when $\sigma_{\omega}^{2} = 0.3^{2}, 0.6^{2},$ and 0.9^{2}, respectively. For the same values of σ_{ω}^{2}, the RMSE for the slope parameters of the GPCM were $0.114, 0.151,$ and 0.194, respectively. This provided additional evidence that the MPCM with ERS performed better.

2.5 BAYESIAN ESTIMATION

Due to the complex nature of multidimensional item response models, a Bayesian approach for estimation is preferred. One favorable advantage of the Bayesian approach is the direct translation of complicated and highly parametrized models. In recent years, software that makes implementation possible such as SAS (Statistical Analysis System) has become rapidly available. SAS has a built in procedure (PROC MCMC) that allows for straightforward translation of response probability models. This makes a Bayesian analysis of multidimensional IRT models more accessible to researchers and scale developers (Stone & Zhu, 2015). PROC MCMC also provides deviance information criteria (DIC) (Spiegelhalter, Best, Carlin, & Angelika van der, 2002).

A frequentist approach to estimation assumes that the unknown parameters of a model are fixed while the sample of data is random. In order to perform hypothesis testing of parameters, hypothetical repeated samples or theoretical sample distributions for parameters under the null are considered. To quantify uncertainty, a likelihood-based approach can be taken where the first and second derivatives are obtained to estimate parameters and quantify their standard errors.

In a Bayesian approach, parameters are random variables and have prior distributions that reflect the uncertainty about the true values before observing the data. This information is then combined with the observed data in order to update the distribution, called the posterior distribution. This posterior distribution for parameters given the observed data is used to estimate the parameters as well as quantify their uncertainty (Congdon, 2002; Fox, 2010; Gelman, 2014).

Stone and Zhu (2015) list out several advantages of using the Bayesian paradigm, particularly in relation to IRT models:

1) Parameter estimates are typically more precise than maximum likelihood estimates.
2) Methods accommodate perfect and imperfect response patterns.
3) Trait parameters are estimated simultaneously with item parameters.
4) Use of prior distributions may provide more stable estimation particularly with small samples or short surveys.

5) Estimation of more complex or highly parameterized models is more accessible.

Furthermore, Meijer and Glas (2003) emphasize two additional advantages of the Bayesian approach:

1) There exists no need to derive the theorectical sampling distirbution of the statistic.
2) Uncertainty of person-fit statistics that depend on unknown quantities such as item and person parameters is explicitly taken into account.

A decision a researcher must make is to the selection of the prior distribution. The prior distribution should reflect any information about the parameters of interest available. However, if there is no prior information available, it is preferable that the prior have minimal influence compared to the likelihood (the data). These priors are called non-informative prior distributions.

The posterior distribution for a parameter can be derived analytically but typically involves complex integrals for most statistical models. In order to alleviate this problem, Markov chain Monte Carlo (MCMC) methods are used. MCMC methods generate a sample of parameter values from the joint posterior distributions for the parameters instead of analytically solving complex mathematics. Once the posterior samples have been obtained, inference may be performed (Gelman, 2014). In other words, point estimates using sample-based inference are as simple as computing the mean over the posterior samples. As an example, Wainer, Bradlow, and Wang (2007) provide a 5 step outline for an MCMC sampling algorithm:

1) An initial starting vector is selected, $\Lambda = \Lambda^{(\mathrm{t=0})}$, where t denotes the iteration number. Set $t = 0$.
2) Select some subset of the parameters, λ_1, and draw updated values $\lambda^{(t+1)}$ from its full conditional distribution $p\left(\lambda_1 \middle| Y, \Lambda_{-\lambda_1}^{(t)}\right)$, where $\Lambda_{-\lambda_1}^{(t)}$ denotes the entire parameter vector Λ excluding parameters λ_1, evaluated at its t −th value, and Y denotes the observed test data.

3) Select some subset of the parameters, λ_2, and draw updated values $\lambda_2^{(t+1)}$ from its full conditional distribution $p\left(\lambda_2 \middle| Y, \Lambda_{-\lambda_1,-\lambda_2}^{(t)}, \lambda_1^{(t+1)}\right)$, where $\Lambda_{-\lambda_1,-\lambda_2}^{(t)}$ denotes the set of all parameters excluding λ_1 and λ_2, where λ_2 is evaluated at its t-th value, and $\lambda_1^{(t+1)}$ is the updated value of λ_1 obtained in step 2.
4) Sample $\Lambda_{-\lambda_1,-\lambda_2}^{(t)}$ from its full conditional distribution $p\left(\Lambda_{-\lambda_1,-\lambda_2} \middle| Y, \lambda_1^{(t+1)}, \lambda_2^{(t+1)}\right)$. Let $t = t + 1$.
5) If $t \leq M$ (a prespecified number of iterations), go to Step 2. If not, then stop.

Wainer et al. (2007) warn that these five steps only provide a general outline of how a MCMC algorithm proceeds in order to build the posterior distribution. Many decisions a researcher must make are generalized over. For instance, consider step 2. There are many MCMC samplers that can cycle through the parameters meaning there does not exist a unique solution.

Another consideration that must be made is how to choose the starting values. Selecting good starting values can be crucial for models that may be multimodal or high-dimensional (Wainer et al., 2007). A third consideration that must be made is how to subset the parameter vector for sampling. Specifying how the parameters will be sampled will lead to either good or poor MCMC mixing properties. In other word, for a $j = 1, \dots, J$ item survey, sampling the J a's (using a 2-Parameter Logistic model) all at once may lead to the Markov chain getting stuck at specified values or moving slowly through the parameter space (Wainer et al., 2007).

The most common method for sampling is known as the Metropolis-Hastings algorithm. This method selects a potential value λ^* from a proposal distribution, $g(\lambda)$. It then either "stays" by setting $\lambda^{(t+1)}$ equal to the previous value $\lambda^{(t)}$ or sets $\lambda^{(t+1)} = \lambda^*$, with probability

$$\min\left(\frac{p\left(\lambda^*\middle|Y,\Lambda_{-\lambda}^{(t)}\right)g(\lambda^*)}{p\left(\lambda^{(t)}\middle|Y,\Lambda_{-\lambda}^{(t)}\right)g(\lambda^{(t)})},1\right) \tag{12}$$

Wainer et al. (2007) caution that the probability of selecting a “good” candidate vector is low for high-dimensional vectors, causing the MCMC sampler to get stuck at values near or exactly equal to $\lambda^{(t)}$. In order to alleviate this problem in the multidimensional space, an alternative sampler may be used. The Gibbs sampler, a special case of the Metropolis-Hastings algorithm, may be used to alleviate this problem.

Fox (2010) describes the Gibbs sampler with IRT models. Values from independent conditional distributions that are in the limit equal to drawing from the joint posterior distribution are drawn. Thus, the joint posterior reduces to a set of one-dimensional conditional distributions that can be sampled. Each parameter is sampled from a vector of parameters, conditioned on the most recent sampled values of the other parameters. After a sequential sampling, the posterior distribution for each parameter is formed. The point estimate and corresponding standard error of the parameters are reported as the mean and standard deviation of the posterior distribution, respectively.

A final consideration that must be made when implementing an MCMC sampler is the choice of the number of iterations. To determine the number of iterations necessary, one must consider both the number of utilized iterations and the number of burn-in iterations. In the case of starting with poor initial values, there is a period of “walking towards” the area of the posterior distribution that has significant probability (Wainer et al., 2007). These iterations should be discarded and are known as the burn-in period. The remaining iterations are used to estimate the posterior quantities of interest. The iterations used should only be after the MCMC sampler has reached a stationary point. To choose the number of iterations utilized and the number of iterations for burn-in, one must balance computation time with Monte Carlo simulation error (Wainer et al., 2007).

2.5.1 Bayesian Model Checking

Gelman, Meng, and Stern (1996) warn that Bayesian analysis can be misleading when the model is far from plausible. They proposed that in Bayesian statistics a model can be checked in at least three ways.

1) Examining the sensitivity of inferences to reasonable changes in the distribution and the likelihood.
2) Checking that the posterior inferences are reasonable, given the substantive context of the model.
3) Checking that the model fits the data.

The first two methods focus on the inferences while the last method is specific only to the model. The third method calls for the use of a check of the posterior predictive distribution for *discrepancy*. This is an extension of classical test statistics to allow dependence on unknown parameters. In other words, the use of simple discrepancy measures reveal useful information about model misfit and when model assumptions are violated (Sinharay, Johnson, & Stern, 2006).

A popular Bayesian diagnostic tool, proposed by Rubin (1984) , is known as Posterior Predictive Model-Checking (PPMC). Sinharay et al. (2006) stated that PPMC is a popular tool for four primary reasons: 1 – intuitive and simple to apply; 2 – strong theoretical basis; 3 – provides graphical checks which are the most natural and easily comprehensible way to perform PPMC;

and 4 – provide numerical evidence about model misfit. PPMC involves comparing observed data with replicated data using several diagnostice measures (discrepancy measures) that are sensitive to model misfit. The replicated data is data that is generated or predicted by the assumed model.

Fox (2010) states the PPMC is advantageous due to the theoretical sampling distirbution of the statistic not needing to be derived. Furthermore, the discrepency measures may depend on unknown parameters but the association uncertiainty is explicitly taken into account when computing posterior predictive probability values (*PPP*-values). Gelman et al. (1996) posit that Bayesian inference is a powerful tool to learn about model defects because a discrepancy measure allows for the examination of any function of the data and the parameters. Additionally, the posterior predicative approach is suitable for assessing the fitness of a *single* model. In other words, it is possible to construct discrepancies to detect the lack of fit of a single model in the absence of explicit alternative models.

2.6 SUMMARY

The multidimensional nominal response model, the IRTree model, and the modified generalized partial credit model are all novel approaches developed to account for extreme response tendencies. Each has been theoretically developed, however, limited studies of application are available. Through recent developments and availability of Bayesian estimation techniques, further study of these highly complex multidimensional IRT models is possible.

3.0 METHODS

The multidimensional nominal response model (MNRM), the IRTree Model, and the modified generalized partial credit model (MPCM) each account for extreme response style (ERS) with conceptual uniqueness. The MNRM treats ERS as an independent latent trait. The IRTree Model deconstructs the respondent's decision making process into two stages in order to separate out the ERS tendencies. The MPCM assumes a random nature to the threshold size that varies across individuals exhibiting different levels of ERS tendencies.

As each method conceptualizes extreme response differently it is difficult to perform a comparison among the models. Furthermore, it may be challenging for researchers to know which model to use in different situations. An empirical study to compare the model parameter estimates and the model fit was performed. Additionally, a simulation study was performed in order to determine which model is appropriate to use when ERS is present.

3.1 PART 1 – EMPIRICAL STUDY

In order to illustrate model differences, an empirical study comparing the MNRM, the IRTree Model, and the MPCM was conducted. The empirical study addressed the following research questions:

RQ (1) How do the trait of interest estimates found using the IRTree model, the MNRM, and the MPCM correlate?

RQ (2) How do the extreme response trait estimates found using the IRTree Model, the MNRM, and the MPCM methods correlate?

RQ (3) How do the IRTree Model, the MNRM, and the MPCM methods compare in model fit?

3.1.1 Data

Bolt and Newton (2011) investigated extreme response style when working with a self-report instrument designed to measure two substantive traits. In this study, only one of those two traits was examined. U.S. respondents ($\mathrm{n} = 5{,}330$) were questioned on the "Value of Science" subscale (VAL) from the Program for International Student Assessment (PISA, 2006). The subscale consisted of 10 items (Table 4) with response options: 1- *Strongly Disagree*, 2-*Disgree*, 3-*Agree*, and 4-*Strongly* A*gree*.

Table 4. Value of Science Subscale Item Questions (Bolt & Newton, 2011)

Item	Prompt
1	"Advances in <broad science and technology> usually improve people's living conditions"
2	"<Broad science> is important for helping us to understand the natural world"
3	"Some concepts in <broad science> help me see how to relate to other people"
4	"advances in <broad science and technology> usually help improve the economy"
5	"I will use <broad science> in many ways when I am an adult"
6	"<Board science> is valuable to society"
7	"<Broad science> is very relevant to me"
8	"I find that <broad science> helps me to understand the things around me"
9	"Advances in <broad science and technology> usually bring social benefits"
10	"When I leave school there will be many opportunities for me to use <broad science>"

An individual's extreme response rate is calculated by dividing the number of items the individual selected *Strongly Disagree* or *Strongly Agree* by the total number of items on the subscale (Greenleaf, 1992b). With only ten items, individual extreme response rates ranged from 0.0 to 1.0 at 0.1 increments (Table 5). Nearly 13% of respondents had an extreme response rate of 80% or above. In other words, 13% of respondents answered at least 8 of the 10 items using either response option *1-Strongly Disagree* or response option *4-Strongly Agree*. Greenleaf (1992b) compared the individual extreme response rates to the proportion of extreme response options on any given item. Out of the four response options, two were considered extreme: *4-Strongly Agree* and *1-Strongly Disagree.* With 50% of the response options considered extreme, the benchmark to compare extreme response rates to is 0.5. 30% of the individuals surveyed selected extreme response options on 50% or more of the items. Overall item 1 and item 2 had the highest tendency of extreme responses. On item 1 42% of respondents selected an extreme response option and on item 2 45% of respondents selected an extreme response option. The proportions of respondents who selected either 1-*Strongly Disagree* or 4-*Strongly Agree* on each item are displayed in Table 6.

Table 5. Individual Extreme Response Rates from PISA responses

Extreme Response Rate	Frequency	Percentage	Cumulative Percentage
0	1552	29.12	29.12
0.1	599	11.24	40.36
0.2	606	11.37	51.73
0.3	544	10.21	61.93
0.4	415	7.79	69.72
0.5	424	7.95	77.67
0.6	279	5.23	82.91
0.7	234	4.39	87.30
0.8	224	4.20	91.50
0.9	169	3.17	94.67
1.0	284	5.33	100.00

Table 6. Item Extreme Response Rates for PISA responses

Item	Proportion Extreme
1	.42
2	.45
3	.21
4	.32
5	.28
6	.36
7	.25
8	.29
9	.25
10	.28

3.1.2 Estimation

The data was fit using the IRTree model, the multidimensional nominal response model, and the modified generalized partial credit model. The three models were estimated using Bayesian methods and SAS PROC MCMC. The MCMC procedure consisted of 25,000 iterations where the first 5,000 iterations were discarded as burn-in and the subsequent 20,000 iterations were retained for estimation.

In order to estimate the MNRM, slope parameters were fixed values set to mirror Bolt and Newton (2011). Slope parameters associated with the VAL trait, $\mathrm{a_{jk1}}$, were fixed at -3, -1, 1, and 3 for categories $k = 1, \ldots, 4$, respectively. The slope parameters associated with the extreme response trait, $\mathrm{a_{jk2}}$, for $k = 1, \ldots, 4$ were fixed at 1, -1, -1, and 1, respectively. The slope parameter values were fixed for all items, $j = 1, \ldots 10$. The intercept parameters, c_{jk}, for categories, $k = 2, 3,$ and 4 were estimated using a normal prior distribution with mean 0 and variance 25. Figure 5 graphically displays the uninformative prior used to estimate the intercept parameters. For model identification purposes, c_{j1} was equal to the $-\sum_{h=2}^{4} c_{jh}$. Trait distributions were estimated by mirroring the techniques from Johnson and Bolt (2010). The VAL trait, θ_1, followed a normal distribution with mean zero and variance 1. The extreme response trait, θ_{ERS}, followed a normal distribution with mean zero and variance 1. The correlation between the traits was fixed to 0. The correlation was fixed to zero as extreme response style is the tendency to endorse extreme response options regardless of the content or the trait being measured (Greenleaf, 1992b).

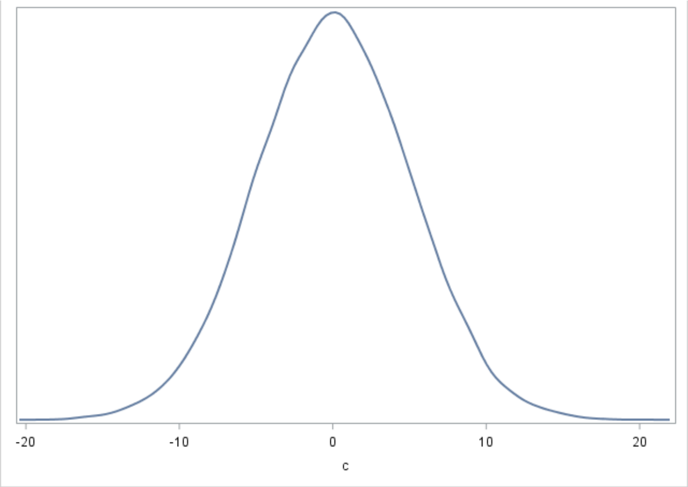


Figure 5. Prior distribution for intercept parameters under the MNRM

To fit the MPCM, discrimination parameters, a_j, used an informative lognormal prior distribution with mean 0 and variance 2. The location parameter, δ_j, was estimated with an informative Normal prior distribution with mean 0 and variance 2. The random category threshold parameters between 1-*Strongly Disagree* and 2-*Disagree*, τ_1, and between 2-*Disagree* and 3-*Agree*, τ_2, were estimated using a less informative normal prior density with mean zero and variance 10. The three prior distributions used for estimation are displayed in Figure 6. The random category threshold between 3-*Agree* and 4-*Strongly Agree*, τ_3 was set to the negative sum of τ_1 and τ_2. The Value of Science trait, θ, was normally distributed with mean zero and variance 1. The random-effects trait associated with extreme response tendencies, ω, followed a lognormal distribution with mean zero and variance σ_{ω}^{2}. The precision of ω, inverse of the variance term, $\psi = \frac{1}{\sigma_{\omega}^{2}}$, was estimated using a gamma prior with shape 1 and scale 10. Figure 7 displays the Gamma distribution with shape 1 and scale 10.

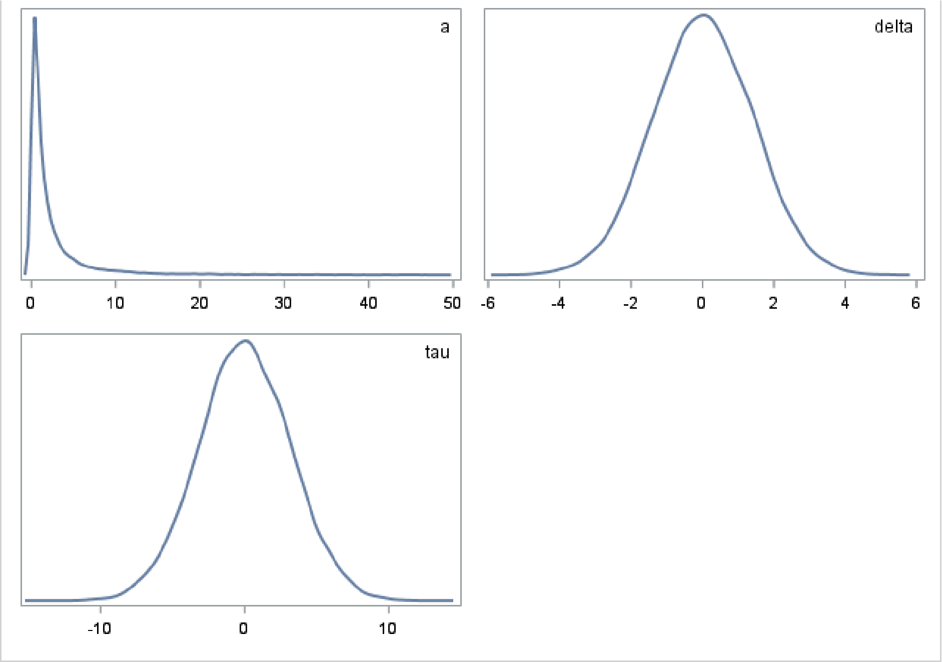


Figure 6. Prior distributions used for discrimination parameters, location parameters, and threshold parameters for the MPCM

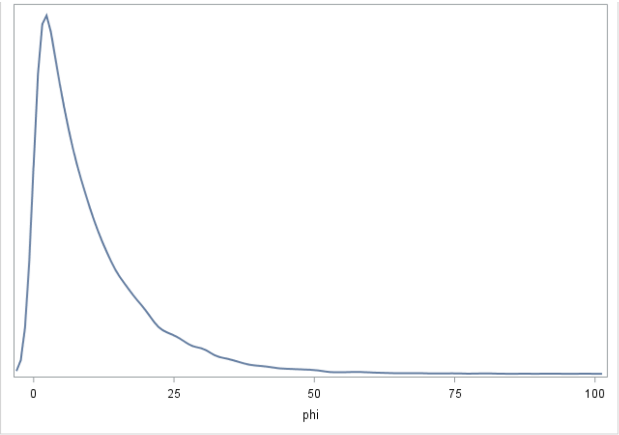


Figure 7. Gamma prior distribution used for the precision of the random effect for extreme response tendencies

To estimate the IRTree model, the slope parameters, a_1 and a_2, used an uninformative half-normal prior with mean 0, variance 16, and lower threshold 0. The intercept parameters, b_1 and b_2, were estimated using a less informative normal prior with mean 0 and variance 16. The shift constant, ν, assumed a normal prior density with mean 0 and variance 16. The prior distributions used to estimate the item parameters for the IRTree Model are displayed in Figure 8. The trait associated with Value of Science, θ_1, and the trait associated with extreme response tendencies, θ_{ERS}, followed a multivariate normal distribution with mean vector zero and variances equal to one. The correlation, ρ, between θ_1 and θ_{ERS} was estimated using the noninformative normal prior with mean 0, variance 0.5, and strict lower (-1) and upper boundaries (1), displayed in Figure 9.

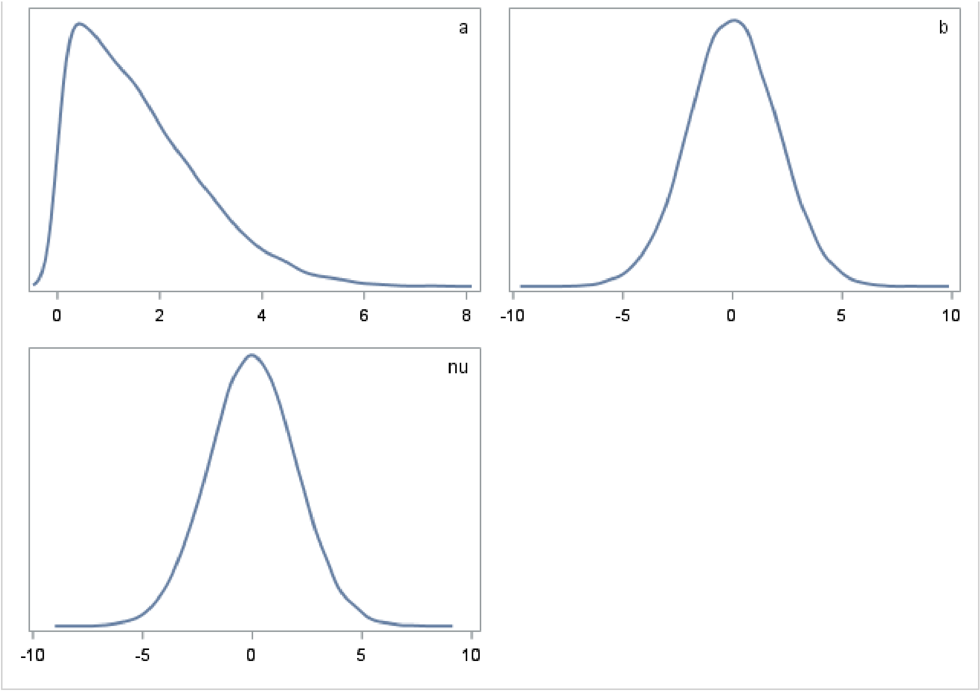


Figure 8. Item parameter prior distributions for the IRTree Model

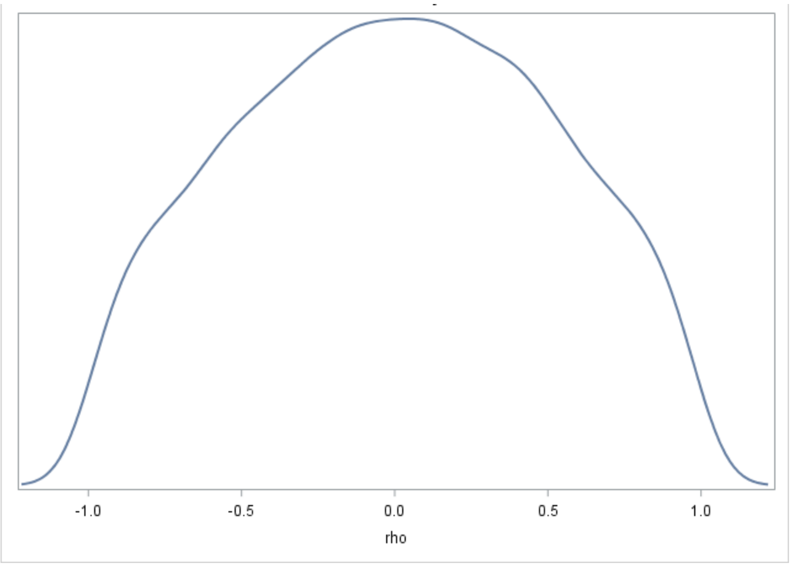


Figure 9. Prior distribution for the correlation between VAL and ERS traits

3.1.3 Convergence Diagnostics

To successfully estimate the parameters using MCMC, the Markov chain must converge to the posterior distribution. Without convergence, any inferences based on the posterior distribution would be invalid. In this case, the chain of random draws would not represent the desired posterior distribution of the parameters of interest.

To assess convergence of the MCMC algorithm two plots were examined. The first plots the iterates of the parameters from the simulation runs and monitors the trends. This plot can be used to detect abnormal or nonstationary behavior in the chain (Fox, 2010). The plot of successive draws is called a trace plot. This plot was investigated to see that the draws successfully traverse the entire parameter space. Traversing the entire parameter space quickly is an indication of good posterior mixing. Two example trace plots are displayed in Figure 10 from Stone and Zhu (2015). The trace plot, or "history plot", on the left shows a high likelihood of convergence. This plot traverses the entire posterior space quickly. The trace plot on the right shows a Markov chain with a low likelihood of convergence. The figure does not appear to show randomly plotted values around the mean of the Markov chain. Additionally, the chain does not traverse the space quickly. Instead the iterates follow a pattern of high posterior values followed by a group of low posterior values.

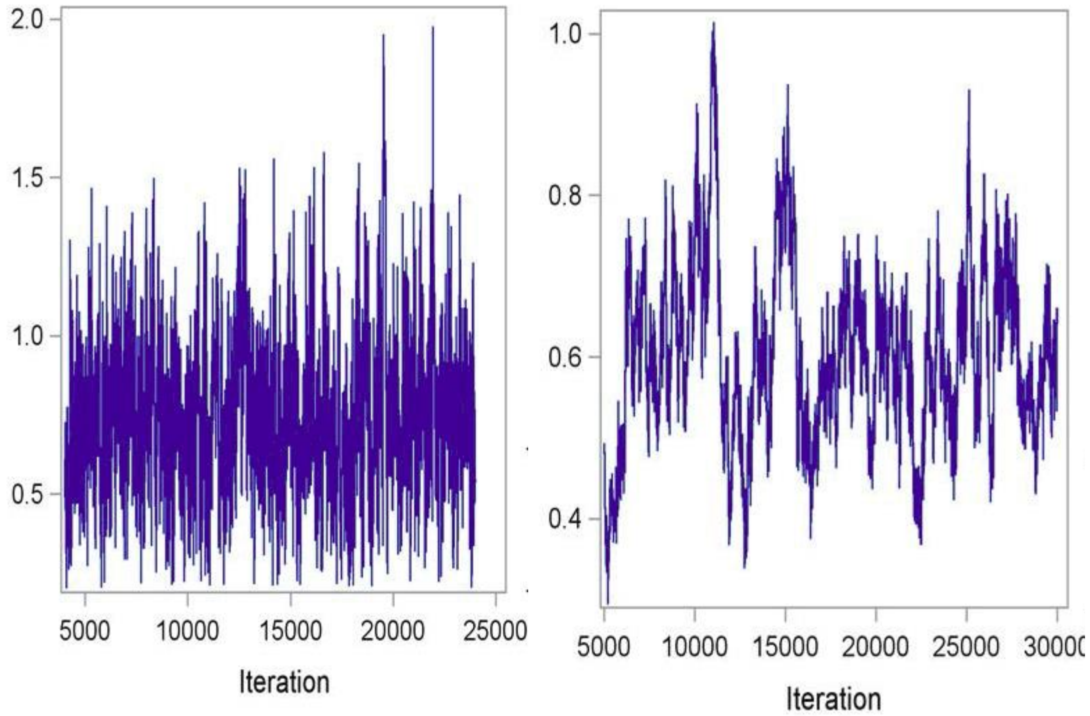


Figure 10. Trace plots displaying evidence of convergence and non-convergence (Stone & Zhu, 2015)

The second plot that was examined to assess convergence was an autocorrelation plot. Fox (2010) defines the autocorrelation at lag h as

$$r_h = \frac{\Sigma_{m=h+1}^{M}(\theta^{(m)} - \hat{\theta}_M)(\theta^{(m-h)} - \hat{\theta}_M)}{\Sigma_{m=h+1}^{M}\left(\theta^{(m)} - \hat{\theta}_M\right)^2} \qquad (\ 13\ )$$

where $\theta^{(m)}, m = 1, \dots, M$ denotes the output from one simulation run and $\hat{\theta}_M$ is the estimate of the posterior mean given dependent samples for the posterior distribution. Using the sample autocorrelation estimates of different orders, the correlation between adjacent iterates of an MCMC sequence is monitored. Two autocorrelation plots are displayed in Figure 11. These plots reflect the amount of autocorrelation among sampled values for different lags. The autocorrelation plot on the left is evidence of an efficient Markov chain. If the MCMC algorithm is mixing well, then this autocorrelation descends rapidly to zero. This is not, however, direct evidence of convergence of the Markov chain. If an MCMC algorithm is generating highly correlated posterior values, many iterations will be required for convergence. This would lead to an inefficient chain. An inefficient chain is evidence of a Markov chain that is not mixing well. The plot on the right of Figure 11 is an example of a chain that has significant lag. This displays a Markov chain that is inefficient. Thinning the chain by keeping every k^{th} simulated draw is one way to relieve inefficiencies of the chain.

