Extreme Response Style: A Simulation Study Comparison of Three Multidimensional Item Response Models

Applied Psychological Measurement
2019, Vol. 43(4) 322–335
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618789392
journals.sagepub.com/home/apm

Brian C. Leventhal¹

Abstract

Several multidimensional item response models have been proposed for survey responses affected by response styles. Through simulation, this study compares three models designed to account for extreme response tendencies: the IRTree Model, the multidimensional nominal response model, and the modified generalized partial credit model. The modified generalized partial credit model results in the lowest item mean squared error (MSE) across simulation conditions of sample size (500, 1,000), survey length (10, 20), and number of response options (4, 6). The multidimensional nominal response model is equally suitable for surveys measuring one substantive trait using responses to 10 four-option, forced-choice Likert-type items. Based on data validation, comparison of item MSE, and posterior predictive model checking, the IRTree Model is hypothesized to account for additional sources of construct-irrelevant variance.

Keywords

MIRT, extreme response style, MCMC, questionnaires

Introduction

Item responses to surveys may be influenced by construct-irrelevant variance factors such as a respondent's 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). Baumgartner and Steenkamp (2001) outlined several response styles that have been identified including acquiescence response style, disacquiescence response style, net acquiescence response style, response range, midpoint responding, and extreme response (ER) style.

ER style occurs when individuals systematically endorse response options at the ends of the rating scale. Individuals exhibiting an ER style may not always select an option at the end of the rating scale. Accordingly, an individual's tendency to select ERs is assumed to reside along a continuum. ER tendency has garnered significant research attention (Cabooter, Millet, Weijters, & Pandelaere, 2016; Greenleaf, 1992b; Hamilton, 1968). It is considered one of the two (along with acquiescence) most problematic response styles in attitude and survey research (Schuman

¹James Madison University, Harrisonburg, VA, USA

Corresponding Author:

Brian C. Leventhal, Department of Graduate Psychology, James Madison University, 298 Port Republic Road, Harrisonburg, VA 22807, USA.
Email: leventbc@jmu.edu

& Presser, 1981; van Herk, Poortinga, & Verhallen, 2004). Tendency to select ERs is associated with characteristics such as gender, education level, age, trait anxiety, and ethnicity (Berg & Collier, 1953; Greenleaf, 1992a, 1992b; Lewis & Taylor, 1955; Marin, Gamba, & Marin, 1992; Moors, 2012). As a result, the associations between the intended substantive trait and other constructs are confounded (Bolt & Newton, 2011).

The goal of this work is to compare, through simulation, three models of item response designed to account for ER: the multidimensional nominal response model (MNRM), the IRTree Model, and the modified generalized partial credit model (MPCM).

Background

Jackson and Messick (1958) stressed the development and evaluation of methods accounting for response styles. Van Vaerenbergh and Thomas (2013) outlined several methods proposed to account for ER style including representative indicators for response styles (RIRS; Baumgartner & Steenkamp, 2001; Greenleaf, 1992a, 1992b), representative indicators for response style means and covariance structure (RIRMACS; Weijters, Schillewaert, & Geuens, 2008), latent-class regression analysis (Moors, 2010; van Rosmalen, van Herk, & Groenen, 2010), latent-class confirmatory factor analysis (LCFA; Kieruj & Moors, 2010; Moors, 2003), and item response theory (IRT) models. RIRS, RIRMACS, and the latent-class methods are less favorable than IRT models (Van Vaerenbergh & Thomas, 2013). Specifically, RIRS and RIRMACS require additional items to be added to the survey to account for ER style. Although latent-class regression analysis and LCFA do not require additional items to account for ER style, they may be difficult to specify and require a familiarity with latent-class analysis limiting their accessibility to researchers. Alternatively, IRT models do not require additional items and are more accessible to psychological researchers (see Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Reeve, 2002; Yen & Fitzpatrick, 2006).

