Title

The Heteroscedastic Skew Graded Response Model: An Answer to the Non-Normality Predicament?

Author

Rodriguez, Anthony

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The Heteroscedastic Skew Graded Response Model: An Answer to the Non-Normality Predicament?

A dissertation submitted in partial satisfaction of the requirement for the degree Doctor of Philosophy in Psychology

by

Anthony Rodriguez

2017

ABSTRACT OF THE DISSERTATION

The Heteroscedastic Skew Graded Response Model:

An Answer to the Non-Normality Predicament?

by

Anthony Rodriguez

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2017

Professor Steven Paul Reise, Chair

As item response theory models are more frequently applied to psychological assessment, understanding the ramifications of failing to account for non-normality is of utmost importance, especially, considering the likelihood of encountering traits that are non-normally distributed in the population (e.g., anxiety, depression). Previous research has established concerns with regard to bias in item parameter and trait score estimates when non-normality is ignored, and, as such developed models to aid in minimizing bias. Once such model is the heteroscedastic GRM with a skewed latent trait (HSGRM). This research provides an in-depth examination of the viability and utility of the HSGRM. Under various degrees of skew and heteroscedasticity, including extreme on both, this research addresses the consequences of ignoring non-normality and how to address it. A simulation study was conducted to evaluate the ability of the HSGRM to provide improved item parameter estimates, and recover the shape of the distribution (i.e., skew). Results support the HSGRM as a major improvement over the traditional GRM when faced with non-normality in data due to skew in the trait and heteroscedastic errors.

The dissertation of Anthony Rodriguez is approved.

Peter M. Bentler

Hongjing Lu

Jose Felipe Martinez

Steven Paul Reise, Committee Chair

University of California, Los Angeles

2017

For Angela and Petrizia

and

In Memory of Patricia Brown

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My deepest and most sincere appreciation and admiration goes to my advisor Dr. Steven Reise. Throughout every stage of my career at UCLA, Steve has been an unbelievable source of inspiration, guidance, mentorship, and, of course, unparalleled humor. I thoroughly enjoyed our lively discussions about research interwoven with topics like Spruce Goose, baseball, cinema, television, our personal collections, or my favorite – classic rock. I was always amazed by how Steve managed to find the link between psychometric theory and the most current episode of Game of thrones – a true skill! I will follow his example of how to be an exceptional researcher, academic, and man throughout the rest of my career and life.

Many thanks to the other member of my dissertation committee as well, Dr. Peter Bentler, Dr. Hongjing Lu, and Dr. Jose Felipe Martinez for their invaluable feedback and critique. I am also extremely grateful to Dr. Dylan Molenaar for his guidance and code, both of which made this dissertation possible. I would also like to thank Dr. Andrew Moskowitz for spending countless hours brainstorming with me ideas for tackling this dissertation – a true friend. To Dale Kim, I am extremely grateful for his assistance in creating extremely efficient code for scraping my output. This saved my sanity! As a whole, I would like to thank my colleagues and friends for their moral support and friendship over the years.

Finally, I am grateful to my family. A very special thanks goes to my grandmother. She always believed in me and was certain I would achieve great things. She inspired me to look ahead and to let no obstacle interfere with my goals. To my parents, I am extremely grateful for the support and love they have given me over the years and for always encouraging my intellectual development. Above all others, I thank my wife, Angela. She has been my rock and foundation throughout this process - a constant source of support, encouragement, love, reassurance, and, above all, patience! My true beloved! And of course, to our precious daughter, Petrizia. Although she is only four months old, she has been so influential. She has compelled me to work faster and harder so I could enjoy the wonders of being her father.

VITA

EDUCATION

2012	California State University, Fullerton M.A., Psychology
2010	California State University, Fullerton B.S., Child and Adolescent Studies *Magna Cum Laude*

SELECT PUBLICATIONS AND PRESENTATIONS

Rodriguez, A., Reise, S. P., Spritzer, K. L., & Hayes, R. D. (in press). Alternative approaches to addressing non-normal distributions in application of IRT model to patient-reported outcomes. *Medical Care*

Incollingo Rodriguez, A. C., **Rodriguez, A**., Callahan, L. C., Saxbe, D., & Tomiyama, A. J. (in press). The buddy system: A randomized controlled experiment of the benefits and costs of dieting in pairs*. Journal of Health Psychology*

Dominguez-Lara, S., & **Rodriguez, A.** (2017). Índices estadísticos de modelos bifactor. Interacciones. Publicación anticipada en línea (Advanced online publication). doi:10.24016/2017.v3n2.51

Reise, S. P., & **Rodriguez, A.** (2016). Item response theory and the measurement of psychiatric constructs: Some empirical and conceptual issues and challenges. *Psychological Medicine, 46(*10), 2025-2039. doi:10.1017/S0033291716000520

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223-237. (2016): Correction to: Applying bifactor statistical indices in the evaluation of psychological measures, *Journal of Personality Assessment*, doi:10.1080/00223891.2015.1117928

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating, interpreting, and applying statistical indices. *Psychological Methods, 21*(2), 137-150. doi:10.1037/met0000045

Marelich, W. D., Grandfield, E., Graham, J., Warstadt, M., & **Rodriguez, A.** (2014). Initial contact on dating websites: Results and strategies from the L.U.R.E. Project. *Electronic Journal of Human Sexuality, 17.*

Gallego, J. C., & **Rodriguez, A.** (2013). Castilian speakers' attitudes towards accents and regional stereotypes in Spain. *Sociolinguistic Studies, 6.3,* 149-178. doi:10.1558/sols.v6i3.543.

Bonifay, W., & **Rodriguez, A**. (2016, July). *A graded response model for items that elicit avoidance of extreme categories.* Paper presented at the Annual Meeting of the Psychometric Society, Asheville, NC.

Incollingo Rodriguez, A. C., **Rodriguez, A.,** Nguyen-Cuu, J.*, Standen, E. C.*, White, M. L.*, Callahan, L. C., Saxbe, D., Tomiyama, A. J. (2016, May). *Debunking the Buddy System: The Unintended Consequences of Dieting in Pairs*. Paper presented at the annual Convention of the Association for Psychological Science, Chicago, IL.

Incollingo Rodriguez, A. C., **Rodriguez, A.,** Callahan, L. C., Saxbe, D., Tomiyama, A. J. (2016, March). *Second-Hand Stress: Physiological and Psychosocial Consequences Associated with Dieting in Pairs.* Poster presented at the American Psychosomatic Society Annual Scientific Meeting, Denver, Colorado.

Tanaka, S. M.*, **Rodriguez, A.**, & Reise, S. P. (2015, April). *Stereotype vulnerability and mindset of women in upper division computer science.* Poster presented at the annual meeting of the Western Psychological Association, Red Rock NV.

Bell, A., & **Rodriguez, A.** (2014, February). *Sociocultural vulnerabilities as mediators of the effects of individual risk factors on body dissatisfaction in college-aged women.* Poster presented at the annual meeting of the Society for Personality and Social Psychology, Austin, TX.

Gottfried, A. W., **Rodriguez, A.,** & Gottfried, A. E. (2013, April). *What mediates reading to young children and reading achievement?* Paper presented at the biennial meeting of the Society for Research in Child Development, Seattle, WA.

Oliver, P., Reichard, R. J., **Rodriguez, A.,** Wray-Lake, L., Gottfried, A. W., & Gottfried, A. E. (2013, April). *Encouragement of Leadership: From Adolescence to Early Adulthood.* Paper presented at the biennial meeting of the Society for Research in Child Development, Seattle, WA.

SELECT AWARDS

Shepard Ivory Franz Distinguished Teaching Assistant Award

Walter Klopfer Award – Best Empirical Paper 2016 in Journal of Personality Assessment

Eugene V. Cota-Robles Fellowship

INTRODUCTION

Item response theory (IRT) has become an increasingly attractive approach for addressing measurement issues. Whereas, classical test theory has historically dominated the field of psychometrics, more and more applications of IRT are implemented in a variety of substantive domains including medical and health outcomes research (e.g., Cella, & Stone, 2015; Garcia, Aryal, & Walters, 2015; Gorter & Fox, 2015; McCracken, Chilcot, & Norton, 2015), cognitive (e.g., Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2015; Shono, Ames, & Stacy, 2015), behavioral genetics (e.g., Murray, Molenaar, Johnson, & Krueger, 2016), psychopathology (e.g., Parent, McKee, Rough, & Forehand, 2016; Olino, 2016; Keeley, Webb, Peterson, 2016), and of course, for quite some time now, in educational assessment (e.g., SAT, GRE, state standardized testing). Moreover, IRT has been applied in the development of item pools to be administered through computer adaptive tests (CAT) as well as in the identification of differential item function. Through mathematical modeling, IRT offers statistical methods for evaluating items and scales, creating and administering psychological measures, and ultimately measuring individuals on psychological traits (Reise, Ainsworth, & Haviland, 2005).

A great deal of work in IRT has been focused on dichotomous models, namely, the one-, two-, and three-parameter logistic models. However, in psychological research, many constructs of interest such as anxiety, attitudes, and personality traits fall along a continuum and thus benefit from response options that allow for improved measurement (i.e., Likert-type formats). For these types of response formats, polytomous IRT models are more appropriately suited. One commonly applied polytomous IRT model is the Graded Response Model (GRM; Samejima, 1969, 1996). For unidimensional models, the GRM estimates a single discrimination (slope) parameter and *c-1* category-boundary thresholds. Each of these thresholds represents the point on the latent trait (θ) where there is a 50% probability of responding in and above a given category. These threshold parameters, in turn, are used to calculate the probability of responding in a specific category conditional on the latent trait. The threshold parameters

determine where the psychometric information (conditional precision of measurement) is located and the slope determines how much. The estimated item parameters, therefore, are critically important in terms of determining how well a measure is functioning; Inaccuracy or bias in item parameter estimation can lead to a faulty interpretation of a scale's precision.

Normality and IRT Parameter Estimation

The basic process of fitting an IRT model is straightforward: That is, data are collected from a polytomously scored measure, an appropriate IRT model is selected (e.g., the GRM), the model parameters are estimated, and trait scores are produced. However, in order for the estimated model parameters to be accurate several critical assumptions have to be met. For example, most IRT models are estimated using marginal maximum likelihood (MML; Bock & Lieberman, 1970). MML, as originally presented, makes certain assumptions, for instance, that individuals are independent, item responses are independent conditional on the latent trait, item response curves are logistic, and that before estimating item parameters, the probability distribution of the population is specified (Bock & Lieberman, 1970). Bock and Aitkin (1981) relaxed the final assumption such that the form of the trait distribution need not be specified *a priori*.

Despite this, though, a normal distribution is generally assumed for the population (by default in many IRT software packages) and models are fit accordingly. That said, for many constructs this might be entirely reasonable, but in psychology, in particular, the normality assumption might be particularly unlikely for many constructs. In psychology, traits like depression or anxiety or other severe disorders would not be expected to be normally distributed in a general adult population. Similarly, in education research, the issue of non-normality might also present concerns. For instance, proficiency may be expected to be non-normally distributed in particular subgroups of interest (e.g., English learners) or perhaps the overall distribution can be characterized as a mixture of normals (Monroe & Cai, 2014). As cleverly noted by Micceri (1989) in an investigation of 440 large-sample achievement and

psychometric measures, the normal distribution, among other fantastical and mythical creatures, might not best characterize particular constructs or traits. Therefore, when the latent trait is not expected to be normally distributed, or even if there is uncertainty, to assume that the latent trait is normally distributed is likely a model misspecification. This can have definite ramifications, as will be seen shortly.

Part of the disregard of normality violations, in the context of IRT, can be attributed to the widespread notion among substantive research that estimation of IRT models is robust to normality violations (e.g., Cooper, Balsis, & Zimmerman, 2010; Greco, Lambert, & Baer, 2008; Kim, Kim, & Kamphaus, 2010; Krueger & Finger, 2001; McGlinchey & Zimmerman, 2007; Meade, 2010; Purpura, Wilson, & Lonigan, 2010; Samuel, Simms, Clark, Livesley, & Wigider, 2010; Thomas & Locke, 2010). On the other hand, a great deal of simulation work has demonstrated that non-normality in the latent trait can have deleterious effects on item discrimination, category parameters, and trait scores. When the latent trait is non-normal, bias has been found to emerge in parameter estimates (Abdel-fattah, 1994; Boulet, 1996; De Ayala & Sava-Bolesta, 1999; DeMars, 2003; Kirisci, Hsu, & Yu, 2001; Stone, 1992; Wollack, Bolt, Cohen, Lee, & Young-Sun, 2002) and specifically for item slopes (Azevedo, Bolfarine, & Andrade, 2011), trait score estimates (Seong, 1990; Ree, 1979; Swaminathan & Gifford, 1983; Woods & Lin, 2009) and in item category parameters (Preston & Reise, 2014) where it has been found that as the true shape of the latent trait deviates from a normal distribution, extreme item threshold parameters become more biased (van den Oord, 2005; Zwinderman & van der Wollenberg, 1990). The findings provide a concrete basis for investigating alternative methods designed to handle non-normality. The message is clear. Failing to account for non-normality can have negative effects and bias item parameter estimates.

Deciding on how to address non-normality in the latent trait, given the negative ramifications, has proven to be a topic of great interest in the psychometric community. To this end, a variety of methods have been proposed and evaluated under extensive simulation work.

