# How Can We use Mixture, Multi-Process, and Other Multi-dimensional Item Response Theory Models to Account for Midpoint and Extreme Response Style Use in Personality Assessment?

Michael Lucci[*] Clement Stone[†] Fritz Ostendorf[‡] Gary Hart[§]
Suzanne Lane[¶]

**Abstract**

Since survey respondents may view item response options differently, accounting for midpoint (MRS) and extreme response style (ERS) use is important to accurately estimate the latent trait. This study investigated how five different IRT models for addressing ERS and MRS performed for three different personality subscales (Anxiety, Openness to Experience Feelings, and Compliance) from the German version of Costa and McCrae's NEO Personality Inventory-Revised. The mixture graded response and mixture partial credit models were compared with three multidimensional IRT models: the Multi-process (MPM), Multidimensional Partial Credit (MPCM), and Multidimensional Nominal Response (MNRM) models.

Response process traits of the MPM differed from response style traits of the other models. The two and three class mixture models, the two and three dimensional MNRM and MPCM, and the two process model for intensity ERS and direction fit better than standard IRT models. ERS accounted for more item response variability than MRS. The MPCM is suggested to questionnaire users to account for ERS and MRS due to the number of estimated parameters and amount of explained variability in item responses.

**Key Words:** Personality Subscales/Assessment, Midpoint/Extreme Response Styles, Multi-process (IRT tree) Model, Mixture Item Response Model, Multi-dimensional Item Response model, K-means clustering

## 1. Surveys and Response Styles

Surveys are a very common research method used to measure attitudes, determine customer satisfaction and evaluate programs. Organizations have increased use of personality measures in making personnel hiring decisions and predicting job performance (Rothstein and Goffin, 2006). Thus, examining personality questionnaire response data for detecting possible response style effects is worthwhile. Traditional scoring of many personality measures uses Likert's method of summed ratings. The instruments often consist of items with equally-spaced response options which indicate graduate degrees of a trait. For example, consider the item "I get worried easily." and the possible response options: $strongly\ disagree$ (0), $disagree$ (1), $neutral$ (2), $agree$ (3), $strongly\ agree$ (4). Each response option has a numerical value. The respondent selects an option for each item and the item responses are summed to get a trait estimate.

Use of the sum score for a trait estimate is not recommended when a response style occurs. A response style occurs if a person tends to prefer a particular category across situations and time, regardless of item content (Van Herk, Poortinga and Verhallen, 2004). Persons with Acquiescence response style (ARS) agree with items regardless of content. Persons with Midpoint response style (MRS) prefer the midpoint regardless of content.

[*]University of Pittsburgh at Geensburg
[†]Emeritus, University of Pittsburgh
[‡]Emeritus, University of Bielefeld
[§]University of Pittsburgh at Geensburg
[¶]University of Pittsburgh

Persons with Extreme Response Style (ERS) prefer one or both extreme options with any content. ERS increases or decreases item means and decreases magnitude of multivariate relationships. MRS brings observed item means closer to the midpoint and increases magnitude of multivariate relationships. Thus, use of response styles can distort item statistics, analyses, and inferences based on the sum score.

Response styles can occur due to mode of survey administration (Jordan, Marcus and Reeder, 1980), differences in demographic variables (Harzing, 2006), or relationships with personality variables (Austin, Deary and Egan, 2006; Wetzel and Carstensen, 2015). Due to the measurement problems caused by response styles (RSs), researchers and practitioners have developed methods to detect and correct for use of response styles using a set of large uncorrelated, heterogeneous items with analysis of covariance models (Reynolds and Smith, 2010), multilevel regression (Baumgartner and Steenkamp, 2001) or correlated factor structure models (Weijters, Geuens and Schillewaert, 2010; Weijters, Schillewaert and Geuens, 2008). The group differences were less significant with the removal of response style effects.

There are some limitations to the methods mentioned. Adding extra items to measure response styles lengthens the survey. The sum score method for detecting RSs does not separate effects due to person and items (De Jong, Steenkamp, Fox and Baumgartner, 2008). The sum score method gives each item equal weight when actually some items may indicate the trait more strongly than others. Persons can also vary in their response style tendencies. The methods do not try to explain a response style process (Zettler, Lang, Hülsheger and Hilbig, 2015). Fortunately, Item response theory (IRT) methods exist to overcome these limitations.

An IRT model gives the probability a person chooses a response option based on the underlying trait (Van Vaerenbergh and Thomas, 2013). An IRT model provides a direct link to a person's response behavior since it includes parameters for the items and the substantive trait of interest (TOI). The IRT model output gives estimates for the substantive TOI of the respondents and for the discrimination and difficulty parameters for the items. The discrimination parameter indicates how well the items relate to the trait. The difficulty or category threshold parameters indicate the level of a trait needed to endorse a particular rating category. We suggest use of an Item Response Theory model to get a trait estimate when response style use is suspected.

Two ways to view response styles exist. If a practitioner has a categorical view of response styles, then a person has a response style or not. The practitioner uses a mixture IRT model to estimate the number of classes in the sample, class membership probabilities, and the item parameters (discrimination and difficulties). The model also provides a substantive trait estimate (adjusted for response style use).

If a practitioner has a continuous view of response styles, then a subject has extreme and midpoint response style traits to some degree. The practitioner uses a multidimensional IRT model to get response style trait level estimates and an adjusted estimate for the TOI.

## 2. Models Compared

Our study compares mixture, multi-process, and other multidimensional IRT models. An example of an IRT model for a binary item is the two parameter logistic model (2PLM) model (Embretson and Reise, 2000) shown in equation 1:

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

This model gives the probability of a person with trait level $\theta$ agreeing with the item content. The $a_i$ gives the item discrimination or slope parameter. This indicates how strongly the item relates to the trait and distinguishes between persons at different trait levels. The $b_i$ gives the difficulty or category threshold and represents the trait level required for a person to have 0.5 probability to endorse the item. A person with a higher trait level has a higher probability of endorsing the item; that is, selecting the *agree* category instead of the *disagree* category. The use of an IRT model allows practitioners to put person trait and item (or category) difficulty parameters on the same measurement scale. This allows for direct comparison between persons and items which is not possible with the sum score (classical test theory) approach.

For items with more than two categories, other models which give the probability of selecting a particular category exist. For example, the partial credit model (PCM) assumes that the scale items relate to the trait of interest (TOI) in the same way. In the graded response model (GRM), the items are assumed to relate to the TOI in different ways. The response categories are assumed to be ordered in the PCM and GRM. The nominal response model (NRM) does not assume that the categories are ordered. The response categories for each item can have different slope (discrimination) and intercept (difficulty) parameters. The PCM, GRM, and NRM are standard, one dimensional models to measure the TOI. To account for response styles, parameters are added to these models to create more complex IRT models.