Three IRT models that account for ER tendencies have been selected in this study: the MNRM, the IRTree Model, and the MPCM. Each model is an extension of a popular model that has garnered significant research attention. The formulaic expressions of the MNRM, the IRTree Model, and the MPCM are presented in Online Appendix A.

The NR model (Bock, 1972) is a popular IRT model for nonordered categorical item responses. The unidimensional NR model has been used to investigate response scale options (Thissen, Cai, & Bock, 2010) in addition to many other applications in psychological research (see discussion in Thissen, 1993). The multidimensional extension of the NR model has been applied to modeling responses to testlets (Revuelta, 2014) as well as modeling response styles (Falk & Cai, 2016). Bolt and Johnson (2009) specified the MNRM to isolate the extreme tendency trait from substantive traits. They theorized the ER trait as a unique and independent trait from all others being measured.

As an alternative model to account for ER tendencies, Thissen-Roe and Thissen (2013) proposed modifying the two-parameter logistic model and the graded response model (Samejima, 1969) through the use of an item response tree. Item response tree applications have recently gained momentum in psychological research. Namely, item response trees have been used to model omitted responses (Jeon & De Boeck, 2016), to model answer change behavior (Jeon, De Boeck, & van der Linden, 2017), and to parse out rater effects in performance assessments (Myers, Ames, Leventhal, & Holzman, 2018). By specifying the IRTree Model, Thissen-Roe and Thissen (2013) sought to account for ER tendencies by modeling the response process. Specifically, they modeled the hypothesized two-stage sequential decision-making process used to answer Likert-type items.

The third model studied to account for ER tendencies extends the partial credit model (PCM; Masters, 1982). For a given item, the PCM models the transition from one category to the next. The PCM was generalized (Muraki, 1992) to allow items to relate unequally to the underlying trait being measured. The generalized partial credit model has many applications in educational and psychological testing. For example, the generalized PCM has been applied to scale polytomous items on the National Assessment of Educational Progress writing assessment (National Center for Education Statistics, 2008) and to model item responses from computerized adaptive assessments (Becker et al., 2005). Jin and Wang (2014) proposed a modified version of the generalized partial credit model (MPCM) to account for ER tendencies. To do this, they treat the distance between response options as individual random effects.

Comparison of the Models

The MNRM, the IRTree Model, and the MPCM all conform to a MIRT paradigm. These methods, however, approach modeling ER tendency uniquely. The MNRM introduces a second latent trait that is related to the response options in order of their extremity. The ER style trait has a fixed relationship with the probability of each response option. The MNRM forces an assumed relationship between the propensity to respond and the ER trait by fixing the category slope parameters. The IRTree Model estimates the slope associated with the ER trait. This model does not constrain the item slopes for each item nor does it fix the ER slopes across items. The IRTree Model allows for all items to have little to no relationship with the ER construct. In addition, the IRTree Model equates the propensity to select a *strong* intensity response option (i.e., ER response) across the initial decisions. This mirroring of ER propensities within an item is also evident in the MNRM. In the MNRM, the slopes of the ER tendencies trait are positive for ER categories and negative for non-ER categories (i.e., 1, $-.25$, $-.25$, $-.25$, $-.25$, 1 for a six-category item). Unlike the MNRM and the IRTree Model, the MPCM does not force a symmetric relationship between the conceptualized ER trait and the propensity for an item response. The MPCM does not fix the random category threshold parameters for ER tendencies nor does it assume equal propensities of a *strong* response regardless of agreement or disagreement. Thus, the MPCM is the least constrained of the three models compared.

Simulation Factors

A simulation study was performed to compare the three models under varying conditions of sample size, survey length, and the number of response options per item.

Sample size. Jin and Wang (2014) determined parameter recovery under the MPCM with small sample size was inadequate. Considering this inadequacy, it is important to note whether the IRTree Model or the MNRM are adequate for small sample sizes. To investigate sample size, small (500) and large (1,000) sample sizes were simulated.