Some of the earliest work approached the issue of non-normality by using the empirical histogram method (EHM; Mislevy, 1984; Mislevy & Bock, 1990; Schmitt, Mehta, Aggen, Kubarych, & Neale, 2006). During the maximization step of the EM algorithm (MML/EM; Bock & Aitkin, 1981), the shape of the latent trait is estimated using quadrature points. This method, however, was problematic as the shape of the distribution could easily change depending on the number of quadratures being used. As the number of quadratures increase so do the number of parameters being estimated, and ultimately, the end results is often a jagged representation of the latent trait. To address this jaggedness, Ramsay-Curve IRT was developed (RC-IRT; Woods, 2006, 2007; Woods & Thissen, 2004, 2006), which estimates the latent trait density using a spline-based approach wherein the density estimate is smooth and also requires fewer estimated parameters. This method was further refined using Davidian Curve IRT (DC-IRT; Woods & Lin, 2009) wherein the attractive features of RC-IRT are retained, but require fewer tuning parameters.

More recently, consideration has been given to the notion of positive or unipolar traits where the trait is not on a continuum from $-\infty\ to\ \infty$ but rather from $0\ to\ \infty$. To capture this, Lucke (2014) proposed a set of unipolar IRT (UIRM) models such as the Log-Logistic and Weibull UIRM, to properly capture these types of traits (e.g., gambling addiction). In this conceptualization, a trait score of 0 reflects “no disorder” while trait scores greater than 0 reflect the presence of the disorder to some degree. As another alternative, Wall, Park and Moustaki (2015) proposed a zero-inflated mixture IRT model. Rather than assume the trait is normally distributed in the overall population, the model assumes that the non-normality stems from a mixture of two populations, that is, a degenerate no-trait population and a traited population for whom the trait is relevant and normally distributed. Thus, the overall distribution is comprised of both a clinical and non-clinical class – a group for whom the trait is present and another for whom it is not. The ***majority*** of the non-clinical group is excluded from analyses such that the estimation of trait scores and item parameters are done in the strictly clinical group (*for a*

detailed discussion see Wall et al., 2015).

In an altogether separate vein, a novel approach, one that forms the basis of this investigation, is to conceptualize the sources of non-normality as stemming from not only a skewed latent trait but also in terms of heteroscedastic errors (Molenaar, Dolan, & De Boeck, 2012; Molenaar, 2015). For the polytomous case, Sameijima's (1969) GRM can be derived by assuming a linear relationship between the latent trait θ and a normally distributed variable U_i that underlies ordinal item responses. This normality assumption on the underlying variable U_i brings with it the assumption of homoscedastic errors. Moreover, the latent trait is also typically assumed to be normally distributed. When data are in fact normal and thus symmetric, the implication is that $U_{i|\theta}$ is also symmetric and thus category response functions are symmetric. When the conditional distribution of $U_{i|\theta}$ is skewed, the result can be asymmetric category response functions. However, as noted by Molenaar et al. (2012) asymmetry in the category response functions may not be simply attributable to skewness in the conditional distribution of $U_{i|\theta}$. In fact, asymmetry in category response functions can occur even when $U_{i|\theta}$ has a symmetrical distribution. That is, when $U_{i|\theta}$ has a symmetric distribution, asymmetry in the category response function can arise from heteroscedastic errors in the regression between U_i and θ. Therefore, to account for skew in the latent trait and/or heteroscedasticity, Molenaar et al. (2012) proposed the heteroscedastic GRM with a skewed latent trait distribution model wherein heteroscedastic errors are modeled and a skew-normal distribution is used rather than the traditional normal distribution (to be described in depth below). This model is flexible and can test for both skew and heteroscedasticity, or either individually, and more importantly, treats the traditional GRM as a special case with no skew in the trait and heteroscedasticity.

To date, there is only the single study that proposed and examined the heteroscedastic GRM with a skewed latent trait. In the original paper (Molenaar et al., 2012), the model performed well under very specific conditions, that is, depending on sample sizes (*N* = 400 and

800), skew and heteroscedasticity parameters were recovered well. As for item discrimination and threshold parameters, when compared to results obtained from the traditional GRM, the model was a significant improvement, especially as heteroscedasticity increased. When heteroscedasticity was small, though, there was little effect. However, the conditions under which simulations were conducted were far less extreme than those faced in psychological and educational data. Specifically, skew was fixed to either 0 or +/- 0.5 and heteroscedastic errors were fixed to 0 or 0.4. In practice, skew routinely exceeds +/-0.5, especially in the context of psychological disorders. Consider Figure 1a where skew is specified to 0.5 with no heteroscedasticity and Figure 1b where skew is set to 0.5 and heteroscedasticity to 0.4. Although clearly there are some violations of normality, consider Figure 2a which is based on simulated data with skew = .99 and 0.8 heteroscedasticity and 2b produced from real impulsivity data taken from the Barratt Impulsiveness Scale (Patton & Stanford, 1995).

Although not identical, the real data is quite similar to the far more extreme simulated data. The question then becomes, how well does the heteroscedastic GRM with a skewed latent trait perform under these types of extreme non-normal conditions? The primary objective of this project is to perform an extensive investigation of the utility and viability of this model under extreme conditions, similar to those faced in psychological and, to a lesser degree, educational research. The objective can be broken into the following key components:

1) What is the consequence of model misspecification, that is, failing to model skew and heteroscedastic errors?

2) Establish the viability of the HSGRM such that the model can correctly recover item parameters when there is no skew or heteroscedasticity present (i.e., normal ogive GRM).

3) In the presence of heteroscedasticity and/or skew, investigate parameter recovery and bias when fitting the true model and HSGRM.

4) Determine the frequency with which the correct model was properly identified and the prevalence of the HSGRM being overly selected as the best fitting model.

In order to further understand the issues involved in the present research, some background information on polytomous IRT models, the GRM, MML estimation, and the heteroscedastic GRM with a skewed latent trait model is necessary and will be discussed in the following sections.

Polytomous IRT models

Polytomous IRT models provide researchers a means to improve measurement by evaluating scales consisting of multi-point Likert-type response options. Whereas in the dichotomous case, parametric models are designed to mathematically model the relation between individual differences on a latent trait and the probability of endorsing an item, polytomous IRT models are interested in the relation between individual differences on the latent trait and the probability of responding in a particular category (Bock, 1972; Drasgow, Levine, Tsien, Williams & Mead, 1995; Thissen, 1976; Thissen & Steinberg, 1984).

Describing this relationship can be accomplished by examining operating characteristic curves (OCCs) also known as threshold response curves (TRCs), category response curves (CRCs), and item response curves (IRCs). However, deciding which to use, in part, is a function of the chosen polytomous IRT model (e.g., Graded Response Model: Samejima, 1969; Generalized Partial Credit: Muraki, 1992; 1993; Nominal Response Model: Bock, 1972). As noted by Embretson and Reise (2000), a set of CRCs offer a particularly useful method for describing polytomous items and, depending on the IRT model, can be produced using different equations. CRCs convey information regarding the probability of responding in a particular

category conditional on the trait level.

Consider an item with *c*=5 category response options (1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Neutral*, 4 = *Agree*, 5 = *Strongly Agree*). The CRCs for this hypothetical item are presented in Figure 3. For the lowest response option (*x = 1*), the probability of endorsing this category is a monotonically decreasing function such that as trait levels increase, the probability of endorsing this category decreases. On the other end, for the highest response option (x = 5), the probability of endorsing the highest category increases monotonically as a function of increases in the latent trait. Unimodal functions characterize the probability of category endorsement for the middle categories (x = 2, 3, or 4). Specifically, the probability of endorsing a particular category increases and then decreases as a function of increases in the trait level.

As previously mentioned, there are a variety of polytomous IRT models. In fact, these models can be categorized into three classes that have been investigated extensively: sequential models, adjacent-category models, and cumulative-boundary models (Andersen, 1977, 1997; Andrich 1978a, 1978b, 1995; Hemker, van der Ark, & Sijtsma, 2001; Masters, 1982; Masters & Wright, 1997; Mellenbergh, 1995; Muraki, 1990, 1992; Samejima, 1969, 1972; Tutz, 1990). Sequential models (Tutz, 1990), also known as continuation ratio models (Molenaar, 1983; Hemker et al., 2001; Mellenbergh, 1995), model the process of responding above a between-category boundary under the assumption that previous boundaries have been passed. Adjacent-category models (Molenaar, 1983) also referred to as divide-by-total models (Thissen & Steinberg, 1986), and Rasch Models (Andrich, 1995) are based on modeling the process of responding in category x versus x-1 (i.e., x = 2 vs. 1; x = 3 vs. 2; x = 4 vs. 3; x = 5 vs. 4) and can be understood as constrained versions of the nominal response model. Cumulative-boundary models, also referred to as cumulative probability models (Molenaar, 1983), difference models (Thissen & Steinberg, 1986), and Thurstone models (Andrich, 1995) extend the standard two parameter logistic IRT model (discussed below) by modeling the process of responding above a between-category boundary (i.e., x = 1 vs 2,3,4,5; x = 1, 2 vs 3,4,5; x = 1,2,

3 vs 4,5; x = 1,2,3,4 vs 5). This last class of models is of particular interest in the current paper and so we direct our attention to the Graded Response Model.

Graded Response Model: Logistic and Normal Ogive

When polytomous data are ordered categorical responses, a suitable option, even when items do not have the same number of response options, is to use the Graded Response Model (GRM; Samejima, 1969; 1996). The logistic GRM is an extension of Birnbaum's (1968) two-parameter logistic (2PL) model where the conditional probability of item endorsement is

$$P(X = 1|\theta) = \frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}} \quad (1)$$

and is determined by a single slope (α) and location (β).

This mathematical form is then generalized in the logistic GRM such that each item is still described by a *single* slope parameter (α_i), but the equation now includes *c-1* $= (m_i)$ between-category threshold parameters (β_{ij}'s). Each of these thresholds, in turn, is associated with its respective operating characteristic curve (OCC), also referred to as threshold response curves (TRCs), similar to the item characteristic curve obtained from a 2PL model. As noted in Figure 4, these OCCs represent a series of dichotomies in which a 2PL is fit to each between-category boundary with the added constraint that, within an item, the slope is equal. Therefore, threshold parameters are easily interpretable such that they represent the trait level required to have a 50% probability of responding in or above a given threshold. Estimating an OCC for each threshold is the first step in a two-step process toward computing the category response probabilities for an item. The OCC is estimated by

$$P_{ix}^{*}(\theta) = \frac{e^{\alpha_i(\theta-\beta_{ij})}}{1 + e^{\alpha_i(\theta-\beta_{ij})}} \quad (2)$$

reflecting the conditional probability of an individual's item response (x) falling above a given threshold ($j = 1 \ldots m_i$). By taking the difference between OCCs, step two, the response probabilities for each category can be computed:

$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{i(x+1)}(\theta) \quad (3)$$

Given that the probability of responding at or above the lowest category is 1.0, and it is impossible to respond above the highest category, we can therefore compute the probabilities for responding in each category conditional on the latent trait. Using our example item with 5 response options, we compute response category probabilities as:

$$\begin{aligned} P_{i0}(\theta) &= 1.0 - P^*_{i1}(\theta) \\ P_{i1}(\theta) &= P^*_{i1}(\theta) - P^*_{i2}(\theta) \\ P_{i2}(\theta) &= P^*_{i2}(\theta) - P^*_{i3}(\theta) \\ P_{i3}(\theta) &= P^*_{i3}(\theta) - P^*_{i4}(\theta) \\ P_{i4}(\theta) &= P^*_{i4}(\theta) - 0 \end{aligned}$$

The resulting CRCs inform on the probability of an individual responding in a particular response category conditional on their standing on the latent trait. Figure 3 demonstrates that responding in the first category (*Strongly Disagree*) is most likely for an individual who has trait levels at or below -1.5SD. Responding in the second category (Disagree) is more likely for an individual with traits estimates between -1.5 and -0.5SD. Individuals most likely to endorse the third category (Neutral) are those with trait levels that fall between -0.5 and 0.5SD. Individuals with trait scores between 0.5 and 1.5SD are more likely to make use of the fourth response option (Agree) while those with trait score estimates greater than 1.5SD are more likely to

endorse the fifth response option (Strongly Agree).

Taken together, it is apparent that the shape and location of CRCs, and therefore also OCCs, is determined by item parameters (α, β_{ij}). More specifically, steep OCCs and peaked narrow CRCs are a function of generally larger slopes and thus are indicative of greater item information. The location and spread of OCCs and CRCs along the latent trait continuum depend on the threshold parameters, points on the trait where item information is maximized.