### 2.1 Mixture Models

With mixture models, we add parameter constraints to ensure that the same trait is measured in each class. Our study compared the constrained mixture partial credit model (mixPCM) and the constrained mixture graded response model (mixGRM) (Rost, 1991; Sawatzky, Ratner, Kopec and Zumbo, 2012; Wetzel, Carstensen and Böhnke, 2013) to each other and the other models. In the mixPCM, the items are assumed to have common discrimination (slope) parameters which means that the items indicate the trait in the same way. The mixPCM has different item difficulties and thresholds. The item locations (sums of the item thresholds or difficulties) are constant across classes to ensure that the same trait is measured in each class.

In the mixGRM, the items have different discriminations and thresholds. Since the items can have different discrimination parameters, the items are assumed to indicate the trait in different ways. That is, some items may be more strongly related to the trait than others. The discrimination (slope) parameters are constant across classes to ensure the measured trait is the same in each class.

The two mixture models and the multidimensional partial credit and multidimensional nominal response models in this study assume that the judgment process the respondent makes is ordinal. The item response captures the trait level only. The Multi-process model (MPM) assumes that the item response captures both the trait level and a response style process. The judgment process in the MPM consists of many discrete decisions (Böckenholt and Meiser, 2017).

### 2.2 Multiprocess and Multi-dimensional Models

Figure 1 shows the three successive processes of the MPM (or IRTree) model: indifference-MRS (1, use of midpoint or not), direction (2, agree or disagree), and intensity-ERS (3, extreme or not) (Böckenholt, 2012). If a person has no distinct opinion about an item's content, the person selects the middle category and the response process ends. If the person has a distinct opinion about the item content, the person uses the direction process to

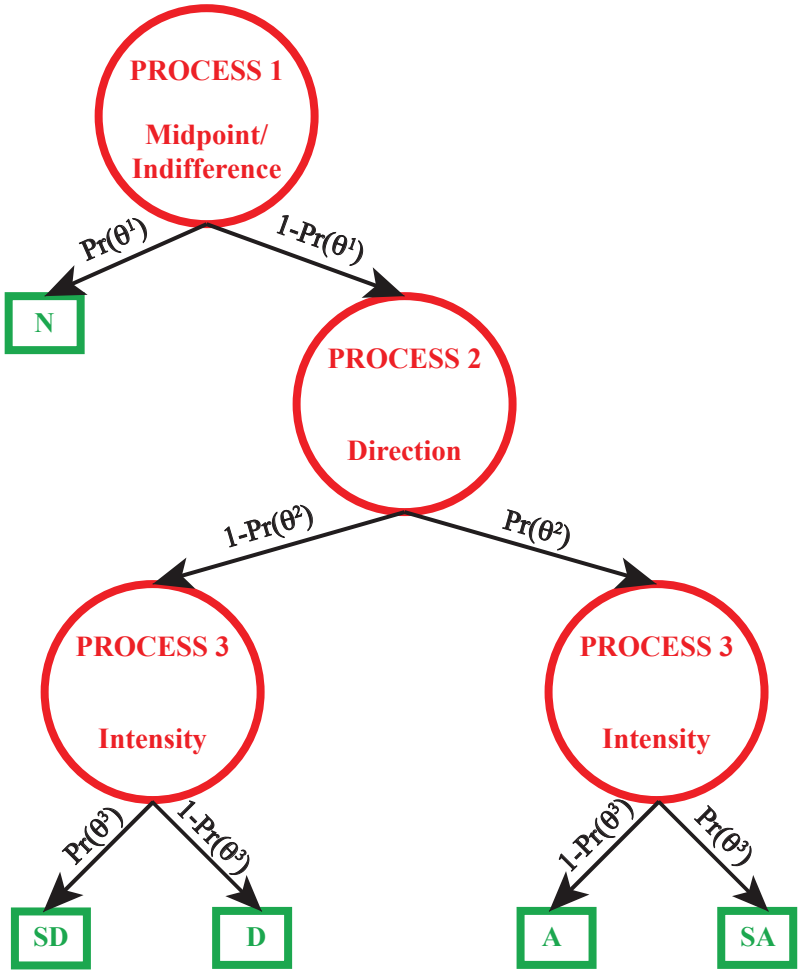


**Figure 1**: Tree structure of Three Successive Processes: Unobserved processes used to respond to a five-point item. $Pr(\theta^h)$ = Probability person with trait level $\theta^h$ uses process h to respond to the item. SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree. Adapted from Böckenholt (2012).

express disagreement or agreement. Then the person uses the intensity process to express how strongly the opinion is held; the person chooses an extreme option or not.

The MPM is a partial compensatory model. The response process traits operate independently (Zhang and Wang, 2020) and do not combine additively. The probability of the chosen response is the product of the probabilities on the branches to the observed response.

The other two multidimensional models compared are compensatory models in which different traits work together and combine additively. A person low on one trait may compensate with a high level of another trait. The Multidimensional Partial Credit Model (MPCM) has a common item discrimination parameter for each dimension: trait of interest, extreme response style, and midpoint response style. The items can have different difficulty threshold parameters for each item in each dimension.

The Multidimensional Nominal Response Model (MNRM) has similar construction to the MPCM; however, there are separate discrimination (slopes) parameters so that some items can relate to the trait more strongly than others. The Multidimensional Nominal Response Model (MNRM) uses different item-discrimination parameters for each category in an item and assumes that the items can relate to the trait differently. The items have different category slopes and intercepts. Estimating the MNRM is similar to performing a

set of logistic regressions which give the log odds of a person choosing one category over another (Falk and Ju, 2020).

The five models in this study do not require additional items to account for response styles. We hypothesized that the five models would fit better to the data than the standard models (PCM, GRM, NRM). The mixGRM, MPM, and MNRM were hypothesized to fit better than the mixPCM and MPCM since the subscale items had varying test biserial correlations (see below).

## 3. Method

### 3.1 Sample

We used 11,407 participants from the nonclinical, standardization sample of the German *NEO Personality Inventory - Revised (NEO-PI-R)* in this study. Our sample was 64% female and the ages ranged between 16 and 60 years (M = 28.88, SD = 10.46). The sample was a subset of a larger sample with 11,724 cases where 317 persons older than 60 were excluded. The sample data was collected in over 50 separate studies from 1992 to 2001 in various places in Austria, Germany, and Switzerland (Wetzel and Carstensen, 2015).

### 3.2 Instrument and Subscale Selection

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

We chose three facet subscales of moderate to high reliability to reflect varying degrees of ERS and MRS for model comparisons. We used one and two dimensional exploratory factor analyses to choose subscales which measured one primary (substantive) trait construct (one dimension) and where there could be a secondary dimension that reflected person differences (that is, due to possible response style use).

We calculated the percentages of extreme options and midpoints used for each scale. The Anxiety (N1) scale had the highest amount of midpoints used (21.37%) and the least amount of extreme options used (13.17%). The Openness to Experience Feelings (O3) scale had the greatest percentage of extreme options used (25.24%) and the least amount of midpoints used (14.08%). The Compliance (A4) scale had 17.90% midpoint use and 17.64% extreme option use.