Survey length. Responses to surveys of varying length have been evaluated using traditional descriptive and more advanced models accounting for ER tendencies. For example, Greenleaf (1992b) examined 16 items. Jin and Wang (2014) investigated a scale measuring interpersonal conflicts by Lo (2001) that had 20 items. Bolt and Newton (2011) considered two subscales from the Programme for International Student Assessment (PISA) exam: five items comprising the “Enjoy Science” subscale and 10 items covering the “Value of Science” subscale. With previous studies considered, responses to short (10 items) and long (20 items) surveys designed to measure a single latent trait were simulated.

Number of response options. The use of a neutral response option for Likert-type items has been the subject of debate. Bishop (1987) offered a neutral response to enable 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 (Johns, 2005; Krosnick et al., 2002). Conversely, studies have shown a significant number of individuals selecting “no opinion” or “neutral” when they truly do have an opinion (Bishop, 1987; Kalton, Roberts, & Holt, 1980). To eliminate any effects of neutrality, item responses were simulated for items that contained no middle or neutral option (an even number of response options). A four-option rating scale considered on the scale 1-*strongly disagree*, 2-*disagree*, 3-*agree*, and 4-*strongly agree* and a six-option rating scale considered as 1-*strongly disagree*, 2-*disagree*, 3-*somewhat disagree*, 4-*somewhat agree*, 5-*agree*, and 6-*strongly agree* were compared. Categories associated with *strong* opinions were classified as ERs.

Data Generation

Data were generated under each model. Item and trait distributions were selected to mimic past studies (Bolt & Johnson, 2009; Bolt & Newton, 2011; Jin & Wang, 2014). The item parameter generation distributions and the trait generating distributions for the MNRM, the IRTree Model, and the MPCM are presented in Table B1 in Online Appendix B. Data generation method was analyzed as an independent variable.

Model Evaluation

Item responses simulated under each model were fit with the IRTree Model, the MPCM, and the MNRM. Item mean squared error (IMSE) was evaluated to account for varying survey length and each model's unique parameterization. IMSE is the squared difference between observed total score, $\sum_{j=1}^J U_j$, and expected total score, $E(U_j|\theta) = \sum_{j=1}^J \sum_{k=1}^{m_j} k \times P_{jk}(\theta)$, divided by the number of items analyzed, J . The mean IMSE over R replications for each simulation condition is

$$\overline{\text{IMSE}} = \frac{\sum_{j=1}^J U_j - \sum_{j=1}^J \sum_{k=1}^{m_j} k \times P_{jk}(\theta)}{J \times R}. \quad (1)$$

Model Estimation

Parameter estimation was performed 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. Convergence to stable posterior distributions was checked using trace plots and autocorrelation plots. No issues with convergence were found. The mean of each parameter from the posterior distribution was used as the parameter estimate. Prior distributions were selected similar to those used in previous studies (Bolt & Newton, 2011; Jin & Wang, 2014; Johnson & Bolt, 2010). For more information regarding prior distributions and fixed values for item parameters, please refer to Online Appendix B. Example SAS syntax is provided in Online Appendix C.

Posterior Predictive Model Checking (PPMC)

A PPMC was performed to evaluate model fit using discrepancy measures (see Levy, Mislevy, & Sinharay, 2009; Stone & Zhu, 2015). Discrepancy measures at the item level and person level were selected to directly assess ER tendencies. The individual ER rate is calculated by dividing

the number of items respondents selected *strongly disagree* or *strongly agree* by the total number of items on the survey (Greenleaf, 1992b). Person-level model fit was evaluated by comparing the frequency of persons with each individual ER rate. The item ER rate, used to evaluate item-level fit, is calculated as the proportion of individuals who selected an ER option for each item (Greenleaf, 1992b). PPMC results are presented graphically in Online Appendix E.