An alternative is to use a normal distribution function instead of the logistic which can be derived from a factor analytic (regression) framework. The resulting model is referred to as the normal-ogive GRM. Under this framework, it is assumed that a normally distributed variable, U_i, underlies each categorical response variable, y_i, a relation expressed via threshold parameters. The underlying variable, U_i, is therefore modeled as a linear function of the latent trait, θ.

$$U_i = \nu_i + \lambda_i \theta + \epsilon_i \tag{4}$$

where ν_i is the intercept, λ_i is the factor loading, and ϵ_i is the measurement error. The marginal distribution of U_i, $g(.)$ is found by

$$g(U_i) = \int_{-\infty}^{\infty} f(U_{i|\theta}) h(\theta) d\theta \tag{5}$$

Here the density function of θ is $h(.)$ and the conditional density function of $U_{i|\theta}$ is $f(.)$, which reflects scores on U_i conditioned on the latent trait. Moreover, the conditional mean and variance are

$$E\left(U_{i|\theta}\right) = \nu_i + \lambda_i \theta$$
$$Var\left(U_{i|\theta}\right) = \sigma_{\epsilon i}^2 \tag{6}$$

Therefore, for an individual at a given trait level, θ, the probability of endorsing a particular category *c* can be found by

$$P\left(Y_{i|\theta} = c\right) = \int_{\tau_{ic}}^{\tau_{i(c+1)}} f(U_{i|\theta}) dU_{i|\theta} \quad for\ c = 0, \ldots, C_i - 1 \tag{7}$$

Here, an individual's observed score on item *i* conditioned on the latent trait is denoted $Y_{i|\theta}$ with C_i indicating the number of response options for a given item *i* and $\tau_{i0} = -\infty$ and $\tau_{i(C_i-1)} = \infty$. Moreover, given the expected value and variance of $U_{i|\theta}$ we can compute the probability of an individual endorsing a given response category given their trait score by

$$P\left(Y_{i|\theta} = c\right) = F\left(\frac{\nu_i + \lambda_i\theta - \tau_{ic}}{\sigma_{\epsilon i}}\right) - F\left(\frac{\nu_i + \lambda_i\theta - \tau_{i(c+1)}}{\sigma_{\epsilon i}}\right) \tag{8}$$

and when substituting $\nu_i = 0$, $\alpha_i = \frac{\lambda_i}{\sigma_{\epsilon i}}$, and $\beta_{ic} = -\frac{\tau}{\sigma_{\epsilon i}}$ we arrive at the normal-ogive GRM

$$P\left(Y_{i|\theta} = c\right) = F(\alpha_i\theta + \beta_{ic}) - F\left(\alpha_i\theta + \beta_{i(c+1)}\right) \tag{9}$$

with the traditional item discrimination (α_i: slope) parameter and category location parameter (β_{ic}: threshold).

The model, as discussed above, brings with it distributional assumptions, specifically, that the latent trait (θ) and errors (ϵ_i) are normally distributed and, through convolution, we know that U_i must also be normally distributed. Moreover, we know that when U_i is normally distributed, by means of Cramer's theorem (Cramer, 1937), the latent trait (θ) and $U_{i|\theta}$ are both normally distributed. Therefore, it logically follows that, if U_i and θ are assumed to be linearly related, then asymmetry in the marginal distribution of U_i is attributable to a variety of sources: 1) heteroscedastic errors ($\sigma^2_{\epsilon i}$) across the latent trait θ; 2) skewness in the latent trait

distribution; and/or 3) skewness in the distribution of $U_{i|\theta}$. Research devoted to examining sources of asymmetry in the marginal distribution of U_i have primarily focused on skewness in the distribution of $U_{i|\theta}$ (Samejima, 1997; 2000; 2008; Bazán et al., 2006; Ramsay & Abrahamowicz, 1989). However, as noted by Molenaar, Dolan, and De Boeck (2012), very little attention has been given to heteroscedasticity despite its effect on category response functions. It is also true that when the latent trait is skewed, non-normality can manifest in U_i. To this end, Molenaar et al. (2012) proposed an extension of the GRM to include both heteroscedastic $\sigma_{\epsilon i}^2$ and a skewed latent trait θ.

Heteroscedastic GRM with a Skewed Latent Trait (HSGRM)

Recall that under the marginal distribution of U_i and given the expected value and variance of $U_{i|\theta}$, the $\sigma_{\epsilon i}^2$ are assumed to be homoscedastic. Therefore, to account for potential heteroscedasticity, Molenaar et al. (2012) propose modeling $\sigma_{\epsilon i}^2$ as a function of the latent trait θ

$$\sigma_{\epsilon i|\theta}^2 = k(\theta; \delta_i) \tag{10}$$

such that *k*(.) can take on any strictly positive function and includes a parameter vector δ_i with elements $\delta_{i0}, \ldots, \delta_{ir}$ where *r* reflects an *r*th degree polynomial function. Molenaar et al. (2012) considered an exponential function such as that proposed by Hessen and Dolan (2009)

$$\sigma_{\epsilon i|\theta}^2 = \exp\left(\delta_{i0} + \delta_{i1}\theta + \delta_{i2}\theta^2 + \cdots + \delta_{ir}\theta^r\right). \tag{11}$$

In its simplest form

$$\sigma_{\epsilon i|\theta}^2 = \exp\left(\delta_{i0} + \delta_{i1}\theta\right) \tag{12}$$

with *r* = 1, baseline parameter $\delta_{i0} \in (-\infty, \infty)$ and heteroscedasticity parameter $\delta_{i1} \in (-\infty, \infty)$, it is clear that homoscedasticity is not violated when $\delta_{i1} = 0$, whereas $\delta_{i1} > 0$ suggests that as

traits levels increase across θ so do error variances, while $\delta_{i1} < 0$ reflects decreases in error variances as trait levels increase. However, the exponential function is problematic for the extended GRM given that the presence of heteroscedasticity in the distribution of U_i results in skewed CRCs and, more importantly, results in upper limits for the highest and lowest category response options. For an in depth discussion see Molenaar et al. (2012).

To address this issue, an alternative function was proposed

$$\sigma^2_{\epsilon i|\theta} = \frac{2\delta_{i0}}{1+\exp\left(-\delta_{i1}\frac{\theta - E(\theta)}{SD(\theta)}\right)} \tag{14}$$

where now the baseline parameter has a lower limit, $\delta_{i0} \in [0, \infty]$ while $\delta_{i1} \in [-\infty, \infty]$. For the lowest and highest response categories, an upper bound has now been set to avoid approaching 0.5 such that when $\delta_{i1} > 0$ and $\theta \to \infty \Rightarrow \sigma^2_{\epsilon i|\theta} \to \delta_0$. A lower limit has been set to 0 such that $\theta \to -\infty \Rightarrow \sigma^2_{\epsilon i|\theta} \to 0$. We also see that when $\delta_{i1} < 0$ and $\theta \to -\infty \Rightarrow \sigma^2_{\epsilon i|\theta} \to \delta_{i0}$ and when $\theta \to -\infty \Rightarrow \sigma^2_{\epsilon i|\theta} \to 0$. Finally, when $\delta_{i1} = 0 \Rightarrow \sigma^2_{\epsilon i|\theta} \to \delta_{i0}$.

The above approach identifies and models one of the sources of non-normality, namely, heteroscedastic error variance. However, skewness in the latent trait also presents an issue and must be modeled and distinguished from heteroscedastic errors. The skew normal distribution (Azzalini 1985, 1986; Azzalini & Capatanio, 1999) has much to offer and has been recently incorporated into the 2-parameter normal-ogive model (Bazán, Branco, & Bolfarine, 2005) as well as the 2PL and extended into a Bayesian framework (Azevedo, Bolfarine & Andrade, 2011; Azevedo, Bolfarine, & Andrade, 2012). Moreover, the added attraction of the skew-normal distribution is that the normal distribution is a special case. The probability density function of the skew-normal for a random variable, x, is given by

$$h(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi(\psi\left(\frac{x-\xi}{\omega}\right)) \tag{15}$$

where $\phi(.)$ denotes the standard normal probability density function, $\Phi(.)$ is the cumulative density function, ξ is a location parameter, $\xi \in (-\infty, \infty)$, ω is a scale parameter, $\omega \in [0, \infty)$, and ψ is a shape parameter, $\psi \in (-\infty, \infty)$. Using these parameters we can easily obtain the expected value, variance, and skew of x by

$$E(x) = \xi + \omega\rho\sqrt{\frac{2}{\pi}} \tag{16}$$

$$Var(x) = \omega^2\left(1 - \frac{2\rho^2}{\pi}\right) \tag{17}$$

$$Skew(x) = \frac{4-\pi}{2}\frac{\left(\rho\sqrt{\frac{2}{\pi}}\right)^3}{\left(1-\frac{2\rho^2}{\pi}\right)^{\frac{3}{2}}} \tag{18}$$

where

$$\rho = \frac{\psi}{\sqrt{1+\psi^2}} \tag{19}$$

The probability density function of the skew normal model can be reduced to the normal density function when the shape (skew) parameter, ψ, is equal to 0.

In order for the model to be identified, we must first consider that in Eq. 8 both the latent trait and U_i are unobserved variables and therefore require constraints for identification. By convention these would be $E(\theta) = 0$, $E(U_i) = 0$, $Var(\theta) = 1$, and $Var(U_i) = 1$. However due to the estimation of the model (as shown shortly), these specifications result in

$$\sigma^2_{\epsilon i|\theta} = Var(U_i) - \lambda_i^2 * Var(\theta) = 1 - \lambda_i^2 \tag{20}$$

$$\nu_i = E(U_i) - \lambda_i * E(\theta) = 0 \tag{21}$$

Thus, the model cannot include heteroscedastic errors given that $\sigma_{\epsilon i}^2$ is not a free parameter such that error variances are a function of factor loadings. Moreover, intercepts (ν_i) are not included in the model. In order to identify the scale and define the unit of measurement of U_i while also being able to estimate $\sigma_{\epsilon i}^2$ and ν_i for all items, a reasonable approach is adopted where constraints are imposed on the first two thresholds thus defining the unit of measurement (Lee, Poon, & Bentler, 1990; Mehta, Neale, & Flay, 2004; Shi & Lee, 2000). Therefore, now that $\sigma_{\epsilon i}^2$ is a free parameter and can be estimated by Eq. 14, and given that Eq. 15 specifies the density function for a skew normal distribution underlying θ, the heteroscedastic GRM with a skewed latent trait (HSGRM) is formed (Molenaar et al., 2012). It is worth noting that an interesting feature of the HSGRM is that, unlike the graded response model, it does not impose symmetric category response curves.

Marginal Maximum Likelihood

Traditional maximum likelihood estimation is problematic for IRT because it attempts to simultaneously estimate both item and person parameters and, as such, will always add additional parameters as subjects are added. In place of maximum likelihood, one of the most frequently applied full information estimation procedures for fitting parametric IRT models is Marginal Maximum Likelihood (MML; Bock & Lieberman, 1970; Bock & Aitkin, 1981). Instead of individual theta estimates, MML imposes a population distribution and then estimates item parameters given the assumed distribution. Moreover, given that trait levels are unknown, MML uses response pattern probabilities as expectations belonging to a population distribution (Bock & Lieberman, 1970). For this reason, observed data are viewed as randomly sampled from the population (Embretson & Reise, 2000). In part, the appeal of MML rest in the fact that data are integrated over the parameter distribution and parameters are estimated by maximum-likelihood in the marginal distribution. In its original form, Bock and Lieberman (1970) presented MML in the context of dichotomously scored items. For a given subject, their response pattern would

assign them to one of 2^n mutually exclusive categories. In a sample with *N* individuals with *j* response patterns, the frequency of the response patterns are denoted r_j , where *j*=1,2, …, *k* and $k \leq \min(N, K)$ with *N* possible response patterns. Therefore, assuming *N* subjects represent a normal random sample ($\mu = 0, \sigma = 1$) from a population, then category response frequencies (r_i) are equal to Np_i and will be multinomially distributed with parameters N and P_i, where $E(p_i) = P_i$. Thus, by means of the multinomial law, item parameters can be used to inform on the probability of the sample, resulting in the likelihood function

$$L = \frac{N!}{\prod_j r_j!} \prod_{j=1}^{2^n} P_j^{r_j} \tag{22}$$

Item parameter likelihood equations for the slope (a_i) and intercept (k_i) are

$$\frac{\partial logL}{\partial a_i} = N \sum_{j=1}^{2^n} \frac{p_j}{P_j} \frac{\partial P_j}{\partial a_i} = 0 \tag{23}$$

and

$$\frac{\partial logL}{\partial k_i} = N \sum_{j=1}^{2^n} \frac{p_j}{P_j} \frac{\partial P_j}{\partial k_i} = 0 \tag{24}$$

Although this procedure can be applied to any IRT model, issues arise due to the optimization procedure. The Newton-Raphson algorithm presents computational concerns as the number of items increases beyond 12 items (Bock & Aitkin, 1981). From a computational standpoint, this method is extremely demanding given that, for *n* items, a *2n x 2n* information matrix must be generated and inverted between four to five times where each element in the information matrix is the sum of 2^n terms. For obvious reasons this is problematic especially in longer tests. Another issue pertains to a priori knowledge of the distributional form of the trait. Given these concerns, Bock and Aitkin (1981) reformulated the Bock-Lieberman likelihood equations (Dempster, Laird, & Rubin, 1977) in order to relax the need to specify the distributional form by

allowing the distributional form to be approximated as a discrete distribution with a set number of points. Therefore, item parameters can be estimated by simply integrating over the marginal distribution. This was accomplished using the EM algorithm's two-step process: expectation (E) and maximization (M).

In the MML/EM estimation procedure, observed response pattern probabilities are estimated at each iteration where each of the probabilities involves an integral. Numerical integration is necessary because the integrals do not have a closed-form solution. This is usually accomplished via Gauss-Hermite quadratures (Davis & Rabinovitz, 1975). The EM algorithm iteratively produces maximum likelihood model parameter estimates. During the E step, a provisional set of item parameters is obtained and treated as the true item parameters. Using these item parameters and the response patterns, the expected proportion of individuals selecting a given category is estimated. During the M-step, item parameters are estimated under the assumption that the response proportions obtained in the E-step are true probabilities. This process is repeated iteratively until a specified convergence criteria is reached.