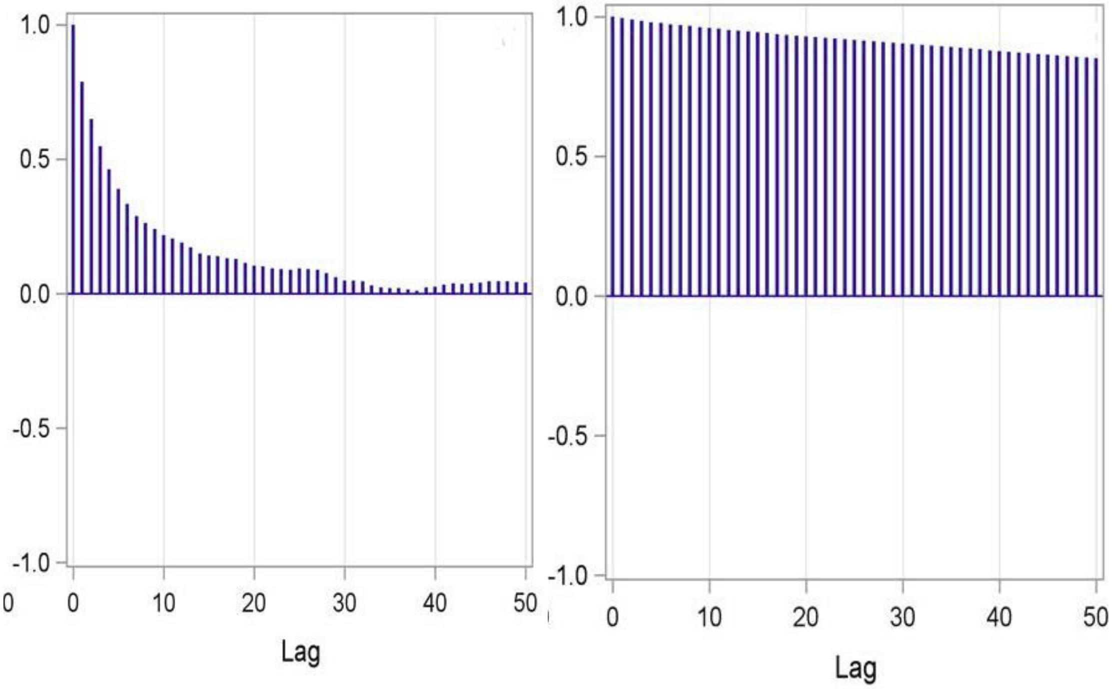


Figure 11. Autocorrelation plots showing evidence of efficient and inefficient Markov chains (Stone & Zhu, 2015)

There are many convergence diagnostics in addition to the two convergence diagnostics outlined previously. Cowles and Carlin (1996) provide an in-depth review of alternative measures of convergence diagnostics. Without the Markov chain converging to the posterior distribution, inferences about parameters based on the distribution are invalid. Therefore, it is crucial to assess the convergence of the Markov chain prior to making any Bayesian inferences.

3.1.4 Posterior Predictive Model Checking (PPMC)

A posterior predictive model check (PPMC) was performed using the MCMC output after estimation. PPMC methods have been proved useful in evaluating model fit for IRT models (see (Levy, Mislevy, & Sinharay, 2009; Sinharay, 2005, 2006; Stone & Zhu, 2015; Zhu & Stone, 2011). The PPMC method draws sampled model parameters from the joint posterior distribution to simulate item responses under the model. For each draw from the posterior distribution, a data set of model-based predicted observations is simulated. Each dataset represents a sample from the posterior distribution. A discrepancy statistic is computed for each replication to form a distribution of the discrepancy statistic under the null condition of model-data-fit. This discrepancy statistic is then computed on the observed data. Any systematic differences between the discrepancy statistic under the null condition and the discrepancy under the observed data is used to evaluate fit (Stone & Zhu, 2015).

To evaluate fit, the PPMC Method makes use of posterior predictive probability values (*PPP*-values). A *PPP*-value is calculated by determining the proportion of sampled iterations where the discrepancy statistic is larger than the same discrepancy statistic calculated under the observed data. In other words,

$$PPP = P\big(T(D^{rep}) \geq T(D)\big) \tag{14}$$

where $T(D^{rep})$ is the distribution of the discrepancy statistics calculated across each iteration (replication) of sampled values, and, $T(D)$ is the value of the discrepancy statistic computed on the observed data (Gelman, 2014). PPP-values near 0.5 indicate that no systematic differences exist between predicated values and observed values of the discrepancy statistic. Thus, PPP-values near 0.5 indicate model fit. PPP-values near 0 or 1 indicate systematic differences in the discrepancy statistic calculated on the observed values and the predicted values. Thus, PPP-values near less than 0.05 or greater than 0.95 indicate inadequate model fit.

To use PPMC, a discrepancy statistic must be selected. Under traditional IRT model, there are assumptions of dimensionality and local independence that are evaluated using pairwise tests. One example is the *odds ratio* (OR). The OR is used to evaluate local dependence for dichotomous items (Chen & Thissen, 1997). OR for dichotomously scored item pairs (j and j^*) are calculated from a 2X2 table by $\frac{n_{oo}n_{11}}{n_{o1}n_{10}}$, where n_{pq} is the observed number of examinees having response p (0 or 1) on Item j and response q (0 or 1) on Item j^*.

An odds ratio can be computed for every combination of item scores for polytomously scored items. It has been found useful to compute global OR for polytomously scored items (Agresti, 2002). For any two items with 4-response options each, the 4x4 can be reduced to a 2x2 table by combining categories. Zhu and Stone (2011) explored the global odds ratio as a discrepancy measure when analyzing 5 response category (0 – 4) Likert-scaled items. They dichotomized the item responses by grouping 0,1, and 2 as incorrect and grouping 3 and 4 as correct. Under this dichotomization, they found the global odds ratio to be a useful discrepancy measure for evaluating model fit with the graded response model.

Discrepancy measures are selected to be related to the inferences and uses of the model. The three models explored in this study are developed to account for extreme response tendencies.

The discrepancy measures analyzed, therefore, incorporated this purpose. The global odds ratio computed for each model (IRTree Model, MNRM, and MPCM) was computed by dichotomizing 1-*Strongly Disagree* and 4-*Strongly Agree* vs. 2-*Disagree* and 3-*Agree.* In other words, the dichotomy evaluated using the global OR was "extreme responses" and "non-extreme responses".

Another way to assess model fit is at the test-level. The observed and predicted total test score distributions were compared. Model fit was also evaluated at the item-level. Instead of using traditional discrepancy measures for item-level PPMC, a new discrepancy measure for item-level model fit was developed specifically for models accounting for extreme response styles. In order to evaluate the presence of extreme response tendencies, Greenleaf (1992b) evaluated the item extreme response rate. Greenleaf (1992b) proposed determining the proportion of individuals who selected an extreme response on a given item. Model fit at the item-level was investigated using this traditional measure of extreme tendencies as a discrepancy statistic. The proportion of extreme responses by item was calculated for each predicted model. The *PPP*-value for each item was equal to the proportion of these values greater than the item extreme response rate in the observed data.

The final PPMC discrepancy statistic used to evaluate model fit was at the person-level. An individual's extreme response rate is the proportion of items the individual selected *1-Strongly Disagree* or *4-Strongly Agree* for on the subscale (Greenleaf, 1992b). With only ten items, individual extreme response rates ranged from 0.0 to 1.0 at 0.1 increments. Person-level model fit was evaluated by comparing the frequency of persons with each individual extreme response rate. The *PPP*-value was calculated by determining the proportion of frequencies of individual extreme response rates from the predicted models that are greater than the frequencies of individual extreme response rates in the observed data.

3.1.5 Baseline of Model Fit

In order to have a baseline comparison of model fit, the unidimensional graded response model (Samejima, 1969) was analyzed. The graded response (GR) model is a model used for assessment applications that have Likert type response scales and are unidimensional. In the GR model, the cumulative probability that a person selects category k or higher on item j is given as

$$P_{jk}^{*}(\theta) = \frac{e^{Da_j(\theta - b_{jk})}}{1 + e^{Da_j(\theta - b_{jk})}} \qquad (15)$$

where a_j is the discrimination parameter for item j, b_{jk} is the threshold parameter for category k on item j, and D is the scaling constant. Given the cumulative probabilities and constraints of $P_{j1}^{*}(\theta) = 1$ and $P_{j(m_j+1)}^{*} = 0$, the probability of selecting category k on item j is the difference between the cumulative probabilities for two adjacent categories. The term $Da_j(\theta - b_{jk})$ is known as the logistic deviate. To improve convergence during the MCMC procedure, the logistic deviate can be reparametrized as $a_j\theta - d_{jk}$ where $b_{jk} = d_{jk}/a_j$ and $D = 1$. This form is known as the slope-intercept form.

Multidimensional models are used when more than one dimension is present. The comparison to a unidimensional model that does not account for extreme response style was used to illustrate the need to account for extreme response style as an alternative dimension. The unidimensional GR model was used to have a baseline comparison as a model that does not exhibit adequate fit.

The GR model was estimated using SAS PROC MCMC. The slope-intercept form of the GR model was estimated. The slope was assumed unknown but held constant across each item. The slope parameter was estimated using a lognormal prior distribution with mean 0 and variance 25. The intercept parameters were estimated with a normal prior distribution with mean 0 and variance 4. The lower bound for each successive category was the resulting estimate of the previous category threshold intercept.

3.1.6 Analysis

The results from the estimation of the MPCM, MNRM, and the IRTree model were compared. To address RQ (1), the pair-wise Pearson correlations between the scores for the trait of interest from all three models were computed. To address RQ (2), the pair-wise correlation between the estimates of the extreme response tendencies of the three models were computed.

In order to compare model fit, RQ (3), posterior predictive model checking, the deviance information criteria (DIC), and the mean square error between expected total score and observed total score were calculated. Spiegelhalter et al. (2002) define the DIC as the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity. The deviance is defined as

$$D(\theta) = -2\log p(y|\theta) + 2\log p(y) \quad (16)$$

The deviance does not discriminate between models because it will always prefer higher-dimensional models (Fox, 2010). Thus, an estimate for the number of effective model parameters is defined as the sum of the posterior mean deviance and the estimated deviance given a posterior estimate of θ as

$$\begin{aligned} p_D &= E(-2\log p(y|\theta)) + 2\log p\left(y\middle|\hat{\theta}\right) \\ &= \overline{D(\theta)} - D\left(\hat{\theta}\right) \end{aligned} \tag{17}$$

The DIC is then defined as

$$\begin{aligned} DIC &= \overline{D(\theta)} + p_D \\ &= D\left(\hat{\theta}\right) + 2p_D \end{aligned} \tag{18}$$

The model with the lowest DIC was considered to have the best fit.

Additionally, the mean squared error (MSE) between expected total score, $E\left(U_j\middle|\theta\right) = \Sigma_{j=1}^{J}\sum_{k=1}^{m_j} k * P_{jk}(\boldsymbol{\theta})$, and observed total, $\sum_{j=1}^{J} U_j$, was computed. The mean square error for J items with m_j response options was calculated as

$$MSE = \left(\sum_{j=1}^{J} U_j - \sum_{j=1}^{J}\sum_{k=1}^{m_j} k * P_{jk}(\boldsymbol{\theta})\right)^2 \tag{19}$$

where U_j is an oberseved item response on item j and $\boldsymbol{\theta}$ is the vector of estimated traits. The model with the lowest MSE was considered the best fitting model.

3.2 PART 2 – SIMULATION STUDY

A simulation study was performed to evaluate how well each model performs under different criteria. Jin and Wang (2014) provided a list of factors to consider when developing a simulation study involving extreme response style. The factors to consider are:

1) The larger the number of items, the better the identification of ERS will be.
2) The larger number of response options, the better the identification will be. The authors point out that it is easier to identify ERS with 7-point scales versus 3-point scales.
3) Identification of which categories are considered extreme may affect estimation. For example, on a 7-point scale, some may argue that options 1 and 7 are the only extreme categories while others may argue options 1, 2, 6, and 7 are extreme opinions.
4) Models used should separate the latent trait of interest from that for extreme response. This leads to "purified" latent traits.

3.2.1 Simulation Factors

To address the list of extreme response tendency simulation criteria provided by Jin and Wang (2014), the number of items, the number of response options, the sample size, and the data generation method were varied as simulation factors.

3.2.1.1 Number of Items

At the survey level, 10 items and 20 items will be simulated. A small set of survey items and a large set of survey items were considered. In a study by Greenleaf (1992b) on extreme response tendencies, 16 items were used. Bolt and Newton (2011) considered two subscales from the PISA exam; the “Enjoy Science” subscale contained 5 items and the “Value of Science” consisted of 10 items. Admittedly, the authors found that combining the two scales to 15 items led to better estimation of the extreme response trait. Jin and Wang (2014) investigated a scale measuring interpersonal conflicts by Lo (2001) that had 20 items. Considering previous studies, this study simulated two survey lengths measuring a single trait: short [10 items] and long [20 items] subscales.

3.2.1.2 Number of Response Options

At the item-level, the number of response options was varied. Research has found that as the number of response options increases, the better the estimation of the ERS tendency is (Jin & Wang, 2014). When developing response options for a Likert-style item, a researcher must determine whether or not to include a neutral option. Bishop (1987) offered a neutral response to avoid false responses. This enabled individuals who were indifferent about a subject to select no opinion instead of being forced to take a side that did not reflect their true beliefs (Edwards & Smith, 2016; Johns, 2005; Krosnick et al., 2002). Conversely, studies have shown a significant number in individuals selecting “no opinion” or “neutral” option when they truly do have an opinion (Bishop, 1987; Edwards & Smith, 2016; Johns, 2005; Kalton, Roberts, & Holt, 1980; Krosnick et al., 2002). To eliminate any effects of neutrality, item responses were simulated for items that contained no middle or neutral option.

To compare different size response scales, 4 and 6 response option items were analyzed. The four-option rating-scale was considered on the scale 1-*Strongly Disagree*, 2-*Disagree*, 3-*Agree*, and 4-*Strongly Agree*. The six-option rating-scale was considered as 1-*Strongly Disagree*, 2-*Disagree*, 3-*Somewhat Disagree*, 4-*Somewhat Agree*, 5-*Agree*, and 6-*Strongly Agree*. Categories associated with *Strong* opinions were classified as extreme.

3.2.1.3 Sample Size

Jin and Wang (2014) considered several different sample sizes in their simulation study and found that small sample size did not have adequate recovery of item parameters under the MPCM. As this study is designed to determine which model is most appropriate, it is important that small sample size be considered. In other words, even though the MPCM was found to be inadequate for small sample sizes in previous studies, it is important to note if either the IRTree Model or the MNRM are adequate under that condition. To investigate sample size, small [500] and large [1,000] sample sizes were simulated.

3.2.1.4 Data Generation Model

To evaluate ERS model recovery, item responses exhibiting an extreme response trait were simulated. To simulate extreme response tendencies, data was generated under each model: the multidimensional nominal response model, the modified partial credit model, and the IRTree Model. In the simulation study the data generation method was treated as an independent variable.

3.2.2 Research Questions

The following research questions were developed to assess the effects of the simulation factors previously described.

> RQ (4) Is the pattern of differences in the item mean square error between the expected total score and true total score among the levels of sample size significantly different among the models of estimation?

RQ (5) Is the pattern of differences in the item mean square error between the expected total score and true total score among the levels of survey length significantly different among the models of estimation?

RQ (6) Is the pattern of differences in the item mean square error between the expected total score and true total score among the levels of number of category response options significantly different among the models of estimation?

3.2.3 Data Generation

The MNRM, the IRTree Model, and the MPCM were used to simulate item responses from individuals who exhibited different degrees of a latent trait and different extreme response tendencies. Table 7 displays the data generation distributions and the fixed values for the item parameters for the MNRM, the IRTree Model, and the MPCM. In order to mirror the study by Bolt and Newton (2011) using the MNRM, the slope parameters associated with the trait of interest were fixed to equally spaced integers. The slope parameter values associated with the ERS trait were set to positive values for the extreme response options and negative values for the intermediate response options. The intercept parameters were randomly sampled from a normal distribution with mean 0 and variance 4.

The parameter values for the MPCM were generated to mimic those from the study by Jin and Wang (2014). The item slope parameters were generated from the lognormal distribution with mean 0 and variance .09. The item difficulties were generated from a uniform distribution ranging from -2 to 2. The threshold parameters were set to equally spaced fixed values.

With no prior simulation study relating to the IRTree Model, parameter values were generated to mirror the other models. The intercept parameters were generated from a uniform distribution ranging from -2 to 2. The slope parameters were created from a half-normal distribution with mean 0 and variance 4. The shift parameter, ν, was generated from a uniform distribution ranging from -1 to 1.

Table 7. Item parameter data generation distributions

Model	Item Parameter	Generation Distribution
MNRM	4 categories: $a_{j11}, \dots, a_{j41}$	Set: $-3, -1, 1, 3$
	6 categories: $a_{j11}, \dots, a_{j61}$	Set: $-5, -3, -1, 1, 3, 5$
	4 categories: $a_{j12}, \dots, a_{j42}$	Set: $1, -1, -1, 1$
	6 categories: $a_{j12}, \dots, a_{j62}$	Set: $1, -.25, -.25, -.25, -.25, 1$
	4 categories: c_{j1}, c_{j2}, c_{j3}	$\sim\mathcal{N}(0,4)$
	6 categories: $c_{j1}, c_{j2}, c_{j3}, c_{j4}, c_{j5}$	
	4 categories: c_{j4}	Set $= -\Sigma_{\mathrm{k}=1}^{3} c_{jk}$
	6 categories: c_{j6}	Set $= -\Sigma_{\mathrm{k}=1}^{5} c_{jk}$
IRTree Model	$a_1, \_a_2$	$\sim half - \mathcal{N}(0,4)$
	b_1, b_2, c_1, c_2	$\sim\mathcal{U}(-2,2)$
	ν	$\sim\mathcal{U}(-1,1)$
MPCM	a	$\sim log\mathcal{N}(0, .3^2)$
	b	$\sim\mathcal{U}(-2,2)$
	4- category: τ_1, τ_2	Set: $-0.6, 0, 0.6$
	6- category: $\tau_1, \tau_2, \tau_3, \tau_4$	Set: $-0.8, -0.4, 0, 0.4, 0.8$

To generate the trait values, distributions were selected to mimic past studies. Generation distributions are displayed in Table 8. For both the MNRM and the MPCM, the substantive traits were generated from a normal distribution with mean 0 and variance 1. The ERS trait for the MNRM was also generated from the normal distribution with mean 0 and variance 1. For the MPCM, the ERS trait, ω, can only take positive values. The ERS random effect trait, ω, was generated from a lognormal distribution with mean 0 and variance .16.

For the IRTree Model, the substantive trait parameter, θ_1, and extreme response tendency parameter, θ_{ERS}, were generated using a multivariate normal distribution with mean vector 0, identity variance matrix, and correlation ρ. The correlation, ρ, was fixed to 0 making the simulation of the trait paramerter and the extreme response tendency parameter equivalent to simulating two univariate normal distributions with mean 0 and variance 1. By definition, an individual's extreme response style is their tendency to select extreme options regardless of the content or trait measured. Therefore, for this study the definition is translated into a correlation of 0 for the IRTree Model generation method. This will also avoid confounding effects of favoring the IRTree Model during analysis. In other words, if the IRTree Model simulates a correlation between the traits and is the only method to account for a correlation between the traits, it may be favorited during the estimation phase.

Table 8. Individual parameter data generation distributions

Model	Person Parameter	Generation Distribution
MNRM	θ_1	$\sim\mathcal{N}(0,1)$
	θ_{ERS}	$\sim\mathcal{N}(0,1)$
IRTree Model	ρ	0
	$\begin{bmatrix}\theta_1\\ \theta_{ERS}\end{bmatrix}$	$\sim\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$
MPCM	θ	$\sim\mathcal{N}(0,1)$
	ω	$\sim log\mathcal{N}(0,.4^2)$

3.2.4 Model Estimation

The MNRM, the IRTree Model, and the MPCM models were estimated using Markov chain Monte Carlo (MCMC) methods and SAS PROC MCMC. To parallel the empirical section of this study, similar prior distributions and fixed values were used for estimation when 4-point Likert items were considered.

To estimate the multidimensional nominal response model with six response categories, the slope parameters associated with the substantive trait were fixed to $-5, -3, -1, 1, 3,$ and 5 for $k = 1, \dots, 6$, respectively. The slope parameter associated with the extreme response tendency trait were fixed at $1, -.25, -.25, -.25, -.25,$ and 1 for $k = 1, \dots, 6$, respectively. The intercept parameters for categories $k = 2, \dots, 6$ were estimated with a normal prior distribution with mean 0 and variance 25. For model identification purposes, the intercept parameter for the category *Strongly Disagree* was equal to the negative sum of the intercept parameters of the other categories.

To estimate the modified partial credit model, the discrimination parameter, $\mathrm{a_j}$, used a lognormal prior distribution with mean zero and variance 2. To estimate the location parameter, δ_j, the prior density was normal with mean 0 and variance 2. For the $K-$category items, the random category threshold parameters, τ_k (ex: τ_1 is between categories 1 and 2), for $k = 1, \dots, K - 2$ thresholds were estimated using a normal prior density with mean zero and variance 10. The threshold parameter (τ_{K-1}) between categories $K - 1$ and K was equal to the negative sum of $\tau_1, \dots \tau_{K-2}$.

To estimate the IRTree Model, the slope parameters, a_1 and a_2, used a half-normal prior with mean zero, variance 16, and lower threshold zero. The discrimination parameters, b_1, b_2, c_1, and c_2 were estimated using a normal prior with mean 0 and variance 16. The shift constant, ν, assumed a normal prior density with mean 0 and variance 16.

The trait distributions for all three models were the same as those used for the empirical study. The trait distributions for 4-category and 6-category items were equivalent as trait distributions are independent of the number of item response options used.

3.2.5 Evaluation Criteria

Each model is parameterized uniquely. Thus, recovery of total score using expected total score, $E\left(U_j \middle| \boldsymbol{\theta}\right) = \sum_{j=1}^{J} \sum_{k=1}^{m_j} k * P_{jk}(\theta)$, was compared. To compare recovery of total score, Item Mean Squared Error (IMSE) between the expected total score and the simulated (true) total score was calculated. The IMSE for a survey with *J* items and m_j response options for item *j* was calculated as

$$IMSE = \frac{\left(\sum_{j=1}^{J} U_j - \sum_{j=1}^{J} \sum_{k=1}^{m_j} k * P_{jk}(\theta)\right)^2}{J} \qquad (20)$$

The IMSE was treated as the dependent variable in the design with all simulation criteria as independent variables.

Mean plots were evaluated to address the research questions. To answer RQ (4), the mean plot displaying the interaction between sample size and analysis model averaged over survey length and number of response categories was examined. To address RQ (5), the mean plot of the interaction between the length of the survey and the model of analysis averaged over sample size and the number of response options was examined. To examine RQ (6), patterns in the mean plot of the interaction between number of response options and the model of analysis averaged over sample size and survey length were investigated.

3.3 SUMMARY

The methodology of this study seeks to answer whether there are differences between the MNRM, the MPCM, and the IRTree Model. Through an empirical study of the data collected on the PISA examination, the correlation between trait of interest estimates and the correlation between extreme response trait estimates determined whether there was a difference in rank ordering of individuals on each trait. The empirical study also investigated the MSE of the difference between observed total score and expected total score calculated under the three models.

The simulation study sought to address which factors affect the estimation and total score recovery under the three models. By investigating the mean plots of the interaction of the method used for analysis and sample size, survey length, and the number of response options, the simulation study helps inform researchers about model performance under various conditions.

4.0 RESULTS

4.1 EMPIRCIAL STUDY

Analysis was performed on 5,330 response sets to 10 items on the "Value of Science" Subscale from the Program for International Student Assessment (PISA, 2006). The 10 items each were presented with response options: 1-*Strongly Disagree*, 2-*Disagree*, 3-*Agree*, and 4-*Strongly Agree*. Based on the item extreme response rates and the individual extreme response rates, there existed evidence of individuals exhibiting extreme response tendencies. The responses were analyzed using the IRTree Model, the MNRM, and the MPCM.

4.1.1 Convergence

When estimating parameters in the Bayesian paradigm it is essential to evaluate convergence to the posterior distribution during the MCMC estimation routine. Trace plots, autocorrelation plots, and posterior density plots for each estimated item parameter under the IRTree Model, the MNRM, and the MPCM were examined to evaluate convergence. For example, consider the item parameters under the IRTree Model, the MNRM, and the MPCM for item 10. Figure 12 displays the trace plots, the autocorrelation plots, and the posterior density plots for the five estimated item parameters for item 10 under the IRTree model. Although the horizontal axis begins at iteration 0 for the trace plots, this iteration represents iteration 5000. Iteration 5000 is the last iteration used for the burn-in period. For each item parameter, the trace plots are randomly scattered. This random scatter indicates good mixing of the posterior distribution.

The autocorrelation plots for parameters a_1 and a_2 do not indicate any potential problem of dependence of sampled values. The autocorrelation plots for parameters $b_1, b_2,$ and ν, indicate a possible (but minimal) problem with expected dependence of sampled values requiring a higher number of iterations to represent the sample space. The density plot of the posterior distribution for each item parameter is unimodal and symmetric under the IRTree Model. This indicates that the mean is a meaningful value for the parameter estimates.

Figure **13** displays the trace plots, the autocorrelation plots, and the posterior density plots for the three estimated intercept parameters under the MNRM model. The fourth intercept parameter is a linear combination of the other three intercepts. The three trace plots indicate proper mixing across the posterior distribution. The values in the trace plots are traversing the parameter space in a random pattern. The autocorrelation plots for c_2 and c_3 converge rapidly to 0 indicating little to no autocorrelation. The autocorrelation plot for c_4 indicates a potential dependence of posterior samples at higher lags. Finally, the density plots of the posterior distributions for all item parameters are unimodal and symmetric. Thus, the mean is a useful estimate of the true parameter values.