The A4 scale sum score had negligible correlation with the proportion of midpoints (r = 0.02) and with the proportion of extreme options (r = -0.10) used. The Anxiety scale sum score had slightly higher, yet still negligible correlation with midpoint proportion used (r = -0.11) and extreme option proportion used (r = 0.15). The Openness to Experience Feelings sum score correlated moderately and negatively with proportion of midpoints used (r = -0.59) and correlated positively and markedly with extreme option use (r = 0.74). These statistics supported using the chosen scales to compare model performance for this study.

### 3.3 Preliminary Analyses

We calculated the proportion of total extreme options (TERS) used and proportion of midpoints (TMRS) used for each person for each subscale. We then used these proportions

with K-means clustering analysis to determine the sizes of groups preferring the midpoint, extreme options, or general (*agree* or *disagree*) categories for each scale. The different group sizes illustrate that some respondents use the response options differently based on scale content.

Table 1 presents the different sizes and respondent characteristics for the K-means analyses with three groups (Midpoint preferred, Extreme options preferred, and General with *disagree* or *agree* preferred) for the chosen scales. For all scales, the General groups are the largest of the three groups. The sizes of the Midpoint and Extreme groups differ from the General groups and differ across scales. For discussion purposes, we assume that less than 25% of the sample represents a small group; between 25% to 50% of the sample is a medium group; and greater than 50% of the sample is a large group.

The Anxiety (N1) scale has a medium sized Midpoint group nearly twice the size of its small Extreme group (26.7% versus 14.2%). The N1 General group is large in size (59.1%). The Openness to Experience Feelings (O3) scale has a medium sized Midpoint group (25.7%) that is slightly larger than its small Extreme group (23.1%). The O3 General group has a large size (51.2%). The Compliance (A4) Midpoint Group is medium-sized (38.6%) and roughly twice as large as its small Extreme Group (18.6%). The A4 General group has a medium size (42.9%).

The mean proportions of midpoints and extremes used across scales further differentiates the Midpoint and Extreme groups. For discussion purposes, we describe the mean proportions with language similar to correlation description (e.g., 0.2-0.4, low; 0.4-0.6, moderate; 0.6-0.8, marked). Table 1 shows the mean proportion of midpoints used per person (M TMRS) and mean proportion of extremes used per person (M TERS) in each scale's response style groups (Midpoint, Extreme, and General). As would be expected due to each group's category preference, the M TERS is always negligible in the Midpoint group for a scale; while the M TMRS is always negligible for the Extreme group. The M TMRS and M TERS are always negligible in the General group due to this group's preference for the *agree* or *disagree* categories.

Comparing the mean proportions across the three groups within each scale illustrates the characteristics of the persons in the groups. For the Anxiety (N1) scale, we believe there is moderate use of MRS since the M TMRS is 0.47 for the Midpoint group. This mean is higher than the M TMRS for the Openness to Experience Feelings (O3) Scale (.38) and Compliance (A4) Scale (.34). The respondents tend to use the midpoint less for these scales than the persons in the Midpoint group for the N1 scale. We believe this implies low use of MRS.

Similarly, persons in the O3 Extreme group use extreme options more than persons in the Extreme groups for the other two scales. We believe this suggests marked use of ERS for the O3 scale. We hypothesized low use of ERS for the N1 and A4 scales.

### 3.4 Mixture Model Analyses

Much previous work with the mixture PCM exists. The PCM has a constant item discrimination (slope). This slope parameter in IRT is analogous to the item test biserial correlation (Yen and Fitzpatrick, 2006). For the three facet scales, the range of biserial correlations for the scale items were N1 (0.49 to 0.65), O3 (0.34 to 0.60), and A4 (0.18 to 0.44). Since these correlations vary widely, we hypothesized that the mixture graded response model with varying item slopes could provide better fit than the mixture PCM. We used *MPlus* (Muthén and Muthén, 1998-2012) software to estimate the standard PCM and GRM. Then we estimated two- and three- class mixture models for the PCM and GRM and examined the classes to interpret them for response style use.

**Table 1**: K-means Cluster Results for Three Different Response Style Groups.
The percentage of the sample (N = $11,407$) assigned to the group which preferred the Midpoint, Extreme options, or General ($agree$ and $disagree$) options. EC = Extreme cluster, MC = Midpoint cluster, G = General cluster, TMRS = proportion of midpoints used by person within scale, TERS = proportion of extremes used by person within scale, M = Mean proportion of midpoints or Extremes used, SEM = Standard error of the mean, N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance.

| Scale | MC Size<br>TMRS, TERS<br>M(SEM) | EC Size<br>TMRS, TERS<br>M(SEM) | GC Size<br>TMRS, TERS<br>M(SEM) |
|---|---|---|---|
| N1 | 26.7%<br>**0.47(0.002)**, 0.04(0.001) | 14.2%<br>0.08(0.003), **0.51(0.004)** | 59.1%<br>0.13(0.001), 0.08(0.001) |
| O3 | 25.7%<br>**0.38(0.003)**, 0.05(0.002) | 23.1%<br>0.04(0.001), **0.67(0.003)** | 51.2%<br>0.07(0.001), 0.17(0.002) |
| A4 | 38.6%<br>**0.34(0.002)**, 0.10(0.002) | 18.6%<br>0.09(0.002), **0.47(0.003)** | 42.9%<br>0.07(0.001), 0.12(0.001) |

To select the best mixture model, we examined four criteria. We sought the model with the lowest Bayesian Information Criterion (BIC) and the model with the least amount of absolute standardized bivariate Pearson residuals greater than three for the best model to data fit.

Additionally, we looked for the greatest scaled entropy (sEn), a measure of classification quality. We also wanted the diagonal probabilities of the mean class assignment probability table to be greater than or equal to 0.8 for high quality classification (solution interpretability).

### 3.5 Multidimensional Model Analyses

We estimated two and three multi-dimensional nominal response and partial credit models for the substantive trait of interest and at least one or both MRS and ERS traits for comparison with the standard NRM and PCM models with *flexMIRT* software (Houts and Cai, 2015). The estimation did not require recoding the data for the compensatory models (MPCM, MNRM).

Estimating the partial compensatory two and three process (MPM) models required recoding the observed data into pseudo-item responses. The pseudo-items captured the hypothesized latent response processes. The three process models involved binary pseudo-items for the indifference, direction, and intensity processes. The two process indifference-MRS model involved a process for indifference and a combined direction-intensity process. The two process intensity-ERS model involved an intensity process and a combined direction-indifference process.

### 3.6 Model Comparison Analyses

To compare all of the models, we examined the Bayesian Information Criterion (BIC). We also checked the Root Mean Square error of approximation (RMSEA) for the MIRT and standard models. The RMSEA was not available for the mixture models in *MPlus*

(Muthén and Muthén, 1998-2012). Better fitting models have lower BIC and lower RMSEA values.