To better understand this procedure, consider the parameterization presented by Samejima (1997) for a normal-ogive graded response model where the marginal likelihood function is expressed as:

$$L(\gamma) = \prod_{p=1}^{P} \int_{-\infty}^{\infty} h(\theta_p) P_{\upsilon_p}(\theta_p) d\theta_p \quad = \quad \prod_{p=1}^{P} P_{\upsilon_j} \quad = \quad \prod_{V} P_{\upsilon}^{r_\upsilon} \tag{26}$$

where $\boldsymbol{\gamma}$ is the vector of item parameters, *P* number of examinees, $h(\theta_p)$ is the trait density function, υ_p is a response pattern corresponding to subject *p*, $P_{\upsilon_p}(\theta)$ is the conditional probability for a given response pattern for subject p, P_{υ_p} is the marginal probability for response pattern υ_p for subject *p*, and r_υ is the frequency of a particular response pattern υ. By discretizing the

continuous latent trait θ by q discrete latent classes such that θ_l is characterized by $l = 1,2, \ldots, q$, P_υ can be approximated by $\widetilde{P_\upsilon}$ such that

$$\widetilde{P_\upsilon} = \sum_{l=1}^{q} P_\upsilon(\theta_l)H(\theta_l) \tag{27}$$

where $H(\theta_l)$ reflects the Gauss-Hermite quadrature weight. Then the expected frequency $\tilde{r}_{c_i k}$, for the graded response c_i for item i in latent class l is obtained by

$$\tilde{r}_{c_i k} = \frac{\sum_\upsilon r_\upsilon x_{\upsilon c_i} P_\upsilon(\theta_l)H(\theta_l)}{\widetilde{P_\upsilon}} \tag{28}$$

where $x_{\upsilon c_i}$ is an indicator variable that takes on a value of 1 if $c_i \in \upsilon$, otherwise gets a value of 0. Subsequently the expected sample size, $\widetilde{P_l}$ for class l is obtained by

$$\widetilde{P_l} = \frac{\sum_\upsilon r_\upsilon x_{\upsilon c_i} P_\upsilon(\theta_l)H(\theta_l)}{\widetilde{P_\upsilon}} \tag{29}$$

Therefore, during the E-step, using temporary item parameters, $P_\upsilon(\theta_l)$ is computed, in order to subsequently obtain $\widetilde{P_\upsilon}, \tilde{r}_{c_i l}$, and $\widetilde{P_l}$. Then in the M-step, updated estimates for item parameters are computed by maximizing the approximated likelihood function by replacing P_υ with $\tilde{P}_\upsilon$. This process happens iteratively until a convergence criteria is met.

A MML/EM procedure, as discussed above, is adapted to estimate the heteroscedastic GRM with a skewed latent trait model. Given equation 8 and σ_ϵ^2 as specified in Equation 14 combined with the latent trait density function in Equation 15, the log-marginal likelihood function is maximized by

$$L(\boldsymbol{\gamma}|\boldsymbol{X}) = \sum_{p=1}^{N} log \int_{-\infty}^{\infty} \prod_{i=1}^{q} P\left(Y_{i|\theta)} = x_{(pi)}\right) h(\theta) d\theta \tag{30}$$

where parameter vector $\boldsymbol{\gamma}$ contains $\delta_{i0}, \delta_{i1}, \lambda_i, \nu_i, \tau's$, and ψ for all $i = 1, \ldots, q$ and the N x q item response data matrix $\boldsymbol{X}$ contains responses for N individuals on q items and where $x_{(pi)}$ refers to a specific elements in the X matrix. Moreover, as established in Eq. 16 and 17, both ξ and ω are fixed parameters in order to identify the latent trait θ such that $E(\theta) = 0$ and $Var(\theta) = 1$. Note that the density function *h*(.) included in the likelihood function is not a standard normal but rather has been modified to be consistent with Eq. 15 such that the skew-normal distribution is specified and its corresponding item parameters are contained in parameter vector $\boldsymbol{\gamma}$.

METHOD

Design

This study examined the utility and viability of the HSGRM model under a broad range of conditions. To this end, heteroscedasticity, skew, sample size, and the number of response category options were manipulated. Heteroscedastic errors varied from the homoscedastic case to highly heteroscedastic, that is, four conditions were included: no heteroscedasticity ($\delta_{i1} = 0$), small ($\delta_{i1} = 0.40$), moderate ($\delta_{i1} = 0.80$), and large ($\delta_{i1} = 1.0$). Skew in the latent trait distribution was simulated under four likely scenarios: no skew (0), small (0.50), moderate (0.75), and large (1.0). Two sample sizes were included to reflect typical scenarios, that is, a small (*N*=500) and large sample (*N*=2,000). The number of response categories also varied to include both 3- and 5-category options. All other parameters were fixed to predetermined values. Factor loadings (unstandardized) were set to 1.0, threshold parameters fixed to -2.5, -1.0, 1.0, and 2.5, and baseline residual (δ_{i0}) to 1.0. All 64 conditions were simulated with 10 items.

Due to complexities in the estimation of the HSGRM, 200 datasets per condition were generated. To clarify, the program used to estimate the model (Mx; Neale, Boker, Xie, & Maes, 2004) allows for a single analysis at a time (no looping) and currently takes approximately 1-2 hours to estimate per dataset. For each condition, four models were estimated and compared via likelihood ratio test (LRT): 1) baseline normal-ogive GRM, 2) skew-normal GRM (no heteroscedastic parameter), 3) het-only GRM (no skew parameter), and 4) HSGRM with both heteroscedasticity and skew estimated. This time intensive procedure was critical in addressing the question of determining the frequency with which correct models are identified and the prevalence of over selecting the HSGRM when not merited.

Likelihood Ratio Test

As stated, LRTs were used to test the various models that contain skewness and/or heteroscedastic errors or neither. For some or all the items, under the H_0, $\nu = 0$ and/or $\delta_{i1} = 0$,

therefore, under the H_A, $\nu \neq 0$ and/or $\delta_{i1} \neq 0$. The LRT test statistic, *T*, is computed by

$$T = -2\,x\,[L(\widehat{\boldsymbol{\gamma}}_0|\boldsymbol{X}) - L(\widehat{\boldsymbol{\gamma}}_A|\boldsymbol{X})] \quad (25)$$

In accordance with restrictions specified under the H_0, $\widehat{\boldsymbol{\gamma}}_0$ is the estimated parameter vector from Eq. 24 whereas $\widehat{\boldsymbol{\gamma}}_A$ is the estimated parameter vector under the H_A. The *T* test statistic is distributed as a central χ^2 under the H_0, with degrees of freedom equal to the number of constraints. The test statistic is distributed as a non-central χ^2 under the H_A with a non-centrality parameter contingent on sample and effect size. Additionally, because the H_0 is nested within the H_A and restrictions under the H_0 are limited to constraining ν and/or δ_{i1} with parameter spaces $(-\infty, \infty)$ to 0 conditions are met for the test statistic to approach the theoretical distributions specified under the H_0 and H_A.

Data Generation

All conditions were simulated in R 3.2.3 (R Development Core Team, 2016). Using the *sn* (Azzalini, 2015) and *MASS* (Venables & Ripley, 2002) libraries, 200 datasets, per condition, were generated under the normal-ogive GRM with the skew and heteroscedastic specifications discussed above.

Model Fitting

In order to identify the scale and define the unit of measurement of U_i while also being able to estimate $\sigma_{\epsilon i}^2$ and ν_i for all items, constraints were imposed on the first two thresholds thus defining the unit of measurement (Lee, Poon, & Bentler, 1990; Mehta, Neale, & Flay, 2004; Shi & Lee, 2000). Therefore, the first two thresholds were fixed to -2.5 and -1, respectively. Each of the simulated datasets were fit to the HSGRM by specifying skew and/or heteroscedasticity using Mx freeware (Neale, Boker, Xie, & Maes, 2004). Mx is a combination of a matrix algebra interpreter and a numerical optimizer that allows for the exploration of matrix

algebra through countless operations and function. To date, there is no software that includes this model, nor are there any R functions available to estimate this model. From each Mx run, all item parameters and model fit indices were compiled for all four models (baseline, skew-only, het-only, and HSGRM) in order to evaluate overall parameter recovery and model selection. Parameter recovery was evaluated by examining the average bias and root-mean square difference (RMSD) averaged across items and replications (Woods, 2006).

The goal of this project was to perform an extensive investigation of the utility and viability of this model under extreme conditions. To that end, this research aimed to understand: a) The consequences of model misspecification, that is, failing to model skew and heteroscedastic errors; b) Establish the viability of the HSGRM such that the model can correctly recover item parameters when there is no skew or heteroscedasticity present (i.e., normal-ogive GRM); c) In the presence of heteroscedasticity and/or skew, investigate parameter recovery and bias when fitting the true model and HSGRM; and, d) Determine the frequency with which the correct model was properly identified and the prevalence of the HSGRM being overly selected as the best fitting model.

To address the consequence of model misspecification, that is, failing to model skew and heteroscedastic errors, all data with varying degrees of skew and heteroscedasticity were fit to a normal-ogive GRM. Average absolute bias in parameters and RMSDs were computed across all conditions. In order to establish the viability of the HSGRM, control conditions, that is, conditions with no skew or heteroscedastic errors, were fit to both a constrained version of the HSGRM (i.e., normal-ogive GRM or baseline model) and the full HSGRM with both skew and heteroscedastic errors estimated. Average bias and RMSDs were compared for all item parameters from both models. In the presence of heteroscedasticity and/or skew, parameter recovery was evaluated in terms of average bias and RMSDs for both the true model and HSGRM. Finally, all four models (baseline, skew-only, het-only, and HSGRM) were estimated for each dataset within each condition. Likelihood ratio tests were performed for each dataset in

every condition. These were compiled to evaluate the frequency with which the correct model was properly identified and the prevalence of the HSGRM being overly selected as the best fitting model.

RESULTS

Simulation results are presented in Tables 1 through 21. As a reminder, the skew parameter value presented is not the actual skew of the distribution but rather the skew parameter used in the computation of the actual skew (see eqs.18-19): skew of 0 (skew parameter = 0), 0.5 (skew parameter = 2.17), 0.75 (skew parameter = 3.641), and 1.0 (skew parameter = 28)[1]. Tables 1 through 10 provide descriptive statistics and are described below. Tables 11 through 21 address questions pertaining to establishing parameter bias due to model misspecification, performance of HSGRM under control conditions, item parameter recovery and evaluation of bias and RMSD, and prevalence or correct model identification. For all tables, simulation conditions (i.e., degree of heteroscedastic errors and skew) are denoted by "h" and "s" for heteroscedastic errors and skew, respectively. Large sample (N=2000) results are presented first then small sample results (N=500). In all cases, results from the 5-category response option simulation are presented first, followed by the 3-category response option findings.

Descriptive Statistics

Tables 1 through 10 include descriptive statistics for all item parameters averaged across items and replications for all 64 conditions. Tables 1 through 6 are for items with 5-category response options. Tables 7 through 10 are for items with 3-category response options. Each table includes means and standard deviations for the item parameter (e.g., factor loading) from each of the four possible models: Baseline GRM (no skew or het) denoted "B", skew only denoted "S", heteroscedastic errors only denoted "H", and heteroscedastic skew model denoted "HS". Tables are divided such that large sample descriptives are included first followed by small sample descriptives. Table 1 and 7 combine skew and heteroscedasticity into a single table. Hyphens are used to indicate item parameters not relevant to the estimated model (i.e., skew

[1] A special thanks to Dylan Molenaar for in depth conversations regarding HSGRM model specification and parameter interpretation.

parameter in a heteroscedastic error only model).

Five category response conditions

Descriptives for skew and heteroscedasticity parameters are presented in Table 1. The baseline model is not included as neither of these parameters were estimated. Skew is considered for the skew-only model and HSGRM, as the het-only model does not estimate skew. Heteroscedastic errors are considered for the het-only model and HSGRM given that the skew-only model does not estimate heteroscedastic errors.

Skew. Broadly speaking, for both the skew-only model and HSGRM, when data only included skew (i.e., h0s0, h0s5, h0s75, h0s1), mean skew parameters and thus actual skew closely reflected the true population skew. However, as heteroscedastic errors were included at increasing magnitudes (i.e., h4, h8, h1) and not modeled in the skew-only model, the mean skew parameters (and actual skew) were consistency underestimated for low and moderate skew (0.5 and 0.75). When heteroscedastic errors were modeled using the HSGRM, mean skew parameters tended to approximate the true population value. At the extremes, that is, skew of 0.00 or 1.00, both the skew-only model and HSGRM tended to have mean skew parameters near the true population values. Standard deviations between models were small-to-moderate and generally very similar. These trends were consistent regardless of sample size. The key distinction between large and small samples sizes can be observed in the magnitude of the standard deviations such that, generally speaking, standard deviations were larger in the small sample yet consistent between the skew-only and HSGRM.

Heteroscedastic Errors. For conditions including only heteroscedastic errors and no skew (i.e., h0s0, h4s0, h8s0, h1s0), both the het-only model and HSGRM produced mean heteroscedastic error values near or at the true population value. As skew increased and was left unmodeled, the het-only model consistently had smaller mean heteroscedastic errors with the most notable effect when skew was 1.0. Regardless of skew, the HSGRM had mean heteroscedastic errors close to or exactly the same as the true pop value. Standard deviations

between models were small and generally very similar. In the small sample the same trend is apparent. As mentioned in the case of skew, the key distinction between large and small samples is noted in the magnitude of the standard deviations such that, generally speaking, standard deviations were larger in the small sample yet consistent between the het-only and HSGRM.