Results

For each of the 72 simulation conditions, the mean and standard deviation of the IMSE are displayed in Table D1 in Online Appendix D. There is no evidence of a significant interaction between data generation, estimator, sample size, and the number of response options ($p = .22$); no evidence of a significant three-way interaction between estimator, sample size, and the number of response options ($p = .46$); nor evidence of an interaction between sample size and the number of response options ($p = .88$) on the pattern of mean IMSE. There is no evidence of a significant interaction between data generation, estimator, sample size, and the survey length ($p = .16$), no evidence of a significant three-way interaction of estimator, sample size, and the survey length ($p = .39$), nor evidence of an interaction between sample size and the survey length ($p = .11$) on the pattern of mean IMSE. Within simulation conditions, a priori simple pairwise comparisons of the mixed ANOVA were conducted.

Sample Size

Figure 1 displays the mean IMSE by data generation model, analysis model, and sample size averaged over the number of response categories and the number of items. For each data generation routine, there are no significant differences in mean IMSE among sample sizes for responses fit with all three models ($ps > .18$). Data fit with the MPCM has the lowest average IMSE, except when responses were simulated with the IRTree Model. Trivial differences exist in mean IMSE between responses fit with the MNRM and responses fit with the IRTree Model among sample sizes ($ps > .43$), except when responses were generated using the IRTree Model ($ps < .01$). Compared to responses fit with the MNRM and the IRTree Model, responses fit with the MPCM have significantly lower mean IMSE when generated under the MNRM ($n = 500, 1,000$; $ps < .02$) and when generated under the MPCM ($ps < .01$). Responses generated with the IRTree Model fit with the MNRM result in the highest mean IMSE across sample sizes.

Survey Length

There is evidence of an interaction between survey length, the number of response options, data generation method, and the estimation model on the pattern of mean IMSE, $p = .03$. Figure 2 displays mean IMSE for items with four response options. Consider data simulated with the IRTree Model displayed in the center panel. Surveys with 10 items have significantly lower mean IMSE than surveys with 20 items ($ps < .01$), except for responses fit with the MNRM ($p = .28$). Responses to 20 items fit with the IRTree Model have lower mean IMSE than those responses fit with the MNRM ($p < .01$) and fit with the MPCM ($p < .01$). Responses to 10 items fit with the MNRM have significantly larger average IMSE than when fit with the MPCM ($p < .01$) and the IRTree Model ($p < .01$). Responses simulated with the MNRM are displayed in the left panel. For each estimation routine, there is no significant difference in mean IMSE for surveys with 10 items compared to surveys with 20 items (IRTree Model, $p = .65$; MNRM, $p = .99$; MPCM, $p = .89$). There are significant differences in mean IMSE between responses fit

Figure 1. Mean IMSE for 500 and 1,000 responses generated under the MNRM, the IRTree Model, and the MPCM estimated by the IRTree Model, the MNRM, and the MPCM.

Note. IMSE = item mean squared error; MNRM = multidimensional nominal response model; MPCM = modified generalized partial credit model.

Figure 2. Mean IMSE for responses to surveys with 10 and 20 items with four response options generated under the MNRM, the IRTree Model, and the MPCM estimated by the IRTree Model, the MNRM, and the MPCM.

Note. IMSE = item mean squared error; MNRM = multidimensional nominal response model; MPCM = modified generalized partial credit model.

with the MPCM and the IRTree Model to surveys with 10 items ($p = .03$) and surveys with 20 items ($p < .01$). Significant mean IMSE differences are not evident between responses fit with the IRTree Model and the MNRM for surveys with 10 items ($p = .54$) nor for surveys with 20 items ($p = .42$). Results for responses simulated with the MPCM are displayed in the right panel. When responses are fit with each model, there is no significant difference in mean IMSE for surveys with 10 items compared to surveys with 20 items (IRTree Model, $p = .15$; MNRM, $p =$

Figure 3. Mean IMSE for responses to surveys with 10 and 20 items with six response options generated under the MNRM, the IRTree Model, and the MPCM estimated by the IRTree Model, the MNRM, and the MPCM.