We also compared mixture (mixPCM, mixGRM) and compensatory MIRT models (MPCM, MNRM) using the amount of explained variability (general $R^2$) beyond the standard models (Nagelkerke, 1991). This $R^2$ was not available for the Multi-process (IRTree) models since the one and two process models were not nested within the three process model. We also considered the number of estimated parameters for each model for model comparison and recommendation.

### 3.7 Response Style Use Comparisons

To compare response style use for each scale, we examined group sizes formed from the three class mixture models. We used the mean proportions of extreme options and midpoints used by persons in each group to describe the groups with the correlation coefficient language as with the above K-means clustering analyses. We assumed that less than 25% of the sample represented a small group; between 25% to 50% of the sample was a medium group; and greater than 50% of the sample was a large group for discussion purposes. To demonstrate that the response style estimates from the MIRT models also produce three different response style groups, we performed K-means clustering analyses using model-based MRS and ERS estimates as the clustering variables.

## 4. Results

### 4.1 Mixture Model Results

Table 2 shows the mixture model mean class assignment probabilities for the three scales. Values on the main diagonal indicate the probability of correct classification for a particular class. Two class models yield an Extreme response style class and a Non-extreme response style class. Three class models yield an Extreme response class, a Midpoint response class, and a General class favoring *agree* or *disagree* categories. The two class mixtures have better classification quality than the three class mixtures since the values on the main diagonal of the mean class assignment probability (MCAP) tables are greater than 0.8 for two class mixtures. Three class mixtures tend to have MCAP values 0.8 or lower.

Table 3 shows the mixture model selection criteria for the three scales. The better classification quality of the two class mixtures than three class mixtures is seen in the scale entropy value (sEn). As scaled entropy approaches one, there is better classification quality. The sEn values are generally higher for the two class mixtures (0.51-0.67) than the three class mixtures (0.40-0.60). The mixture PCM has better classification quality than the mixture GRM for both two class and three class models. Although the three class mixture GRM has better fit as indicated by the lowest BIC value and least amount of large absolute bivariate standardized residuals (1.1%-5,7%), the two class mixture PCM has better classification quality (sEn) which is important to consider when identifying persons for a specific response style and interpreting the model results.

### 4.2 Multi-dimensional Results and Model Comparisons

Table 4 shows the Absolute and Relative model fit and explained variability in item responses for all models. The Bayesian Information Criterion (BIC) is lowest for the three dimensional models for TOI, ERS, and MRS and the three class mixture models. The RMSEA tends to be smallest for the three dimensional -PCM and -NRM. It also tends to

**Table 2**: Mixture Model Mean Class Assignment Probabilities.
E = Extreme class, N = Non-extreme class, M = Midpoint class, G = General class. Probabilities in bold indicate persons are classified with high probabilities in the respective class N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance. 2mixGRM = Two Class Graded Response Model, 3mixPCM = Three class Partial Credit Model.

| N1 2mixGRM | | | N1 3mixGRM | | | |
|---|---|---|---|---|---|---|
| | E | N | | E | G | M |
| E | **0.82** | 0.18 | E | **0.78** | 0.14 | 0.08 |
| N | 0.13 | **0.87** | G | 0.11 | **0.69** | 0.20 |
| | | | M | 0.06 | 0.20 | **0.74** |
| N1 2mixPCM | | | N1 3mixPCM | | | |
| | N | E | | E | M | G |
| N | **0.90** | 0.10 | E | **0.82** | 0.04 | 0.15 |
| E | 0.15 | **0.85** | M | 0.03 | **0.78** | 0.19 |
| | | | G | 0.08 | 0.18 | **0.74** |
| O3 2mixGRM | | | O3 3mixGRM | | | |
| | E | N | | G | E | M |
| E | **0.84** | 0.16 | G | **0.68** | 0.10 | 0.22 |
| N | 0.13 | **0.87** | E | 0.13 | **0.79** | 0.08 |
| | | | M | 0.19 | 0.08 | **0.73** |
| O3 2mixPCM | | | O3 3mixPCM | | | |
| | E | N | | G | M | E |
| E | **0.88** | 0.12 | G | **0.79** | 0.09 | 0.13 |
| N | 0.08 | **0.92** | M | 0.11 | **0.87** | 0.02 |
| | | | E | 0.17 | 0.03 | **0.80** |
| A4 2mixGRM | | | A4 3mixGRM | | | |
| | E | N | | M | G | E |
| E | **0.81** | 0.19 | M | **0.73** | 0.18 | 0.09 |
| N | 0.13 | **0.87** | G | 0.23 | **0.67** | 0.11 |
| | | | E | 0.10 | 0.13 | **0.77** |
| A4 2mixPCM | | | A4 3mixPCM | | | |
| | E | N | | G | E | M |
| E | **0.82** | 0.18 | G | **0.68** | 0.10 | 0.22 |
| N | 0.11 | **0.89** | E | 0.13 | **0.79** | 0.09 |
| | | | M | 0.19 | 0.07 | **0.79** |

**Table 3**: Mixture Model Selection Criteria for Facet Subscales.
K = Number of classes, p = number of parameters , BIC = Bayesian Information Criterion, sEn = scaled entropy, MMCAP = Minimum Diagonal value of Mean Class Assignment Probabilities table. ASBPR = Percent of Absolute Bivariate Standardized Pearson residuals greater than three. 2(mixGRM) = two class constrained graded response model. 3(mixPCM) = three class constrained partial credit model.

| Anxiety Facet | | | | | |
|---|---|---|---|---|---|
| K (Model) | p | BIC | sEn | MMCAP | ABSPR $> 3$ |
| 1(GRM) | 40 | 230,152 | — | 1.00 | 38.9% |
| 1(PCM) | 33 | 232,563 | — | 1.00 | 51.4% |
| 2(mixGRM) | 73 | 227,645 | 0.51 | 0.82 | 14.1% |
| 2(mixPCM) | 58 | 229,001 | **0.61** | **0.85** | 25.3% |
| 3(mixGRM) | 106 | **227,003** | 0.44 | 0.69 | **5.7%** |
| 3(mixPCM) | 83 | 228,065 | 0.50 | 0.74 | 14.9% |
| Openness to Experience Feelings Facet | | | | | |
| K (Model) | p | BIC | sEn | MMCAP | ABSPR $> 3$ |
| 1(GRM) | 40 | 198,870 | — | 1.00 | 36.0% |
| 1(PCM) | 33 | 203,391 | — | 1.00 | 48.1% |
| 2(mixGRM) | 73 | 196,413 | 0.53 | 0.84 | 10.1% |
| 2(mixPCM) | 58 | 198,035 | **0.67** | **0.88** | 23.0% |
| 3(mixGRM) | 106 | **196,041** | 0.43 | 0.68 | **4.9%** |
| 3(mixPCM) | 83 | 197,002 | 0.60 | 0.79 | 12.0% |
| Compliance Facet | | | | | |
| K (Model) | p | BIC | sEn | MMCAP | ABSPR $> 3$ |
| 1(GRM) | 40 | 232,554 | — | 1.00 | 28.4% |
| 1(PCM) | 33 | 233,657 | — | 1.00 | 35.1% |
| 2(mixGRM) | 73 | 230,204 | 0.53 | 0.81 | 4.9% |
| 2(mixPCM) | 58 | 230,762 | **0.57** | **0.82** | 12.1% |
| 3(mixGRM) | 106 | **229,888** | 0.40 | 0.67 | **1.1%** |
| 3(mixPCM) | 83 | 230,339 | 0.41 | 0.68 | 7.6% |

be slightly smaller for the two dimensional PCM-ERS and NRM-ERS than the PCM-MRS and NRM-MRS. Thus, these models have improved fit over the standard models.