Baseline residual. Table 2 displays the descriptives for baseline residuals. In the absence of heteroscedastic errors, all four models (baseline, skew-only, het-only, and HSGRM) had similar mean baseline residual values (0.90 – 1.02) as well as standard deviations (0.09 – 0.12). Once heteroscedastic errors were introduced, noticeable differences emerged. As heteroscedastic errors increased, mean baseline residuals gradually decreased in both the baseline and skew-only models, to 0.56 and 0.61, respectively. Note, data were simulated with a baseline residual of 1.0. Between the het-only model and HSGRM, there were no noticeable differences in mean baseline residual when there was no skew (e.g., h0s0, h4s0, h8s0, h1s0) and ranged from 1.0 to 1.01. When skew was introduced, the HSGRM consistently had mean baseline residual values approximating the true population value. The het-only model, on the other hand, saw a decrease in mean baseline residual as a function of increased skew. That said, under the most extreme het/skew combination (h1s1), the mean residual was 0.82 – still far better than that from the baseline GRM or skew-only model. Across all models, standard deviation tended to be small but gradually increased with model complexity (baseline to HSGRM). The same trend was found in the small sample with regard to mean baseline residuals and mean standard deviations. The only noticeable difference was in the magnitude of the standard deviations such that they were larger in the small sample. It is apparent that, regardless of sample size, failing to model heteroscedastic errors negatively impacted baseline residuals as is seen in the baseline and skew-only models. When accounted for, mean residual baseline residuals were exact or close to true population values and when modeling both heteroscedastic errors and skew, the HSGRM mean residuals where closely approximated.

Factor loadings. As seen in Table 3, across all four models, when only skew was present, there were no real differences in mean factor loading with the exception of slightly lower values (0.91 – 1.00) for the het-only model. For the baseline and skew-only models, as heteroscedastic errors increased and went unmodeled, mean factor loadings gradually decreased to 0.71 and 0.78, respectively. For the het-only and HSGRM, mean factor loadings were consistently close to the population parameter with the HSGRM being closest. In the most extreme case of heteroscedasticity and skew, the het-only model, despite not modeling skew, still had a mean factor loading of 0.87. Standard deviations were generally small ranging from 0.04 to 0.10 with the larger deviations found in the HSGRM. The same trend emerged in the small sample with the only noticeable difference pertaining to larger, but still small, standard deviations in the small sample.

Thresholds. Tables 4 and 5 display the descriptives for thresholds 3 and 4. As a reminder, to identify the scale and define the unit of measurement of U_i while also being able to estimate $\sigma_{\epsilon i}^2$ and ν_i for all items, the first two thresholds were fixed to -2.5 and -1, respectively. In conditions with only skew, across all models, mean threshold values are relatively close to the true population values with relatively small standard deviations. That said, mean thresholds for the skew-only model and HSGRM tended to be equal to or extremely close to the true population value. For the baseline and het-only models, there is a slight decrease in threshold means as a function of skew increasing but is negligible. For the baseline and skew-only models, mean threshold values dramatically decrease as a function of increased heteroscedasticity and dwindle down to 0.52 and 0.59 for threshold 3 and 1.4 and 1.56 for threshold 4, respectively. For conditions with only heteroscedastic errors, both the het-only and HSGRM have very similar mean thresholds close to the true population value. Instances where both skew and heteroscedastic errors were included, there was a slight decrease in mean threshold values in het-only models as a function of increased skew whereas the HSGRM maintained consistently close mean threshold values to those of the population. These same

results emerged for the small sample case, however, with slightly larger standard deviations. As seen previously, failing to model heteroscedastic errors seems to have the most deleterious effects, in this case, resulting in drastically smaller mean thresholds.

Intercepts. Table 6 presents descriptives for intercepts. Similar to previous results, mean intercepts for conditions with only skew tend to be relatively consistent with the true population value across models with the exception of the het-only model that had slightly larger mean intercept values. As heteroscedastic errors are introduced and gradually increase, mean intercepts for the baseline and skew-only model once again are most notably affected with substantial decreases in mean intercept values. When only heteroscedastic errors were present, the het-only model and HSGRM both performed similarly with mean intercept values generally close to the true population value. However, as skew is introduced, there is a slight decrease in mean intercept values for the het-only model. Once again, the HSGRM yielded mean intercept values closely approximating the true population value. As with previous parameters, the same results emerged in the small sample.

Three-category response conditions

Skew. Descriptives for skew parameters are presented in Table 7. For conditions only containing skew, both the skew-only model and HSGRM had extremely similar mean skew parameters closely resembling the true population value. As heteroscedastic errors were included and increased, the mean skew parameters for the skew-only model shifted to the left and were smaller, specifically for skew of 0.0, 0.5 and 0.75. When heteroscedastic errors were modeled, mean skew parameters from the HSGRM tended to approximate the true population value. In the extreme, skew of 1.00, both the skew only model and HSGRM had mean skew parameters near the true population values. Standard deviations between models were similar and tended to be moderate in magnitude with a few larger ones emerging when heteroscedastic errors were at a maximum. These trends were consistent regardless of sample size. The most notable difference was in the standard deviations which tended to be larger in the small sample.

Heteroscedastic Errors. Mean heteroscedastic errors for all conditions are also presented in Table 7. Conditions containing only heteroscedastic errors produced similar means for both the het-only and HSGRM and were close to or equal to the true population value. When both skew and heteroscedasticity were present, the HSGRM had mean values either close to or exactly that of the population. Failing to model skew negatively affected the het-only model such that across all conditions, mean heteroscedastic errors were consistently underestimated. Standard deviations were small-to-moderate for both models. Similarly, small sample results show that under conditions with no skew, both models performed similarly as indicated by mean heteroscedastic error values similar to the true population value. However, in the presence of any degree of skew, mean heteroscedastic errors were substantially smaller in the het-only model whereas they were closer to the true population value in the HSGRM.

Baseline residuals and Factor Loadings. Baseline residuals are presented in Table 8. Interestingly, across all conditionals, all of the models had mean baseline residuals close to or equal to the true population value. This was true for both large and small sample sizes. The ranges for each model in both large and small samples were: baseline model = 0.96 – 1.01, skew-only = 0.96 – 1.01, het-only = 0.96 – 1.00, and HSGRM = 0.98-1.00. The only difference was that standard deviations were larger in the small sample (range 0.11-0.13) than in the large sample (range 0.05-0.06). Although there was an apparent smaller range in the HSGRM, it is clear that there was no real difference in mean baseline residual across conditions or sample size. A very similar trend emerged for factor loadings as displayed in Table 9. Mean factor loadings across conditions generally cluster around 1.0. For the baseline model in the large sample, mean loadings ranged from 0.94 – 1.01 and in the small sample from 0.94 – 1.04. For the skew-only model, the range in the large sample was 0.98-1.05 and in the small sample from 0.98-1.05. In the het-only model, loadings in the large sample ranged from 0.93-1.0 and in the small sample from 0.94-1.0. Lastly, in the HSGRM, loadings in the large sample ranged from 0.99-1.02 and from 0.98-102 in the small sample. Although there is some distinction, again it

appears that mean factor loadings were not dramatically impacted by the degree of skew or heteroscedasticity and whether they were modeled. As has been the case thus far, the key distinction between sample sizes is noticed in the magnitude of the standard deviations such that in the large sample standard deviations were smaller than those of the small sample.

Intercept. Table 10 includes descriptives for the intercepts across models and conditions. In models containing only skew, mean intercepts were generally similar and close to the true population value. That said, mean intercepts for the het-only models were slightly larger and increased as a function of skew, however, the differences were negligible. Generally speaking, mean intercepts from the HSGRM were either equal to or close to the true population value and when only heteroscedastic errors were included, the het-only model also had mean intercepts equal to or close to the true population value. For the baseline model and skew-only model, failing to model for heteroscedasticity resulted in slightly more negative mean intercept values. Standard deviations were small and ranged from 0.03-0.04. Similar mean intercepts and trends emerged for the small sample, however, with slightly larger standard deviations ranging from 0.07-0.08).

Consequence of Model Misspecification

The first inquiry focused on identifying the consequences of model misspecification, that is, failing to model skew and heteroscedastic errors. To that end, all data containing varying degrees of skew and/or heteroscedasticity were fit to a normal GRM (skew and heteroscedastic errors were not estimated and fixed to zero) when the true heteroscedastic errors and skew were non-zero. Consequences were evaluated in terms of average absolute bias and RMSD. As a note, the term “bias” throughout is used interchangeably with “absolute bias.”

Average absolute bias: Parameter estimate bias is presented in Tables 11 and 12 for all conditions with a combination of skew and/or heteroscedastic errors, collapsed across items and replications. Table 11 pertains to 5-category response conditions and Table 12 to 3-category response conditions. The degree of model misspecification can be seen in the

magnitude of average bias such that values near zero reflect no bias. The 5-category response conditions are discussed first. For conditions where only skew was included (i.e., h0s5, h0s75, h0s1), average bias was small for baseline residuals (0.00 to -0.06), factor loadings (0.00 to -0.04), intercepts (0.00 to -0.04), and the third thresholds (0.00 to -0.06) with minor gradual increases in average bias as skew increased from 0 to 1. In the case of the fourth threshold, mean bias was slightly larger (0.00 to -0.16) again with increases tied to greater skew. When sample size was small, similar results emerged such that average bias was small for baseline residuals (0.01 to -0.05), factor loadings (0.01 to -0.04), intercepts (0.01 to -0.05), and the third thresholds (0.01 to -0.03) with minor gradual increases in average bias as skew increased from 0 to 1. Mean bias in the fourth threshold was slightly larger (0.03 to -0.14) and increased as a function of skew. Generally speaking, in the presence of only skew, regardless of sample size, there was minimal bias in item parameters, and more specifically, in slopes and thresholds.

To graphically demonstrate the effect, consider Figure 5, which presents CRCs and TRCs for the condition with no heteroscedastic errors and skew of 1.0 (i.e., h0s1). The CRCs are practically identical with only a slight difference on the lower end of the trait pertaining to category options 1 and 2. Moreover, the TRCs show that with the exception of scores on the lower end of the trait tending to be upwardly biased under the Baseline GRM, expected scores are virtually identical. This would suggest that skew alone left unmolded might not present as great a concern.

The inclusion of heteroscedasticity resulted in substantially more pronounced bias. For conditions with heteroscedastic errors of 0.4, mean bias for baseline residuals (-0.20 to -0.23), factor loadings (-0.10 to -0.15), third threshold (-0.21 to -0.23), fourth threshold (-0.48 to -0.57), and intercepts (-0.13 to -0.14) were noticeably larger across all parameters and increased as a function of combined skew. The same was true for the small sample such that mean bias for baseline residuals (-0.20 to -0.23), factor loadings (-0.10 to -0.15), third threshold (-0.22 to -0.23), fourth threshold (-0.48 to -0.57), and intercepts (-0.13 to -0.15) were also larger for all

parameters and increased as a function of combined skew. Increasing heteroscedasticity to 0.8 resulted in even greater mean bias in baseline residuals (-0.38 to -0.39), factor loadings (-0.20 to -0.24), third threshold (-0.40 to -0.43), fourth threshold (-0.92 to -0.96), and intercepts (-0.24 to -0.26) and once again increased with skew. In the small sample, mean bias was equally as substantial for baseline residuals (-0.39 to -0.40), factor loadings (-0.21 to -0.25), third threshold (-0.41 to -0.45), fourth threshold (-0.94 to -0.97), and intercepts (-0.25 to -0.26). When heteroscedasticity was at a maximum of 1.0, mean bias for baseline residuals (-0.44 to -0.47), factor loadings (-0.25 to -0.29), third threshold (-0.48 to -0.54), fourth threshold (-1.09 to -1.15), and intercepts (-0.29 to -0.32) were largest. Similarly, for small samples, mean bias peaked for baseline residuals (-0.45 to -0.47), factor loadings (-0.25 to -0.29), third threshold (-0.49 to -0.55), fourth threshold (-1.10 to -1.15), and intercepts (-0.29 to -0.32). Taken together, it is quite apparent that simply fitting a normal GRM and failing to model heteroscedastic error and skew, even in moderate cases, will result in noticeable parameter bias. In more extreme cases, the bias is severely problematic.

To illustrate the consequence of model misspecification, consider a large sample 5-category item with heteroscedasticity of 0.8 and no skew. Figure 6 translates the observed bias in slopes and thresholds into a visual display, specifically, in terms of CRCs and TRCs for the misspecified baseline GRM and HSGRM. Note the how the curves for the lowest and highest response options have shifted toward the extremes. As seen in the TRCs for both models, at the lower and upper extremes of the latent trait, the baseline GRM tends to have larger expected scores compared to the HSGRM. That said, in the middle trait range, there does not appear to be any difference.

Now compare this to an item with the same conditions however changing skew to 0.5 (i.e., h8s5). As seen, noticeable downward bias emerged for both the slope and thresholds, among the other item parameters. To visually see the impact, consider Figure 7, which presents CRCs and TRCs for the misspecified baseline GRM and the correct full HSGRM. Given the

downward bias in thresholds, it is not surprising to see the CRCs also shift down on the latent trait such that the probability of an individual responding in a particular response category requires less of the trait. Moreover, examining the test response curve, it is also apparent that expected scores were upwardly biased such that under the baseline GRM expected scores are consistently higher across the latent trait than under the correctly specified full HSGRM. Comparing the models with and without skew (i.e., h8s0 versus h8s5), it is apparent that the inclusion of skew to high heteroscedasticity resulted in a more pronounced bias in expected scores such that when skew is added, expected scores tend to be consistently larger under the misspecified model. Figures 8 - 13 display CRCs and TRC for other key conditions.