Note. IMSE = item mean squared error; MNRM = multidimensional nominal response model; MPCM = modified generalized partial credit model.

.98; MPCM, $p = .65$). Responses fit with the MPCM have lower mean IMSE than when fit using the IRTree Model (10, 20 items: $p < .01$). There are no significant differences in mean IMSE among responses fit with the MNRM and fit with the MPCM (10 items: $p = .88$; 20 items: $p = .53$).

Consider items with six response options as seen in Figure 3. When fit with the MNRM, responses to 10 items have significantly lower mean IMSE than responses to 20 items across data generation routines ($ps < .01$). Responses fit with the IRTree Model display a significant effect of survey length on mean IMSE (responses generated with IRTree Model: $p < .01$; MPCM: $p < .01$), except when simulated under the MNRM ($p = .75$). Responses to surveys consisting of 10 and 20 items fit with the MPCM only exhibit significant mean IMSE differences when generated under the IRTree Model, $p < .01$ (data generated under MNRM, $p = .81$; MPCM, $p = .54$). Responses to 20 items fit with the MPCM have lower mean IMSE compared to responses fit with the MNRM ($ps < .01$) and responses fit with the IRTree Model ($ps < .01$), except when simulated under the IRTree Model. Responses generated under the IRTree Model to 20 items fit with the IRTree Model reveal significantly lower mean IMSEs than responses fit with the MPCM ($p < .01$) and responses fit with the MNRM ($p < .01$). Responses to 10 items fit with the MPCM and the MNRM result in insignificant mean IMSE differences (data generated under MNRM, $p = .98$; MPCM, $p = .30$), except when responses were simulated using the IRTree Model ($p < .01$). For responses to 10 items generated under the IRTree Model, no significant difference in mean IMSE exists between fit with the MPCM and fit with the IRTree Model, $p = .13$. Under this data generation routine, the IRTree and the MPCM have significantly lower mean IMSE than the same responses fit with the MNRM ($ps < .01$).

Number of Response Options

Mean IMSE by data generation model, analysis model, and number of responses options averaged over sample size for 10-item surveys is displayed in Figure 4. Mean IMSE is greater for

Figure 4. Mean IMSE for responses to 10 items with four and six response options generated under the MNRM, the IRTree Model, and the MPCM estimated by the IRTree Model, the MNRM, and the MPCM. Note. IMSE = item mean squared error; MNRM = multidimensional nominal response model; MPCM = modified generalized partial credit model.

six-option items compared with four-option items across data generation models ($ps < .02$). Consider data simulated under the IRTree Model (center panel). Trivial mean IMSE differences exist for responses fit with the IRTree Model compared to responses fit with the MPCM ($ps > .12$). Fit with the MNRM results in significantly larger mean IMSE compared to fit with the IRTree Model (four options; six options, $p < .01$) and fit with the MPCM ($ps < .01$). Responses simulated using the MNRM are presented in the left panel. There is no discernible difference in mean IMSE between responses fit with the MNRM and fit with the MPCM ($ps > .97$). Responses fit with the MPCM have significantly lower mean IMSE than the IRTree Model (four options, six options; $p < .01$) for data simulated using the MPCM (right panel).

Figure 5 displays mean IMSE by number of response options for 20-item surveys. The mean IMSE for six-option items is significantly greater than for four-option items ($ps < .01$). Data simulated using the IRTree Model are presented in the center panel. In this case, fit with the IRTree Model results in the smallest mean IMSE. For six-option surveys, the MNRM has the greatest mean IMSE ($ps < .01$). There are no discernible differences in mean IMSE between responses to four-option items fit with the MNRM and fit the MPCM, $p = .62$. Mean IMSE by number of response options for data simulated using the MNRM is presented in the left panel. There are no ostensible differences in mean IMSE between fit with the three models when responses to items with four response options are considered. Responses to items with six response options fit with the MNRM have the highest mean IMSE ($ps < .01$). The same response fit with the MPCM have significantly lower mean IMSE compared to fit with the IRTree Model ($p < .01$). A similar pattern is evident for responses simulated under the MPCM, visualized in the right panel, except no significant difference in mean IMSE exists between responses to items with six-option fit with the IRTree Model and fit the MNRM, $p = .10$.