The RMSEA for the three-process (MPM) model and the two-process Mid-PM is much larger than that for the two- process Ext-PM. So, only the two process model for intensity extreme process is recommended as having better fit than the standard GRM. The RMSEA is not available for the mixture models.

The explained variability in item responses beyond the standard IRT models is quantified with the general coefficient of determination, $R^2$. This $R^2$ is larger for two-dimensional PCM-ERS and NRM-ERS than it is for two-dimensional PCM-MRS and NRM-MRS. The $R^2$ is larger for the two class mixtures for the extreme and non-extreme response styles than the increase in $R^2$ for the three class mixtures. Thus, ERS explains more variability in item responses than MRS. $R^2$ is largest for the three class PCM and three class GRM. So modeling both MRS and ERS explains the most variability in item responses. The general coefficient of determination is not available for the two and three process models which are not nested in a standard model. Thus, we cannot use $R^2$ to compare these models with the others.

### 4.3 Response Style Use Comparisons

#### *4.3.1 Mixture Model Groups*

To examine response style use for each scale, we compare group sizes formed from the three class mixture models first. The group sizes in each mixture model output differ depending on scale content. Table 5 shows class sizes for the three scales under the three class mixture PCM (3mixPCM). The Anxiety (N1) scale has a medium-sized group (32.5%) of persons with moderate MRS. Persons in this MRS class have a moderate mean proportion of midpoints used (M TMRS = 0.43). A small class (17.9%) prefers extreme options over the midpoint and the respondents have a moderate mean proportion of extremes used (M TERS =0.45). Thus, based on the relative class sizes and mean proportions of midpoints and extreme options used, there tends to be medium use of midpoint response style (MRS) and low to moderate use of extreme response style (ERS) for the Anxiety scale.

For the 3mixPCM, the Openness to Experience Feelings (O3) scale has the smallest MRS class and the largest ERS class. The Midpoint class is small (21.4%) and the Extreme size class is medium-sized (31.0%). With the low mean proportion of midpoints used in the Midpoint group (0.39) and moderate mean proportion of extremes used in the Extreme group (0.58), the O3 scale has medium level of ERS and low level of MRS.

The Compliance (A4) scale has two medium-sized Midpoint and Extreme response style groups; however, the midpoint class is larger than the Extreme class (34.7% vs. 25.6%). The mean proportion of midpoints used for the Midpoint class is low (0.34) and the mean proportion of extremes used for the Extreme group is low (0.41). Thus, the A4 scale has both low MRS and low ERS use under the 3mixPCM.

For all three scales, the General class which prefers *disagree* and *agree* is the largest of the three groups determined by the 3mixPCM. The General class always has mean proportions of midpoint and extreme options used that are no larger than 0.11. Thus, the General class has negligible use of MRS and ERS.

Table 5 also shows the different class sizes for the scales under the three class mixture GRM (3mixGRM). For the Anxiety (N1) scale, a medium-sized class prefers the midpoint over extremes (38.3%) and this class has low use of MRS (M TMRS = 0.38). A small class prefers extreme options over midpoints (20.3%) and this Extreme class has low use of ERS (M TERS = 0.40). The N1 scale tends to evoke low use of MRS and ERS.

The Openness to Experience Feelings (O3) scale has a medium-sized class that prefers extremes over midpoints (31.8%) and this Extreme Group has moderate use of ERS (M TERS = 0.56). The O3 scale also has a medium-sized class that prefers the midpoint over extremes (30.3%). This Midpoint response style class has low use of MRS (M TMRS = 0.32). Thus, O3 scale has moderate use of ERS and low use of MRS.

The Compliance (A4) scale has a medium-sized Midpoint response style class (37.9%) with a low mean proportion of midpoints used (M TMRS = 0.33). The A4 ERS class is smaller (26.0%) and the mean proportion of extremes used was low (M TERS = 0.41). Thus, the A4 scale tends to have low use of MRS and ERS under the mixGRM.

In summary, the use of midpoint and extreme options for both mixture models validates the interpretation of the groups. For example, the Midpoint group uses the midpoint option more than the other groups and the extreme options less than the other groups do. For each scale, the Extreme group uses extreme options more than the other groups do.

**Table 4**: Absolute and Relative Model Fit and Explained Variability in Item Responses.
p = number of parameters, BIC = Bayesian Information Criterion, RMSEA = Root Mean Square Error Approximation, $R^2$ = general coefficient of determination (Nagelkerke, 1991), MPCM = Multi-dimensional Partial Credit model for trait, ERS, and MRS dimensions, PCM-ERS =Two dimensional Partial Credit model for trait and ERS dimensions, MNRM = Multidimensional Nominal Response Model with freely estimated overall item category slopes (FEOIC) on all three dimensions, NRM-ERS = Two dimensional Nominal Response Model for substantive and ERS traits with FEOIC slopes on both dimensions, NRM-MRS = Two Dimensional NRM for substantive and MRS traits with FEOIC slopes on both dimensions, MPM = Multi-Process model with varying item slopes for all binary pseudoitems for indifference, direction, and intensity processes. Ext-PM = Two process model for Intensity and Direction, Mid-PM = Two process model for Indifference and Direction, Absolute fit statistics such as RMSEA are not available for the mixture models in *Mplus*. Bayesian Information criterion (BIC) is presented for all models for relative fit comparisons. 3mixGRM = three class mixture GRM.