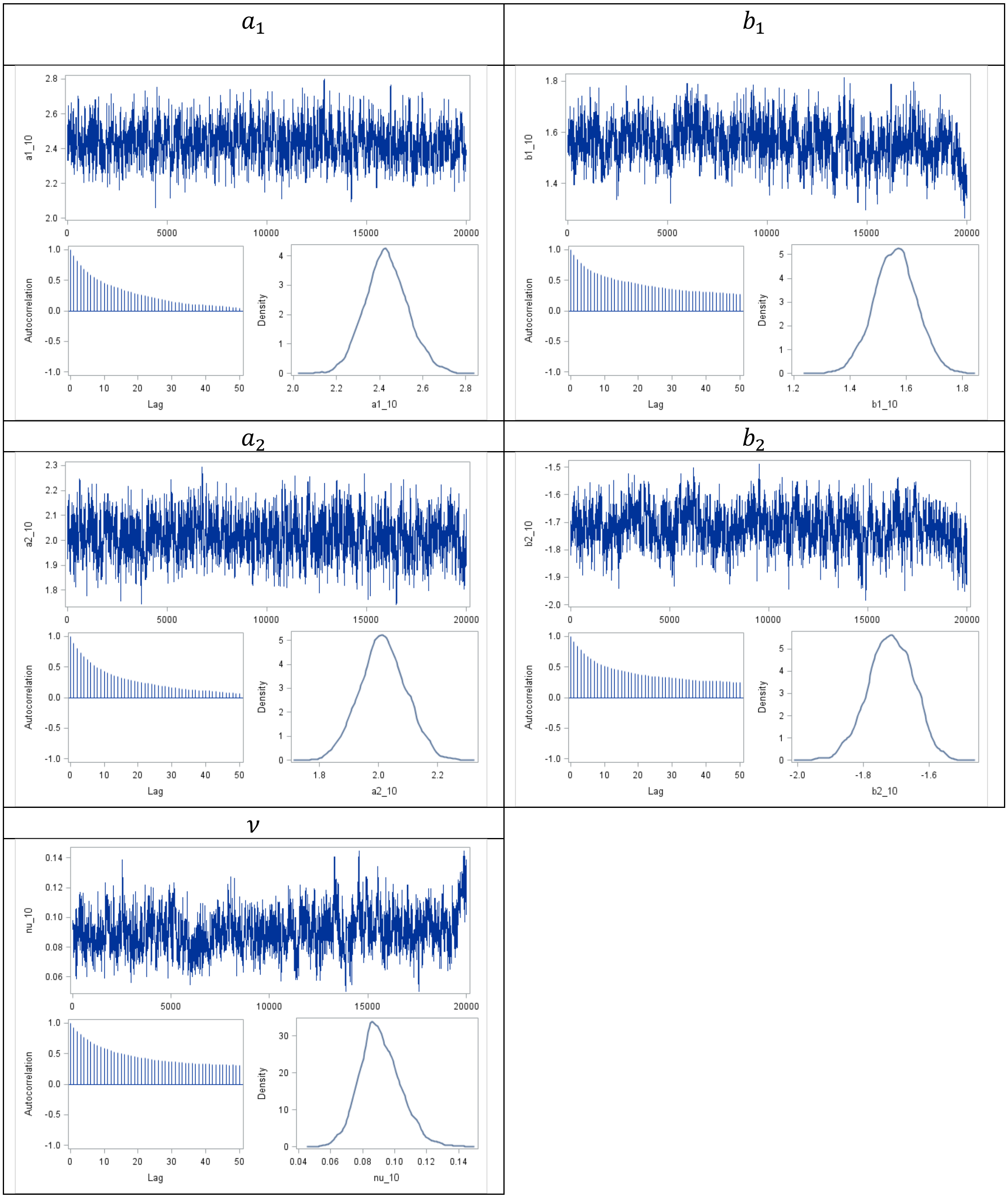
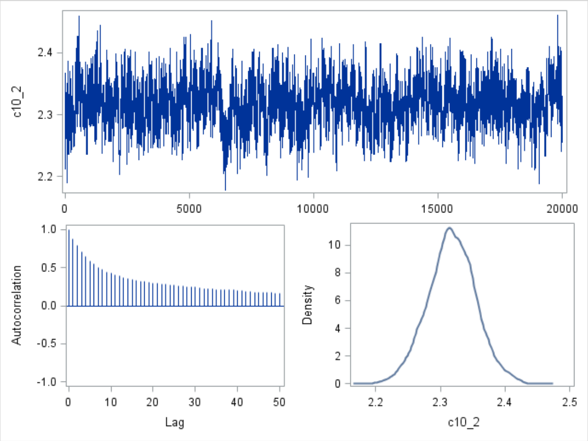


Figure 12. Trace plots, autocorrelation plots, and posterior density plots for item 10 parameters under the IRTree Model

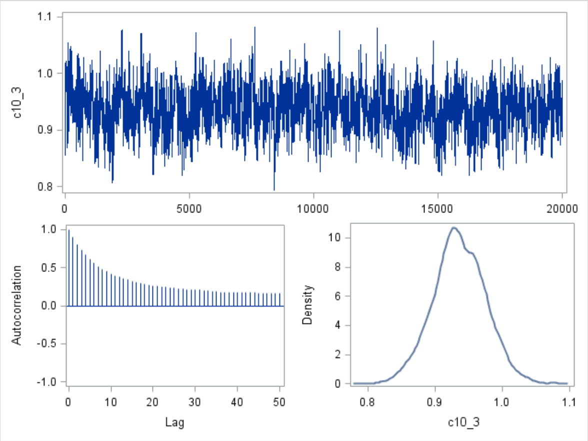
Figure 14 shows the trace plots, the autocorrelation plots, and the posterior density plots of estimated parameters $a, b, \tau_1,$ and τ_2 under the MPCM model. Because τ_3 is a linear combination of τ_1 and τ_2 a density plot, a trace plot, and an autocorrelation plot were not generated. For each parameter, the trace plot represents proper mixing and convergence of the posterior distribution. The density plot of the posterior distribution for each parameter is unimodal and symmetric. The autocorrelation plot for a and τ_2 decrease rapidly to 0 indicating little to no autocorrelation. The autocorrelation plots for b and τ_1 decrease towards 0 but not as swiftly as a and τ_2 indicating minor autocorrelation at higher lags.

Analysis of the trace plots, the autocorrelation plots, and the posterior density plots for each item under each model showed a similar pattern to those exhibited by item parameters for item 10. Trait parameters also displayed similar patterns when the trace plots, the autocorrelation plots, and the posterior density plots were investigated. Overall, the MCMC estimation routine converged for all estimated parameters.

c_2



c_3



c_4

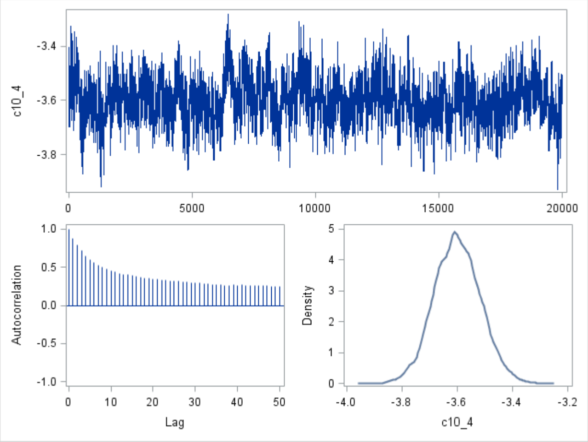


Figure 13. Trace plots, autocorrelation plots, and posterior density plots for item 10 parameters under the MNRM

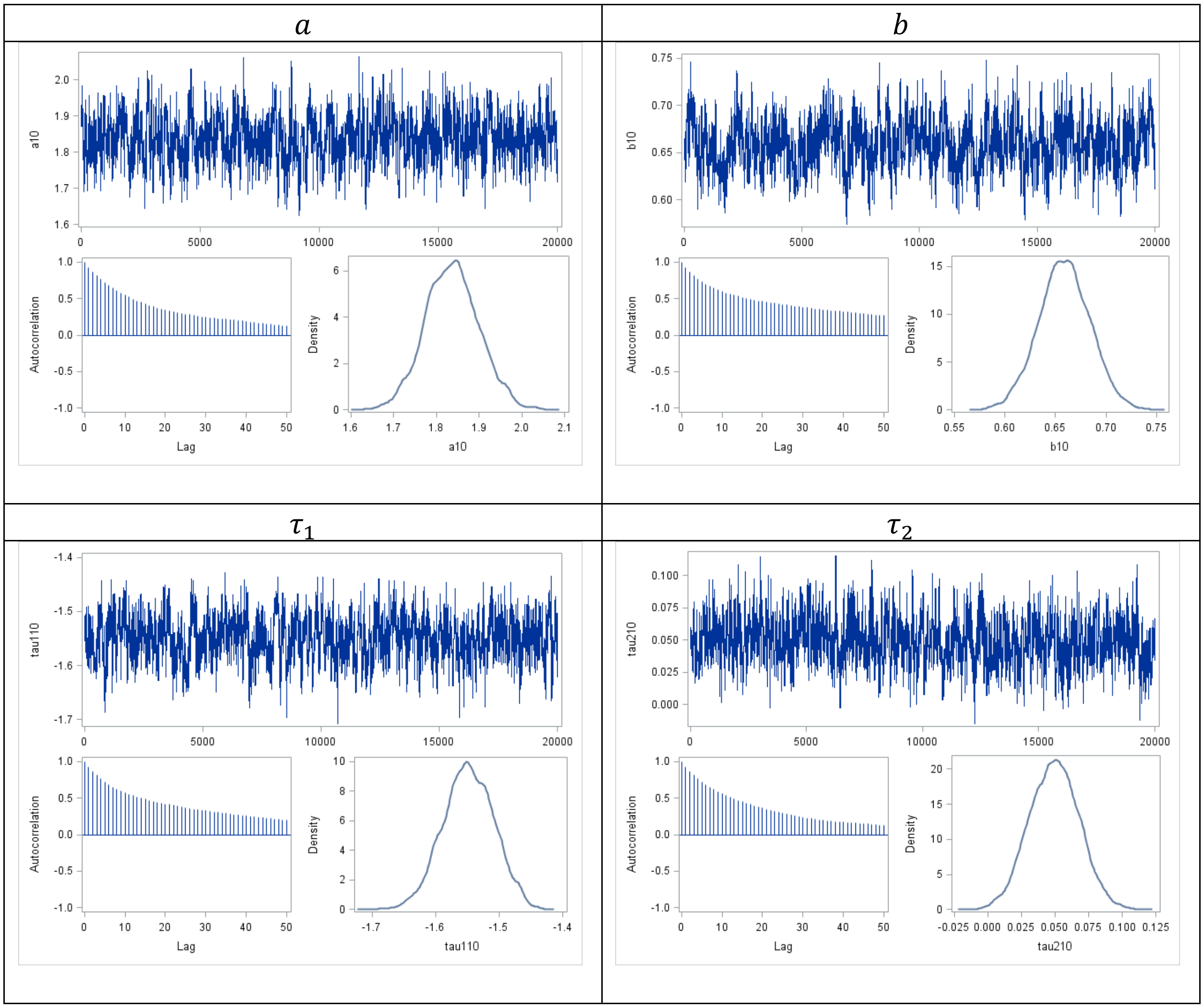


Figure 14. Trace plots, autocorrelation plots, and posterior density plots for item 10 parameters under the MPCM

4.1.2 Relationship of Substantive Trait Estimates

The relationship between substantive trait estimates across the three models was explored after estimation was completed and convergence was determined. Figure 15 displays the scatterplot matrix of substantive trait estimates among the three models. The scatterplot matrix indicates all substantive traits have a strong, positive, linear correlation. This is quantified with the Pearson correlations presented in Table 9. The rank-ordering of individuals on the respective substantive trait estimates are relatively the same (high correlations) across the three models.

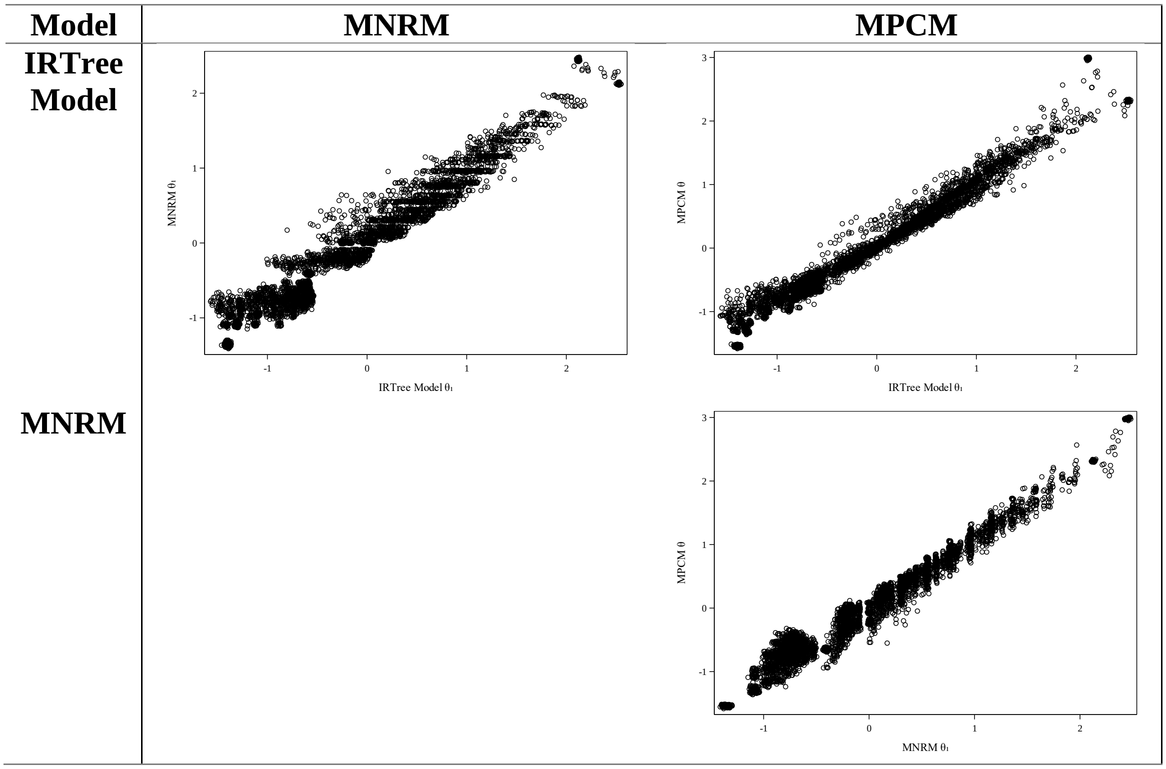


Figure 15. Scatterplots between the Substantive Trait Estimates from the IRTree Model, the MNRM, and the MPCM

Table 9. Correlations between the Substantive Trait Estimates from the IRTree Model, the MNRM, and the MPCM

Methods	IRTree Model	MNRM	MPCM
IRTree Model	1		
MNRM	0.97406	1	
MPCM	0.98419	0.98218	1

4.1.3 Relationship of Extreme Response Tendency Estimates

The relationship of extreme response tendency estimates was investigated. The extreme response trait estimates ($\hat{\omega}$) from the modified generalized partial credit model were transformed to be comparable with the extreme response estimates from the IRTree Model and the MNRM. The extreme response trait (ω) was assumed to follow a log-normal density. In other words, the trait follows a normal distribution after the log of the values is calculated. The Pearson correlation was found between $\log(\omega)$, θ_{ERS} from the MNRM, and θ_{ERS} from the IRTree Model. The correlations (after log transformation) are displayed in Table 10 and the scatterplot matrix (after log transformation) is shown in Figure 16.

Table 10. Correlations between the Extreme Response Estimates from the IRTree Model, the MNRM, and the MPCM

Methods	**IRTree Model**	**MNRM**	**Log (MPCM)**
IRTree Model	1		
MNRM	0.82502	1	
Log (MPCM)	-0.83824	-0.97578	1

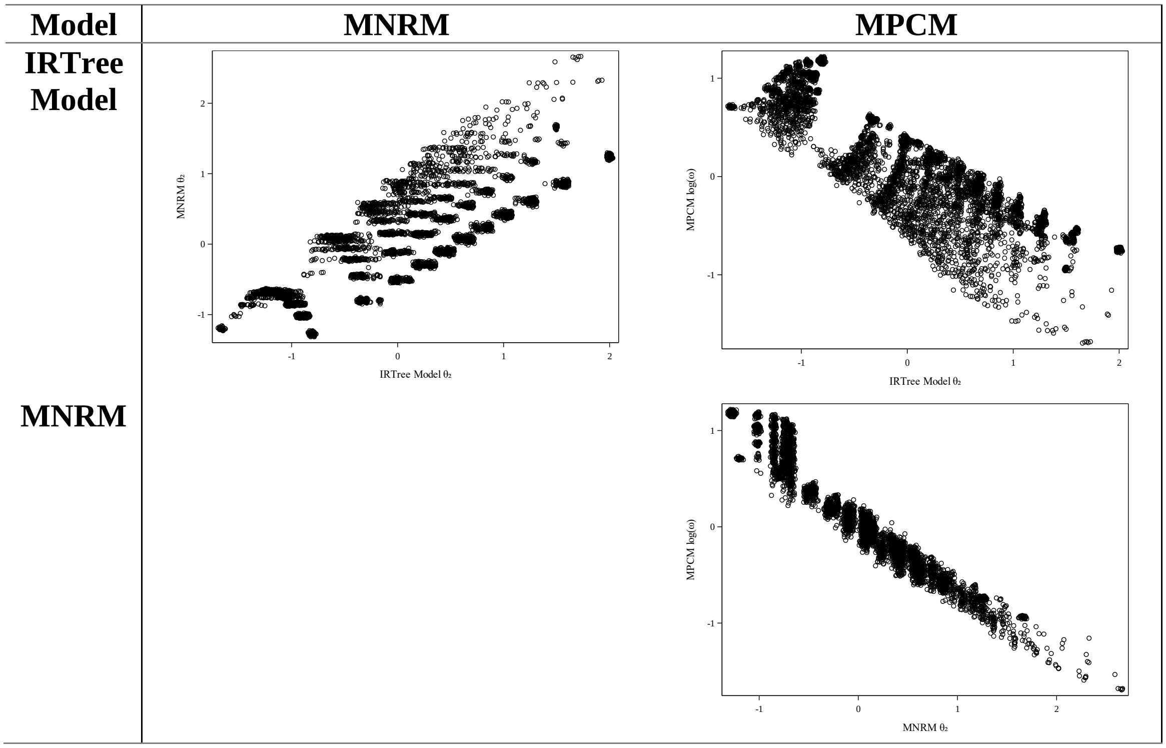


Figure 16. Scatterplots between the Extreme Response Estimates from the IRTree Model, the MNRM, and the MPCM

There is a moderately strong, positive, linear relationship between the extreme tendency trait estimates from the IRTree Model and the extreme tendency trait estimates from the MNRM. Similarly, there is a moderately strong, negative, linear relationship between the IRTree Model extreme tendency trait estimates and the transformed extreme response estimates from the MPCM. There is a strong, negative, linear relationship between the extreme response tendency trait estimates from the MNRM and the transformed extreme response tendency estimates from the MPCM. The two negative correlations exist due to the definition of the extreme tendencies effect under the MPCM. As the extreme tendencies effect estimate increases, an individual has a lower tendency to select the extreme categories. Extreme response traits are defined in the competing direction under the MNRM and the IRTree Model.

Weaker correlations exist when the IRTree Model is paired with the MPCM and the MNRM compared to when the MNRM is paired with the MPCM. This may be due to the definition of the trait. The trait is being treated as extreme response tendencies, however, the IRTree Model is modeling a response process. The response process may be different than an individual's extreme response tendencies. Furthermore, of the three models, the IRTree Model is the only model that assumes the two traits are compensatory. The shift term, $\nu a_2(b_1 + a_1\theta_1)$, in the model introduces a compensatory nature between the substantive trait and the extreme response tendencies trait. This property is unique to the IRTree Model compared to the MNRM and the MPCM. The correlation of the ERS trait estimates is still strong among the IRTree Model with the MPCM and the MNRN but not as strong as the correlation of the ERS trait estimates between the MNRM and the MPCM. The correlation between the ERS trait estimates, however, is not as strong as the correlations between the substantive trait estimates among the IRTree Model, the MPCM, and the MNRM.

4.1.4 Posterior Predictive Model Checking

Posterior Predictive Model Checking (PPMC) techniques were used to evaluate model fit for the IRTree Model, the MNRM, and the MPCM. Recall, PPMC compares an observed discrepancy measure to the distribution of the same discrepancy measure computed in replicated data under the null of model fit. The first method examined the global odds ratio. The second method examined test-level fit by analyzing the distribution of total scores. The third method explored item-level fit by investigating the proportion of extreme responders for each item. The fourth, and final method, analyzed person-level fit through exploration of the frequency of individuals who selected extreme response options.

4.1.4.1 Global Odds Ratio

The first discrepancy measure used to evaluate model fit was the global odds ratio (OR). The global OR discrepancy measure was used to evaluate local item dependence. The odds ratio has been used to evaluate local dependence for dichotomous items (Chen & Thissen, 1997). Item responses were grouped to extend this method for use with polytomous items. Once item responses are group into a dichotomy, the resulting odds ratio is known as a global odds ratio.

Items responses for the 10 items on the Value of Science subscale were 1-*Strongly Disagree,* 2-*Disagree,* 3-*Agree,* and 4-*Strongly Agree*. Without grouping item responses, there exists 45 4x4 item pairwise contingency tables. When using PPMC methods, it is appropriate to evaluate discrepancy measures that are relevant to both the attributes the method is modelling and the resulting inferences. The IRTree model, the MNRM, and the MPCM were developed to account for extreme response tendencies. Thus, the odds ratio was modified to account for extreme response style. Specifically, the 45 4x4 contingency tables were adapted to account for ERS to calculate the global OR. To do this, the item responses were dichotomized as 1-*Strongly Disagree* with 4-*Strongly Agree* and 2-*Disagree* with 3-*Agree*. In other words, the dichotomy became "extreme responses" vs. "non-extreme responses".

For each model, the 45-pairwise global ORs were calculated under the 2000 replications of simulated data from the posterior distribution. The global OR was also calculated using the observed data. The *PPP*-value for each item pairing was calculated by counting the number of global ORs under the simulated data that were greater than the global OR in observed data. The *PPP*-values are visualized using pie plots. Pie Plots for item pairs that are entirely black represent

PPP-values of 1. A *PPP*-value of 1 occurs when the observed global OR is less than the global ORs computed in each replicated dataset. Pie plots for item pairs that are entirely white represent *PPP*-values of 0. A *PPP*-value of 0 occurs when the observed global OR is greater than the OR computed in each replicated dataset. Pie plots that are equal parts black and white result from *PPP*-values near 0.5. These pie plots represent adequate model fit.

PPP-value pie plots for all item pairings calculated under the IRTree model, the MNRM, the MPCM, and the GR model are displayed in Figure 17. None of the three models of interest have adequate model fit for all item pairings when the global OR is used to investigate pairwise item dependence. The global odd ratios for each item pairing from the simulated data under the GR model consistently underestimated the global odd ratios calculated from the observed data. This can be seen by almost every item pairing resulting in a pie plot being entirely white. The GR model was expected to have inadequate fit since the individuals analyzed exhibited extreme response tendencies. Thus, the GR should not be analyzed for absolute fit. The GR model was instead used as a baseline for comparison. The results from the IRTree Model, the MNRM, and the MPCM were evidence of better model fit compared to the pie plots generated under the GR model. Although this fit is still not adequate for the three models of interest, there was improvement in model fit when compared the GR model.

The results from the global OR are sensitive to the selection of the dichotomy. For items with the response scale 1-*Strongly Disagree,* 2-*Disagree,* 3- *Agree,* and 4- *Strongly Agree*, it can be argued that the response options should be dichotomized by pairing *1* with *2* and *3* with *4*. This would create two arbitrary categories of those who agree to any degree versus those who disagree to any degree. Results for the three primary models of interest using the “Agree” versus “Disagree” groups showed similar *PPP*-values to those obtained using the “extreme” versus “non-extreme" dichotomy. The GR model showed improvement under this dichotomy but was still inadequate. This result was expected because the goal of the GR model is to capture order responses closer aligned to "Agree" and "Disagree" than "extreme" and "non-extreme".

The inadequacy of model fit under each dichotomy for the IRTree Model, the MPCM, and the MNRM was not unexpected. Critical information was lost in the responses when the categories were collapsed. Specifically, the extreme tendencies of an individual who would select *Strongly Agree* over *Agree* or an individual who would select *Strongly Disagree* over *Disagree* were lost. In other words, the models are trying to determine a trait that would push an individual from *Agree* to *Strongly Agree.* This information is lost in both dichotomization methods described.

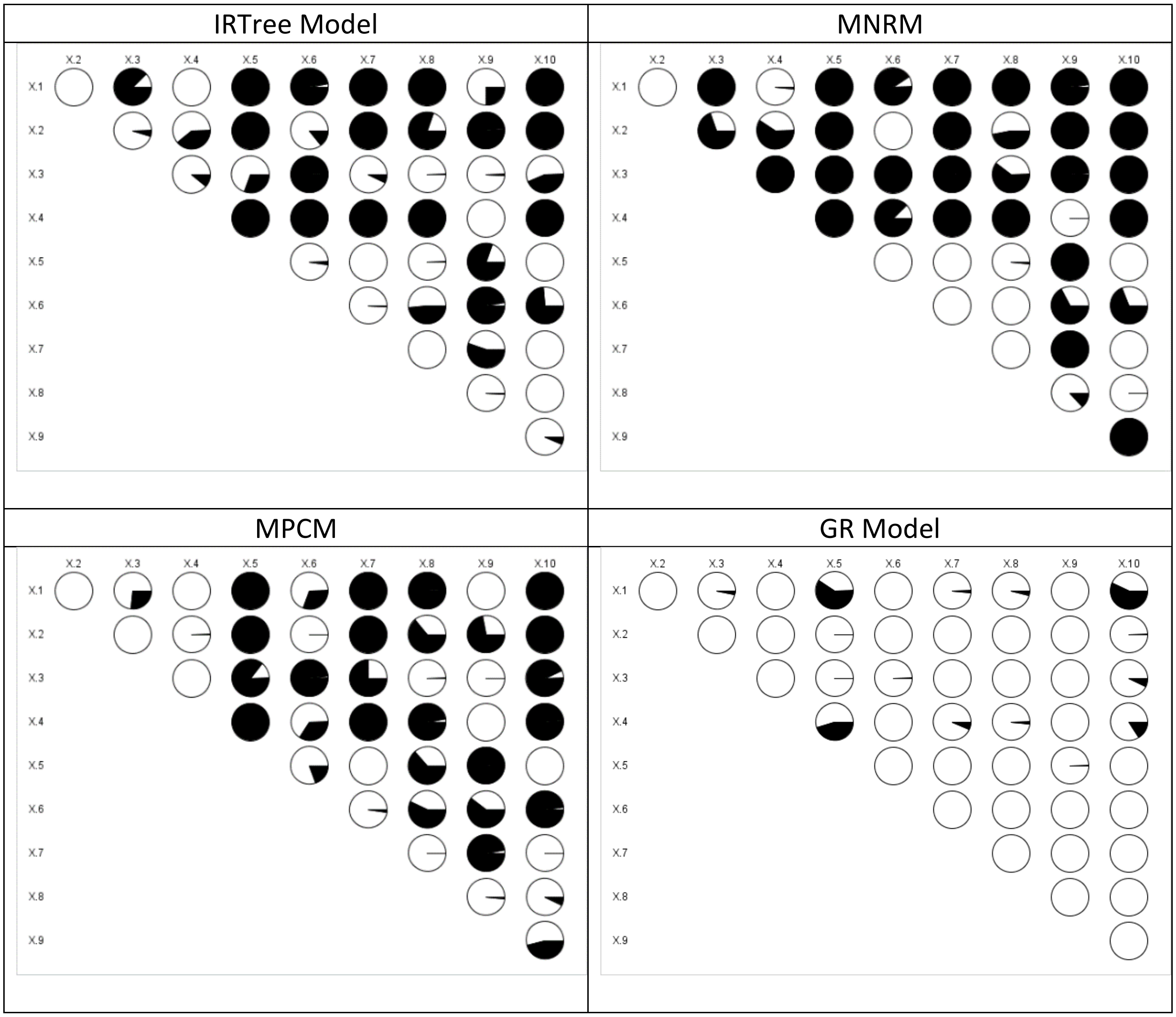


Figure 17. Global odds ratios of "extreme" vs. "non-extreme" groupings under the IRTree Model, the MNRM, and the MPCM

4.1.4.2 Total Score Distribution

The IRTree Model, the MNRM, and the MPCM were analyzed for test-level model fit using discrepancy measures related to the total score. The total score may capture the pattern of extreme response tendencies. Individuals with a high presence of the substantive trait and a high tendency to select extreme responses are likely to select 4-*Strongly Agree*. These individuals will tend to have extremely high total scores. Individuals with a low presence of the substantive trait and a high tendency to select extreme responses are likely to select 1-*Strongly Disagree.* These individuals will tend to have very low total scores. Respondents with moderate degrees of extreme response tendencies will have total scores near the middle of the total score distribution. Overall, a model that captures the patterns of the total score will likely capture some properties of extreme response tendencies.

Comparison of the observed and model-predicted total test score distributions is a test-level method of analyzing model fit. The mean and the standard deviation of the total score distribution are discrepancy measures used to analyze test-level fit. The *PPP*-value for each is the proportion of instances in which the statistic (mean or standard deviation) is greater in the replicated datasets than the value of the statistic in the observed total score distribution.