Discussion

Through simulation, the MNRM, the MPCM, and the IRTree Model were compared. The sample size (500, 1,000), survey length (10 items, 20 items), and the number of response options

Figure 5. Mean IMSE for responses to 20 items with four and six response options generated under the MNRM, the IRTree Model, and the MPCM estimated by the IRTree Model, the MNRM, and the MPCM. Note. IMSE = item mean squared error; MNRM = multidimensional nominal response model; MPCM = modified generalized partial credit model.

(four, six) varied. The three models are proposed to account for ER style. The IRTree Model offers modeling the two-stage decision-making process; the MNRM treats ER tendencies as an independent dimension; and the MPCM proposes a random effects model. IMSE was assessed between expected and true total score.

Responses were generated using each of the three models. Responses generated with the IRTree Model resulted in average IMSE markedly different than the other two models. During data validation, item ER rates and individual ER rates were averaged over replications. Responses simulated using the MNRM and the MPCM exhibited similar individual and item ER rates. Responses simulated using the IRTree model had distinctly different rates compared with the other two models.

Responses simulated with the IRTree Model fit with IRTree Model outperformed the same responses fit with both the MNRM and the MPCM. This pattern was consistent across all simulation conditions. PPMC results were consistent with data validation and model fit for the IRTree Model. Inadequate model fit was evidenced when responses generated with the MNRM and MPCM were fit with the IRTree Model. The IRTree Model is designed to model a two-step decision-making response process. The results suggest that the IRTree Model may not be appropriate for modeling ER style. It is likely that, in addition to ER style, the response process model is accounting for other construct-irrelevant variance sources. To account for additional construct-irrelevant variance, such as other response styles, a generalized IRTree model (Böckenholt, 2017; Böckenholt & Meiser, 2017) has been proposed. This generalized IRTree reduces the constraints from the IRTree Model presented here. Specifically, the generalized version does not force decision two probabilities to be symmetric across decision one options. Relaxing this symmetry assumption allows the model to account for multiple response style effects.

For all three models compared in this study, there was no difference in mean IMSE across sample sizes. There was evidence of an interaction between survey length and the number of response options on mean IMSE. Responses fit with the MPCM consistently resulted in the lowest mean IMSE across simulation conditions. For 10-item surveys, however, there was no

difference in mean IMSE between the MPCM and the MNRM, except for data simulated under the IRTree Model.

Responses to items with six response options resulted in larger mean IMSE than items with four response options. Jin and Wang (2014) noted that an item response scale should be lengthened to obtain a more precise estimate of the ER tendency trait. They did not, however, have to consider the inherent differences between four and six response option items when evaluating trait parameter recovery. The difference observed in the current study may be due to the outcome measure, IMSE. There is more variability in responses to items with six options compared to items with four options. As a result, more statistical noise exists between expected total score and observed total score. Conclusions, therefore, are only made about model comparison and should not be made about model parameter recovery.

In this study, longer surveys tended to have higher mean IMSE than shorter surveys. This study is limited to having only considered surveys with at most 20 items. Simulation studies examining item and person parameter behavior for alternative models accounting ER style have explored surveys with higher lengths (see de Jong, Steenkamp, Fox, & Baumgartner, 2008). In addition, the items on the surveys in this simulation are assumed to be homogeneous. This is unlike alternative studies with longer surveys that may be deliberately constructed with heterogeneous items (Greenleaf, 1992b). Therefore, the reader is cautioned when extrapolating the results of this study to surveys longer than 20 items.