| | | Anxiety | | | Openness to Experience Feelings | | | Compliance | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | p | BIC | RMSEA | $R^2$ | BIC | RMSEA | $R^2$ | BIC | RMSEA | $R^2$ |
| PCM | 33 | 232,563 | 0.04 | — | 203,391 | 0.05 | — | 233,657 | 0.04 | — |
| PCM-ERS | 35 | 229,328 | 0.03 | 0.25 | 198,226 | 0.03 | 0.38 | 230,867 | 0.02 | 0.22 |
| PCM-MRS | 35 | 231,028 | 0.06 | 0.13 | 202,203 | 0.05 | 0.12 | 232,866 | 0.04 | 0.07 |
| MPCM | 38 | 227,812 | 0.03 | **0.34** | 196,901 | 0.02 | **0.44** | 230,227 | 0.02 | **0.26** |
| NRM | 64 | 231,414 | 0.04 | — | 198,276 | 0.04 | — | 232,598 | 0.03 | — |
| NRM-ERS | 73 | 228,322 | 0.03 | 0.24 | 196,148 | 0.02 | 0.18 | 230,434 | 0.02 | 0.18 |
| NRM-MRS | 73 | 229,945 | 0.03 | 0.13 | 197,543 | 0.03 | 0.07 | 231,876 | 0.03 | 0.07 |
| MNRM | 83 | **226,858** | 0.02 | 0.34 | **195,445** | 0.01 | 0.23 | **229,774** | 0.01 | 0.23 |
| Ext-PM | 41 | 229,876 | **0.05** | — | 197,108 | **0.04** | — | 232,027 | **0.05** | — |
| Mid-PM | 49 | 230,890 | 0.13 | — | 198,062 | 0.19 | — | 232,762 | 0.12 | — |
| MPM | 51 | 228,813 | 0.15 | — | 196,414 | 0.15 | — | 231,723 | 0.20 | — |
| 2mixPCM | 58 | 229,001 | — | 0.28 | 198,035 | — | 0.39 | 230,762 | — | 0.24 |
| 3mixPCM | 83 | 228,065 | — | 0.35 | 197,002 | — | 0.45 | 230,339 | — | 0.28 |
| GRM | 40 | 230,152 | 0.03 | — | 198,526 | 0.03 | — | 232,260 | 0.03 | — |
| 2mixGRM | 73 | 229,645 | — | 0.22 | 196,413 | — | 0.22 | 230,204 | — | 0.21 |
| 3mixGRM | 106 | 227,003 | — | 0.28 | 196,041 | — | 0.26 | 229,888 | — | 0.25 |

#### *4.3.2 Examining Response Style Groups from Multidimensional Model Estimates*

Table 6 shows the different group sizes from K-means clustering with the Multidimensional model based estimates for ERS and MRS. For the MPCM K-means results, the Anxiety scale has a medium-sized Midpoint group (29.9%) with moderate MRS use (M TMRS = 0.44) and a small Extreme group (23.2%) with low use of ERS (M TERS = 0.40). The Open to Experience Feelings scale has a medium-sized Midpoint group (25.8%) with low use of MRS (M TMRS = 0.38) and a medium-sized Extreme group (35.7%) with moderate use of ERS (M TERS = 0.56) that was larger than the Midpoint group. The A4 scale has a medium-sized Midpoint group (34.9%) with low use of MRS (M TMRS = 0.35) and a smaller medium-sized Extreme group (31.5%) with low use of ERS (M TERS = 0.38). The General groups for the three scales show negligible ERS and MRS use with mean proportions of extremes and midpoints used that were less than or equal to 0.13.

Next we describe the K-means groups from the MNRM ERS and MRS estimates. The N1 scale has a medium-sized Midpoint group (29.7%) with moderate MRS use (M TMRS = .44) and a medium sized Extreme group (25.6%) with low ERS use (M TERS = 0.38). The O3 scale has a small K-MNRM Midpoint group (24.4%) with low MRS use and a medium-sized Extreme group (34.2%) with moderate ERS use (M TERS = 0.51). The A4 scale has a medium-sized K-MNRM Midpoint (29.3%) group with low MRS use (M TMRS = 0.37) and a medium-sized Extreme group (29.4%) with low ERS use (M TERS = 0.39). Though medium in size, the general K-MNRM groups are larger than the extreme and midpoint groups for all scales. They have negligible use of MRS and ERS with mean proportions of midpoints and extremes used which are all less than or equal to 0.15.

Now we consider the K-means groups based on the Multiprocess model estimates. In Table 6, the Anxiety scale shows moderate use of indifference-MRS (M TMRS = 0.42) in a medium sized K-MPM Midpoint group (31.4%). For the N1 scale, there is also moderate use of intensity-ERS (M TERS = 0.42) in a small K-MPM Extreme group (21.4%). The Open to Experience Feelings scale has a medium sized K-MPM Midpoint group (28.6%) with low use of indifference MRS (M TMRS = 0.33). There is a medium sized K-MPM Extreme group (27.8%) with moderate use of intensity ERS (M TERS = 0.61) for the O3 scale. The Compliance scale has a medium sized K-MPM Midpoint group (30.5%) with low indifference MRS (M TMRS = 0.33) and a medium sized K-MPM Extreme group (22.5%) with moderate use of intensity ERS (M TERS = 0.43). The General groups for all three scales showed negligible indifference-MRS and intensity-ERS use with mean proportions of midpoints and extremes used less than or equal to 0.17.

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

## 5. Discussion

To choose a model, a practitioner can consider the nature of response styles, the assumptions about the attitudinal judgement process, and the implementation and estimation of the models (Böckenholt and Meiser, 2017). If s/he views response styles as discrete variables, then a mixture model is suggested. If s/he sees responses styles as continuous variables, then one of the multi-dimensional (MIRT) models is appropriate.

Additionally, the practitioner might consider the assumptions about the latent judgement process. If it is assumed to be ordinal, then a mixture model or compensatory MIRT

**Table 5**: Mixture Model Class Sizes of Three Different Response Style Groups.
The percentage of the sample (N = $11,407$) assigned to the class designated as Midpoint, Extreme, or General based on preferred option(s). EC = Extreme class, MC = Midpoint class, G = General class, TMRS = proportion of midpoints used used by persons in class, TERS = Mean proportion of extremes used used by persons in class, M = Mean of the proportion of midpoints (extremes) used, SEM = Standard Error of the Mean. N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance.

| | | Three class mixture PCM | |
|---|---|---|---|
| Scale | MC Size<br>TMRS, TERS<br>M(SEM) | EC Size<br>TMRS, TERS<br>M(SEM) | GC Size<br>TMRS, TERS<br>M(SEM) |
| N1 | 32.5%<br>**0.43**(0.002), 0.03(0.001) | 17.9%<br>0.11(0.003), **0.45**(0.004) | 49.6%<br>0.11(0.001), 0.08(0.001) |
| O3 | 21.4%<br>**0.39**(0.003), 0.05(0.002) | 31.0%<br>0.06(0.002), **0.58**(0.003) | 47.6%<br>0.08(0.001), 0.13(0.002) |
| A4 | 34.7%<br>**0.34**(0.002), 0.09(0.002) | 25.6%<br>0.11(0.003), **0.41**(0.003) | 39.6%<br>0.08(0.001), 0.09(0.001) |
| | | Three class mixture GRM | |
| Scale | MC Size<br>TMRS, TERS<br>M(SEM) | EC Size<br>TMRS, TERS<br>M(SEM) | GC Size<br>TMRS, TERS<br>M(SEM) |
| N1 | 38.3%<br>**0.38**,(0.003), 0.04(0.001) | 20.3%<br>0.12(0.003), **0.40**(0.004) | 41.5%<br>0.11(0.002), 0.09(0.002) |
| O3 | 30.3%<br>**0.32**(0.003), 0.10(0.002) | 31.8%<br>0.07(0.002), **0.56**(0.004) | 37.9%<br>0.06(0.001), 0.12(0.002) |
| A4 | 37.9%<br>**0.33**(0.002), 0.10(0.002) | 26.0%<br>0.11(0.002), **0.41**(0.003) | 36.1%<br>0.07(0.001), 0.09(0.001) |