For the 3-category response conditions, when only skew was included, mean bias was negligible for baseline residuals (0.00 to 0.01), factor loadings (0.00 to -0.03), and intercepts (0.00 to -0.02). In small samples, the average bias for baseline residuals (0.00 to 0.00), factor loadings (0.00 to -0.03), and intercepts (0.00 to -0.02) were equally tiny. Mean bias for baseline residuals remained low (0.00 to -0.04) for both large and small samples despite increasing heteroscedastic errors, even to a maximum of 1.0. The same was true for factor loadings such that mean bias (0.00 to -0.06) changed minimally as heteroscedasticity increased. Mean bias for intercepts (0.00 to -0.07) in large samples and small samples (0.00 to -0.08) tended to remain on the “higher” end once heteroscedastic errors were 0.8 and larger. Regardless, mean bias for intercepts were generally low. Across all conditions, minimal bias was evident for baseline residuals, factor loadings and intercepts.

RMSD. Tables 11 and 12 also contain RMSD’s of the item parameters for each of the conditions, collapsing across items and replications with smaller values indicating greater accuracy. RMSDs in Table 11 pertain to 5-category response conditions and Table 12 to 3-category response conditions. For conditions with only skew (i.e., h0s5, h0s75, h0s1), RMSDs for baseline residual (0.08 to 0.09), factor loadings (0.05 to 0.06), third threshold (0.10 to 0.11), fourth threshold (0.16 to 0.21), and intercepts (0.06 to 0.07) were relatively small and only

increases slightly as skew increased. When heteroscedasticity was 0.4, RMSDs were larger for baseline residual (0.19 to 0.21), factor loadings (0.10 to 0.14), third threshold (0.21 to 0.22), fourth threshold (0.45 to 0.53), and intercepts (0.13 to 0.14). When heteroscedasticity was 0.8 RMSDs increased for baseline residuals (0.34 to 0.36), factor loadings (0.18 to 0.22), third threshold (0.37 to 0.40), fourth threshold (0.83 to 0.87), and intercepts (0.22 to 0.24). Raising heteroscedastic errors to 1.0 produced the largest RMSDs for baseline residual (0.40 to 0.42), factor loadings (0.23 to 0.26), third threshold (0.44 to 0.49), fourth threshold (0.98 to 1.03), and intercepts (0.26 to 0.28). It is important to note that for each degree of heteroscedasticity (i.e., h4, h8, h1), there were only slight increases in RMSD as a function of skew increasing. In other words, for a condition with heteroscedasticity of 0.8 and no skew, the change in RMSD when adding skew of 0.5, 0.75, and 1.0 was negligible. This is an important point. It highlights that the greatest effect on RMSDs, as with mean bias, rests on heteroscedastic errors not being modeled. In small samples, with the exception of skew only conditions, which tended to have larger RMSDs overall compared to the larger sample, ranges of RMSDs were relatively similar to those from the large sample.

Given the average bias results for the 3-category response conditions, it is not surprising that RMSDs tended to be rather small across all conditions and regardless of sample size. When only skew was included, RMSDs for baseline residuals (0.05), factor loadings (0.03 to 0.04), and intercepts (0.03 to 0.04) were noticeably small. RMSDs for baseline residuals remained low (0.05 to 0.06) despite increasing heteroscedastic errors, even to a maximum of 1.0. The same was true for factor loadings such that RMSDs (0.03 to 0.07) changed minimally as a function of increased heteroscedastic errors. RMSDs for intercepts (0.03 to 0.08) also changed marginally as a function of increased heteroscedasticity. The same general trend occurred in small samples albeit with slightly larger RMSDs. Across conditions, RMSDs for baseline residuals (0.10 to 0.11), factor loadings (0.07 to 0.08), and intercepts (0.06 to 0.09) were hardly affected by skew and/or heteroscedasticity.

Collectively, bias and RMSDs were most pronounced in the 5-category response conditions and, more specifically, as skew and heteroscedasticity increased. In the most extreme combinations of skew and heteroscedastic errors, the greatest bias and RMSDs emerged. As noted, even in moderate cases of skew and/or heteroscedasticity, failing to account for these resulted in clear parameter bias. These effects surfaced regardless of sample size. In the case of the 3-category response conditions, average bias and RMSDs suggest very little impact on parameter estimates regardless of degree of skew and/or heteroscedasticity and sample size.

Viability of HSGRM under Control Conditions

The viability of the HSGRM was evaluated in two ways by using data from control conditions, that is, conditions with no skew or heteroscedasticity for both 3- and 5-categories and in both large and small samples (i.e., h0s0) – clean, normal data. First, a constrained version of the HSGRM was estimated for data from each control condition with both skew and heteroscedasticity fixed to zero (Baseline GRM model). Second, the full HSGRM with both skew and heteroscedasticity estimated was fit to the same data. Parameter recovery was evaluated for both models using average absolute bias and RMSDs. Results for the constrained HSGRM are presented in Tables 13 and 14 and for the full HSGRM in Tables 15 and 16.

Average absolute bias: For the constrained HSGRM, in the large sample case with 5-category response options, parameter recovery was quite good, that is, they were recovered accurately. Across all items, the average absolute bias for the baseline residual, factor loadings, third and fourth thresholds, and intercept was less than 0.001. Similar parameter recovery accuracy was found in the small sample such that mean bias in baseline residuals, factor loadings, third threshold, and intercepts were all less than or equal to 0.01 and for the fourth threshold 0.026. Parameter recovery for the 3-category response condition was also particularly good. Mean bias in the baseline residuals, factor loadings, and intercepts were all ostensibly zero in both the large and small samples.

For the full HSGRM, parameter recovery was also quite good in the large sample with 5-category response options. Mean bias across items for the baseline residual, factor loadings, third and fourth thresholds, and intercept ranged from -0.013 to 0.009, indistinguishable from zero. In the small sample however, mean bias tended to be slightly larger across item parameters. Mean bias for item parameters ranged from 0.01 to 0.07. Although not particularly concerning, there is an apparent effect of modeling skew and heteroscedasticity when none is present. Parameter recovery for the 3-category response condition was also quite good. In the large sample, across all item parameters, mean bias ranged from -0.007 to 0.002, clearly small and indistinguishable from zero. Similar parameter recovery accuracy was found in the small sample such that mean bias across all item parameters ranged from -0.022 to 0.005, again ostensibly zero. Therefore, it appears that generally speaking, both the constrained and full HSGRM were able to recover item parameters well with the constrained HSGRM performing slightly better than the full HSGRM. Based on the magnitude of mean bias, it is clear that under the full HSGRM, the mean bias was slightly larger, especially for the 5-category response option in small samples and most notably in the fourth threshold.

Although some bias emerged under the full HSGRM, Figure 7 makes clear that the bias is ostensibly zero and resulted in no meaningful differences. Consider the CRCs and test response curves (TRC) for the 5-category condition in the large sample. The CRC from both the baseline GRM and full HSGRM are identical. Moreover, the TRCs for both models completely overlap. This suggests that fitting either model to this data made no noticeable difference and that the emergent bias did not matter. Moreover, this graphically demonstrates the viability of the HSGRM and its constrained baseline model.

RMSD: Tables 13 and 14 contain RMSD's for item parameters under the constrained HSGRM with smaller values indicating greater accuracy. As was the case with the average absolute bias, for the 5-category response condition, mean RMSD for baseline residuals (M = 0.08), factor loadings (M = 0.05), thresholds 3 (M = 0.10) and 4 (M = 0.16), and the intercept (M

= 0.06) were generally small. In the small sample, mean RMSD for baseline residuals (M = 0.178), factor loadings (M = 0.099), thresholds 3 (M = 0.208) and four (M = 0.339), and the intercept (M = 0.132) were noticeably larger but by no means an issue. Although this is not particularly concerning, it should be noted that with smaller sample sizes, the RMSD's for the thresholds were most dramatically impacted. In the case of the 3-category response conditions, mean RMSD's were particularly small for the baseline residuals (M = 0.054), factor loadings (M = 0.034), and intercepts (M = 0.031) in the large sample. Similarly, for the small sample, mean RMSD's were all small for the baseline residuals (M = 0.105), factor loadings (M = 0.067), and intercepts (M = 0.062). As noted, with fewer categories, sample size effects on RMSD's were less dramatic when compared to the more notable effect with 5-category response options. That said, for both sample sizes and varying category response options, RMSD's were relatively small.

The 5-category condition, under the full HSGRM, produced mean RMSDs that were consistently larger across all item parameters, albeit generally small. The most notable changes are in the RMSDs for the skew parameter and heteroscedastic errors. Whereas previously not estimated, under this model, mean RMSDs for the skew parameter was 0.36 and for heteroscedastic errors, 0.14. Moreover, in the small sample, these effects are more pronounced with mean RMSD of 0.53 for the skew parameter and 0.31 for heteroscedastic errors. Moreover, the fourth threshold also saw a sizeable increase such that the mean RMSD went from 0.34 under the constrained model to 0.50 under the full. In the 3-category condition, the real only difference emerged in the mean RMSDs fro the skew parameter (0.44) and for heteroscedastic errors (0.18). RMSDs for the remaining item parameters remained unchanged. Similarly, in the small sample, the mean RMSDs for the skew parameter (0.84) and heteroscedastic errors (0.39) were the only noticeable changes. The remaining parameters remained unchanged when compared to the constrained HSGRM. That said, it is apparent that RMSDs for the skew parameter and heteroscedastic errors tend to increase as a function of fewer response

categories and smaller sample sizes.

Collectively, it is apparent that when item response data does not include skew or heteroscedastic errors, both the constrained and full HSGRM are able to recover item parameters particularly well, regardless of sample size and number of categories. The key difference is apparent in the magnitude of RMSDs specifically, for non-existing skew and heteroscedastic errors.

Item parameter recovery for 5-category response conditions

The third question of interest focused on the performance of the HSGRM in the presence of heteroscedasticity and/or skew, specifically with regard to item parameter recovery via absolute bias and RMSD. This was accomplished in two ways. First, for each condition, the true model was estimated (e.g., skew-only model for data containing only skew) and the full HSGRM (e.g., HSGRM for data containing only skew). In each case, absolute bias and RMSD were computed and evaluated for all item parameters. This spoke to the performance of both a constrained version and full version of the HSGRM for conditions with varying degrees of skew and heteroscedasticity.

The results below are presented and discussed within the context of mean bias and RMSDs across items and replications. Item level results are not discussed throughout. However, Figures 13 through 44 provide box plots for each relevant item parameter for each item within a given condition. For instance, a condition with only skew present, the relevant item parameters would be skew, baseline residual, factor loadings, thresholds, and intercepts – heteroscedastic errors are not included as they are not estimated in the correct model (skew-only).

Absolute bias: Table 17 displays the average absolute bias in item parameter recovery for 5-category response conditions. For each condition, mean item parameter bias was first evaluated by comparing the true population values to those obtained from the correct model. That is, models containing both skew and heteroscedastic errors used HSGRM results, models

with only skew were evaluated in terms of the skew-only model, models with only heteroscedastic errors were evaluated in terms of the het-only model. Subsequently, for each of these conditions, item parameter estimates obtained from the HSGRM were also evaluated for mean bias as a point of comparison and is discussed below.

For models only including skew, item parameter recovery was quite good. Across all item parameters, average bias was either zero or near zero. In fact, the largest mean bias of 0.04 was in fourth threshold and when skew was at its maximum. Parameter recovery was also extremely good for het-only models such that regardless of the magnitude of heteroscedasticity, all item parameters were recovered with great precision. In fact, the range of mean bias across all het-only models and across all item parameters was from 0.00 to 0.02. With regard to models containing both skew and heteroscedastic errors, models generally recovered item parameters very well. There are a few exceptions where the skew parameter tended to suffer most, although perhaps not meaningfully. When heteroscedastic errors were 0.8 and skew at 0.75, the mean bias was -0.23. Recall this mean bias reflects bias in the skew parameter not the actual skew of the distribution (see Eqs. 18 and 19). Therefore, a mean bias of -0.23 translates into an average skew of 0.72 versus the true skew of 0.75. Similarly when heteroscedastic errors were 1.0 and skew was 0.5 and 0.75, mean bias were -0.22 and -0.32, respectively. Again, putting this into perspective, this average bias translates into actual skew of 0.44 versus 0.50 and 0.71 versus 0.75, respectively. It is important to note that in these conditions where bias emerged in the skew parameter, there were also slight increases in mean bias for heteroscedastic errors (-0.02 to -0.06) and the fourth threshold (-0.02 to -0.08). However, despite these “larger” mean bias, overall the HSGRM recovered item parameters well.

In the case of small samples, across all conditions, item parameters were also generally recovered quite well. For conditions with only skew, small mean bias was observed across all item parameters (0.00 to 0.07) with one exception. When skew was 0.75, mean skew bias was 0.24.Transforming this value into actual skew yields a value of 0.77 as opposed to the true

population value of 0.75. Item parameter recovery for models including only heteroscedastic errors was excellent across all item parameters. Mean bias across all parameters ranged from -0.01 to 0.03. For conditions containing both skew and heteroscedastic errors, item parameter recovery was also quite good. That said, similar to the large sample results, when heteroscedastic errors were 0.8 and skew 0.5 and 0.75, mean bias in the skew parameter was -0.21 and -0.16 respectively. Reflecting a difference of 0.44 versus 0.50 and 0.73 versus 0.75. When heteroscedasticity was 1.0 and skew 0.5 and 0.75, mean bias in the skew parameter was -0.49 and -0.29, respectively. A distinction of 0.36 versus 0.50 and 0.72 versus 0.75. In this case the former mean bias perhaps merits some attention.