PPMC results were generally consistent with the pattern of mean IMSE comparing the MNRM with the MPCM. For the condition with large differences of mean IMSE, the item ER rates indicated inadequate model fit for the MNRM and adequate model fit for the MPCM. The item ER rates exhibited no differences in model fit in the condition with minimal differences in mean IMSE across models of estimation. Person-level PPMC methods showed differences in the MNRM and MPCM for the condition when there were large differences of mean IMSE. The results of the PPMC methods deviated from the IMSE results in the condition with small IMSE differences. This may be a result of the low sample size. Observed and predicted frequencies of the 500 individuals were scarce or nonexistent in some of the 21 unique individual ER rates analyzed. This scarcity may provide misleading results when interpreting the individual ER rates as a discrepancy statistic.

Conclusion

Overall, the MPCM resulted in the lowest mean IMSE among the three models across sample size, survey length, and the number of category response options. Evidence suggests that the IRTree Model is measuring a unique process compared with the MNRM and the MPCM. Based on the results of this study, the MPCM model is recommended to use when modeling ER style. To account for ER style, the MNRM is adequate when modeling responses to short surveys with few response options. The IRTree Model is not recommended to account strictly for ER tendencies. Further evaluation of the IRTree Model should be conducted to provide insight in the appropriate applications of the model.

Recall that the IRTree Model assumes equal propensity for the intensity decision and the MNRM fixes the category slope parameters of the ER trait's relationship with the probability of item response. Unlike these two models, the MPCM does not put a symmetric constraint on model parameters associated with the ER trait. This may explain why the MPCM results in the lowest mean IMSE in most simulation conditions, specifically across data generation routines, compared with the other models.

In this simulation study, when the data are simulated under each model, it is assumed to reflect the truth. The true model and approach to account for ER style is unknown. This study

is intended to help researchers determine which model is most appropriate for their use. The MPCM is recommended due to its less constrained nature and performance in the simulation study. This simulation indicated that the MPCM resulted in lower mean IMSE under more conditions compared with both the IRTree Model and the MPCM.

The simulation conditions were chosen to represent a variety of practical applications. The simulation conditions, however, are still limited to those selected. 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. It may be evident that the IRTree Model is more suitable for items with an odd number of response categories. Questions about individuals drawn to ERs versus individuals drawn to neutral or middle categories should be addressed.

The use of Bayesian Markov chain Monte Carlo (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. 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 ER tendencies (see Jin & Wang, 2014).

The prior distributions used in the study were based on previous studies of the three models. Further evaluation of the prior distributions effect on the estimation is required. For example, when a category intercept parameter is assumed to follow a normal prior distribution with mean 0, variance 25, it is relatively flat with approximately 50% of the values falling between -3.5 and 3.5 . This prior is considered weakly informative. As the variance of the prior decreases, however, the distribution becomes more peaked and dominant relative to the data. The prior distribution effect on the estimation of the model parameters requires further evaluation.

ER style and the substantive trait were assumed to be independent. There are several traits that are correlated with ER style such as self-concept clarity and simplistic thinking (Cabooter, Millet, et al., 2016; Naemi, Beal, & Payne, 2009). Recent research suggests that ER style may be domain specific (Cabooter, Weijters, De Beuckelaer, & Davidov, 2016). Further investigation of ER tendencies should incorporate correlated substantive traits and ER 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 ER is still developing. For example, Thissen-Roe and Thissen (2013) presented a reparametrized version of the proportional thresholds model (PTM, Rossi, Gilula, & Allenby, 2001) to account for ER style. Besides models within the IRT paradigm, there also exist several alternative models as previously mentioned. The models in this study were selected as they are extensions to common IRT models. Supplementary evaluation of alternative models accounting for ER should be performed and compared with the IRTree Model, the MNRM, and the MPCM.