Histograms are used to visualize the replicated total score distribution means and standard deviations. The mean histogram plots the frequency of total score means from the 2000 replicated datasets. A vertical line representing the mean of the observed total score distribution is overlaid on the histogram. The standard deviation histogram plots the standard deviation of total scores from the 2000 replicated datasets. The vertical line overlaid on the histogram signifies the standard deviation of the observed total score distribution.

Figure 18 displays the total score mean and total score standard deviation histograms produced using replicated data under the IRTree Model. The means of the predicted (replicated) total scores are consistently greater than the observed mean total score. This pattern is quantified with a *PPP*-value equal to 1. In other words, 100% of the predicted total score distributions had a mean greater than the mean of observed total scores. This is an indication that the data do not adequately fit the IRTree Model. The standard deviation of observed total scores is located on the upper tail of the distribution of the predicted total score standard deviations. This is enumerated by a *PPP*-value of 0.01. 99% of predicted total score distributions had a standard deviation less than the observed total score standard deviation. This indicates the data does not adequately fit the IRTree Model. Overall, based on these discrepancy measures, the IRTree Model overestimated the mean total score while underestimating the spread.

Figure 19 displays the histogram of predicted mean total scores and the histogram of predicted total score standard deviations computed under the MNRM. The observed mean total score is located at the center of the histogram of the predicted mean total score distribution. This evidence of test-level adequate fit is quantified by the *PPP*-value of 0.49. Approximately half of predicted mean total scores were above the observed total score mean. 100% of the predicted total score distributions had a standard deviation higher than the standard deviation of the observed total scores. The MNRM properly captured the pattern of the total score mean but overestimated the standard deviation of the total scores.

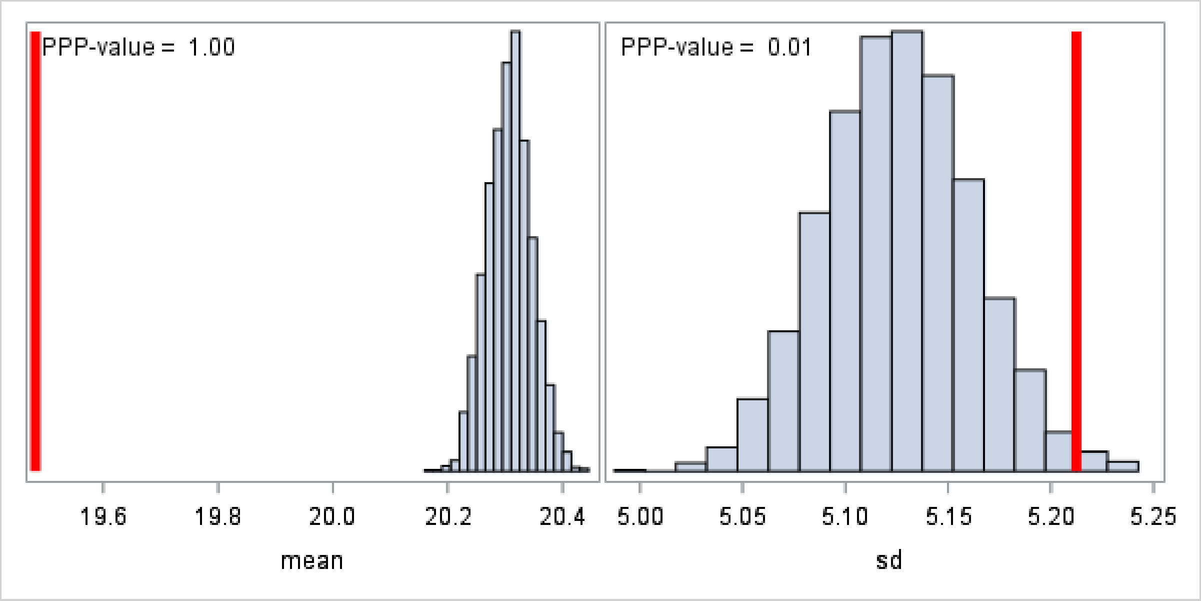


Figure 18. Predicted mean total score distribution and predicted standard deviation of total score distribution under the IRTree Model

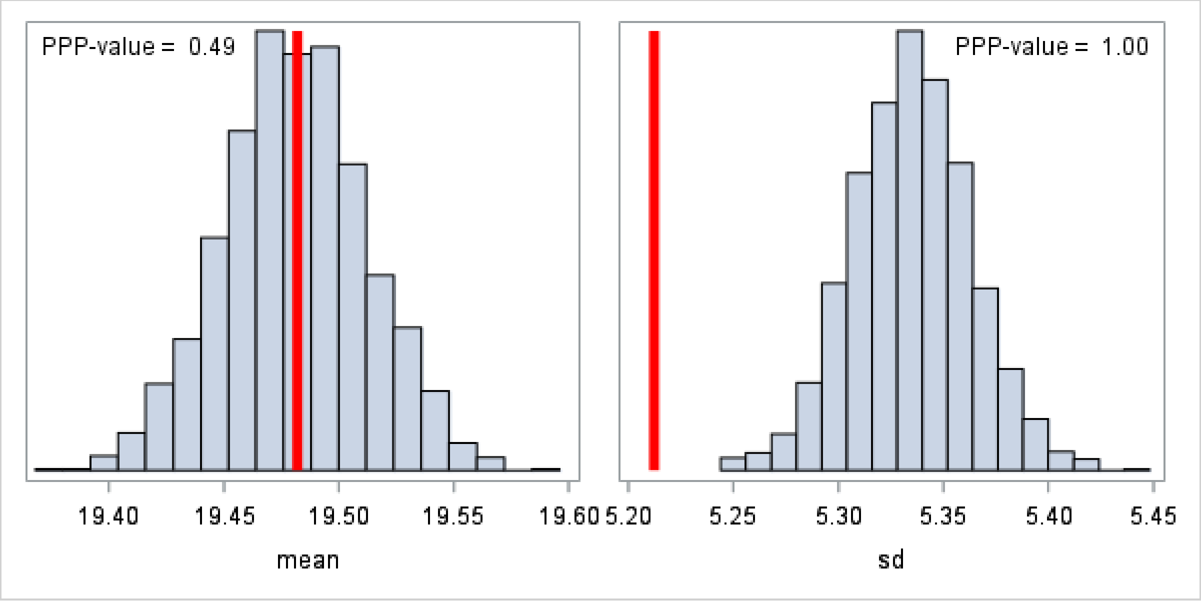


Figure 19. Predicted mean total score distribution and predicted standard deviation of total score distribution under the MNRM

Figure 20 displays the distribution of total score means from the predicted responses and the distribution of total score standard deviations from the predicted responses under the MPCM. The observed total score mean is positioned in the center of the predicted total score distribution means. A *PPP*-value equal to 0.51 is evidence of adequate test-level model fit. The standard deviation of the observed total scores is neighboring the center of the distribution of predicted total score standard deviations. 61% of predicted total score distributions had a standard deviation greater than the observed total score standard deviation. The MPCM provided adequate fit related to the mean total score and the standard deviation of the total scores.

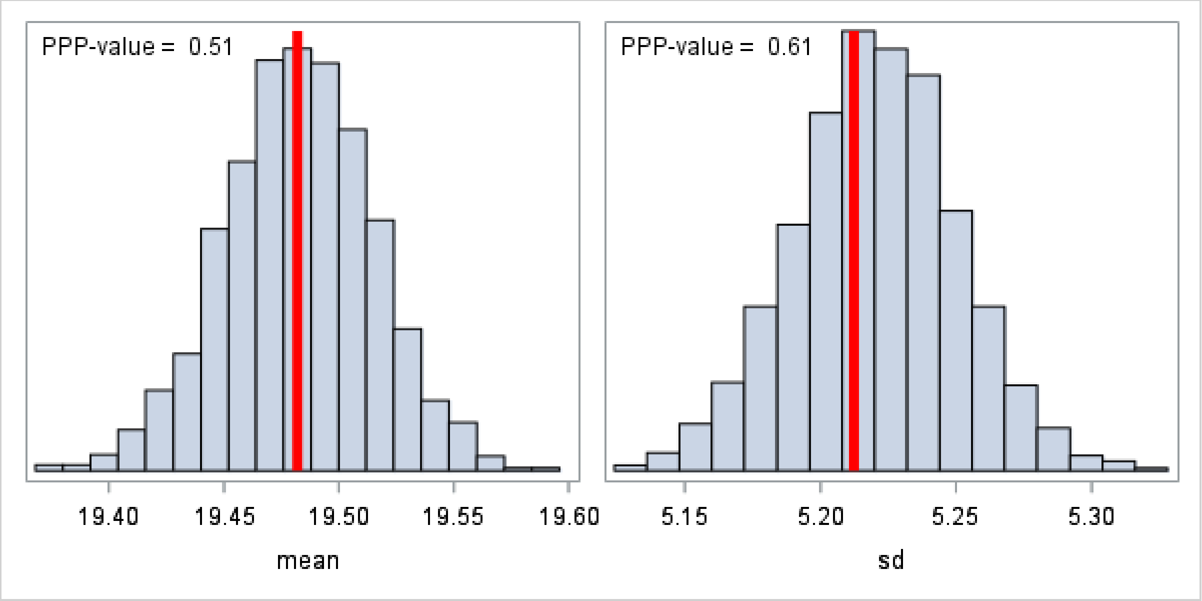


Figure 20. Predicted mean total score distribution and predicted standard deviation of total score distribution under the MPCM

A histogram of predicted total score means and a histogram of predicted total score standard deviations were developed under the GR model for baseline comparison. Figure 21 displays the two total score statistic distributions developed using predicted datasets under the GR model. The GR models captured the mean of the total score distribution (*PPP*-value = 0.37) but not the standard deviation (*PPP*-value = 0.99). The *PPP*-value for mean total score calculated under the MNRM and the *PPP*-value for mean total score calculated using the MPCM were closer to 0.50 than the *PPP*-value for mean total score calculated under the GR model. The GR model captured observed total score mean and observed total score standard deviation more effectively than the IRTree Model. The MPCM captured observed total score standard deviation better than the GR model.

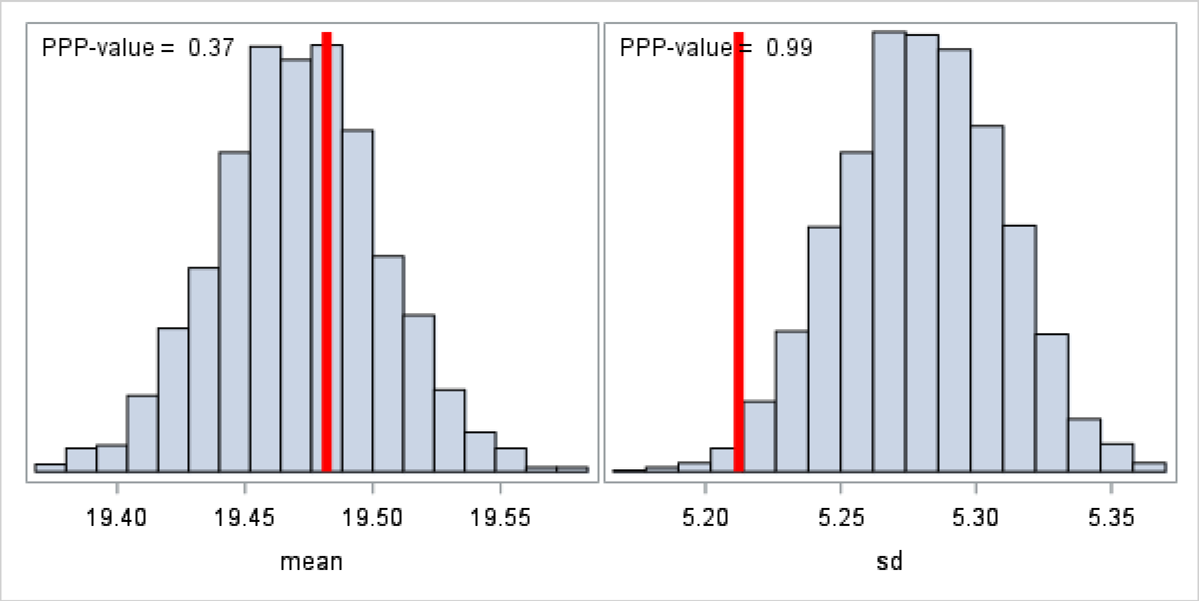


Figure 21. Predicted mean total score distribution and predicted standard deviation of total score distribution under the GR Model

Overall, the MNRM and the MPCM provided adequate test-level model fit when the mean total score is used as a discrepancy measure. When the standard deviation of total scores was used as a discrepancy measure only the MNRM provided adequate model fit. The IRTree model did not provide any advantage of model fit over the unidimensional GR model when the mean total score and the standard deviation of total scores were considered as discrepancy measures.

4.1.4.3 Observed Total Score Groups

The total score mean and the total score standard deviation are useful discrepancy measures for capturing observed mean total score and observed total score standard deviation. The mean and standard deviations, however, are two point estimates that attempt to quantify the entirety of the total score distribution. An investigation of the total score continuum was done as an alternative to analysis involving just the mean total score and the standard deviation of total scores. The total score continuum can be investigated by exploring the frequency of individuals with each observed total score. The frequency of individuals with each observed total score can be compared to the frequency predicted to be at each total score in the replicated datasets. For the Value of Science subscale there were 31 unique total scores (10 to 40) present. Therefore, it was not feasible to investigate all the unique scores. Instead, similar scores were grouped.

The choice of groups was made to group similar patterns of extreme response tendencies. Individuals were classified as being extreme response averse, having low/moderate extreme response tendencies, or having a high rate of selecting extreme response options. Individuals with a high rate of selecting extreme response options had high total scores (tended to select 4) and low total scores (tended to select 1). Individuals averse to selecting extreme response options fell near the middle of the total score continuum. Individuals who had a low/moderate rate of extreme response selections fell between the two previous categorizations.

Table 11 displays the five groups created. High total scores in the sample were sparse. As a result, group 5 had a larger range of scores than the other four groups. Group 1 included the 1,232 individuals with total scores between 10 and 15 (inclusive). Group 1 respondents averaged

7.5 extreme response options selected out of the 10 possible responses. A parallel group was formed on the high end of the total score continuum. The 104 respondents in group 5 obtained total scores ranging from 31 to 40 with an average of 7.1 extreme response options selected. The low/moderate extreme response groups included individuals with total scores between 16 and 20 (group 2) and individuals with total scores between 26 and 30 (group 4). The 1,931 individuals who had scores between 16 and 20 selected an average of 2.4 extreme response options. The 436 respondents with total scores between 26 and 30 selected an average of 1.4 extreme response options. The remaining 1,627 respondents (group 3) had total scores ranging from 21 to 25 with less than 1 extreme response option selected on average.

Table 11. Total score group frequency and extreme response average for PISA responses

Group	Total Score	Frequency	Average Number of Extreme Responses
1	[10,15]	1232	7.5
2	[16,20]	1931	2.4
3	[21,25]	1627	0.9
4	[26,30]	436	1.4
5	[31,40]	104	7.1

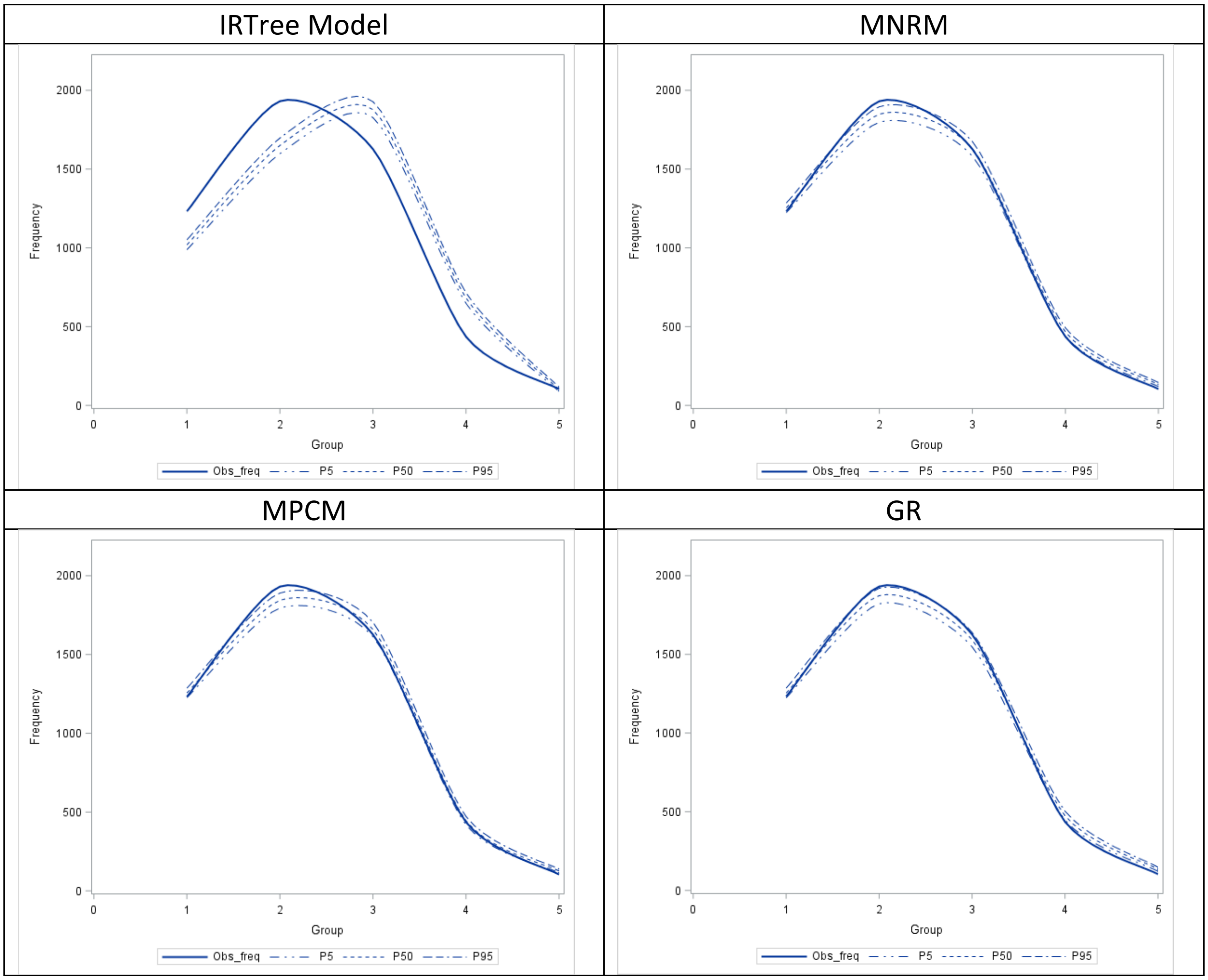
Figure 22 plots the frequencies for the observed total score groups and the 5^{th}, 50^{th}**,** and 95^{th} percentiles for frequencies in each total score group from the replicated data under the IRTree Model, the MNRM, the MPCM, and the GR Model.   The observed frequency (solid line) for the IRTree Model does not fall within the 5th and 95th percentiles for any group except group 5. This indicates that the data did not adequately fit the IRTree Model at the test-level. This is further evidenced by the *PPP*-values for groups 1, 2, 3, and 4 being outside the credible interval (0.05 – 0.95) seen in Table 12.

The resulting plot developed using the MNRM showed a similar pattern to the plot developed using the MPCM. For each, the observed total score group frequency fell within or close to the 90% credible interval (between 5th and 95th percentiles) of the total test score group frequency distribution. Based on the results of the MNRM, the observed frequency fell within the credible interval for group 1 (*PPP*-value=0.89) but not for group 5 (*PPP*-value = 1). Similarly, the results for the MPCM show the observed frequency within the credible interval for group 1 (*PPP*-value = 0.91) but not group 5 (*PPP*-value = 1).

Comparatively, for the GR model (bottom right of Figure 22) the observed data did not fall within the credible interval for groups 2, 4, or 5. The observed data fell just within the credible interval for group 1 and group 3. The GR model, the MNRM, and the MPCM had similar total score group frequency test-level fit. The GR model, however, has better total score group frequency test-level fit compared to the IRTree Model.

Table 12. Total score group frequency *PPP*-values.

Group	IRTree Model	MNRM	MPCM	GR Model
1	0	0.89	0.91	0.88
2	0	0	0	0.02
3	1	0.55	0.87	0.11
4	1	0.97	0.71	0.99
5	0.46	1	1	1



IRTree Model

MNRM

MPCM

GR

Figure 22. Predicted total score group frequency percentiles under the IRTree Model, the MNRM, the MPCM, and the GR Model

4.1.4.4 Item-Level Discrepancy Measure

A new discrepancy measure is presented to investigate the item-level fit of the IRTree Model, the MNRM, and the MPCM. Greenleaf (1992b) posited an item-level descriptive measure of extreme response tendencies. The proportion of extreme responses selected by the individuals was computed for each item. For each item, the number of individuals who selected either *1* or *4* is divided by the total number of respondents. The resulting proportion is the extreme response rate for a given item. The item extreme response rates were calculated for each replication of predicted values. The *PPP*-value for each item is the proportion of item extreme response rates greater than the observed item extreme response rate. The observed item extreme response rates are shown in Table 6.

The item extreme response rate *PPP*-values calculated under the IRTree Model, the MNRM, the MPCM, and the GR Model are displayed in Table 13. Consider item 1 using the IRTree Model with a *PPP*-value equal to .2875. This indicates that 28.75% of the predicted sample datasets resulted in more than 42% (Table 6) of extreme responses on item 1. The IRTree Model resulted in three item-level *PPP*-values >0.95 or <0.05 (inadequate item-level fit). The MNRM and the MPCM did not have any *PPP*-values indicating inadequate item-level fit. The item-level *PPP*-values calculated under the MNRM were closer to 0.5 compared to the corresponding *PPP*-values calculated under the MPCM. This indicates that the data had better item-level fit under the MNRM.

Zero item extreme response rate *PPP*-values indicated inadequate fit for the GR model. In other words, the observed data, with relation to the proportion of extreme responses on each item, fit the GR model adequately. Overall, *PPP*-values for the MNRM were closer to 0.5 compared to those calculated under the GR model. This indicates the observed data fit the MNRM better than the GR model. The *PPP*-values for the MPCM model exhibited no consistent pattern when compared to the *PPP*-values for the GR model. The MNRM and the MPCM showed adequate item-level model fit while the IRTree Model did not. Furthermore, the data fit the IRTree Model worse than the data fit the GR model at the item-level.

Table 13. Item extreme response rate *PPP*-values under the IRTree Model, the MNRM, the MPCM, and the GR Model. (Note: * indicates *PPP*-value outside .05 - .95 range)

Item	IRTree Model	MNRM	MPCM	GR Model
1	0.2875	0.474	0.26	0.6145
2	0.4455	0.511	0.4445	0.541
3	0.658	0.51	0.478	0.8885
4	0.185	0.532	0.305	0.7805
5	0.962*	0.491	0.6455	0.5075
6	0.476	0.502	0.6235	0.4575
7	0.9935*	0.5075	0.345	0.417
8	0.8345	0.4865	0.422	0.4985
9	0.5855	0.48	0.1925	0.7675
10	0.9755*	0.4905	0.4645	0.658

4.1.4.5 Person-Level Discrepancy Measure

Greenleaf (1992b) also proposed a person-level descriptive statistic for extreme response tendencies. An individual's extreme response rate is calculated by dividing the number of items the individual selected *Strongly Disagree* or *Strongly Agree* by the total number of items on the subscale. The frequency of individuals with each proportion of extreme response selections is then computed. This is only feasible for surveys consisting of a small number of items. The Value of Science subscale (10 items) contains 11 possible extreme response rates. Extreme response rates in the observed data ranged from 0 to 1 at increments of 0.1.

The *PPP*-values calculated using the individual extreme response rate discrepancy measure are displayed in Table 14. For example, consider an extreme response rate of 0 under the MPCM with a *PPP*-value equal to 0.1980. In 19.80% of the predicted datasets there existed a higher proportion of respondents who selected 0 (out of 10) extreme response options than the 29.12%, the observed frequency (Table 5) that selected 0 extreme response options.

Of the 11 *PPP*-values calculated under the IRTree Model, 8 indicated inadequate person-level fit (>0.95 or <0.05). Two of the remaining three were borderline to the cut-offs of inadequate fit (0.086 and 0.935). Five person-level *PPP*-values from the MNRM indicated inadequate fit. Two (0.086 and 0.909) of the remaining six person-level *PPP*-values from the MNRM were borderline to the credible interval. Eight person-level *PPP*-values under the MPCM indicated inadequate model fit.

Overall, the data conformed to the person-level extreme response properties of the IRTree Model the least. The data fit the person-level extreme response rates of the MPCM and the MNRM better than the IRTree Model. However, there was still mixed results. There is not a substantial difference in person-level fit between the MNRM and the MPCM.

The results of person-level fit for the three models accounting for extreme response style were compared to the results of person-level fit using the GR model. The GR model resulted in 7 out of 11 individual extreme response rate *PPP*-values that indicated inadequate fit. The observed data fit the person-level extreme response rates of the GR worse than the MNRM. The observed data fit the individual extreme response rates of the GR model better than the IRTree Model. For individuals with high extreme response rates (i.e. 70%, 80%, 90%, and 100%), the observed data fit the person-level characteristics of the MNRM the best. The MNRM had three of four *PPP*-values within the credible interval, the IRTree Model only had 1 value in the credible interval, the MPCM only had 1 value in the credible interval, and the GR model had 2 values in the credible interval.

Table 14. Individual extreme response rate *PPP*-values under the IRTree Model, the MNRM, the MPCM, and the GR Model.

Proportion	IRTree Model	MNRM	MPCM	GR
0.0	0.000*	0.006*	0.198	0000*
0.1	1.000*	1.000*	1.000*	1.000*
0.2	0.086	0.086	0.023*	1.000*
0.3	0.000*	0.000*	0.003*	0.9965*
0.4	0.334	0.109	0.737	0.7345
0.5	0.000*	0.000*	0.000*	0.000*
0.6	0.970*	0.746	0.978*	0.2295
0.7	0.993*	0.909	0.990*	0.5025
0.8	0.935	0.649	0.929	0.2415
0.9	1.000*	1.000*	1.000*	0.9985*
1.0	0.000*	0.117	0.000*	0.000*

4.1.5 Deviance Information Criterion

In addition to PPMC, model fit was investigated using more traditional Bayesian methods. The deviance information criteria (DIC) is the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity (Spiegelhalter et al., 2002). For model comparison purposes, the DIC can be compared for nested and non-nested models. Dbar, Dmean, pD, and the DIC are provided for each of the three models in Table 15. Dmean measures how well the “best” model describes the data. The “best” model is based on point estimates for the model parameters (Stone & Zhu, 2015). The Dbar value measures how well the model describes the data, on average. DIC takes into account both of these values along with model complexity (pD). Dbar, Dmean, and DIC were all greatest for the IRTree Model and least for the MPCM. This indicates that the data fit the MPCM best and the data fit the IRTree Model the poorest.

Dbar, Dmean, pD, and DIC were also computed for the GR model. Dbar, Dmean, and the DIC were all significantly higher for the GR Model when compared to the models accounting for extreme response style. In other words, the observed data fit the IRTree Model, the MNRM, and the MPCM better compared to the GR model.

Lunn, Jackson, Best, Thomas, and Spiegelhalter (2012) provide general guidelines for interpreting the DIC differences across models. Differences of more than 10 rule out the model with higher DIC; differences between 5 and 10 reflect "substantial" differences in favor of the model with the small DIC; and, it may be misleading to select a preferred model for differences that are less than 5 (Stone & Zhu, 2015). Thus, based on the results for the three models provided in Table 15, the MPCM is preferred over the MNRM and the IRTree Model. Additionally, the MNRM is preferred over the IRTree Model. The IRTree Model is not preferred over the two other methods that model extreme response tendencies. The IRTree Model, however, is preferred over the unidimensional GR model.

Table 15. Deviance information criterion for the IRTree Model, the MNRM, the MPCM, and the GR Model

Method	Dbar	Dmean	pD	DIC
IRTree Model	73262.47	65580.29	7682.175	80944.64
MNRM	72368.90	65209.46	7159.437	79528.34
MPCM	71399.03	64013.49	7385.535	78784.56
GR Model	80481.95	75659.80	4822.156	85304.11

4.1.6 Mean Squared Error

The final way to compare the model fit of the IRTree Model, the MNRM, and the MPCM was using the mean squared error (MSE) for the expected total score and observed total score. The mean squared error was calculated at each iteration of the MCMC process that built the parameter posterior distributions. A distribution of MSE values under each model was developed. The mean, standard deviation, minimum, and maximum values of the MSE for each of the three models are displayed in Table 16. The MNRM had the lowest average MSE with the least variation. The IRTree Model had the highest average MSE with the highest variation. Additionally, the MSE was calculated for the GR model. The GR model had a MSE average greater than the MNRM and the MPCM but less than the IRTree Model.