**Table 6**: K-means groups from Multi-dimensional Model Response Style Trait Estimates. Percentage of the sample (N = 11, 407) assigned to group designated as Midpoint, Extreme, or General based on preferred option(s). M TMRS = Mean proportion of midpoints used by persons in group, M TERS = Mean proportion of extremes used in group. SEM = standard error of the mean, N1 = Anxiety, O3 = Openness to Experience Feelings, A4 = Compliance.

| Scale | Midpoint Group Size<br>M TMRS, TERS (SEM) | Extreme Group Size<br>M TMRS, TERS (SEM) | General Group Size<br>M TMRS, TERS (SEM) |
|---|---|---|---|
| | Multi-dimensional Partial Credit Model Groups | | |
| N1 | 29.9%<br>**0.44**(0.002), 0.03(0.001) | 23.2%<br>0.09(0.002), **0.40**(0.004) | 46.9%<br>0.13(0.001), 0.06(0.001) |
| O3 | 25.8%<br>**0.38**(0.003),0 .05(0.002) | 35.7%<br>0.05(0.001), **0.56**(0.003) | 38.5%<br>0.06(0.001), 0.10(0.001) |
| A4 | 34.9%<br>**0.35**(0.002), 0.08(0.001) | 31.5%<br>0.11(.002),**0.38**(0.002) | 33.7%<br>0.07(0.001), 0.08(0.001) |
| | Multi-dimensional Nominal Response Model Groups | | |
| N1 | 29.7%<br>**0.44**(0.002), 0.02(0.001) | 25.6%<br>0.10(0.002),**0.38**(0.004) | 44.7%<br>0.13(0.001), 0.06(0.001) |
| O3 | 24.4%<br>**0.39**(0.003), 0.07(0.002) | 34.2%<br>0.06(0.001),**0.51**(0.004) | 42.8%<br>0.06(0.001), 0.15(0.002) |
| A4 | 29.3%<br>**0.37**(0.002), 0.09(0.002) | 29.4%<br>0.10(0.002), **0.39**(0.003) | 41.2%<br>0.10(0.001), 0.08(0.001) |
| | Multi-process Model Groups | | |
| N1 | 31.4%<br>**0.42**(0.003), 0.01(0.001) | 21.4%<br>0.07(.002), **0.42(0.004)** | 47.2%<br>0.14(0.002), 0.08(0.001) |
| O3 | 28.6%<br>**0.33**(0.003),0 .03(0.001) | 27.8%<br>0.03(0.001), **0.61(0.003)** | 43.6%<br>0.09(0.001), 0.17(0.002) |
| A4 | 30.5%<br>**0.33**(0.002), 0.06(0.001) | 22.5%<br>0.10(0.002), **0.43(0.003)** | 46.9%<br>0.12(0.002), 0.13(0.001) |

model (such as MPCM or MNRM) is appropriate. If the latent judgement process is believed to consist of discrete decisions, then a multi-process model may be appropriate.

A practitioner would also want to consider what is needed for implementation and estimation. The mixture IRT models involve an exploratory approach where the response styles are not specified beforehand. The implementation can be time consuming since the practitioner must estimate the models and examine the classes and interpret them for response style use. The multidimensional models involve a confirmatory approach where the response styles are specified beforehand and the estimation and interpretation usually take less time than the mixture modeling does.

For most models in this study, adding parameters to standard models to account for responses improves model to data fit. The mixture models fit better than the standard PCM and GRM. The three class mixture models can classify persons in Midpoint, Extreme, and General response style classes. The three class GRM fit better than the three class PCM, but the three class PCM explains more variability in item responses and has less parameters than the three class GRM. The two class models have better classification quality than the three class models which is important in interpreting the solution and identifying persons with a particular response style. Since the classification quality is higher for the two class mixture PCM, it is suggested for this study's scales. It could identify persons with Extreme and Non-extreme response styles, but not Midpoint response styles.

For the partial compensatory MIRT (multi-process) models, the two process intensity-ERS model fit better than the standard models. The two process indifference-MRS and three process models did not fit well; so, they cannot be recommended for the scales in this study.

With the compensatory MIRT models, the three dimensional nominal response model for TOI, ERS, and MRS fit better than the three dimensional partial credit model (MPCM). The MPCM is preferred over the multi-dimensional nominal response model (MNRM) since the MPCM has fewer parameters to estimate. The MPCM also explains more of the variability in responses than the MNRM. Thus, we suggest the MPCM for these scales.

The results from the five models in our study show that ERS explains more variability in responses than MRS. This is consistent with previous work (Wetzel and Carstensen, 2015). The different sizes of the Midpoint, Extreme, and General Response style groups for different subscales supports the idea that "scale specific" response styles are detected with the five models examined in this study.

The value in the approaches with the models in this study is that using them requires no additional items. However, this is a also limitation since the models only account for "scale specific" response styles. There is not a way to measure general response style tendencies with these particular models alone. With mixture models, one way to examine consistency of general response style use is to use a second order latent class analysis (Wetzel et al., 2013). The person class membership assignments for several traits from the mixture model analyses are used in a latent class analysis. The latent classes are then examined to determine if persons in the classes are consistently using a response style across traits. For such an analysis, we would want to use a larger number of scales than the three scales used here.

Our study has other limitations. We used real data where the truth is unknown. The ERS and MRS use was inferred. It is possible, for example, that the persons using extreme response style were truly high on the trait of interest and persons with midpoint response style had average trait levels. Thus, the given recommendations are limited possibly to the real data used.

A third limitation is that we used the MPM and other MIRT model estimates in K-means clustering to form response style groups. We do not suggest this in practice since

the K-means algorithm is not sensitive to the variable standard errors of the point estimates used in the clustering. Using MIRT model point estimates in K-means analyses could have affected the size of the groups and their interpretation.

This fourth limitation is that the size of the Midpoint and Extreme response style groups was small compared to the General response style groups. If we reduced the size of the General group by randomly deleting subjects, this might increase response style group impact and then increase a model's ability to capture response style tendencies.

Another way to reduce the size of the groups is to randomly split the data in half. This would allow us to check for model overfitting which can occur if the modeling approach is exploratory and the models have a large number of estimated parameters (McCrea, 2013). For the exploratory models in this study, the mixture graded response model had a larger number of estimated parameters than the mixture partial credit model. Thus, one possibility for future research is to randomly divide the data and estimate the models on each half to see if similar results are found for the mixture and multidimensional models on each half.