HSGRM item parameter estimates were also obtained for conditions with only skew or heteroscedastic errors and evaluated for bias. Results are presented in Table 19. For conditions including only skew, item parameter recovery was quite good with mean bias ranging from -0.01 to 0.04 across all item parameters. Recovery was also good when heteroscedasticity was 0.4 with mean bias ranging from -0.01 to 0.02 across all item parameters. However, with larger heteroscedastic errors (0.8 and 1.0), some mean bias emerged in the skew parameters, specifically, -0.13 and -0.09, granted these transform into actual skew of ostensibly zero which make sense given that in the het-only models, there is no skew. However, in small samples, there is consistent mean bias across all skew parameters ranging from -0.36 to 0.31. In fact, across all item parameters, there are apparent increases in mean bias that were not present previously when running the proper model. Therefore, in larger samples, item parameter recovery under the HSGRM was very similar to results obtained from correct models. However, in small samples, there does appear to be some consequence to over-fitting.

RMSD. Table 17 presents RMSDs for all conditions across items and replications. Generally speaking, across all conditions RMSDs were quite good and small. Conditions with only skew had small RMSDs for the skew parameters (0.00 to 0.39), baseline residuals (0.08), factor loadings (0.05 to 0.06), third threshold (0.10), fourth threshold (0.16 to 0.17), and

intercepts (0.06). Models with only heteroscedastic errors also performed well as indicated by small RMSDs for baseline residuals (0.11), heteroscedastic errors (0.13 to 0.14), factor loadings (0.05 to 0.06), third threshold (0.12), fourth threshold (0.22 to 0.23), and intercepts (0.07 to 0.08). For conditions with both skew and heteroscedastic errors, RMSDs were also generally small for baselines residuals (0.11 to 0.14), heteroscedastic errors (0.15 to 0.22), factor loadings (0.06 to 0.09), third threshold (0.12 to 0.15), fourth threshold (0.23 to 0.29), and intercepts (0.06 to 0.09). Skew parameters on the other hand tended to have larger RMSDs that corresponded with conditions previously addressed for having larger mean bias. Interestingly, regardless of het-skew combinations, RMSDs for skew of 0 and 1.0 were 0.00. Broadly speaking, RMSDs increased for all parameters as a function of increased heteroscedasticity with the exception of the skew parameter. Compared to the large sample, RMSDs from the small samples tended to be noticeably larger across all item parameters but followed the same trend such that increases in heteroscedastic errors were linked to increases in RMSDs. Thus, not surprising, there was greater precision in large samples.

RMSDs were also computed from HSGRM item parameter estimates obtained for conditions with only skew or heteroscedastic errors. Generally speaking, RMSDs were similar or slightly larger than those from the correct models. Obviously, due to the fact that skew and heteroscedastic errors were now estimated for models not including them, the emergence of RMSDs is not surprising. For instance in conditions with only skew, RMSDs for heteroscedastic errors were 0.14. Similarly, for conditions with only heteroscedastic errors, RMSDs for skew parameters ranged from 0.27 to 0.41. In small samples, the same trends emerge such that RMSDs were similar or slightly larger to those obtained from the correct models. Again, given that skew and heteroscedastic errors were estimated for conditions containing neither, it is not surprising that RMSDs emerged. RMSDs for skew parameters were computed between 0.46 and 0.49 for conditions without skew. In conditions with only skew, RMSDs for heteroscedastic errors ranged from 0.31 to 0.32. Therefore, for the skew parameter, RMSDs although larger in

small samples, were still relatively low. On the other hand, RMSDs for heteroscedastic errors tended to be consistently higher in smaller samples and particularly large for conditions containing no true heteroscedasticity.

Item parameter recovery for 3-category response conditions

Absolute bias: Tables 18 displays the average absolute bias in item parameter recovery for 3-category response conditions. For each condition, item parameter bias was first evaluated by comparing the true population values to those obtained from the correct model. That is, models containing both skew and heteroscedastic errors used HSGRM results, models with only skew were evaluated in terms of the skew-only model, models with only heteroscedastic errors were evaluated in terms of the het-only model. Subsequently, for each of these conditions, item parameter estimates obtained from the HSGRM were also evaluated for bias as a point of comparison and is discussed below.

Item parameter recovery was quite good across all conditions, with a few exceptions. Generally, mean bias in the skew parameter was most noticeable when skew was 0.5 and 0.75 combined with heteroscedasticity of 0.8 and 1.0. This same trend emerged previously in the 5-category response conditions. When heteroscedasticity was 0.8, mean bias in the skew parameters was -0.28 and -0.23, respectively which translates into actual mean skew bias of 0.42 versus 0.50 and 0.73 versus 0.75. When heteroscedastic errors were 1.0, mean bias in the skew parameters were -1.08 and -0.30 which when transformed reflect average skew of 0.16 versus 0.5 and 0.72 versus 0.75. In this case, mean bias in the skew parameter for h1s5 is concerning and consistent with results from the 5-category response condition. Mean bias for baseline residuals (-0.02 to 0.00), factor loadings (-0.01 to 0.02), and intercepts (-0.05 to 0.00) were exceptionally small. For models including heteroscedastic errors, mean bias ranged from -0.11 to 0.09 and tended to be the largest when combined with greater skew. That said, they presented no real concern.

Not surprisingly, the same mean bias results emerged in the small sample but tended to be slightly more dramatic. For the most part, item parameter recovery was quite good across all conditions, except in the case of the skew parameter. Again, mean bias in the skew parameter was greatest when skew was 0.5 and 0.75 and combined with heteroscedasticity of 0.8 and 1.0. However, slight increases even emerged when heteroscedastic errors were not present (e.g., h0s5 and h0s75) and when small (h4s5 and h4s75). At skew of 0.5 and 0.75, for conditions with no heteroscedastic errors, mean bias in the skew parameter was 0.18 and 0.09 (actual skew of .54 and .76) and when heteroscedastic errors were 0.4, mean bias in the skew parameters was 0.11 and 0.17 (actual skew of .53 and .77). Therefore, although there was some bias, skew parameters were still recovered relatively well. Note that unlike the 5-category conditions, skew was positively biased for these conditions. When heteroscedasticity was 0.8, however, mean bias in the skew parameters was -0.76 and -0.18, respectively which translates into actual mean skew of 0.27 versus 0.50 and 0.73 versus 0.75. When heteroscedastic errors were 1.0, mean bias in the skew parameters were -1.26 and -0.30 which when transformed reflect average skew of 0.11 versus 0.5 and 0.72 versus 0.75. In this case, mean bias in the skew parameter for h8s5 and h1s5 suggests skew is substantially underestimated and concerning. Mean bias for baseline residuals (-0.02 to 0.00), factor loadings (-0.02 to 0.02), and intercepts (-0.05 to 0.01) were exceptionally small. For models including heteroscedastic errors, mean bias generally ranged from -0.06 to 0.10, however, when heteroscedasticity was 0.8 and 1.0 and combined with maximum skew, mean bias was 0.22 and 0.33 respectively. It is apparent that in small samples with 3-category response items, greatest mean bias was found in skew and certain larger heteroscedastic conditions.

For conditions with only skew or heteroscedastic errors only, HSGRM item parameter estimates were also obtained and evaluated for bias. Results are presented in Table 20. Across all conditions, mean bias for baseline residuals (0.00), factor loadings (0.00 to 0.02), heteroscedastic errors (-0.01 to 0.02), and intercepts (-0.03 to 0.00) were all small and near

zero. Mean bias in skew parameters for models containing only skew were small (-0.02 to 0.00), however, tended to increase as conditions contained greater heteroscedastic errors (-0.20 to -0.25). That said, mean bias of this magnitude translates back to actual skew of essentially zero. In the smaller sample, mean bias was still small for baseline residuals (0.00), heteroscedastic errors (-0.1 to 0.05), factor loadings (0.00 to 0.01), and intercepts (-0.03 to 0.01). Comparatively larger bias emerged for skew parameters (-0.02 to -0.45), which again once converted translates into actual mean skew of 0.02 – nothing to really be concerned about. Generally speaking, bias was not improved by fitting the HSGRM to these conditions and if anything resulted in bias in the skew parameter, albeit negligible.

RMSD. Table 18 presents RMSDs for all conditions across items and replications. Generally speaking, across all conditions RMSDs were quite good and small for baseline residuals (0.05 to 0.06), factor loadings (0.03 to 0.04), and intercepts (0.03 to 0.06). On the other hand, when skew was 0.5 and 0.75, RMSDs were generally larger for the skew parameter ranging from 0.25 to 1.35 increasing as a function of heteroscedastic errors. More specifically, with heteroscedasticity of 0.8 and 1.0, skew of 0.5 had the largest RMSDs of 0.72 and 1.35, respectively. For heteroscedastic errors, RMSDs ranged from 0.18 to 0.39 and were generally larger when combined with greater skew such that at maximum heteroscedastic errors and skew, the RMSD was 0.39. When sample size was small, RMSDs were larger across all item parameters and conditions. RMSDs for baseline residual (0.10 to 0.12), factor loadings (0.07 to 0.08), and intercepts (0.06 to 0.08) were still relatively small. For the skew parameter, there was a noticeably increase in range (0.62 to 1.54) but affected the same conditions previously addressed in the large sample results. Regardless of het-skew combinations, RMSDs for skew of 0 and 1.0 were 0.00. For heteroscedastic errors, RMSDs ranged from 0.40 to 0.95, a rather drastic increase from the large sample results. Moreover, the largest RMSDs were found in conditions with greater heteroscedastic errors and, of course, large skew.

In conditions with only skew or heteroscedastic errors, RMSDs computed from HSGRM

item parameter estimates were evaluated. Generally speaking, RMSDs were similar or equivalent for item parameters from the correct models discussed above. Given that skew and heteroscedastic errors were now estimated for models not including them, larger RMSDs were expected. For instance in conditions with only skew, RMSDs for heteroscedastic errors were 0.18. Conditions with only heteroscedastic errors produced RMSDs for skew parameters ranging from 0.42 to 0.53. In small samples, RMSDs for baseline residuals, factor loadings, and intercepts were similar to or equivalent to those computed from the correct models discussed previously. However, the greatest shift, not surprisingly, pertains to RMSDs for skew parameters and heteroscedastic errors for conditions including one and not the other. For conditions with only skew, RMSDs for heteroscedastic errors were moderate (0.39 to 0.43), in fact, they were on par with RMSDs for models containing only heteroscedastic errors (0.40 to 0.44). In conditions with only heteroscedasticity, RMSDs for skew parameters (0.58 to 0.69) were comparable to those for conditions with only skew (0.00 to 1.19). Taken together, for models with only skew or only heteroscedastic errors, RMSDs for baseline residuals, factor loadings, and intercepts, seem to be relatively small and unchanged regardless of whether a correct model is estimated or if the HSGRM is used. Other than the obvious presence of RMSDs for parameters that should not be estimated, it does appear that in small samples and for conditions with only skew, there is a slight increase in the magnitude of RMSDs for the skew parameter when both heteroscedasticity and skew are estimated.

Correct model identification and over selection of HSGRM

The fourth question pertained to determining the frequency with which the correct model was properly identified and the prevalence of the HSGRM being overly selected as the best fitting model. For all conditions, likelihood ratio tests were performed comparing competing models with the correct model. Of particular interest is the percentage of times the full HSGRM was preferred over the correct model. For cases where the full HSGRM was the correct model, percentage of times that competing models were selected as the best fitting model is also of

interest. That is, for cases with both skew and heteroscedastic errors, how often was the skew-only or het-only models chosen. Results from these compiled likelihood ratio tests are presented in Table 21. Aside from evaluating model selection via likelihood ratio tests, it was also investigated via information criteria, namely, AIC. All conditions can be grouped into 4 types: no skew or heteroscedasticity, skew-only, het-only, and both heteroscedasticity and skew. Therefore, knowing what the true model was (i.e., het-only), Table 22 provides the percentages of times each of the four models was selected given the true model. For example, when the true model generating the data was a het-only model, how many times was the baseline, skew-only, het-only (correct), and HSGRM models chosen as the best fitting model? Note, similar conditions were combined such that all skew-only conditions (i.e., h0s5, h0s75, h0s1) were evaluated together, all het-only conditions (i.e., h4s0, h8s0, h1s0) were evaluated together, and all full HSGRM conditions (i.e., h4s5, h8s75, h1s1) were evaluated together.

First consider the control conditions with no skew and no heteroscedastic errors with 5-category response options. When sample size was large, the full HSGRM was chosen over the baseline constrained GRM only 6% of the time whereas for small samples, this occurred 8% of the time. (Those would be Type 1 error rates when null is true) For conditions where only skew was included, 3-7% of the time the full HSGRM was the preferred model and in the small samples this occurred 4-7% of the time. When only heteroscedastic errors were present, the full HSGRM was preferred between 1-11% of the time and only 2-6% of the time in smaller samples. When both skew and heteroscedastic errors were present, the full HSGRM was chosen 99-100% of time over skew-only or het-only models. In smaller samples, the full HSGRM was chosen 86-100% of the time over het-only models, and 100% of the time over skew-only models with one exception. When heteroscedastic errors were 0.4 and combined with any degree of skew, full HSGRM was preferred only 70-72% of the time.

The control conditions for the 3-category response items, when sample size was large, identified the full HSGRM as the best fitting model 4% of the time whereas when sample size

was small, this occurred 8% of the time. When only skew was present, the full HSGRM was chosen over the skew-only model 4-6% of the time and in the small sample 5-9% of the time. For conditions with only heteroscedastic errors, the full HSGRM was chosen between 1-3% of the time in both the large and small samples. When both skew and heteroscedastic errors were present, the full HSGRM was chosen over het-only models 89-100% of the time and 99-100% of the time over skew-only models. In small samples, this trend was not as clean. For most models, the full HSGRM was chosen over the het-only model 97-100% of the time. However, when skew of 0.5 was paired with heteroscedastic errors or 0.8 and 1.0, the full HSGRM was chosen over the het-model only 62% and 43% of the time, respectively. Note, these were conditions particularly problematic in terms of mean bias and large RMSDs. Moreover, in small samples the full HSGRM was generally preferred over the skew-only model such that 98-100% of the time it was the better fitting model. However, as was the case with 5-category response and small samples, when heteroscedastic errors were 0.4 and combined with skew, the full HSGRM was only preferred 49-56% of the time.