Table 16. Mean squared error of expected total score and observed total score for the IRTree Model, the MNRM, the MPCM, and the GR Model

Method	Mean	Std Dev	Min	Max
IRTree Model	3.92	.11	3.5	4.3
MNRM	2.58	.06	2.4	2.9
MPCM	2.78	.07	2.5	3.0
GR Model	2.86	.06	2.0	3.1

4.1.7 Summary of Model Fit

Model fit was assessed using PPMC, deviance information criteria, and the mean squared error between the observed and expected total score. The global odds ratio was calculated using the "extreme" vs. "non-extreme" dichotomy. The observed data did not fit the IRTree Model, the MNRM, or the MPCM with respect to the global odds ratio. The total score mean discrepancy measure provided evidence of adequate test-level fit for the MNRM and the MPCM models. The total score standard deviation discrepancy measure provided evidence of adequate test-level fit for the MPCM model only. The total score group frequency discrepancy measure provided additional evidence of test-level fit for both the MNRM and the MPCM. The proportion of extreme responses by item discrepancy measure provided evidence of item-level fit for the MNRM and the MPCM. The individual extreme response rate discrepancy measure provided mixed results for the MNRM and the MPCM. There was a clear lack of model fit evidenced for the IRTree Model when the individual extreme response rate discrepancy measure was considered. The MPCM was the favored model based on the DIC and the MNRM was the favored model based on the MSE between expected total score and observed total score.

Overall the MNRM and the MPCM outperformed the GR model with respect to the discrepancy measures. The IRTree Model had worse or similar fit compared to the GR model with respect to the discrepancy measures. Based on the DIC, all three models accounting for extreme response style were preferred over the GR model. The IRTree Model resulted in a higher MSE compared to the GR Model. Both the MNRM and MPCM displayed lower MSE of expected and observed total score compared to the GR model.

4.1.8 Individual Response Sets

Five selected individual response sets and trait values estimated under the IRTree Model, the MNRM, and the MPCM are displayed in Table 17. Individual 5 and individual 4426 selected an extreme response on every item. Individual 5 selected 1-*Strongly Disagree* in response to each item and individual 4426 selected 4-*Strongly Agree* on each item. Individuals 2148 and 5210 selected a non-extreme response to all ten items. Individual 2148 selected responses 2-*Disagree* in response to each item while individual 5210 had an equal number of 3-*Agree* and 2-*Disagree* selections. Individual 4600 responded to only 5 items with an extreme response option.

Table 17. Selected individual PISA response sets and trait estimates under the IRTree Model the MNRM, and the MPCM

		IRTree Model		**MNRM**		**MPCM**	
Individual	**Response Set**	$\hat{\theta}_1$	$\hat{\theta}_{ERS}$	$\hat{\theta}_1$	$\hat{\theta}_{ERS}$	$\hat{\theta}$	$\hat{\omega}$
5	1111111111	-1.38 (.71)	1.99 (.53)	-1.36 (.74)	1.24 (.76)	-1.53 (.65)	0.48 (.25)
2148	2222222222	-0.57 (.54)	-0.81 (.54)	-0.40 (.50)	-1.27 (.69)	-0.64 (.64)	3.28 (1.91)
4426	4444444444	2.12 (.43)	1.49 (.40)	2.45 (.55)	1.66 (.69)	2.98 (.52)	0.39 (.20)
4600	1122112321	-0.29 (.46)	0.63 (.30)	-0.25 (.44)	0.55 (.53)	-0.23 (.43)	0.68 (.29)
5210	2222323333	1.14 (.34)	-1.28 (.61)	0.96 (.31)	-0.65 (.73)	1.24 (.40)	2.27 (1.29)

Individual 5 selected 1-*Strongly Disagree* for each item. For the IRTree Model, this individual was estimated to have the lowest (among the five selected respondents) substantive trait estimate ($\hat{\theta}_1 = -1.38$) and the largest extreme response tendency trait estimate ($\hat{\theta}_{ERS} = 1.99$). Individual 4426 selected 4-*Strongly Agree* for each item. Individual 4426 was estimated to have the highest substantive trait estimate $\left(\hat{\theta}_1 = 2.12\right)$ and the second highest extreme response tendency trait estimate $\left(\hat{\theta}_{ERS} = 1.49\right)$ under the IRTree Model. Although individual 5 and individual 4426 selected an equal number of extreme response options (10), individual 5 was estimated to have higher extreme tendencies when the IRTree Model was used for analysis.

Under the MNRM, the substantive trait for individual 5 was estimated to be $\hat{\theta}_1 = -1.36$ and the substantive trait for individual 4426 was estimated as $\hat{\theta}_1 = 2.45$. Individual 5 had an estimated extreme response trait of $\hat{\theta}_{ERS} = 1.24$ while individual 4426 had an extreme response tendency trait estimate of $\hat{\theta}_{ERS} = 1.66$. When the MPCM was used to fit the 5,330 responses on the Value of Science subscale, individual 5 had a substantive trait estimate of $\hat{\theta} = -1.53$ and individual 4426 had a substantive trait estimate of $\hat{\theta} = 2.98$. The extreme response tendency estimate, $\widehat{\omega}$, was least for individual 4426 ($\widehat{\omega} = 0.39$). Individual 5 had the subsequent smallest extreme response tendency estimate ($\widehat{\omega} = 0.48$).

The rank ordering of trait estimates can be compared across the three models. The MPCM and the MNRM both estimated individual 4426 with a higher extreme response tendency than individual 5. The IRTree Model estimated individual 5 to have a higher extreme response tendency than individual 4426.

Only 104 individuals obtained total scores greater than 30 while there were 1,232 individuals who had total scores between 10 and 15 (Table 11). The total score provides information about the trait estimates from the MNRM and the MPCM even though it is not a sufficient statistic. An individual with a high total score (40) was estimated to have a higher extreme response tendency than an individual with a low total score (10). This may be due to a lower frequency of individuals at the top of the score scale. With few individuals obtaining high total scores, a high presence of a secondary trait, besides the substantive trait, may be advancing individuals to high total scores. In this sample, a high frequency of individuals has low total scores. Therefore, it may be the low presence of the substantive trait having an impact greater than a high presence of the secondary, extreme response trait.

This explanation, however, does not explicitly explain why there are differences between the IRTree Model and the other two models. The weighted sum of the item responses, $\Sigma a_j X_{ij}$, is a sufficient statistic for estimating the location of a person's trait estimate in a 2-Paramter logistic (2-PL) model (de Ayala, 2013). The IRTree Model is made up of one 2-PL model and one modified 2-PL model with compensatory effects. The second trait, θ_{ERS}, is associated with the modified 2-PL model with compensatory effects.

For an individual to select 4-*Strongly Agree* on an item, they must score a 1 on the first decision (select *Agree*) and they must score a 1 on the second decision (select *Strongly*). Thus, individual 4426 had a high total score on decision 1 and a high total score on decision 2. For an individual to select *1-Strongly Disagree* on an item, they must score a 0 on the first decision (select *Disagree*) and they must score 1 on the second decision (select *Strongly*). Thus, individual 5 had a low total score on decision 1 and a high total score on decision 2. The pattern of total scores for decision 1 and decision 2 help explain the pattern of trait estimates for individual 5 and individual 4426. Furthermore, the compensatory nature of the second decision influences the difference in extreme response trait estimates under the IRTree Model. In other words, individual 5 has a low total score on the first decision. In turn, the ERS trait compensates for the low presence of the substantive trait. Individual 4426 has a high total score on the first decision. Thus, the ERS trait does not compensate for the substantive trait. As a product of this compensation, the ERS trait estimate for individual 5 is greater than the ERS trait estimate of individual 4426.

For an equal number of extreme options selected, analysis of individual 2148 and individual 5210 provided additional evidence that higher total scores results in a higher tendency of extreme response selection than low total score for the MNRM and MPCM but not the IRTree Model. Individual 2148 and individual 5210 both selected 0 extreme response option. Individual 2148 had a lower total score (20) than individual 5210 (25). Under the MNRM, individual 2148 had a lower extreme response trait estimate ($\hat{\theta}_{ERS} = -1.27$) than individual 5210 ($\hat{\theta}_{ERS} = -0.65$). Under the MPCM, individual 2148 had a lower extreme response tendency ($\hat{\omega} = 3.28$) than individual 5210 ($\hat{\omega} = 2.27$). The individual (5210) with the higher total score had a higher extreme response tendency estimate than the individual (2148) with lower total score even though they both selected the same number of extreme response categories. The IRTree Model provided opposing results. Individual 2148 is estimated to have a higher extreme response trait than individual 5210. This may be a result of the ERS trait compensating for a low presence of the substantive trait for individual 2148. This result emphasizes the difference between the IRTree Model and the MPCM as well as the difference between the IRTree Model and the MNRM.

Estimation of the traits for individual 4600 resulted in a consistent pattern across the three models. Individual 4600 selected five extreme responses and 5 non-extreme response options. For the IRTree Model ($\hat{\theta}_{ERS} = -0.63$), the MNRM $\left(\hat{\theta}_{ERS} = 0.55\right)$ and the MPCM ($\hat{\omega} = 0.68$), individual 4600 had a moderate estimate of extreme tendencies compared to the other 4 selected individuals. Individual 4600 was ranked third in degree of extreme response tendencies under the IRTree Model, the MNRM, and the MPCM.

The rank ordering of trait estimates of five individuals was analyzed to help understand the relationship between the IRTree Model, the MNRM, and the MPCM. Based on the results of the first research question, the correlation between extreme response traits was very strong between the MPCM and the MNRM. The five selected individuals displayed this strong correlation. The correlation of extreme tendency traits between the IRTree Model and the MNRM as well as the correlation between the IRTree Model and the MPCM were moderately strong. The five selected individuals illustrated this weaker correlation between extreme tendency estimates of the IRTree Model and of the MNRM and between extreme tendency estimates of the IRTree Model and of the MPCM compared to the correlation between the MNRM and the MPCM extreme tendency estimates.

4.1.9 Item Trace Lines

The three approaches to accounting for extreme response style attempt to model the interaction an individual has with the response scale for Likert type items. This interaction can be visualized using category response curves. Category response curves, also known as trace lines, are used to plot the propensity to endorse each response option across the substantive trait continuum. Category response curves plot the continuum of one trait at a time. The category response curves of propensity versus the substantive trait given a value of ERS can be analyzed for multidimensional models with one substantive trait and one ERS trait. The resulting trace lines indicate the probability to select each response option along the substantive trait continuum for individuals with the same ERS tendency.

Trace lines can be generated for every item and for every value of ERS tendencies for each model. For example, the interaction of individuals with a high extreme response tendency with item 10 and the interaction of individuals with a low extreme response tendency with item 10 were considered. The extreme response trait estimates for individual 5 were used to represent high extreme response tendencies and the extreme response trait estimates for individual 5210 were used to represent low extreme response tendencies.

Item parameters and person parameters were estimated simultaneously using PROC MCMC in SAS. The item parameter estimates for items 1 – 10 are displayed in Appendix A. The item parameter estimates were derived by taking the mean of the post burn-in iterations of the MCMC procedure estimating the posterior distribution. Table 18 displays the item parameter estimates for item 10. For comparison purposes, the item parameters estimates for the MNRM found by Bolt and Newton (2011) when two scales were considered (value of science and enjoyment of science) were .291, 2.146, 1.005, and -3.444 for intercepts $c_1, \dots, c_4$, respectively.

The item parameters are not equivalent due to different estimation techniques; however, the estimates are approximately equal for the same item and model.

Table 18. Item 10 parameter estimates under the IRTree Model, the MNRM, and the MPCM

Method	**Parameter**	**Estimate**
IRTree Model	a_1	2.4318
	b_1	1.5603
	a_2	2.0147
	b_2	-1.7175
	ν	.0910
MNRM	c_1	.3465
	c_2	2.3176
	c_3	.9360
	c_4	-3.6001
MPCM	a	1.8370
	b	.6589
	τ_1	-1.5487
	τ_2	.0494
	τ_3	1.4993

Figure 23 displays the trace plots of item 10 for an individual with an extreme response tendency trait of $\theta_{ERS} = 1.99$ (IRTree Model), $\theta_{ERS} = 1.24$ (MNRM), and $\omega = .48$ (MPCM). The vertical axis is the propensity to select each category for item 10. The horizontal axis represents the substantive trait continuum. Response option trace lines are labeled with their quantitative value (1- *Strongly Disagree,* 2-*Disagree,* 3-*Agree,* and, 4-*Strongly Agree*). The vertical line within each plot represents the placement of individual 5's substantive trait estimate.

Across the substantive trait continuum for the IRTree Model, the response option with the highest probability of selection is either *1* or *4*. An individual with an extreme response trait of 1.99 for the IRTree Model has very low chance of selecting a non-extreme category on item 10 regardless of the value of their substantive trait. Individuals with substantive traits above 0.5 are likely to select 4-*Strongly Agree* and individuals with substantive traits below 0.5 are likely to select 1-*Strongly Disagree.*

Under the MNRM, individuals with an extreme response trait of 1.24 and substantive traits above 1 are most likely to select 4-*Strongly Agree* on item 10. Individuals with an extreme response trait equal to 1.24 and a substantive trait less than 0 are most likely to select 1-*Strongly Disagree* on item 10. Unlike the results from the IRTree Model, there exists a group of individuals who are most likely to endorse a non-extreme category. Individuals with substantive trait values between 0 and 1 are most likely to endorse one of the middle categories (2 and 3) more than any other option on item 10.

The MPCM was analyzed for individuals with extreme response tendencies of .48. For the MPCM, individuals with substantive traits below 0 or above 1.5 are most likely to endorse an extreme category. A non-extreme category has the highest probability of selection for individuals with substantive trait estimates between 0 and 1.5.

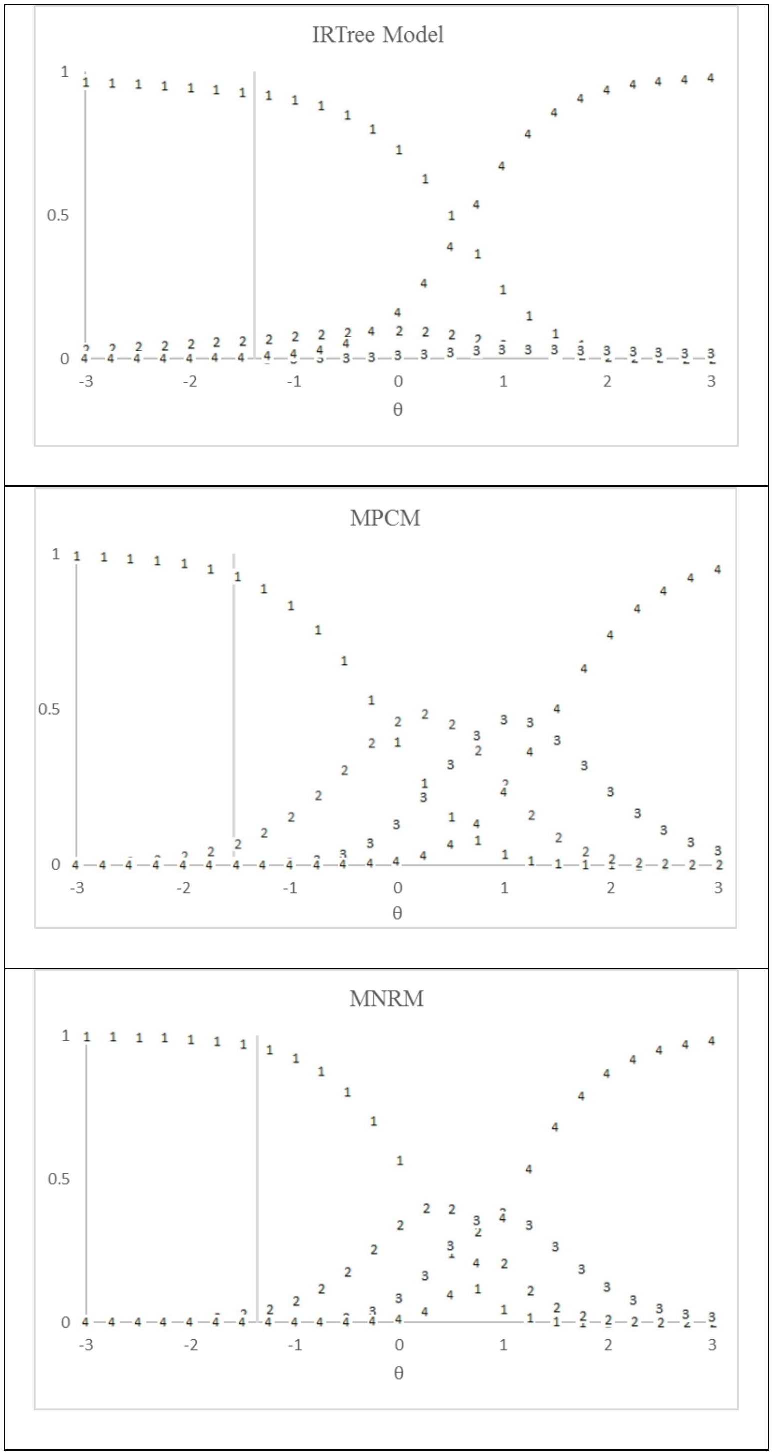


Figure 23. Item 10 trace plots for an individual with an extreme response tendency trait of $\boldsymbol{\theta_{ERS} = 1.99}$ (IRTree Model), $\boldsymbol{\theta_{ERS} = 1.24}$ (MNRM), and $\boldsymbol{\omega = .48}$ (MPCM)

The trace plots for item 10 related to individuals with extreme response tendencies of $\theta_{ERS} = -1.28$ (IRTree Model), $\theta_{ERS} = -0.65$ (MNRM), and $\omega = 2.27$ (MPCM) are displayed in Figure 24**.** The vertical line on each plot represents the trait estimate of individual 5210. The plots are used to illustrate individuals who have a very low tendency for extreme responses. The trace lines for item 10 under the IRTree Model show, regardless of the substantive trait value, either 2-*Disagree* or 3-*Agree* has the highest propensity of endorsement.

Individuals with low extreme response tendencies ($\omega = 2.27$) under the MPCM, tend to select either 3-*Agree* or 2-*Disagree* with the highest probability along the substantive trait continuum. The exception is for substantive trait values near -3. The substantive trait overpowers the tendency to be extreme response averse for individuals with very low presence of the substantive trait. In other words, even though the individual has a very low extreme response tendency (2.27), an extreme category (1-*Strongly Disagree*) has highest probability of selection.

The trace plot developed under the MNRM for individuals with $\theta_{ERS} = -0.65$ exhibits a unique pattern compared to the trace plot using $\omega = 2.27$ for the MPCM and the trace plot using $\theta_{ERS} = -1.28$ for the IRTree Model. Individuals with substantive trait values greater than 3 endorse category *4* with the highest probability even though these individuals do not have a high extreme response tendency. Individuals with substantive trait estimates greater than 0.8 but less than 3 have the highest propensity to select 3-*Agree*. Individuals with substantive trait estimates between -1.5 and 0.8 have the highest propensity towards selecting 2-*Disagree* on item 10. Even with a low tendency to select extreme responses, individuals with substantive trait values less than -1.5 tend to select 1-*Strongly Disagree*.

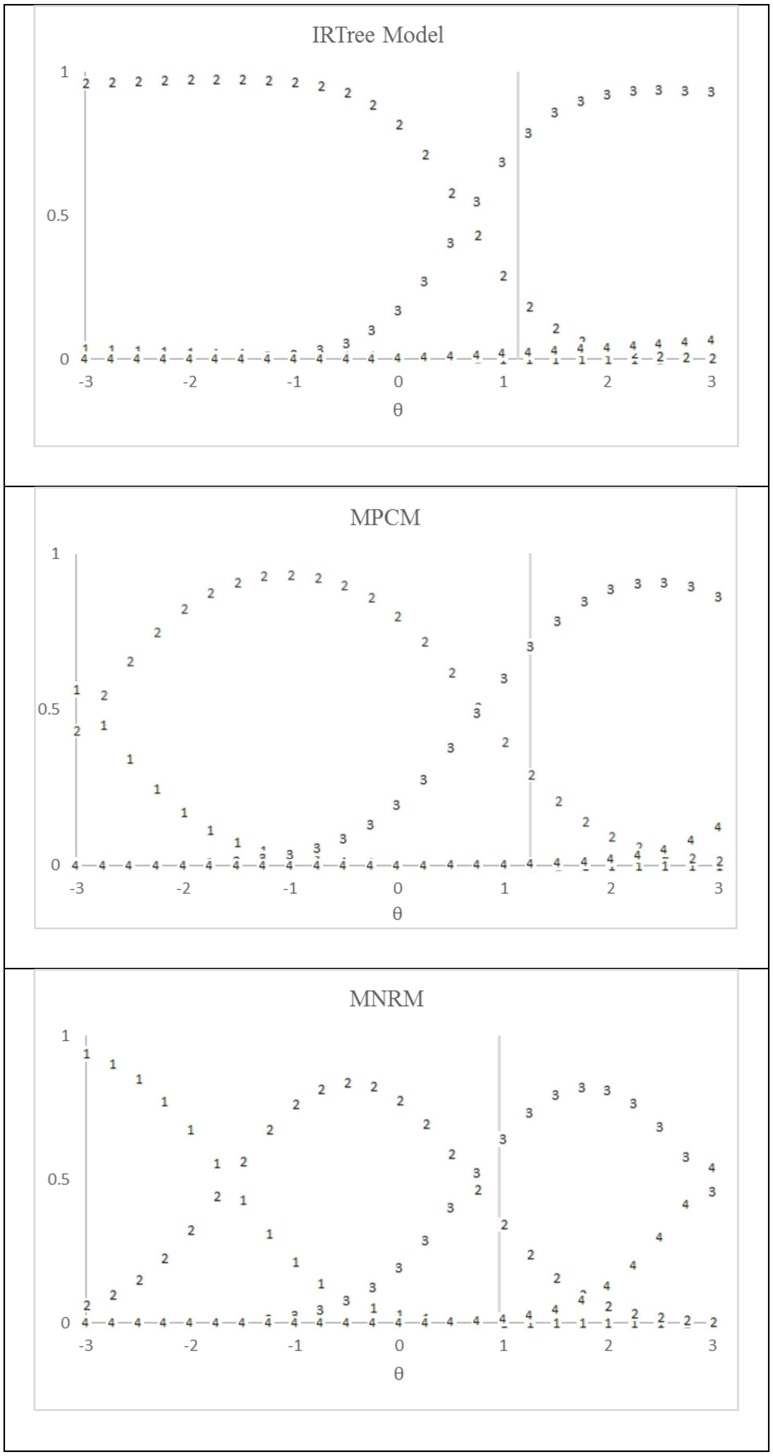


Figure 24. Item 10 trace plots for an individual with an extreme response tendency trait of $\boldsymbol{\theta_{ERS} = -1.28}$ (IRTree Model), $\boldsymbol{\theta_{ERS} = -0.65}$ (MNRM), and $\boldsymbol{\omega = 2.27}$ (MPCM)

4.1.10 Summary of Empirical Study

The IRTree Model, the MPCM, and the MNRM exhibit both similar and dissimilar patterns when analyzing the 5,330 responses on the Value of Science subscale. The IRTree Model models a response process. The MNRM treats extreme response tendencies as an additional non-compensatory trait. The MPCM assumes an individual's extreme response tendency is a random person effect. The Value of Science trait estimates across the three models were highly correlated. The extreme tendencies trait was strongly correlated between the MNRM and the MPCM. The extreme tendencies trait had a moderate relationship between estimates obtained from the MNRM and the IRTree Model and between estimates from the MPCM and the IRTree Model.

The IRTree Model had inadequate model fit at the test-level, person-level, and item-level. The MNRM had varied results of test-level model fit but adequate person-level and item-level model fit. The MPCM had adequate test-level, person-level, and item-level model fit. The MPCM and MNRM outperformed the unidimensional GR model, however, the IRTree Model had similar lack of model fit to the GR model with respect to the PPMC discrepancy statistics. Based on the DIC, the MPCM had the best fit, and based on the MSE, the MNRM had the best model fit. Of course, comparison of models using real data does not provide an underlying true model that reflects response style use. Therefore, to further understand the differences in model fit among the three models, a simulation study was performed.

4.2 SIMULATION STUDY

4.2.1 Data Generation

Item responses were simulated using the IRTree Model, the MNRM, and the MPCM. Item responses were simulated for 500 and 1000 individuals responding to 10 and 20 items with 4 and 6 response options. The IRTree Model, the MNRM, and the MPCM were used to simulate the responses from individuals who exhibit extreme response tendencies. The descriptive statistics proposed by Greenleaf (1992b) were used to validate the simulated responses. Greenleaf (1992b) proposed calculating the extreme response rates of individuals as well as the proportion of extreme response selections for each item as descriptive measures of extreme response tendencies. The resulting two rates were compared to the proportion of extreme response options on any given item. Consider an item with 6 response options ranging from *1-Strongly Disagree* to *6-Strongly Agree*. Two response options (*1* and *6*) out of 6 (33%) are considered extreme options. Individuals taking a survey are exhibiting extreme response tendencies if: 1) the proportion of extreme response selections on each item is greater than 0.33 and 2) the percentage of individuals selecting an extreme response option on more than 33% of their responses is high.

The proportion of extreme response selections were calculated under each condition of sample size (500 and 1000), number of response options (10 and 20), and number of items (10 and 20) for the 25 iterations under the data generation models (IRTree Model, MNRM, and MPCM). The average proportion of extreme responses by item over the 25 iterations are summarized in Table 19 for 4-category items and Table 20 for 6-cateogy items. Detailed results are presented in Appendix B.

Table 19. Number of items with greater than 50% of respondents selecting an extreme response option for four-category items

Survey Length	Sample Size	IRTree Model	MNRM	MPCM
10 items	500	5	10	10
	1000	5	10	10
20 items	500	5	20	20
	1000	11	20	20

Table 20. Number of items with greater than 33.33% of respondents selecting an extreme response option for six-category items

Survey Length	Sample Size	IRTree Model	MNRM	MPCM
10 items	500	10	10	10
	1000	6	10	10
20 items	500	18	20	20
	1000	19	20	20

The individual extreme response selection rate is calculated by summing the number of extreme responses options selected (*Strongly Agree* or *Strongly Disagree*) by the total number of items. A survey with ten items results in 11 unique possible proportions (0.0 to 1.0 by 0.1) and a twenty-item survey has 21 unique proportions possible (0.0 to 1.0 by 0.05). The average (over the 25 iterations) cumulative percentage of simulated individuals who selected each proportion of extreme responses are summarized in Table 21 for 4-cateogy items and Table 22 for 6-cateogy items. Detailed results are presented in Appendix B.

Table 21. Average percentage of individuals selecting 50% or more extreme response options for 4-category items.

Survey Length	Sample Size	IRTree Model	MNRM	MPCM
10 items	500	59.55%	77.95%	76.15%
	1000	59.18%	79.5%	75.69%
20 items	500	57.27%	73.08%	78.61%
	1000	60.91%	73.7%	78.99%

Table 22. Average percentage of individuals selecting 33.33% or more extreme response options for 6-category items.

Survey Length	Sample Size	IRTree Model	MNRM	MPCM
10 items	500	57.22%	80.28%	73%
	1000	42.65%	82.88%	69.92%
20 items	500	61%	90.12%	82.27%
	1000	63.29%	88.41%	83.69%

As seen in Table 19, the item response rates are borderline to the minimum quality when data was generated under the IRTree Model. For items with four response options, at least 50% of item responses should be 1-*Strongly Disagree* or 4-*Strongly Agree*. Five out of the 10 4-category items satisfy this criterion when responses were simulated for 500 individuals. Only 5 out of the 20 4-category items meet the 50% threshold when responses for 500 individuals were simulated. The 15 items that did not meet the criteria have a proportion slightly under 0.5. Five out of 10 4-category items met the criteria when responses were simulated for 1000 individuals. Only 41% of the 1000 responses to item 10 were 1-*Strongly Disagree* or 4-*Strongly Agree.* Eleven out of the 20 4-category response options had over 50% of the 1000 simulated respondents select an extreme response.

All 10 6-category items met the criteria (≥ .33) when responses for 500 individuals were generated using the IRTree Model. Only 2 out of 20 6-category items failed to meet the 33% threshold when responses for 500 individuals were generated. Four out of the 10 6-category items failed to reach the threshold for the 1,000 individual response sets simulated. 19 out of the 20 6-category items met the .33 threshold when responses for 1000 individuals were simulated.