Acknowledgment

The author thanks Dr. Clement A Stone for his assistance in this work.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 143-156.
- Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., Rose, M., & Walter, O. B. (2005). Development of a Computer-Adaptive Test for Depression (D-CAT). *Quality of Life Research*, 14, 2277-2291.
- Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement*, 13, 164-169.
- Bishop, G. F. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly*, 51, 220-232.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, 22, 69-83.
- Böckenholt, U., & 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, 159-181.
- Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, 33, 335-352.
- Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, 71, 814-833.
- Cabooter, E., Millet, K., Weijters, B., & Pandelaere, M. (2016). The "I" in extreme responding. *Journal of Consumer Psychology*, 26, 510-523.
- Cabooter, E., Weijters, B., De Beuckelaer, A., & Davidov, E. (2016). Is extreme response style domain specific? Findings from two studies in four countries. *Quality & Quantity*, 51, 2605-2622.
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494.
- Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, 10, 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, NJ: Lawrence Erlbaum.
- de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, 45, 104-115. doi:10.1509/jmkr.45.1.104
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, 21, 328-347. doi:10.1037/met0000059
- Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29, 176-188.
- Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, 56, 328-351.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.
- Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, 69, 192-203.
- Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, 55, 243-252.
- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, 48, 1070-1085.
- Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, 42, 467-490. doi:10.3102/1076998616688015
- Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, 74, 116-138.

- Johns, R. (2005). One size doesn't fit all: Selecting response scales for BES attitude items. *Journal of Elections, Public Opinion and Parties*, 15, 237-264.
- Johnson, T. R., & Bolt, D. B. (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, 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, 65-78.
- Kierul, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research*, 22, 320-342. doi: 10.1093/ijpor/edq001
- 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 satisfy? *Public Opinion Quarterly*, 66, 371-403.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33, 519-537. doi: 10.1177/0146621608329504
- Lewis, N. A., & Taylor, J. A. (1955). Anxiety and extreme response preferences. *Educational and Psychological Measurement*, 15, 111-116.
- Lo, K.-Y. (2001). *Interpersonal harmony and the values of forbearance: Understanding generation gap through interpersonal conflicts* (NSC Research Report No. NSC90-2413-H-031-006-SSS). Taipei, Taiwan: National Science Council.
- 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, 498-509.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality & Quantity*, 37, 277-302. doi:10.1023/A:1024472110002
- Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research*, 22, 93-119. doi: 10.1093/ijpor/edp036
- Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, 21, 271-298. doi:10.1080/1359432X.2010.550680
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176. doi:10.1177/014662169201600206
- Myers, A. J., Ames, A. J., Leventhal, B. C., & Holzman, M. A. (2018, April). *An item response tree model for validating rubric scoring processes*. Paper presented at the Annual meeting of the National Council of Measurement in Education, New York, NY.
- Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, 77, 261-286. doi:10.1111/j.1467-6494.2008.00545.x
- National Center for Education Statistics. (2008). *NAEP technical documentation: The generalized partial credit model*. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_models_gen.aspx
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1: Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.
- Reeve, B. B. (2002). *An introduction to modern measurement theory*. National Cancer Institute. Retrieved from <https://pdfs.semanticscholar.org/d6b1/0ae949ff4d89b2bfe27a36e40fd83e7aeb6e.pdf>
- Revuelta, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, 38, 549-562. doi:10.1177/0146621614536272
- Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*, 96, 20-31.

- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 35, 139.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York, NY: Academic Press.
- Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. Cary, NC: SS Institute.
- Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 79-97). Hillsdale, NJ: Lawrence Erlbaum.
- Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. N. R. Ostini (Ed.), *Handbook of polytomous item response theory models* (pp. 43-75). New York, NY: Routledge.
- Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type Items. *Journal of Educational and Behavioral Statistics*, 38, 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 6 EU countries. *Journal of Cross-Cultural Psychology*, 35, 346-360.
- van Rosmalen, J., van Herk, H., & Groenen, P. J. F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, 47, 157-172. doi: 10.1509/jmkr.47.1.157
- Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, 25, 195-217. doi:10.1093/ijpor/eds021
- Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, 36, 409-422. doi:10.1007/s11747-007-0077-6
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: American Council on Education.