Another limitation to our current study is the relatively small number of items per scale (eight) which may have affected MIRT model convergence and estimation with *flexMIRT* software. A future study could check model convergence by using different random start values in software. This would have to be done manually as there is no direct method to do this with *flexMIRT* as there is with *Mplus*. Note also that specifically the multidimensional nominal response model results and interpretation could have been affected. Previous work with the MNRM, such as Falk and Cai (2015), examined scales with at least 10 items. Scales with a larger number of items could provide more response data and better parameter estimates.

Related to the number of items is the number of response options. The items in this study had five. Perhaps items with seven response options could identify midpoint response style more than five point items. Our study found extreme response style to have a greater effect than midpoint response style. This could be due to the relatively smaller number of response options. Only additional research with items with seven response options could inform us if ERS would still have a greater effect than MRS.

Another limitation is that we examined only two response styles with the Multiprocess Model; yet other response styles exist. It may be possible to conduct a future study similar to Levanthal and Zigler (2021) who used a multi-process model and anchoring vignettes to account for Acquiescence response style (ARS) along with Extreme and Midpoint Response styles. They found that estimating ARS produced different estimates for the substantive (TOI), ERS, and MRS traits. Since ARS can possibly affect survey score interpretations in other contexts, Levanthal and Zigler (2021) suggest future work with their method.

With the MIRT models in our study, future research could also involve a more complex multi-dimensional model with homogeneous items for two or more correlated traits of interest and heterogeneous items to measure general Extreme and Midpoint response style use. For example, Falk and Cai (2015) used the multidimensional nominal response model to measure six substantive traits and two response style traits. It may be possible to use the MNRM with the data in our study by choosing heterogeneous items to measure general response style traits.

Lastly, our current study is limited in that the mixture models do not allow for individual variation within classes (Zhang and Wang, 2020). A future study could analyze the data with mixture IRTree modeling (Kim and Bolt, 2021) which allows for a mixture of persons with different underlying response processes (between respondents) and uncertainty at the person level (within respondents). By combining mixture and multiprocess models, we could test if a mixture IRTree model fits better than the chosen models in this study.

Only further research with other scales can tell us how the five models in our study account for response styles for a given scale. Additional research can also tell us if we can use extensions to these models to assess general response style tendencies.

## References


Austin, E. J., Deary, I. J., and Egan, V. (2006), "Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items," *Personality and Individual Differences*, 40(6), 1235–1245.

Baumgartner, H., and Steenkamp, J. B. E. M. (2001), "Response styles in marketing research: A Cross-National investigation," *Journal of Marketing Research*, 38(2), 143–156.

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

Böckenholt, U., and Meiser, T. (2017), "Response style analysis with threshold and multi-process IRT models: A review and tutorial," *British journal of mathematical and statistical psychology*, 70(1), 159–181.

Costa, P. T., and McCrae, R. R. (1992), *Professional manual: revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI)*, Odessa, FL: Psychological Assessment Resources.

Costa, P. T., and McCrae, R. R. (2014), "The NEO Inventories," in *Personality Assessment*, eds. R. P. Archer, and S. R. Smith, New York, NY: Routledge, pp. 229–260.

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

Embretson, S., and Reise, S. (2000), *Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Falk, C. F., and Cai, L. (2015), "A Flexible Full-Information Approach to the Modeling of Response Styles," *Psychological Methods*, 21(3), 328–347.

Falk, C. F., and Ju, U. (2020), "Estimation of Response Styles Using the Multidimensional Nominal Response Model: A Tutorial and Comparison With Sum Scores," *Frontiers in Psychology*, 11, 72.
**URL:** *https://www.frontiersin.org/article/10.3389/fpsyg.2020.00072*

Harzing, A. (2006), "Response Styles in Cross-national Survey Research A 26-country Study," *International Journal of Cross Cultural Management*, 6(2), 243–266.

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

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

Kim, N., and Bolt, D. M. (2021), "A Mixture IRTree Model for Extreme Response Style: Accounting for Response Process Uncertainty," *Educational and Psychological Measurement*, 81(1), 131–154.
**URL:** *https://doi.org/10.1177/0013164420913915*

Levanthal, B., and Zigler, C. K. (2021), "A tree-based approach to identifying response styles with anchoring vignettes," , . Paper presented at the 2021 Annual Meeting of the National Council on Measurement in Education.

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

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

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

Ostendorf, F., and Angleitner, A. (2004), *Neo-PI-R: Neo-Persönlichkeitsinventar nach Costa und McCrae*, : Hogrefe.

Reynolds, N. L., and Smith, A. (2010), "Assessing the impact of response styles on cross-cultural service quality evaluation: a simplified approach to eliminating the problem," *Journal of Service Research*, 13(2), 230–243.

Rost, J. (1991), "A logistic mixture distribution model for polychotomous item responses," *British Journal of Mathematical and Statistical Psychology*, 44(1), 75–92.

Rothstein, M. G., and Goffin, R. D. (2006), "The use of personality measures in personnel selection: What does current research support?," *Human Resource Management Review*, 16(2), 155–180.

Sawatzky, R., Ratner, P. A., Kopec, J. A., and Zumbo, B. D. (2012), "Latent variable mixture models: a promising approach for the validation of patient reported outcomes," *Quality of Life Research*, 21(4), 637–650.

Van Herk, H., Poortinga, Y. H., and Verhallen, T. M. (2004), "Response styles in rating scales evidence of method bias in data from six EU countries," *Journal of Cross-Cultural Psychology*, 35(3), 346–360.

Van Vaerenbergh, Y., and Thomas, T. D. (2013), "Response styles in survey research: A literature review of antecedents, consequences, and remedies," *International Journal of Public Opinion Research*, 25(2), 195–217.

Weijters, B., Geuens, M., and Schillewaert, N. (2010), "The individual consistency of acquiescent and extreme response styles," *Applied Psychological Measurement*, 34(2), 105–121.

Weijters, B., Schillewaert, N., and Geuens, M. (2008), "Assessing response styles across modes of data collection," *Journal of the Academy of Marketing Science*, 36(3), 409–422.

Wetzel, E., and Carstensen, C. H. (2015), "Multidimensional Modeling of Traits and Response Styles," *European Journal of Psychological Assessment*, 33(5), 1–13.

Wetzel, E., Carstensen, C. H., and 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.

Yen, W. M., and Fitzpatrick, A. R. (2006), "Item Response Theory," in *Educational Measurement*, ed. R. Brennan, Vol. 4, Westport, CT: American Council on Education and Praeger Publishers, pp. 111–153.

Zettler, I., Lang, J. W., Hülsheger, U. R., and Hilbig, B. E. (2015), "Dissociating Indifferent, Directional, and Extreme Responding in Personality Data: Applying the Three-Process Model to Self- and Observer Reports," *Journal of personality*, .

Zhang, Y., and Wang, Y. (2020), "Validity of Three IRT Models for Measuring and Controlling Extreme and Midpoint Response Styles," *Frontiers in Psychology*, 11, 271. **URL:** *https://www.frontiersin.org/article/10.3389/fpsyg.2020.00271*