Table 21 displays the cumulative percentage of individuals who selected each potential proportion of extreme response options by simulation condition averaged over the 25 iterations for data generated using the IRTree Model for 4-category items. On average, 59.55 % of the 500 simulated respondents selected either 1-*Strongly Disagree* or 4-*Strongly Agree* on 5 or more items out of the 10 items. On average, 59.18% of the 1000 simulated individuals selected 1-*Strongly Disagree* or 4-*Strongly Agree* on at least 5 out of the 10 items. On average 57.22% of the 500 simulated respondents and 42.65% of the 1000 simulated respondents selected 1-*Strongly Disagree* or 6-*Strongly Agree* on at least 33% of the 10 items.

Consider simulated responses to a 20-item survey using the IRTree Model. On average, 52.78% of the 500 simulated respondents and 60.91% of the 1000 simulated respondents selected 1-*Strongly Disagree* or 4-*Strongly Agree* on at least 10 of the 20 items. On average 61% of the 500 simulated respondents and 63.29% of the 1000 simulated respondents selected either 1-*Strongly Disagree* or 6-*Strongly Agree* on 33% or more of the 20 items.

Table 19 displays the extreme response rate by item for responses simulated under the MNRM. Each item has an extreme response rate over 50% for all simulation conditions related to items with 4 response options. Furthermore, most of these item extreme response rates are above 0.60. Each item with 6-reponse categories (Table 20) has an extreme response rate above .33 across sample sizes and survey lengths. Furthermore, most items with 6 response options have an extreme response rate greater than .5.

For data generated under the MNRM (Table 21). 77.95% of the 500 simulated respondents and 79.5% of the 1000 simulated respondents, on average, selected either 1-*Strongly Disagree* or

4-*Strongly Agree* on at least 5 of the ten item. On average, 80.28% of the 500 simulated respondents and 82.88% of the 1000 simulated individuals selected 1-*Strongly Disagree* or 6-*Strongly Agree* to at least 33% of the 10 items (Table 22). On average 73.08% of the 500 simulated respondents and 73.7% of the 1000 individuals simulated selected an extreme response option on at least 10 out of the 20 4-category items. On average, 90.12% of the 500 respondents and 88.41% of the 1000 respondents were simulated to select either 1-*Strongly Disagree* or 6-*Strongly Agree* on 33% or more of the 20 6-category items.

Table 19 displays the item extreme response rates for responses generated under the MPCM. All items with 4 response categories were simulated to have a proportion of extreme response selections equal to .5 or more, on average. All items with 6 response options (Table 20) were simulated to have an average extreme response rate greater than .33. Furthermore, most items with 6 response options were simulated to have a well-above criteria average extreme response rate, with between 40% and 50% of respondents selecting an extreme response option.

Table 21 displays the average cumulative percentage of individual extreme response rates exhibited under each of the simulated criterion when the MPCM was used to generate the data. On average, 76.15% of the 500 simulated respondents and 75.69% of the 1000 simulated respondents selected either 1-*Strongly Disagree* or 4-*Strongly Agree* on at least 5 of the 10 items. When 10 items were simulated with 6 response options (Table 22), an average of 73% of the 500 respondents and an average of 69.92% of the 1000 respondents were simulated to select extreme response options on over 33% of the items. On average, 78.61% of the 500 simulated individuals and 78.99% of the 1000 simulated individuals selected either 1-*Strongly Disagree* or 4-*Strongly Agree* on at least 10 of the 20 items. On average, 82.27% of the 500 simulated individuals and 83.69%

of the 1000 simulated individuals selected extreme response options on 33% or more of the 20 items with 6 response options.

Based on the descriptive criteria presented by Greenleaf (1992b) data generated under the MPCM and data generated under the MNRM have adequate qualities to represent individuals exhibiting extreme response tendencies. Responses generated under the IRTree Model produce opposing results for each of the two-criterion considered. In all simulation conditions for the IRTree Model generated data, there exist items with extreme response rates that fall below the rate threshold. The individual extreme response rate threshold is marginally exceeded under each condition for data generated under IRTree Model. Data generated under the MNRM and MPCM, however, greatly exceed the item extreme response rate criteria.

4.2.2 Item Mean Squared Error

Item responses simulated under the IRTree Model, the MPCM, and the MNRM were fit with the IRTree Model, the MPCM, and the MNRM. Each model was used to analyze each dataset in the fully crossed design. The models were estimated using PROC MCMC in SAS. For each of the 25 replications, 20,000 iterations were used to build the posterior distribution after a 5,000 iteration burn-in period. The mean of each parameter over the 20,000 iterations of the posterior distribution was used as the parameter estimate. The item mean squared error (IMSE) was calculated as the squared difference between observed total score, $\Sigma_{j=1}^{J} U_j$, and expected total, $E\left(U_j \middle| \boldsymbol{\theta}\right) = \Sigma_{j=1}^{J} \Sigma_{k=1}^{m_j} k * P_{jk}(\theta)$, divided by J, the number of items analyzed. The IMSE was used to account for inherent differences of total scores due to survey length between 20 item surveys and 10 items surveys.

The mean IMSE over R replications for each simulation condition was calculated as

$$\overline{IMSE} = \frac{\Sigma_{j=1}^{J} U_j - \Sigma_{j=1}^{J} \Sigma_{k=1}^{m_j} k * P_{jk}(\theta)}{J * R} \qquad (21)$$

Table 23, Table 24, and Table 25 display the mean IMSE and the standard deviation of IMSE for each of the 72 simulation conditions. Consider data generated under the IRTree Model in Table 23. Across all 8 combinations of sample size, number of items, and the number of response options, the estimated IRTree Model produced the lowest mean IMSE. The mean IMSE was greatest for each of the three models when 1000 responses to 20 items with 6 response options were considered. The MNRM resulted in a higher mean IMSE across all conditions compared to the mean IMSE when the IRTree Model simulated data was fit using MPCM. For each model used for analysis, the average IMSE was least when fitting 500 responses to 10 4-category items.

Consider data generated under the MNRM in Table 24. Across all simulation conditions, nearly equivalent mean IMSE and standard deviation IMSE values were produced when fit using the MPCM and fit using the MNRM except when 20 items with 6 response options were considered. When 500 simulated response sets to 20 6-category items were fit, the MNRM resulted in the highest mean IMSE and the MPCM resulted in the lowest mean IMSE. Analysis using the IRTree Model resulted in the highest mean IMSE values across all simulation conditions except when 20 items with 6 response options were considered.

Analysis using the MPCM resulted in the lowest mean IMSE when data was generated under the MPCM (Table 25). For data generated using the MPCM, the IRTree Model analysis method resulted in the greatest mean IMSE across all simulation conditions except when 20 6-category items were analyzed. For responses to 20 6-category items generated under the MPCM, the mean IMSE was greatest when fit with the MNRM.

Table 23. Mean IMSE by sample size, survey length, and number of response options for data generated under the IRTree Model

	Data Generation Model							
	IRTree Model							
	Sample Size							
	500				1000			
	Number of Items							
	10		20		10		20	
	Response Options							
Data Analysis Model	4	6	4	6	4	6	4	6
IRTree Model	.44 (.14)	.83 (.48)	.82 (.25)	1.38 (.59)	.44 (.15)	.63 (.35)	.85 (.37)	1.37 (.55)
MNRM	1.02 (.17)	1.99 (.32)	1.25 (.16)	3.19 (2.39)	1.06 (.19)	1.72 (.33)	1.18 (.13)	4.11 (2.69)
MPCM	.45 (.16)	.85 (.37)	1.17 (.34)	2.05 (.77)	.51 (.17)	.70 (.49)	1.14 (.40)	2 (.72)

Table 24. Mean IMSE by sample size, survey length, and number of response options for data generated under the MNRM

	Data Generation Model							
	MNRM							
	Sample Size							
	500				1000			
	Number of Items							
	10		20		10		20	
	Response Options							
Data Analysis Model	4	6	4	6	4	6	4	6
IRTree Model	.45 (.08)	.95 (.11)	.44 (.06)	.94 (.10)	.44 (.06)	.89 (.12)	.48 (.08)	.93 (.09)
MNRM	.37 (.04)	.77 (.08)	.36 (.04)	1.56 (1.12)	.37 (.04)	.74 (.08)	.38 (.03)	1.25 (.39)
MPCM	.38 (.04)	.77 (.09)	.36 (.04)	.77 (.07)	.38 (.04)	.74 (.08)	.38 (.04)	.77 (.05)

Table 25. Mean IMSE by sample size, survey length, and number of response options for data generated under the MPCM

	Data Generation Model							
	MPCM							
	Sample Size							
	500				1000			
	Number of Items							
	10		20		10		20	
	Response Options							
Data Analysis Model	4	6	4	6	4	6	4	6
IRTree Model	.59 (.05)	1.17 (.14)	.65 (.06)	1.28 (.13)	.59 (.05)	1.17 (.13)	.66 (.05)	1.38 (.14)
MNRM	.57 (.08)	.96 (.10)	.54 (.05)	1.48 (.90)	.55 (.07)	.96 (.11)	.57 (.04)	1.58 (1.02)
MPCM	.46 (.06)	.83 (.08)	.47 (.03)	.86 (.07)	.45 (.04)	.82 (.1)	.49 (.03)	.87 (.06)

4.2.3 Sample Size

Based on the descriptive statistics, there was no evidence of a significant interaction between sample size and the number of response options on the pattern of mean IMSE. Additionally, there was no evidence of a significant interaction between sample size and the survey length on the pattern of mean IMSE.

Figure 25 displays the mean IMSE for each data generation model and analysis model by sample size averaged over the number of response categories and the number of items. Overall, the horizontal lines indicate the pattern of average IMSE is unaffected by sample size. The MPCM (dotted lines) resulted in the lowest average IMSE across sample sizes when responses were generated under the MNRM and under the MPCM. The IRTree model (dashed line) resulted in a lower mean IMSE than the mean IMSE estimated by the MNRM (solid line) for 500 responses simulated under the MNRM. The MNRM analysis method and the IRTree Model analysis method resulted in equal mean IMSE when 1000 responses were generated under the MNRM. The MNRM analysis model resulted in the highest mean IMSE and the IRTree Model analysis method resulted in the lowest mean IMSE across sample sizes when data was generated under the IRTree Model.

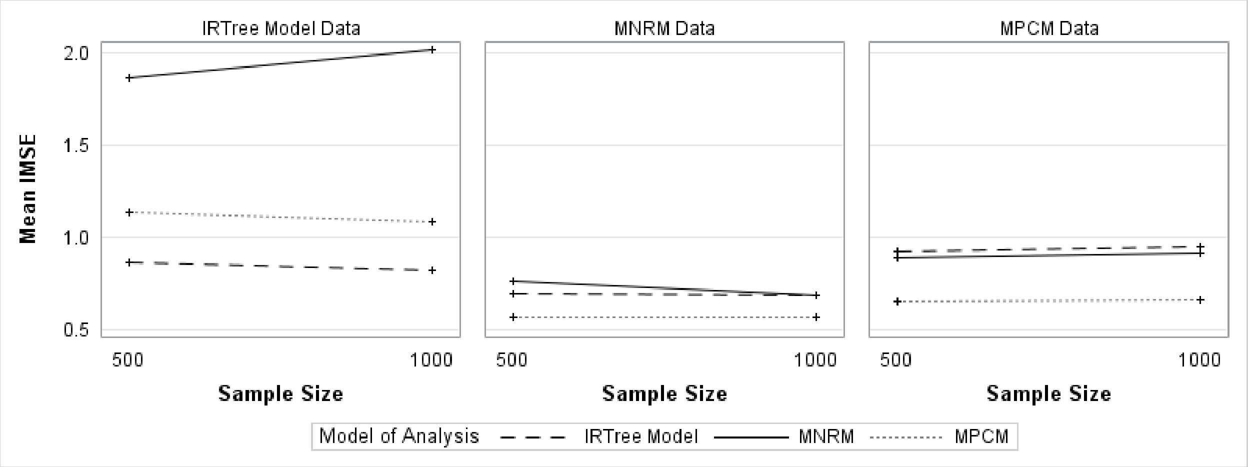


Figure 25. Mean IMSE for 500 and 1000 responses generated under the IRTree Model, the MNRM, and the MPCM and analyzed by the IRTree Model, the MNRM, and the MPCM

4.2.4 Survey Length

There was evidence of an interaction between survey length and the number of response options on the pattern of mean IMSE. Figure 26 displays the mean IMSE for each data generation model and analysis model by survey length for items with 4 response options averaged over sample size. Figure 27 displays the mean IMSE for each data generation model and analysis model by survey length for items with 6 response options averaged over sample size. Responses generated under the IRTree Model to 10 item and 20 item surveys with 4 and 6 responses options fit with the

MNRM resulted in the largest mean IMSE. When responses were generated under the IRTree Model, within model mean IMSE was lower for surveys with 10 items than surveys with 20 items for the IRTree Model, the MNRM, and the MPCM. This pattern of mean IMSE differences was consistent for surveys with items containing 4 response options and surveys with items containing 6 response options.

There was no effect of survey length on the mean IMSE for responses to items with 4 options generated under the MNRM and the MPCM as evidenced by the parallel lines. The mean IMSE was equal for responses generated under the MNRM to surveys with 4 response options fit with the MPCM and responses generated under the MNRM to surveys with 4 response options fit with MNRM. The MPCM analysis method resulted in the lowest mean IMSE for responses to 4-option surveys consisting of 10 and 20 items when responses were generated under the MPCM. The IRTree Model resulted in the highest mean IMSE values for 10 item surveys with 4 response options and 20 item surveys with 4 response options when data was generated under the MNRM and when responses were generated using the MPCM.

Surveys with 20 6-category items fit using the MNRM resulted in the largest mean IMSE within each data generation routine. The mean IMSE was equal for responses generated under the MNRM to 10 6-response option item surveys fit with the MNRM and fit with the MPCM. The MPCM analysis method resulted in the lowest mean IMSE for responses to 10 6-option item surveys generated under the MPCM. The IRTree Model analysis method resulted in the highest mean IMSE for responses to 10 6-response option surveys generated under the MPCM.

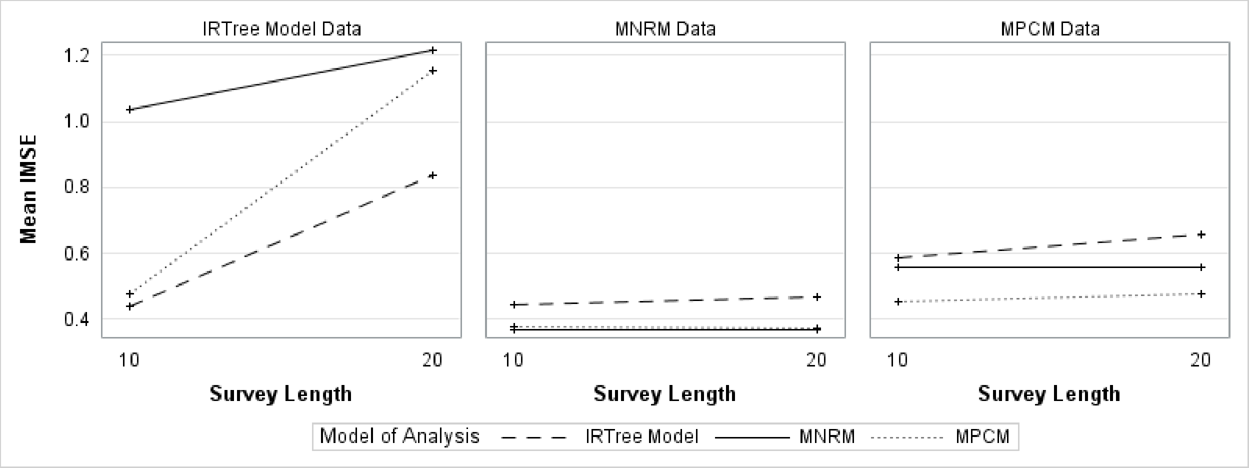


Figure 26. Mean IMSE for responses to surveys with 10 and 20 items with 4 response options generated under the IRTree Model, the MNRM, and the MPCM and analyzed by the IRTree Model, the MNRM, and the MPCM.

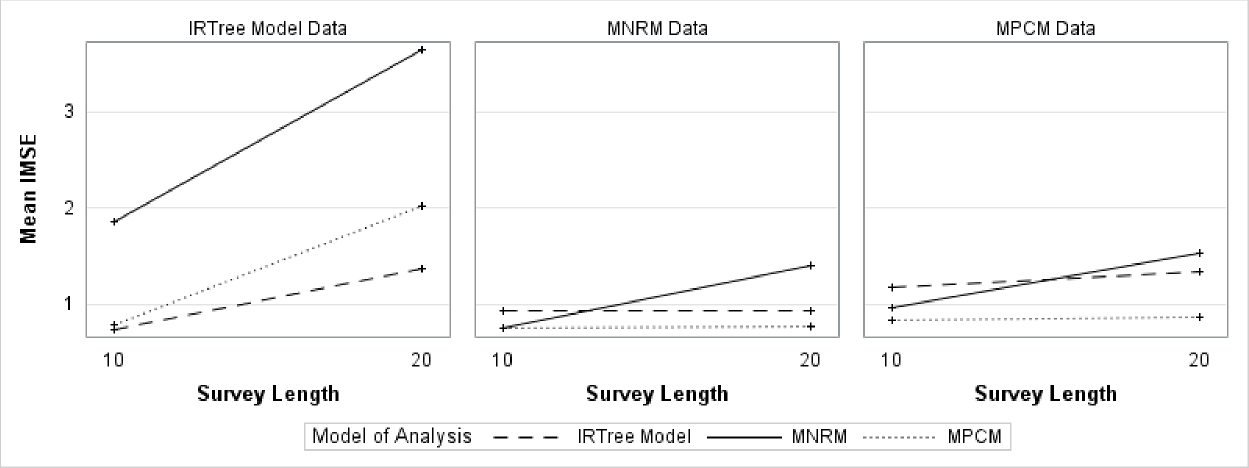


Figure 27. Mean IMSE for responses to surveys with 10 and 20 items with 6 response options generated under the IRTree Model, the MNRM, and the MPCM and analyzed by the IRTree Model, the MNRM, and the MPCM.

4.2.5 Number of Response Options

Figure 28 displays the mean IMSE for each data generation model and analysis model for surveys of 10 items with 4 and 6 response options averaged over sample size. Figure 29 displays the mean IMSE for each data generation model and analysis model for surveys of 20 items with 4 and 6 response options averaged over sample size. Overall, the lines slope up to the right indicating the mean IMSE was greater for items with 6 response options compared to items with 4 response options. The mean IMSE under the MPCM was approximately equal to the mean IMSE using the IRTree Model for responses to 10 item surveys generated under the IRTree Model across the levels of response options. When items responses were simulated under the IRTree Model, the MNRM resulted in a higher mean IMSE for surveys of 10 items with either 4 or 6 response options compared to the other two methods. There was no systematic difference in mean IMSE between analysis performed using the MNRM and analysis performed using the MPCM on MNRM simulated responses to 10 items with 4 and 6 response options. Analysis of MPCM generated responses to 10 items with 4 and 6 response options using the MPCM resulted in the lowest mean IMSE compared to the other two analysis methods. Fitting MNRM and MPCM simulated responses to 10 items with 4 and 6 response options with the IRTree Model led to the largest mean IMSE compared to the same simulated responses analyzed with the MNRM or MPCM.

For 20 items with 6 response options, data generated under the MNRM and data generated under the MPCM fit with the MPCM resulted in the lowest mean IMSE. The mean IMSE was larger when responses to 20 items with 6 categories generated under the MPCM were analyzed with the MNRM compared to the mean IMSE when analysis was performed with the IRTree Model. The mean IMSE was smaller when responses to 20 items with 4 categories generated under the MPCM were analyzed with the MNRM compared to mean IMSE when analysis was completed with the IRTree Model.

Analysis using the MPCM and analysis using the MNRM of responses to 20 item with 4 response options simulated under the MNRM resulted in the lowest mean IMSE. When 20 items with 6 response options generated under the MNRM were considered, the mean IMSE for the MPCM was the least and the mean IMSE for the MNRM was the greatest. Responses to 20 items with 4 and 6 categories generated under the IRTree Model fit with the IRTree Model resulted in the lowest mean IMSE compared to the same responses fit with the MPCM or the MNRM. For responses to 20 items 4 response options simulated under the IRTree Model, there was no difference in mean IMSE when fit with the MNRM and fit with the MPCM. A mean IMSE difference is evident, however, when responses to 20 items with 6 categories simulated under the IRTree Model were fit using the MNRM and the MPCM. Analysis performed using the MNRM resulted in a higher mean IMSE than analysis performed with the MPCM.

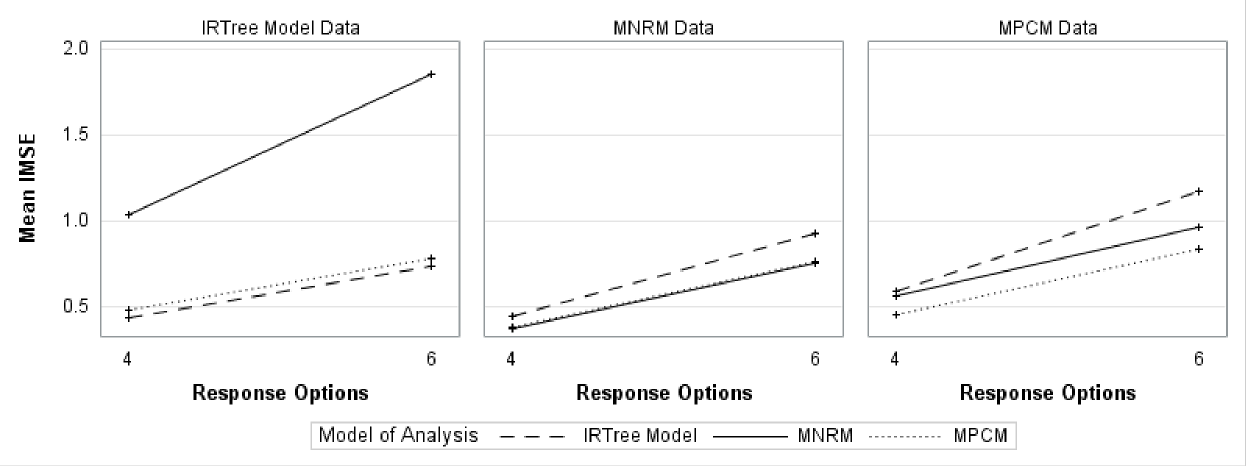


Figure 28. Mean IMSE for responses to 10 items with 4 and 6 response options generated under the IRTree Model, the MNRM, and the MPCM and analyzed by the IRTree Model, the MNRM, and the MPCM.

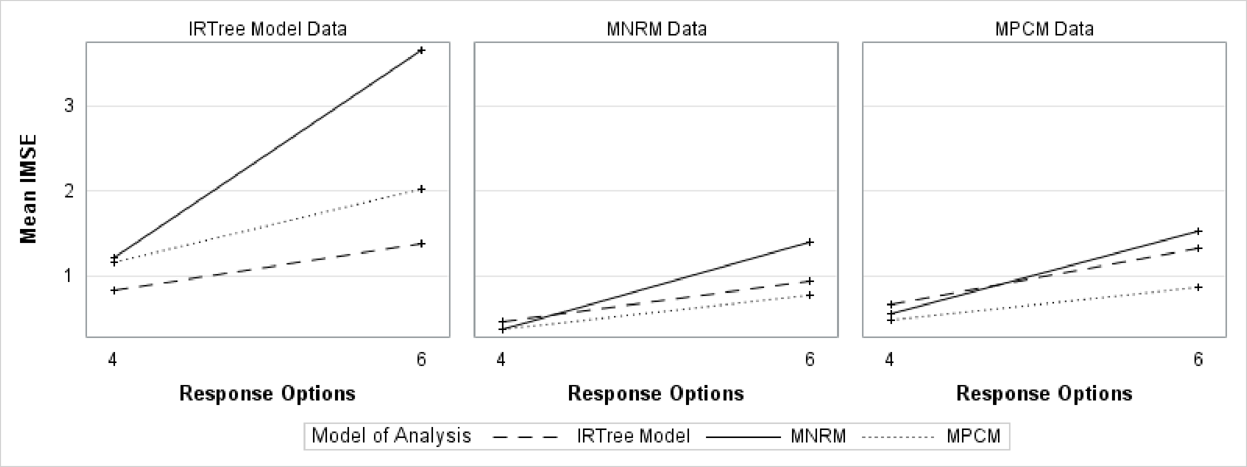


Figure 29. Mean IMSE for responses to 20 items with 4 and 6 response options generated under the IRTree Model, the MNRM, and the MPCM and analyzed by the IRTree Model, the MNRM, and the MPCM.

4.2.6 PPMC

Posterior predictive model checks (PPMC) were performed for the IRTree Model, the MNRM, and the MPCM for 2 of the 24 combinations of survey length, response options, and sample size. PPMC method were used to investigate a cell with large mean IMSE differences across the three models and a cell with minimal mean IMSE differences across the three models. The PPMC methods explored the practical significance of the model differences in terms of item-level and person-level fit. Item-level fit and person-level fit were explored as the discrepancy measures relate specifically to extreme response tendency. The first simulation condition explored was 500 responses to 20 items with 6 response options generated under the MNRM. In the condition, the average IMSE was 0.77 for the MPCM estimation, 0.94 for the IRTree Model estimation, and 1.47 when the MNRM was used for analysis. When 500 responses to 10 items with 4 categories were simulated under the MPCM, the resulting mean IMSE values were 0.46, 0.59, and 0.57 for analysis completed with the MPCM, the IRTree Model, and the MNRM, respectively.

The item extreme response rate *PPP*-values and the individual extreme response rate *PPP*-values were computed for 1 replication in each of the two simulation conditions. Plots were generated to visually compare the resulting *PPP*-values. *PPP*-values near 0.5 provide evidence of adequate model fit. *PPP*-values greater than 0.95 or less than 0.05 are an indication of inadequate model fit.

Figure 30 displays the item extreme response rate *PPP*-values using parameter estimates from the IRTree Model, the MNRM, and the MPCM for 500 simulated responses generated under the MNRM to 20 items with 6 categories. Values near the horizontal reference line, *PPP*-value = 0.5, indicate model recovery of the item extreme response rate. Values above 0.95 indicate overestimation of the item extreme response rate and values below 0.05 indicate underestimation of the item extreme response rate. The results from the PPMC measure show that the MNRM did not capture the pattern of item extreme responses for most items. The *PPP*-values from the IRTree Model and the *PPP*-values from the MPCM all fall within the Bayesian credible interval (0.05 to 0.95). The *PPP*-values calculated under the MPCM fall closer to 0.5 than those calculated under the IRTree Model.

Figure 31 displays the individual extreme response rate *PPP*-values using parameter estimates from the IRTree Model, the MNRM, and the MPCM for 500 simulated responses generated under the MNRM to 20 items with 6 categories. Compared to the *PPP*-values calculated using the item extreme response rate discrepancy statistic, the *PPP*-values for the individual extreme response rate discrepancy statistic have more scatter for each of the three models. This may be a result of the low sample size. Observed and predicted frequencies of individuals at a few of the 21 unique extreme response rates are scarce. There were zero individuals predicted to have an extreme response rate of 0.10 or below in any of the 2,000 replicated data sets for the MPCM.

Additionally, there were no individuals predicted to get an extreme response rate of 0.05 or below in any of the 2,000 replicated datasets under the MNRM and the IRTree Model. The MNRM had inadequate fit across most response rates. The IRTree Model and the MPCM, however, had adequate fit across most individual extreme response rates. The pattern of *PPP*-values across the IRTree Model and the MPCM are indistinguishable in terms of overall person-level model fit.

For 500 responses simulated under the MNRM to 20 items with 6 categories, *PPP*-values from the item extreme response rate discrepancy statistic and *PPP*-values from the individual extreme response rate discrepancy statistic indicated that the MNRM had the worst model fit and the MPCM had the best fit. This result provides evidence that the difference in mean IMSE seen in Table 24 relate to meaningful differences of model fit.

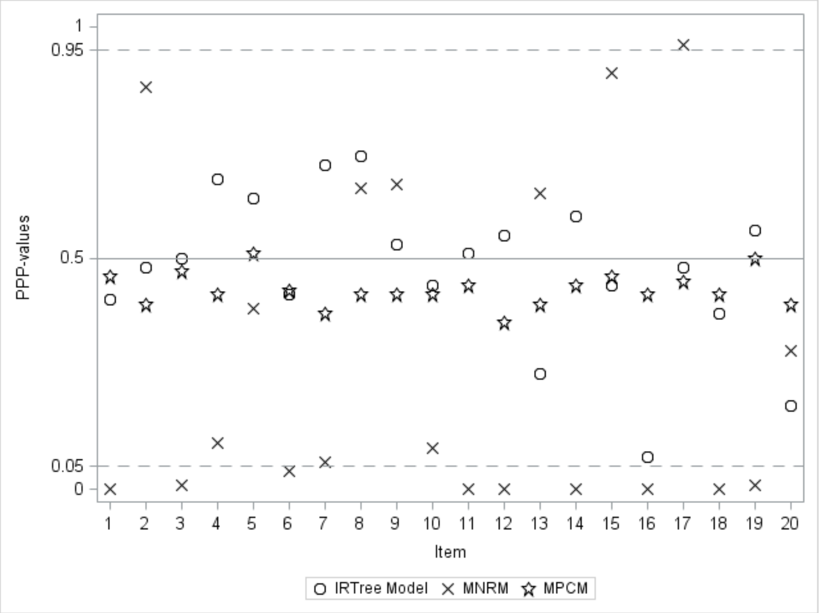


Figure 30. Item extreme response rate *PPP*-values for 500 simulated responses to 20 items with 6 categories under the MNRM

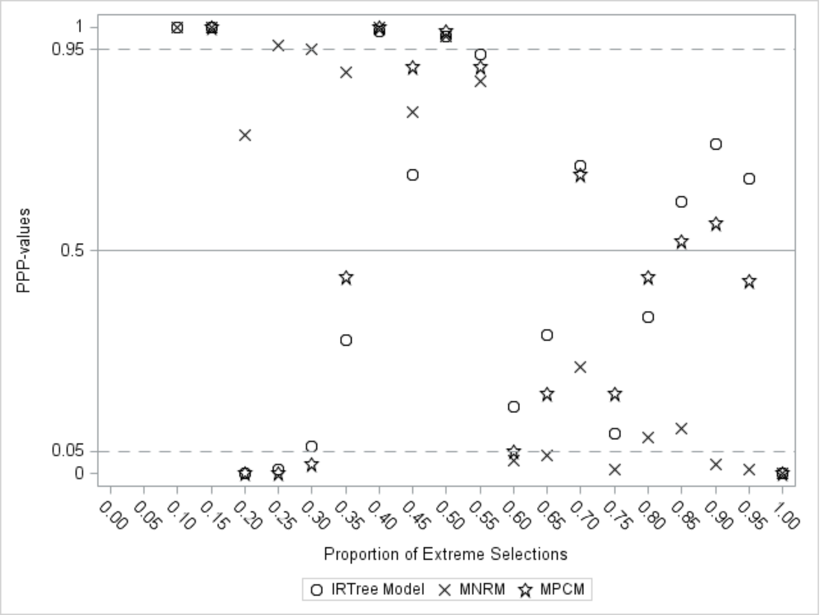


Figure 31. Individual extreme response rate *PPP*-values for 500 simulated responses to 20 items with 6 categories under the MNRM

The second simulation condition considered was 500 responses simulated under the MPCM to 10 items with 4 categories. Figure 32 displays the *PPP*-values for the item extreme response rate discrepancy statistic computed under the IRTree Model, the MNRM, and the MPCM. The *PPP*-values associated with the MNRM and the *PPP*-values associated with the MPCM closely hover around 0.5. This result is an indication of adequate model fit with respect to the item extreme response rates across all items. The *PPP*-values produced under the IRTree Model contain more scatter across the items, however, they all are within the credible interval.

Figure 33 displays the *PPP*-values for the individual extreme response rate discrepancy statistic computed under the IRTree Model, the MNRM, and the MPCM. The *PPP*-values for the MNRM indicate inadequate person-level fit. Most *PPP*-values for the MPCM and the IRTree Model fall within the credible interval but have significant scatter. Three *PPP*-values fall outside the credible interval for the IRTree Model and two *PPP*-values lie outside of the credible interval for the MPCM.   Individual extreme response rates are susceptible to scare frequencies when only 500 responses are analyzed.

The results of the posterior predictive model checking performed using parameter estimates from the IRTree Model, the MNRM, and the MPCM for 500 responses simulated under the MPCM to 10 items with 4 categories were inconsistent. The MPCM and the MNRM were determined to have adequate item-level model fit. The IRTree Model was determined to have a lower degree of item-level fit than the other two models. The item-level model fit results match the pattern of the mean IMSE seen in Table 25**.** The mean IMSE was slightly larger for the IRTree Model than the relatively equal mean IMSE values determined using the MNRM and the MPCM.

The IRTree Model and the MPCM had the best person-level fit when the individual extreme response rate discrepancy statistic was analyzed. The IRTree Model and the MPCM have better quality person-level fit compared to the MNRM. However, this fit is not within the credible interval for all individual response rates. The pattern of person-level fit is inconsistent with the pattern of item-level fit and mean IMSE values. The pattern disagreement may be due to the low sample size considered.

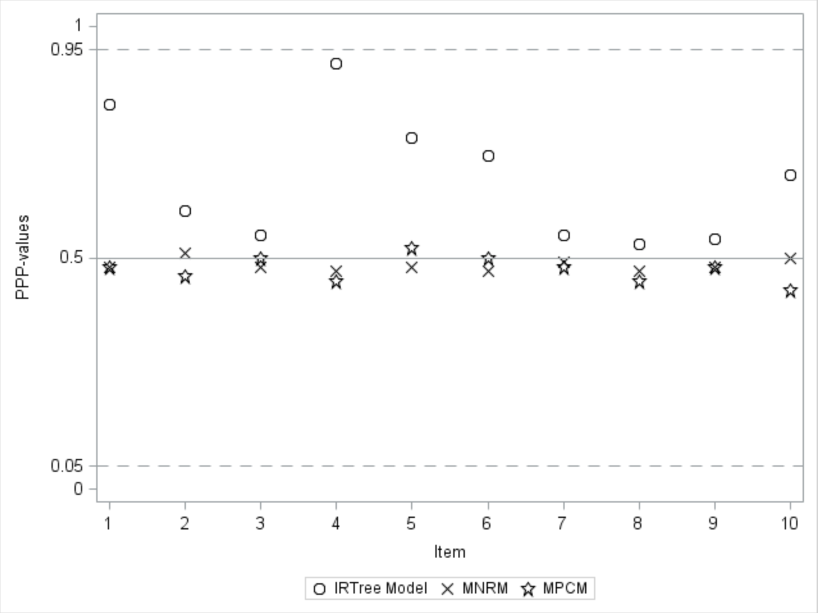


Figure 32. Item extreme response rate *PPP*-values for 500 simulated responses to 10 items with 4 categories under the MPCM

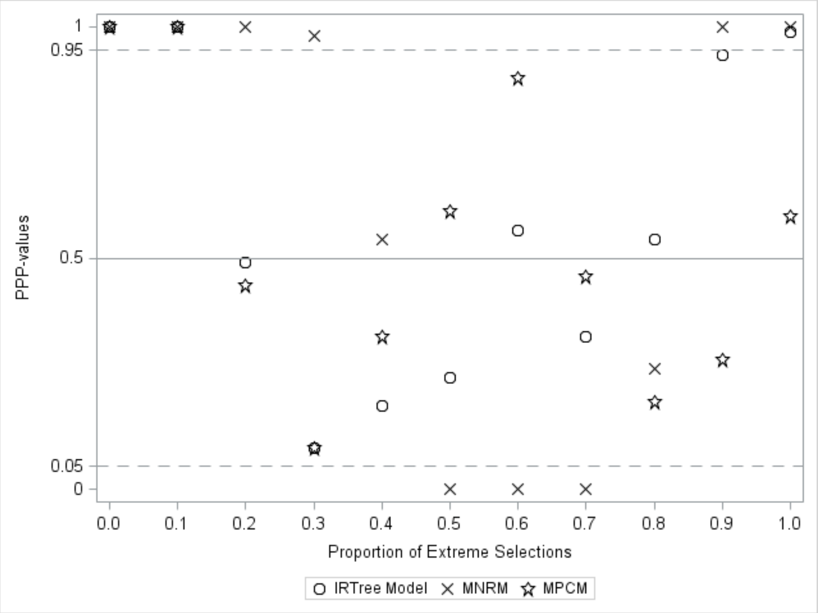


Figure 33. Individual extreme response rate *PPP*-values for 500 simulated responses to 10 items with 4 categories under the MPCM

4.2.7 Summary of Simulation Study

For survey responses simulated using the IRTree Model, the IRTree Model analysis method resulted in the lowest average IMSE for sample sizes 500 and 1000, category response options 4 and 6, and surveys of length 10 items and 20 items. This result was not consistent with results for item responses generated under the MNRM and item responses simulated under the MPCM. The IRTree Model data generation technique resulted in simulated responses that did not exhibit adequate item extreme response rates and marginally exhibited adequate individual extreme response rates using the criteria from Greenleaf (1992b). The IRTree model is designed to model a two-step decision making response process. This accounts for extreme response tendencies differently than the MPCM and the MNRM. This is evident through the data generation and analysis results. The MNRM does fit data generated under the IRTree Model in any conditions. The MPCM fits the data generated under the IRTree Model better (lower mean IMSE) than the MNRM but does not fit the data adequately in the majority of simulation conditions.

Consider data generated under the MPCM and the MNRM. These data exhibit individuals with high extreme response tendencies based on an evaluation of Greenleaf (1992b)'s descriptive statistic conditions. The MPCM resulted in the lowest average IMSE when 500 and 1000 individuals were considered averaged across the number of items on the survey and the number of response options for each item.

Based on the results of the simulation study, the MPCM should be used when responses to surveys exhibit extreme response tendencies. The MPCM will result in an equal to or lower mean IMSE compared to the MNRM and the IRTree Model. The MNRM may be used when responses to 10 items with 4 options are considered. Based on the results of the simulation study, the IRTree Model is not recommended for modelling extreme response tendencies under any condition.

5.0 DISCUSSION

The current study, through an empirical example and a Monte Carlo simulation, evaluated the effectiveness of three models designed to account for extreme response tendencies: the IRTree Model, the Multidimensional Nominal Response Model, and the Modified Generalized Partial Credit Model.

5.1 SUMMARY OF MAJOR FINDINGS - EMPIRICAL STUDY

An empirical study was performed to investigate the relationship between trait estimates and model fit of the IRTree Model, the MNRM, and the MPCM. The data used for analysis was selected to mirror a study performed by Bolt and Newton (2011). 5,330 U.S. respondents were questioned on the "Value of Science" subscale from the Program for International Student Assessment (PISA, 2006). The subscale consisted of ten items with response options: 1-*Strongly Disagree*, 2-*Disagree,* 3-*Agree,* and 4-*Strongly Agree*. The responses exhibited evidence of extreme response tendencies based on the criteria provided by Greenleaf (1992b) using the item extreme response rate and the individual extreme response rate. The model parameters were estimated in the Bayesian paradigm.

5.1.1 Relationship between trait estimates

The empirical study first answered the research question, “How do the trait of interest estimates found using the IRTree model, the MNRM, and the MPCM correlate?”. There was a strong, positive, linear relationship found between the substantive trait estimates from the IRTree Model, the MNRM, and the MPCM.

5.1.2 Relationship between extreme response tendencies

Through the empirical study, the correlation between extreme response tendency estimates found using the IRTree Model, the MNRM, and the MPCM was examined. There was a moderately strong, positive linear relationship between the extreme tendencies estimates from the IRTree model and the MNRM. There was a moderately strong negative relationship between the extreme response trait estimate from the IRTree Model and the natural log of the extreme response tendency estimate under the MPCM. Higher estimates of the extreme tendencies effect under the MPCM coincided with less extreme tendencies, resulting in the negative correlation. There was a very strong negative linear correlation between the extreme response trait estimates from the MNRM and the natural log of the extreme tendencies effect estimated under the MPCM.

5.1.3 Model Fit

The third research question addressed using empirical data was related to model fit. Model fit was evaluated using three broad methods: posterior predictive model checking (PPMC), evaluation of the deviance information criteria, and comparison of mean squared error between observed total score and expected total score. In addition to comparing the fit of the three models accounting for extreme response tendencies, the model fit results from the IRTree Model, the MPCM, and the MNRM were compared to model fit of the graded response model. This was used to evaluate the benefit of including an extreme tendencies trait/effect versus analyzing the responses as unidimensional (GR model).

5.1.3.1 PPMC

Posterior Predictive Model Checking involves comparison of observed discrepancy statistics to discrepancy statistics computed using a simulated distribution of model fit. The simulated distribution of model fit uses sampled model parameters from the estimated joint posterior distribution outputted during the MCMC procedure. The global odds ratio, the mean total score, the standard deviation of the total score, total score group frequencies, item extreme response rates, and individual extreme response rates were discrepancy statistics used to evaluate model fit.

1. Global OR

The global odds ratio was computed for the dichotomy of “extreme categories” versus “non-extreme categories”. In other words, the odds of selecting either *1-Strongly Disagree* or *4-Strongly Agree* over *2-Disagree* or *3-Agree* were evaluated under the observed and predicted datasets. There was inadequate model fit based on the resulting *PPP*-value pie plots for the IRTree Model, the MNRM, and the MPCM. The resulting *PPP*-value pie plots generated using the GR model showed inferior fit than the model fit the IRTree Model, the MNRM, and the MPCM exhibited.

2. Total Score Distribution

The test-level model fit discrepancy statistics evaluated the recovery of the mean and standard deviation of the total score distribution. The IRTree Model overestimated the mean total score and underestimated the standard deviation of the observed total score. The MNRM displayed adequate recovery of the mean total score but overestimated the standard deviation of the observed total score. The MPCM showed adequate recovery of the mean total score and the standard deviation of the total score distribution. Compared to the MNRM, the GR model displayed inferior, but still adequate, recovery of the mean total score and similar lack of recovery of the standard deviation of the observed total score. The GR model recovered the mean total score better than the IRTree Model and similarly did not recover the standard deviation of the total score distribution. The MPCM showed superior model recovery compared to GR model for both the mean total score and standard deviation of the total score.

3. Total Score Groups

Recovery of the frequency of total score groups was investigated to enhance the assessment of model fit with respect to the total score distribution. Total scores that were the result of similar extreme response selection frequencies were grouped. For example, low total scores that exhibited a high frequency of extreme response selections were grouped. Five groups were formed across the total score continuum that represented high, moderate, and low frequencies of extreme response selections. Model fit was evaluated by investigating the recovery of the frequency of individuals falling within each group under the replicated data compared to the observed data. The IRTree Model showed inadequate recovery of frequencies for all five groups. This was significantly worse than the unidimensional GR model. The MNRM and the MPCM showed adequate recovery of the frequencies of individuals in the total score groups.

4. Item-Level

An item-level PPMC discrepancy statistic was developed specifically for models evaluating extreme response tendencies. The recovery of item extreme response rates was evaluated. The proportion of extreme response selections on each item were computed in the observed and replicated response sets. The IRTree Model only adequately recovered the observed extreme response rate for 7 out of 10 items. The extreme response rate was recovered adequately for all ten items under the GR model. Based on the resulting *PPP*-values there was no difference in model fit for the GR model and the MPCM (both displayed adequate fit). The *PPP*-values computed on the item extreme response rates for the MNRM provided more evidence of recovery than the *PPP*-values for both the MPCM and the GR model.

5. Person-Level

A person-level discrepancy statistic was also developed to evaluate models designed to account for extreme response tendencies. The individual extreme response rate was calculated for the observed and replicated predicted datasets. Both the IRTree Model and the GR model showed inadequate model fit with respect to the individual extreme response rate discrepancy statistic. The MNRM showed poor person-level fit and the MPCM showed moderate person-level fit using the individual extreme response rate discrepancy statistic.

5.1.3.2 DIC

A traditional measure of model fit was used in addition to PPMC methods. The deviance information criteria (DIC) was computed for each of the three models along with the GR model to evaluate model fit. The GR model had the highest DIC indicating the worse fit of the four models.

The MPCM displayed the lowest DIC and the IRTree Model resulted in the highest DIC of the three models that account for extreme response tendencies.

5.1.3.3 Mean Squared Error

Analysis of the mean square error between expected total score and observed total score was the final way the IRTree model, the MNRM, the MPCM, and the GR model were evaluated for model fit. The MSE was computed at each iteration of the MCMC to develop a distribution of MSE values for each model. The MNRM had the lowest mean MSE between expected total score and observed total score. The mean MSE calculated using the MPCM was slightly greater than the mean MSE under the MNRM. The IRTree Model had a significantly higher mean MSE compared to the MSE computed under the MNRM and the MSE computed under the MPCM. The GR Model had a lower mean MSE than the IRTree Model but a greater mean MSE than the mean MSE calculated under the MPCM and the MNRM.

5.1.4 Conclusion

Overall the MPCM and the MNRM outperformed the IRTree Model with respect to the PPMC procedures, the DIC, and the MSE between observed total score and expected total score. The MNRM showed better fit than the MPCM using some discrepancy measures but not others. The MPCM had a lower DIC than the MNRM but the MNRM had a lower mean MSE than the MPCM. Based on the evaluation of model fit in the empirical study, there was not a clear choice of the MPCM or the MNRM. Both the MNRM and the MPCM, however, appear to be favored over the IRTree Model.

5.2 SUMMARY OF MAJOR FINDINGS - SIMULATION STUDY

A simulation study was conducted to evaluate model fit using the MSE between expected total score and true total score under different conditions of sample size, survey length, and a varying number of category response options. To compare surveys of varying lengths, the item mean squared error (IMSE) was computed by averaging the MSE over the number of items on the survey evaluated.

5.2.1 Data generation

Data was generated under each of the three models: the IRTree Model, the MNRM, and the MPCM. The item extreme response rates and the individual extreme response rates were used to evaluate the quality of the simulated data with respect to representing individuals who exhibit extreme response tendencies. Responses simulated using the MNRM and the MPCM far exceeded the criteria for representing individuals who exhibit extreme response tendencies. Responses simulated using the IRTree model did not adequately meet the criteria for the individual extreme response rates, but, marginally met the criteria for the item extreme response rates. The results from data generated under the IRTree Model were interpreted with caution due to the IRTree Model exhibiting poor extreme response tendency qualities.

The item parameters of the IRTree Model did not yield as high extreme response tendencies among simulated individual responses as the MNRM and the MPCM. This may be a result of the item parameters selected. The item extreme response rates and individual extreme response rates may yield alternative properties for a different set of item parameters. Even with this expected difference in item and individual extreme response rates, the pattern of item mean squared error is not expected to change.

5.2.2 Effect of Sample Size

The first research question the simulation study addressed related to the pattern of differences in the mean item mean squared error between the expected total score and true total score among the levels of sample size. The pattern of differences of mean IMSE across the levels of sample size was the same for the IRTree Model, the MPCM, and the MNRM. Furthermore, there were no differences in mean IMSE for a sample size of 500 versus a sample size of 1000.

5.2.3 Effect of Survey Length

The second research question addressed through the simulation study explored the pattern of differences in the mean item mean squared error between the expected total score and true total score among the levels of survey length. The pattern of differences of mean IMSE across the levels of survey length interacted with the number of response categories and the model of analysis. The MNRM showed greater average IMSE for surveys consisting of 20 items with 6 response options than for surveys with 10 items with 6 response options. The MNRM showed no difference in mean IMSE for surveys with 20 items with 4 response options and surveys of 10 items with 4 response options. The IRTree Model showed no difference in mean IMSE for surveys of 10 and 20 items.

The MPCM resulted in the lowest mean IMSE across simulation conditions. Moreover, there were no difference in mean IMSE for 10 versus 20 items surveys. This result is not consistent with those found by Jin and Wang (2014). They found that, in general, the longer the test, the better the estimation of person paramters under the MPCM. Jin and Wang (2014) considered survey lenghts of twenty versus 40 items. Thus, differences in IMSE between 20 item surveys versus 40 items surveys analyzed with the IRTree Model, the MNRM, and the MPCM may have been found if this study was performed using similar conditions.

5.2.4 Effect of number of response options

The third research question the simulation study explored related to the pattern of differences in the mean item mean squared error between the expected total score and true total score among the levels of number of category response options. The pattern of differences of mean IMSE across the number of category response options interacted with the number of items. Across survey lengths, for each of the three models, the IMSE was greater when items with 6-response options were evaluated compared to the mean IMSE when items with 4-response options were evaluated. The MNRM and the MPCM resulted in the lowest mean IMSE for 10 items with 4 response categories under MNRM simulated data. The MNRM, however, resulted in the largest mean IMSE for 20 items with 6 response options.

There exist inherent differences in IMSE of expected total score and observed total score for surveys with items containing 4-response options versus surveys with items containing 6-response options. As a result, higher IMSE values for items with 6-response options may be due to these inherent differences or may be due to worse model fit. Jin and Wang (2014) stated that a survey may either be lengthened or the number of response options in each item should be increased to obtain a more precise estimate of the extreme response tendency trait. This result may seem inconsistent with the results presented in this study, however, Jin and Wang (2014) specifically focused on the extreme response tendency trait in their conclusions. They did not have to consider the inherent differences between 4 and 6 response options when considering recovery of the parameter value. In turn, they can make concrete conclusions about the effects of the response options on a single parameter. When considering the MSE in expected and observed total score to make a comparison between uniquely parameterized models, the inherent differences of MSE between items with 4 and 6 response options must be considered.

5.2.5 Conclusion

Overall, the MPCM resulted in the lowest mean IMSE among the IRTree Model, the MPCM, and the MNRM averaged across sample size, survey length, and the number of category response options. Based on the data generated under the MPCM and the MNRM, the MPCM resulted in the lowest mean IMSE. The mean IMSE from MPCM generated responses evaluated with the MNRM was equal to or greater than the mean IMSE of the same responses analyzed with the MPCM. For data generated using the IRTree Model, analysis conducted with IRTree Model resulted in the lowest mean IMSE for all combinations of sample size, survey length, and number of response options.

Consider when the IRTree model was used for analysis for item responses simulated under the MPCM and item responses simulated under the MNRM. The IRTree Model resulted in the highest mean IMSE in most simulation conditions. This, combined with the results that analysis using the MPCM and the MNRM resulted in high IMSE when analyzing data generated under the IRTree Model, provides evidence that the IRTree Model is modelling something unique compared to the MNRM and the MPCM. The IRTree Model is modelling a response process while accounting for extreme response tendencies. The MNRM and the MPCM are capturing the extreme response dimension and extreme response random effect, respectively. In other words, the MNRM and the MPCM are trying to capture extreme response tendencies only. The IRTree Model captures the extreme response tendency and additional information about the sequential-decision response process. Thissen-Roe and Thissen (2013) investigated the correlation between the substantive trait and the ERS trait under the IRTree Model and the MNRM. Similarly, they concluded that the second decision trait may not have been measuring the same construct that the ERS trait in the MNRM was measuring.

Overall, to model extreme response tendencies, the MPCM model is recommended. The MNRM results in approximately equal mean IMSE to the MPCM when 10 item surveys with 4 response options were analyzed and may alternatively be used. Further evaluation of the IRTree Model should be conducted. Although the MNRM and MPCM are recommended, only three models were compared.

5.3 LIMITATIONS AND FUTURE RESEARCH

An empirical study and simulation study were used to evaluate the six proposed research questions. Although the study was carried out after an extensive review of the literature, the study still demonstrated limitations.

5.3.1 Empirical Study

The empirical study made use of an existing dataset used to examine extreme response tendencies. Each of the ten items on the subscale had 4-response options. Preston and Colman (2000) found that four-point scales perform poorly on indices of reliability, validity, and discriminating power. Empirical responses to surveys consisting of items with more than four response options should be explored. Furthermore, Bolt and Newton (2011) found lower bias in extreme response trait estimates when using more than one subscale for analysis at a time. The current study used only one subscale, “Value of Science”. Further analysis and comparison of the three models when multiple subscales are measured at the same time should be completed. This should include independent substantive traits but common ERS traits across the subscales.

Two new discrepancy measures were used during the posterior predictive model checking phase of the empirical study. The item extreme response rate and the individual extreme response rate should be evaluated further. The two discrepancy measures may be evaluated in terms of distinguishing models that account for extreme response tendencies from those that do not when the responses exhibit high extreme response patterns.

5.3.2 Simulation Study

The simulation conditions were chosen to represent a variety of practical applications. The simulation conditions, however, are still limited to those selected. It may be a limiting factor that the results from the current study cannot be generalized beyond the conditions explored. In other words, the patterns exhibited by the mean IMSE for sample sizes 500 and 1000, survey lengths of 10 and 20 items, and number of response options 4 or 6, may not extend to alternative conditions.

The study was limited to looking only at items with an even number of response options. A further investigation should be conducted to determine the effects of an odd number of categories and the middle response style effect. Questions about individuals drawn to extreme responses versus individuals drawn to neutral or middle categories should be addressed.

The use of Bayesian MCMC procedures is both a strength and limitation of the study. The use of Bayesian MCMC procedures allows the direct translation and estimation of the multidimensional models. The MCMC procedure, however, is a time expensive endeavor. Due to estimation time, only 25 replications were used for each simulation condition. Estimation of the IMSE would have benefitted from an increased number of replications per condition, although, previous studies have had fewer replications per condition when estimating models accounting for extreme response tendencies (see Jin and Wang (2014)).

Extreme response tendencies and the substantive traits were assumed to be independent for both the empirical study and the simulation study. This assumption was validated through an extensive literature review. Recent research, however, suggests that extreme response style may be domain specific (Cabooter, Weijters, de Beuckelaer, & Davidov, 2016). Further investigation of extreme response tendencies should incorporate correlated substantive traits and extreme response tendencies.

There exist new and alternative models in addition to the IRTree Model, the MNRM, and the MPCM as development of MIRT models accounting for extreme response style is still in its infancy. For example, Thissen-Roe and Thissen (2013) presented a reparametrized version of the proportional thresholds model (PTM, Rossi, Gilula, & Allenby, 2001) to account for extreme response style. Supplementary evaluation of alternative models accounting for extreme response should be performed and compared to the IRTree Model, the MNRM, and the MPCM.

APPENDIX A

EMPIRICAL STUDY ITEM PARAMETER ESTIMATES

Table 26. Posterior mean item parameter estimates under the IRTree Model, the MNRM, and the MPCM using responses from the Value of Science Subscale

	IRTree		MNRM		MPCM	
Item 1	a_1	1.6767	c_1	2.5786	a	1.1331
	b_1	3.3064	c_2	3.2516	b	1.8107
	a_2	2.0561	c_3	-0.1687	τ_1	-1.9507
	b_2	-0.9267	c_4	-5.6615	τ_2	0.3462
	ν	-0.0645			τ_3	1.6045
Item 2	a_1	2.0029	c_1	3.0081	a	1.5340
	b_1	4.1165	c_2	3.4506	b	1.7299
	a_2	2.4106	c_3	-0.4792	τ_1	-1.7251
	b_2	-0.6549	c_4	-5.9795	τ_2	0.3953
	ν	-0.0274			τ_3	1.3298
Item 3	a_1	1.3741	c_1	0.1641	a	1.1733
	b_1	1.1823	c_2	2.7005	b	0.8359
	a_2	1.8549	c_3	1.3128	τ_1	-2.4160

Table 26 (continued).

	IRTree		MNRM		MPCM	
	b_2	-2.0695	c_4	-4.1773	τ_2	0.0777
	ν	0.0231			τ_3	
Item 4	a_1	1.2908	c_1	1.6466	a	1.1206
	b_1	2.3647	c_2	3.1427	b	1.4540
	a_2	2.1385	c_3	0.3964	τ_1	-2.2451
	b_2	-1.5240	c_4	-5.1857	τ_2	0.3224
	ν	-0.0601			τ_3	1.9227
Item 5	a_1	2.7561	c_1	0.4075	a	2.1126
	b_1	1.6845	c_2	2.3342	b	0.6622
	a_2	2.2318	c_3	0.9910	τ_1	-1.5075
	b_2	-1.8457	c_4	-3.7327	τ_2	0.0205
	ν	0.0764			τ_3	1.487
Item 6	a_1	2.3524	c_1	2.0798	a	2.1919
	b_1	3.6234	c_2	3.2141	b	1.2519
	a_2	3.2451	c_3	0.0407	τ_1	-1.6617
	b_2	-1.3890	c_4	-5.3346	τ_2	0.2534
	ν	-.00418			τ_3	
Item 7	a_1	3.0017	c_1	-0.1150	a	2.4901
	b_1	1.2521	c_2	2.1882	b	0.4938
	a_2	2.3958	c_3	1.2156	τ_1	-1.5370
	b_2	-2.6268	c_4	-3.2888	τ_2	-0.0194
	ν	0.1477			τ_3	1.5564

Table 26 (continued).

	IRTree		MNRM		MPCM	
Item 8	a_1	2.3605	c_1	0.8954	a	2.0779
	b_1	2.3585	c_2	2.7069	b	0.8523
	a_2	2.6021	c_3	0.6006	τ_1	-1.6437
	b_2	-1.7263	c_4	-4.2029	τ_2	-0.0194
	ν	0.0428			τ_3	1.6631
Item 9	a_1	1.6147	c_1	0.5402	a	1.3460
	b_1	1.5442	c_2	2.6652	b	0.8970
	a_2	2.1070	c_3	0.9913	τ_1	-2.0759
	b_2	-1.8395	c_4	-4.1967	τ_2	0.1085
	ν	0.0910			τ_3	1.9674
Item 10	a_1	2.4318	c_1	0.3465	a	1.8370
	b_1	1.5603	c_2	2.3176	b	0.6589
	a_2	2.0147	c_3	0.9360	τ_1	-1.5487
	b_2	-1.7175	c_4	-3.6001	τ_2	0.0494
	ν	0.0910			τ_3	1.4993

APPENDIX B

SIMULATION STUDY DATA GENERATION STATISTICS

Table 27. Average item extreme response rates for responses generated under the IRTree Model

	Sample Size							
	500				1000			
	Response Options							
	4		6		4		6	
	Number of Items							
Item	10	20	10	20	10	20	10	20
1	0.45	0.46	0.37	0.37	0.54	0.53	0.32	0.33
2	0.51	0.48	0.35	0.35	0.55	0.50	0.35	0.31
3	0.52	0.52	0.39	0.34	0.51	0.48	0.34	0.37
4	0.50	0.46	0.41	0.35	0.42	0.51	0.30	0.33
5	0.48	0.49	0.45	0.39	0.52	0.51	0.31	0.40
6	0.46	0.50	0.37	0.34	0.45	0.47	0.34	0.40
7	0.52	0.51	0.37	0.31	0.48	0.47	0.35	0.34
8	0.45	0.48	0.37	0.36	0.48	0.44	0.35	0.39
9	0.43	0.47	0.36	0.38	0.52	0.55	0.33	0.41
10	0.56	0.48	0.41	0.37	0.41	0.49	0.30	0.34
11		0.47		0.32		0.52		0.33
12		0.48		0.41		0.47		0.33
13		0.44		0.32		0.51		0.42
14		0.48		0.34		0.48		0.34
15		0.53		0.35		0.56		0.39
16		0.53		0.36		0.55		0.38
17		0.48		0.33		0.45		0.33
18		0.49		0.41		0.55		0.39
19		0.48		0.38		0.49		0.43
20		0.47		0.42		0.53		0.41

Table 28. Average individual extreme rate cumulative percentage for responses generated using the IRTree Model

	Number of Items								
	10					20			
	Sample Size								
	500		1000			500		1000	
	Response Options								
Prop	4	6	4	6	**Prop**	4	6	4	6
0.0	0.12	0.74	0.04	2.30	0.0	0.01	0.03	0.00	0.01
0.1	1.62	5.39	0.96	11.73	0.05	0.01	0.28	0.00	0.13
0.2	7.38	19.70	5.63	31.93	0.1	0.04	1.50	0.01	0.91
0.3	20.11	42.78	18.63	57.35	0.15	0.17	4.86	0.08	3.27
0.4	40.45	66.87	40.82	79.53	0.2	0.95	12.19	0.36	9.49
0.5	64.14	84.93	65.82	92.52	0.25	3.34	23.50	1.58	20.80
0.6	83.95	94.84	85.16	98.32	0.3	8.27	39.00	4.92	36.71
0.7	94.89	98.70	95.77	99.74	0.35	17.18	56.32	11.49	53.67
0.8	99.08	99.88	99.39	100.00	0.4	30.22	71.54	23.07	70.05
0.9	99.92	100.01	99.94	100.01	0.45	47.22	83.57	39.09	82.82
1.0	100.00	100.01	100.00	100.01	0.5	63.51	91.48	57.29	91.28
					0.55	77.95	96.49	74.30	96.28
					0.6	88.36	98.62	87.22	98.65
					0.65	94.94	99.64	94.70	99.59
					0.7	98.16	99.90	98.36	99.88
					0.75	99.51	99.97	99.65	99.96
					0.8	99.92	99.98	99.96	99.99
					0.85	99.98	99.98	99.99	99.99
					0.9	100.00	99.98	99.99	99.99
					0.95	100.00	99.98	99.99	99.99
					1.0	100.00	99.98	99.99	99.99

Table 29. Average item extreme response rates for responses generated under the MNRM

	Sample Size							
	500				1000			
	Response Options							
	4		6		4		6	
	Number of Items							
Item	10	20	10	20	10	20	10	20
1	0.66	0.64	0.55	0.59	0.66	0.75	0.56	0.58
2	0.73	0.67	0.61	0.53	0.52	0.66	0.55	0.57
3	0.65	0.65	0.53	0.51	0.71	0.69	0.69	0.56
4	0.59	0.66	0.55	0.55	0.65	0.68	0.53	0.61
5	0.68	0.58	0.63	0.61	0.73	0.64	0.66	0.54
6	0.70	0.69	0.53	0.57	0.65	0.56	0.58	0.58
7	0.60	0.64	0.49	0.55	0.65	0.69	0.57	0.53
8	0.63	0.57	0.55	0.57	0.64	0.67	0.51	0.53
9	0.57	0.68	0.56	0.61	0.60	0.65	0.66	0.64
10	0.58	0.55	0.56	0.57	0.64	0.55	0.49	0.60
11		0.61		0.58		0.50		0.51
12		0.61		0.57		0.54		0.61
13		0.55		0.62		0.63		0.47
14		0.64		0.57		0.62		0.60
15		0.58		0.60		0.64		0.57
16		0.60		0.55		0.63		0.58
17		0.62		0.55		0.54		0.60
18		0.63		0.60		0.70		0.64
19		0.58		0.63		0.55		0.49
20		0.57		0.68		0.59		0.57

Table 30. Average individual extreme rate cumulative percentage for responses generated using the MNRM

	Number of Items								
	10					20			
	Sample Size								
	500		1000			500		1000	
	Response Options								
Prop	4	6	4	6	**Prop**	4	6	4	6
0.0	0.56	0.46	0.78	0.60	0.0	0.12	0.02	0.21	0.04
0.1	2.37	2.96	2.77	3.67	0.05	0.56	0.11	0.66	0.24
0.2	5.86	8.98	6.20	8.54	0.1	1.28	0.49	1.64	0.75
0.3	12.29	19.72	12.04	17.12	0.15	2.75	1.38	2.92	1.85
0.4	22.05	33.39	20.50	29.08	0.2	4.83	3.02	4.89	3.68
0.5	34.88	49.97	32.18	43.61	0.25	7.74	5.93	7.64	7.02
0.6	48.91	65.49	46.10	59.51	0.3	11.07	9.88	10.96	11.59
0.7	63.77	78.23	62.28	75.41	0.35	15.43	15.44	15.19	17.62
0.8	78.70	89.74	78.39	87.40	0.4	20.90	22.06	20.42	24.89
0.9	90.64	96.00	91.58	95.21	0.45	26.92	30.76	26.30	32.98
1.0	99.99	99.98	100.00	99.99	0.5	33.60	40.70	32.89	42.23
					0.55	40.58	50.12	40.07	51.26
					0.6	48.64	59.66	47.40	60.42
					0.65	56.79	68.67	55.08	68.99
					0.7	64.89	76.57	62.92	76.86
					0.75	73.08	82.97	71.09	83.56
					0.8	80.72	88.55	78.61	89.30
					0.85	87.34	92.93	85.68	93.74
					0.9	92.92	96.12	91.71	96.80
					0.95	97.04	98.32	96.73	98.93
					1.0	99.98	99.99	99.99	100.00

Table 31. Average item extreme response rates for responses generated under the MPCM

	Sample Size							
	500				1000			
	Response Options							
	4		6		4		6	
	Number of Items							
Item	10	20	10	20	10	20	10	20
1	0.56	0.58	0.45	0.47	0.58	0.58	0.46	0.46
2	0.58	0.58	0.50	0.49	0.58	0.60	0.42	0.49
3	0.59	0.58	0.46	0.46	0.57	0.58	0.44	0.51
4	0.56	0.58	0.47	0.49	0.59	0.60	0.43	0.50
5	0.58	0.57	0.50	0.49	0.59	0.58	0.48	0.45
6	0.57	0.61	0.47	0.44	0.57	0.60	0.47	0.49
7	0.58	0.59	0.47	0.46	0.57	0.58	0.44	0.47
8	0.58	0.58	0.47	0.48	0.59	0.57	0.46	0.49
9	0.55	0.55	0.44	0.42	0.55	0.60	0.47	0.45
10	0.59	0.59	0.45	0.47	0.56	0.56	0.46	0.47
11		0.55		0.44		0.57		0.51
12		0.59		0.47		0.59		0.44
13		0.59		0.48		0.59		0.45
14		0.60		0.47		0.58		0.40
15		0.57		0.47		0.58		0.49
16		0.56		0.45		0.57		0.45
17		0.59		0.41		0.57		0.45
18		0.62		0.44		0.61		0.47
19		0.57		0.46		0.55		0.46
20		0.62		0.50		0.60		0.45

Table 32. Average individual extreme rate cumulative percentage for responses generated using under the MPCM

	Number of Items								
	10					20			
	Sample Size								
	500		1000			500		1000	
	Response Options								
Prop	4	6	4	6	**Prop**	4	6	4	6
0.0	0.13	0.54	0.07	0.94	0.0	0.01	0.08	0.00	0.10
0.1	0.79	3.70	0.76	4.84	0.05	0.02	0.27	0.01	0.34
0.2	3.41	12.11	3.31	14.22	0.1	0.06	1.00	0.04	0.96
0.3	10.43	27.00	10.41	30.08	0.15	0.23	2.63	0.16	2.30
0.4	23.85	46.97	24.31	50.78	0.2	0.52	5.40	0.47	4.84
0.5	44.07	66.64	44.33	70.17	0.25	1.48	10.20	1.23	9.45
0.6	66.13	83.27	65.89	84.80	0.3	3.25	17.73	2.98	16.31
0.7	84.19	93.21	83.76	93.78	0.35	6.56	27.28	6.43	25.95
0.8	95.20	98.18	94.62	98.00	0.4	12.66	39.58	12.37	38.31
0.9	99.28	99.75	98.83	99.59	0.45	21.39	53.40	21.01	51.99
1.0	100.00	100.01	100.00	100.01	0.5	33.49	66.78	32.92	65.12
					0.55	47.67	77.49	47.34	77.04
					0.6	61.66	85.97	61.88	86.00
					0.65	75.16	92.00	75.36	91.98
					0.7	85.55	96.09	85.60	95.87
					0.75	92.33	98.08	92.71	98.03
					0.8	96.30	99.13	96.73	99.06
					0.85	98.56	99.63	98.74	99.56
					0.9	99.66	99.85	99.56	99.88
					0.95	99.92	99.96	99.91	99.97
					1.0	100.00	99.98	100.00	100.01

APPENDIX C

EXAMPLE SAS SYNTAX

C.1 MNRM – ITEMS WITH 4 CATEGORIES

```
proc mcmc data=analyzeset outpost =analyzeMNRM seed = 23 nbi=5000
            nmc=20000 diagnostics=none plots = all nthreads=-1 DIC;
array theta[2];
array c[10,4]     c1_1-c1_4
                        c2_1-c2_4
                        c3_1-c3_4
                        c4_1-c4_4
                        c5_1-c5_4
                        c6_1-c6_4
                        c7_1-c7_4
                        c8_1-c8_4
                        c9_1-c9_4
                        c10_1-c10_4;
array i[10];
array z[4];
array p[4];
parms c1_2-c1_4 0;
parms c2_2-c2_4 0;
parms c3_2-c3_4 0;
parms c4_2-c4_4 0;
parms c5_2-c5_4 0;
parms c6_2-c6_4 0;
parms c7_2-c7_4 0;
parms c8_2-c8_4 0;
parms c9_2-c9_4 0;
parms c10_2-c10_4 0;
c1_1 = -(c1_2+c1_3+c1_4);
c2_1 = -(c2_2+c2_3+c2_4);
c3_1 = -(c3_2+c3_3+c3_4);
c4_1 = -(c4_2+c4_3+c4_4);
c5_1 = -(c5_2+c5_3+c5_4);
c6_1 = -(c6_2+c6_3+c6_4);
```

```
c7_1 = -(c7_2+c7_3+c7_4);
c8_1 = -(c8_2+c8_3+c8_4);
c9_1 = -(c9_2+c9_3+c9_4);
c10_1 = -(c10_2+c10_3+c10_4);

prior c1_2-c1_4 c2_2-c2_4 c3_2-c3_4 c4_2-c4_4 c5_2-c5_4
            c6_2-c6_4 c7_2-c7_4 c8_2-c8_4 c9_2-c9_4 c10_2-c10_4
            ~normal(0, var=25);

random theta1~normal(0,var=1) subject=_obs_;
random theta2~normal(0,var=1) subject=_obs_;
llike=0;
do j = 1 to 10;
            z[1] =exp((-3)*theta[1]+(1)*theta[2]+c[j,1]);
            z[2] =exp((-1)*theta[1]+(-1)*theta[2]+c[j,2]);
            z[3] =exp((1)*theta[1]+(-1)*theta[2]+c[j,3]);
            z[4] =exp((3)*theta[1]+(1)*theta[2]+c[j,4]);
      do k = 1 to 4;
            p[k] = z[k]/(z1+z2+z3+z4);
      end;
      llike = llike+log(p[i[j]]);
end;
model general(llike);
run;
```

C.2 MPCM – ITEMS WITH 4 CATEGORIES

```
      proc mcmc data=analyzeset outpost =analyzeMPCM seed = 23 nbi=5000
nmc=20000   diagnostics=none plots = all nthreads=-1 DIC;
            array i[10];
            array a[10];
            array b[10];
            array p[10,4];
            array tau1[10];
            array tau2[10];
            array Q[10,4];
            array denom[10];
            array tau3[10];**;
            parms a1 b1 tau11 tau21 1;
            parms a2 b2 tau12 tau22 1;
            parms a3 b3 tau13 tau23 1;
            parms a4 b4 tau14 tau24 1;
            parms a5 b5 tau15 tau25 1;
            parms a6 b6 tau16 tau26 1;
            parms a7 b7 tau17 tau27 1;
            parms a8 b8 tau18 tau28 1;
            parms a9 b9 tau19 tau29 1;
            parms a10 b10 tau110 tau210 1;
            parms psi 1;

            prior psi ~gamma(1,scale=10);
            sigma = 1/psi;
            prior a: ~lognormal(0, var=2);
            prior b: ~normal(0,var = 2);
            prior tau1: ~normal(0, prec=.1);
            prior tau2: ~normal(0,prec=.1);
            lprior=lpdflnorm(omega, 0,sigma);

            random theta~normal(0,var=1) subject = _obs_;
            random omega~general(lprior) subject = _obs_ init=(1);
            llike=0;
            do j = 1 to 10;
                  Q[j,1] = 1;
                  Q[j,2]=Q[j,1]*exp(a[j]*(theta-(b[j]+omega*tau1[j])));
                  Q[j,3]=Q[j,2]*exp(a[j]*(theta-(b[j]+omega*tau2[j])));
                  tau3[j] = -(tau1[j] + tau2[j]);
                  Q[j,4]=Q[j,3]*exp(a[j]*(theta-(b[j]+omega*(tau3[j]))));
                  denom[j] = Q[j,1]+Q[j,2]+Q[j,3]+Q[j,4];
                  do k = 1 to 4;
                        p[j,k]=Q[j,k]/denom[j];
                  end;
                  llike=llike +log(p[j,i[j]]);
            end;
            model general(llike);
      run;
```

C.3 IRTREE MODEL – ITEMS WITH 4 CATEGORIES

```
proc mcmc data=analyzeset outpost= analyzeIRTree seed = 23 nbi=5000 nmc=20000
diagnostics=none plots = all nthreads=-1 DIC;
      array theta[2]; array mu[2]; array sigma[2,2];
array a1[10] a1_1-a1_10;
array b1[10] b1_1-b1_10;
array a2[10] a2_1-a2_10;
array b2[10] b2_1-b2_10;
array nu[10] nu_1 - nu_10;
array i[10] i1-i10;
array p[4];
*initialize params for ability 1;
parms a1_1 1 b1_1 0;
parms a1_2 1 b1_2 0;
parms a1_3 1 b1_3 0;
parms a1_4 1 b1_4 0;
parms a1_5 1 b1_5 0;
parms a1_6 1 b1_6 0;
parms a1_7 1 b1_7 0;
parms a1_8 1 b1_8 0;
parms a1_9 1 b1_9 0;
parms a1_10 1 b1_10 0;
*initialize params for ability (ERS);
parms a2_1 1 b2_1 0 nu_1 0;
parms a2_2 1 b2_2 0 nu_2 0;
parms a2_3 1 b2_3 0 nu_3 0;
parms a2_4 1 b2_4 0 nu_4 0;
parms a2_5 1 b2_5 0 nu_5 0;
parms a2_6 1 b2_6 0 nu_6 0;
parms a2_7 1 b2_7 0 nu_7 0;
parms a2_8 1 b2_8 0 nu_8 0;
parms a2_9 1 b2_9 0 nu_9 0;
parms a2_10 1 b2_10 0 nu_10 0;


prior a1_1-a1_10 ~normal(0, var=16,lower=0);
prior a2_1-a2_10 ~normal(0,var=16,lower=0);
prior b1_1-b1_10 ~normal(0,var = 16);
prior b2_1-b2_10 ~normal(0,var = 16);
prior nu_1-nu_10 ~normal(0, var = 16);
parms rho;
begincnst;
sigma[1,1]=1;
sigma[2,2]=1;
mu[1]=0; mu[2]=0;
endcnst;
sigma[1,2]=rho; sigma[2,1]=rho;
prior rho~normal(0, var=.25, lower=-1, upper=1);
random theta~mvn(mu, sigma) subject=_obs_;

llike=0;
do j = 1 to 10;
    p_agree=exp(a1[j]*theta[1] - b1[j])/(1+exp(a1[j]*theta[1] - b1[j]));
      p_disagree = 1-p_agree;
      p_ext1 = 1/(1+exp(-(b2[j]+a2[j]*theta[2]-
            nu[j]*a2[j]*(b1[j]+a1[j]*theta[1])))));
    p_nonext1 = 1-p_ext1;
      p_ext4 = 1/(1+exp(-
            (b2[j]+a2[j]*theta[2]+nu[j]*a2[j]*(b1[j]+a1[j]*theta[1])))));
      p_nonext4 = 1-p_ext4;
    p[1] = p_disagree*p_ext1;
    p[2] = p_disagree*p_nonext1;
    p[3] = p_agree*p_nonext4;
    p[4] = p_agree*p_ext4;
    llike = llike+log(p[i[j]]);
end;
model general(llike);
run;
```

C.4 IRTREE MODEL – ITEMS WITH 6 CATEGORIES

```
proc mcmc data=analyzeset outpost=analyzeIRTree seed = 23 nbi=5000
            nmc=20000 diagnostics=none plots = all nthreads=-1 DIC;
      array theta[2]; array mu[2]; array sigma[2,2];
array a1[10] a1_1-a1_10; *10 items on theta 1;
array b1[10] b1_1-b1_10;
array a2[10] a2_1-a2_10; *all items have ERS trait;
array c1[10] c1_1-c1_10;
array c2[10] c2_1-c2_10;
array nu[10] nu_1-nu_10;
array i[10] i1-i10;
array p[6];
*initialize params for ability 1;
parms a1_1 1 b1_1 0;
parms a1_2 1 b1_2 0;
parms a1_3 1 b1_3 0;
parms a1_4 1 b1_4 0;
parms a1_5 1 b1_5 0;
parms a1_6 1 b1_6 0;
parms a1_7 1 b1_7 0;
parms a1_8 1 b1_8 0;
parms a1_9 1 b1_9 0;
parms a1_10 1 b1_10 0;


*initialize params for ability (ERS);
parms a2_1 1 c1_1 1 c2_1 0 nu_1 0;
parms a2_2 1 c1_2 1 c2_2 0 nu_2 0;
parms a2_3 1 c1_3 1 c2_3 0 nu_3 0;
parms a2_4 1 c1_4 1 c2_4 0 nu_4 0;
parms a2_5 1 c1_5 1 c2_5 0 nu_5 0;
parms a2_6 1 c1_6 1 c2_6 0 nu_6 0;
parms a2_7 1 c1_7 1 c2_7 0 nu_7 0;
parms a2_8 1 c1_8 1 c2_8 0 nu_8 0;
parms a2_9 1 c1_9 1 c2_9 0 nu_9 0;
parms a2_10 1 c1_10 1  c2_10 0 nu_10 0;




prior a1_1-a1_10 ~normal(0, var=16,lower=0);
prior a2_1-a2_10 ~normal(0,var=16,lower=0);
prior b1_1-b1_10 ~normal(0,var = 16);
prior c2_1-c2_10 ~normal(0,var = 16);
prior c1_1 ~normal(0,var=16, lower=c2_1);
prior c1_2 ~normal(0,var=16, lower=c2_2);
prior c1_3 ~normal(0,var=16, lower=c2_3);
prior c1_4 ~normal(0,var=16, lower=c2_4);
prior c1_5 ~normal(0,var=16, lower=c2_5);
prior c1_6 ~normal(0,var=16, lower=c2_6);
prior c1_7 ~normal(0,var=16, lower=c2_7);
prior c1_8 ~normal(0,var=16, lower=c2_8);
prior c1_9 ~normal(0,var=16, lower=c2_9);
prior c1_10 ~normal(0,var=16, lower=c2_10);


prior nu_1-nu_10 ~normal(0, var = 16);
parms rho;
begincnst;
sigma[1,1]=1;
sigma[2,2]=1;
mu[1]=0; mu[2]=0;
endcnst;
```

```
sigma[1,2]=rho; sigma[2,1]=rho;
prior rho~normal(0, var=.25, lower=-1, upper=1);
random theta~mvn(mu, sigma) subject=_obs_;
llike=0;
do j = 1 to 10;
      p_agree=exp(a1[j]*theta[1] - b1[j])/(1+exp(a1[j]*theta[1] - b1[j]));
      p_ext3 = 1/(1+exp(-(c2[j]+a2[j]*theta[2]-
            nu[j]*a2[j]*(b1[j]+a1[j]*theta[1])))));
      p_ext2 = 1/(1+exp(-(c1[j]+a2[j]*theta[2]-
            nu[j]*a2[j]*(b1[j]+a1[j]*theta[1]))))
            -(1/(1+exp(-(c2[j]+a2[j]*theta[2]-
            nu[j]*a2[j]*(b1[j]+a1[j]*theta[1]))))));
      p_ext1 = 1-(1/(1+exp(-(c1[j]+a2[j]*theta[2]-
            nu[j]*a2[j]*(b1[j]+a1[j]*theta[1]))))));
      p_ext4 = 1/(1+exp(-
            (c2[j]+a2[j]*theta[2]+nu[j]*a2[j]*(b1[j]+a1[j]*theta[1])))));
      p_ext5 = 1/(1+exp(-
            (c1[j]+a2[j]*theta[2]+nu[j]*a2[j]*(b1[j]+a1[j]*theta[1]))))
            - (1/(1+exp(-
            (c2[j]+a2[j]*theta[2]+nu[j]*a2[j]*(b1[j]+a1[j]*theta[1]))))));
      p_ext6 = 1-(1/(1+exp(-
            (c1[j]+a2[j]*theta[2]+nu[j]*a2[j]*(b1[j]+a1[j]*theta[1]))))));
      p[1]= (1-p_agree)*p_ext1;
      p[2]=(1-p_agree)*p_ext2;
      p[3]=(1-p_agree)*p_ext3;
      p[4]=p_agree*p_ext4;
      p[5]=p_agree*p_ext5;
      p[6]=p_agree*p_ext6;
    llike = llike+log(p[i[j]]);
end;
model general(llike);
run;
```

BIBLIOGRAPHY

Agresti, A. (2002). *Categorical data analysis* (Vol. 2nd). New York: Wiley-Interscience.

Baumgartner, H. (2001). Response styles in marketing tesearch: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156.

Berg, I. A. (1953). The reliability of extreme position response sets in two tests. *The Journal of Psychology, 36*(1), 3-9.

Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement, 13*(2), 164-169.

Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *The Public Opinion Quarterly, 51*(2), 220-232.

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

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665-678.

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*(5), 335-352.

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

Cabooter, E., Weijters, B., de Beuckelaer, A., & Davidov, E. (2016). Is extreme response style domain specific? Findings from two studies in four countries. *Quality and Quantity*, 1-18.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Congdon, P. (2002). Bayesian statistical modelling. *Measurement Science and Technology, 13*, 643.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology, 60*(2), 151-174.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*(434), 883-904.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*(4), 475.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3-31.

Cronbach, L. J., Snow, R. E., & Wiley, D. E. (1991). *Improving inquiry in social science: a volume in honor of Lee J. Cronbach*. Hillsdale, N.J: Lawrence Erlbaum Associates.

de Ayala, R. J. (2013). *The theory and practice of item response theory*: Guilford Publications.

Edwards, M. L., & Smith, B. C. (2016). The effects of the neutral response option on the extremeness of participant responses. *Journal of Undergraduate Scholarship, 6*.

Fox, J.-P. (2010). *Bayesian item response modeling: theory and applications* (Vol. 1. Aufl.). New York, NY: Springer.

Gelman, A. (2014). *Bayesian data analysis* (Vol. Thirdition.). Boca Raton: CRC Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733-807.

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

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

Hart, H. N. (1923). *Progress report on a test of social attitudes and interests*. Iowa City: The University.

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

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296-309.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological bulletin, 55*(4), 243-252.

Javaras, K. N., & Ripley, B. D. (2007). An "Unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*(478), 454-463.

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

Johns, R. A. (2005). One size doesn't fit all: Selecting response scales for BES attitude items. *Journal of Elections, Public Opinion and Parties, 15*(2), 237-264.

Johnson, T. R., & 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), 92-114.

Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Journal of the Royal Statistical Society. Series D (The Statistician), 29*(1), 65-78.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., . . . Conaway, M. (2002). The impact of "No Opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *The Public Opinion Quarterly, 66*(3), 371-403.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519-537.

Lewis, N. A., & Taylor, J. A. (1955). Anxiety and extreme response preferences. *Educational and Psychological Measurement*, *15*(2), 111-116.

Liu, M., Conrad, F. G., & Lee, S. (2016). Comparing acquiescent and extreme response styles in face-to-face and web surveys. *Quality and Quantity*, 1-18.

Lo, K.-Y. (2001). *Interpersonal harmony and the values of forbearance: Understanding generation gap through interpersonal conflicts*. Retrieved from NSC90-2413-H-031-006-SSS. Taipei, Taiwan: NSC Reserach Report:

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS nook: A practical introduction to Bayesian analysis*: Taylor & Francis.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology, 23*(4), 498-509.

Martin, J. (1964). Acquiescence—Measurement and Theory*. *British Journal of Social and Clinical Psychology, 3*(3), 216-225. doi:10.1111/j.2044-8260.1964.tb00430.x

Meijer, R. R., & Glas, C. A. W. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27*(3), 217-233.

Merrens, M. (1970). Generality and stability of extreme response style. *Psychological Reports, 27*(3), 802-802.

Messick, S. (1966). The psychology of acquiescence: An interpretation of research evidence. *ETS Research Bulletin Series, 1966*(1), 1-44.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of personality, 77*(1), 261-286.

Paulhus, D. L. (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* (pp. 17-59). San Diego, CA, US: Academic Press.

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.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15.

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *The Journal of Social Psychology, 121*(1), 81-96.

Rorer, L. G. (1965). The great response-style myth. *Psychological bulletin, 63*(3), 129-156.

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.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151-1172.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Vol. no. 17): University of New Brunswick.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: experiments on question form, wording, and context*. New York: Academic Press.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*(4), 375-394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*(2), 429-449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298-321.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Angelika van der, L. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 64*(4), 583-639.

Stanley Budner, N. Y., & Budner, S. (1962). Intolerance of ambiguity as a personality variable. *Journal of personality, 30*(1), 29-50.

Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. Cary, NC, USA: SS Institute Inc.

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.

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data From Six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346-360.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Watkins, D., & Cheung, S. (1995). Culture, gender, and response bias: An analysis of responses to the self-description questionnaire. *Journal of Cross-Cultural Psychology, 26*(5), 490.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*(2), 105-121.

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods, 15*(1), 96-110.

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.

Wyer, R. S. (1969). The effects of general response style on measurement of own attitude and the interpretation of attitude-relevant messages. *British Journal of Social and Clinical Psychology, 8*(2), 104-115.

Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models Using Bayesian methods. *Journal of Educational Measurement*, *48*(1), 81-97.