Generally speaking, the skew-only models, het-only models, and full HSGRMs were correctly identified in both 3- and 5-category response data when sample size was large. When sample size was small, the 5-category response conditions still performed well overall with some issues arising with heteroscedastic errors of 0.4 combined with varying skew. The same conditions were problematic in the 3-category response data in addition to skew of 0.5 combined with large heteroscedastic errors. That said, overall performance was promising.

Table 22 displays the results for model selection given the true model as determined by AIC. That is, given the true model, how frequently was each of the four possible models chosen as the best fitting model? For 5-category conditions with a large sample, when the true model was a baseline GRM with no skew or heteroscedasticity, it was correctly chosen as the best fitting model 92% of the time. When the true model was a skew-only model, 97% of the time it was identified correctly as the best fitting model. In the presence of only heteroscedastic errors,

83% of the time it was correctly chosen, with the other 17% choosing the HSGRM as the best fitting model. Finally when the true model was the full HSGRM, it was chosen as the best fitting model 99.99% of the time. In the small sample with 5 categories, when the baseline GRM was the true model, it was correctly chosen as the best fitting model 95.5% of the time. When the skew-only model was the true model, 96.8% of the time it was chosen as the best fitting model. In the case of the het-only model being the true model, it was chosen as the best fitting model 81% of the time with the full HSGRM being chosen 11% of the time. Finally, when the true model was a full HSGRM, it was correctly chosen 86% of the time and 12% of the time.

For 3-category conditions with large samples, when the true model was a baseline GRM with no skew or heteroscedasticity, it was chosen as the best fitting model 93.5% of the time. When the skew-only model was the true model, it was chosen as the best fitting model 97% of the time. For models with heteroscedastic errors only, the het-only model was selected as the best fitting model 91% of the time with the HSGRM being the best fitting model 9% of the time. When the true model was the full HSGRM, it was correctly chosen as the best fitting model 94% of the time with the het-only model accounting for the other 6 percent. When sample size was small, when the true model was the baseline GRM, it was chosen as the best fitting model 87% of the time. When the skew-only model was the true model, it was identified as the best fitting model 95% of the time. In the case where the het-only model was the true model, 77% of the time is was chosen as the best fitting model. Finally when the true model was the HSGRM, it was chosen only 70% of the time.

Taken together, in both the 3- and 5-category conditions with large samples, it appears that the best fitting model overwhelmingly corresponded to the true model. In small samples, the baseline GRM and skew-only models still tended to do relatively well in terms of being selected as the best fitting model. However, the het-only and full HSGRM saw a decrease in terms of the frequency with which the best fitting model corresponded to the true model. More specifically, in the 3-category conditions we saw the greatest decrease in frequency for the het-only and full

HSGRM models. However, it should be restated that these results are based on combining all similar conditions. Within the full HSGRM conditions, there were some conditions that performed better than others. This corresponds to the same conditions discussed previously that tended to be problematic. Generally speaking, the HSGRM, and its constrained versions, perform reasonably well.

DISCUSSION

With the rise in popularity of item response theory models across a plethora of academic domains, and specifically polytomous models, it is entirely likely that models are, unknowingly, being incorrectly applied to constructs that are generally non-normally distributed in the population. Especially, within the area of psychology, constructs such as depression, anxiety, stress, anger, and impulsivity are likely to take on non-normal distributions. The issue is that failing to account for non-normality can have deleterious effects on item parameter. This is by no means a novel concept. When the latent trait is non-normal, bias has been found to emerge in item parameter estimates (Boulet, 1996; De Ayala & Sava-Bolesta, 1999, DeMars, 2003; Stone, 1992; Wollack et al., 2002) and specifically in item category parameters (Preston & Reise, 2014; Zwinderman & van der Wollenberg, 1990), item slopes (Azevedo, Bolfarine, & Andrade, 2011, Drasgow, 1989), and trait score estimates (Seong, 1990, Ree, 1979, Swaminathan & Gifford, 1983, Woods & Lin, 2009). Results from this simulations study, as highlighted in Figures 5-13, confirm findings from these previous studies. Moreover, as the field of psychology continues trending toward greater implementation of modern measurement models, it is imperative that methods for handling non-normality be developed and scrutinized, and more importantly disseminated to substantive researchers. Such research can help combat some of the widespread beliefs regarding the robustness of IRT models to normality violations.

To that end, this research provided an extensive investigation into the viability and utility of a recently proposed alternative graded response model (HSGRM) designed to handle not only skew in the latent trait but also heteroscedastic errors. This research pushed the performance of this model to the extreme by varying large and small samples, 3- and 5-category response options and a wide range of heteroscedastic errors and skew which were all combined to examine overall functioning. Although a computationally demanding task, a necessary one. Ultimately, this work endeavored to examine item parameter recovery for the key parameters (intercept, factor loadings, and thresholds) but also to examine how well skew and

heteroscedastic errors could be recovered. Moreover, this research focused on investigating the consequences of failing to model skew and heteroscedasticity as this relates directly to increased use of item response models without consideration given to distributional form.

As highlighted in Figures 5 - 13, a need for a model that could handle both skew in the latent trait and heteroscedastic errors was made abundantly clear when fitting a standard normal GRM to data that contained varied degrees and combinations of skew and heteroscedastic errors. Interestingly, this would be a standard procedure for a researcher failing to consider potential non-normality in the data. In both large and small sample sizes with 5-category items, the effect of model misspecification was quite alarming, with the exception of conditions with only skew present, which were still downwardly biased. Factor loadings, thresholds, intercepts and residuals were all drastically downwardly biased and bias increased as a function of greater heteroscedastic errors especially when combined with large skew. Similarly, RMSDs across all item parameters were quite large, again increasing as a function of heteroscedastic errors combined with large skew. It was quite apparent that failing to account for heteroscedastic errors and, to a lesser degree, skew in the latent trait was problematic – an issue not easily ignorable. Interestingly, for the 3-category conditions, model misspecification did not seem to significantly impact residuals, factor loadings, or intercepts. Mean bias did still tend to be downward, however, it was not particularly problematic. RMSDs also tended to be small and of no concern. The deleterious effects of model misspecification, that is, failing to account for heteroscedastic errors and skew, can be primarily observed in the bias associated with slopes and thresholds. Such effects are unpacked in the following.

With regard to slopes, the most notable effects are apparent in the 5-category response conditions. When skew alone was present, even at a maximum (skew =1), there was very little impact on the slopes as seen in Figure 5. In this most extreme condition, although some downward bias emerged, it remained relatively negligible. When only heteroscedastic errors were present, there was a clear impact on the slopes such that they were downwardly biased

and bias increased as a function of heteroscedastic errors. That is, even when heteroscedastic errors were low, slopes were notably impacted. In cases where heteroscedastic errors were at a maximum, slopes were extremely affected, that is, slopes were downwardly biased by 0.25 from a slope value of 1.0. The effect was visually depicted in Figures 8 – 10. Interestingly, combining skew to heteroscedastic errors did not dramatically change the effect of just heteroscedastic errors in terms of slope bias. Adding skew did increases the degree of bias, however, not by much. For instance, when heteroscedastic errors were at a maximum and resulted in a mean bias of -0.25 in slopes, the effect of gradually increasing skew to a maximum resulted in a -0.04 change in mean bias. So although there was an effect of skew, the major issue was failing to account for the heteroscedastic errors. This trend was consistent for all degrees of heteroscedastic errors. Moreover, this effect emerged regardless of sample size. In the 3-category conditions, a similar trend emerged, however, nowhere near to the same degree. In all cases, the bias was still downward. That said, in the presence of only skew, very little bias emerged. As heteroscedastic errors increased, so did the bias in slopes. The added effect of skew did not really change the degree of bias – only an incremental contribution. Bias tended to remain relatively similar regardless of the degree of skew combined to heteroscedastic errors. Similar effects were found in small samples. Taken together, regardless of sample size and number of categories, the primary culprit contributing to bias in slopes was the heteroscedastic errors. Although skew contributed to some degree, it was the lesser of two factors.

With regard to the impact on thresholds, this was only relevant to the 5-category conditions given that for identification purposes, the first two thresholds were fixed and in the case of the 3-category conditions, these were the only two thresholds. Although both the third and fourth thresholds were clearly affected by model misspecification and exhibited downward bias, both were differentially impacted such that the fourth threshold was more negatively affected. With regard to similar trends between both thresholds, when only skew was present, as seen in Figure 5, some bias emerged but by no means presented any serious concerns. In

conditions were only heteroscedastic errors were present, bias in both thresholds increased as a function of heteroscedasticity, that is, the largest bias in thresholds was found in conditions with the largest heteroscedastic errors and the smallest for conditions with none. This effect is highlighted in Figures 8 – 10. As was found in the case of slopes, adding varying degrees of skew to heteroscedastic errors minimally impacted threshold bias. Once again, the primary concern revolves around failing to account for heteroscedasticity. These trends were true for both the third and fourth thresholds, however, the key distinction pertains to the magnitude of bias. The bias for the fourth threshold was generally twice the size of the bias found in the third threshold such that in the most extreme case, bias in the fourth threshold was as large as -1.15. All of these trends were true regardless of sample size.

It is worth noting that in conditions with only heteroscedastic errors, when comparing expected scores between the misspecified baseline GRM and HSGRM, expected scores in the extremes of the latent trait tended to be upwardly biased for the baseline GRM. However, once introducing skew, expected scores tended to be consistently upwardly biased in the baseline GRM versus the HSGRM. Therefore, although the bias in threshold parameters did not change much numerically in terms of mean bias, the effects can be observed visually in terms of CRCs and TRCs as depicted in Figures 11-13. Clearly, although failing to model skew alone had some biasing effect on both slopes and thresholds, failing to account for heteroscedastic errors, in fact, had the most damaging impact and was even worse, when combined with greater skew.

Under controlled conditions, the constrained version of the HSGRM (baseline or normal-ogive GRM) performed exceptionally such that item parameter recovery was great, mean bias essentially zero, and small RMSDs. This was true across all conditions including large and small samples and for 3- and 5-category response formats. The HSGRM also performed well such that mean bias was generally small as were RMSDs. In small samples, we tended to see slightly larger mean bias and RMSDs, however, this presented no obvious concerns. Under controlled conditions, both models performed well, slightly favoring the constrained HSGRM.

Given the severe consequences of model misspecification, primarily for the 5-category conditions and tentative good performance of the HSGRM under controlled conditions, it was then applied to all conditions with varying degrees of skew and heteroscedasticity. For each condition, there was a correct model, for instance, conditions with only heteroscedastic errors should be appropriately modeled by a het-only model. However, to test the ability of the HSGRM to properly recover only what is was supposed to, both the correct model and full HSGRM were fit to each dataset within a condition and mean bias and RMSDs were inspected for both and compared. When fitting the appropriate model to each condition, the improvement in item parameter recovery, minimization of mean bias, and reduction in RMSDs was unquestionable. Across all conditions, item parameter estimates, and sample sizes, the performance of the model was outstanding. Even for cases with extreme skew and heteroscedastic errors, mean bias and RMSDs were small and, substantially smaller than those obtained from the model misspecification results. When the HSGRM was fit to conditions with only skew or only heteroscedastic errors, mean bias and RMSDs tended to be larger than those obtained from a correctly specified models but were by no means problematic.

In 3-category response conditions, mean bias was relatively unchanged compared to the original misspecified model. However, now with skew and heteroscedastic errors estimated, some larger mean bias emerged and was particularly problematic with larger heteroscedastic errors and moderate skew. This was true in both large and small samples. However, broadly speaking, parameter recovery was still quite good here. When fitting the HSGRM to conditions with only skew or heteroscedastic errors, the same previously seen issue emerged such that mean bias in skew parameters and heteroscedastic errors and RMSDs tended to be larger than in the correct model results. However, they did not present any real concern.

Overall, the performance of the HSGRM and its constrained models (skew-only and het-only) was undoubtedly a major improvement in item parameter recovery and thus mean bias and RMSDs. Admittedly, the results were most noticeable for 5-category response data. It was

also noted that although the HSGRM generally yielded similar results, the proper models tended to more accurately recover item parameters. This in mind, the overall sensitivity of the HSGRM was evaluated by examining the frequency with which the HSGRM was chosen over the correctly specified model. Moreover, given the true model, the percentage of times each of the four possible models was chosen as the best fitting model was also considered.

With 5-category response conditions, when both heteroscedastic errors and skew were present, the HSGRM was appropriately preferred practically 100% of the time over the more constrained skew-only and het-only models. When only skew or heteroscedasticity was present, less than 10% of the time was the HSGRM chosen, usually as low as 3 percent. In smaller samples, the vast majority of the time the HSGRM was chosen as the best model when it was the correct model. Of course, there were instances where model selection favored a more reduced model but tended to hit very specific conditions, namely those with moderate heteroscedasticity and moderate skew. These were the same conditions generally affected in the 3-category conditions. Similarly, in the 3-category condition, when sample size was large, for the most part, the HSGRM was chosen as the best model when it was appropriately so. In cases where only skew was present or heteroscedastic errors, only a handful of times was the HSGRM incorrectly chosen. For small samples sizes, the HSGRM was generally chosen when appropriate. Interestingly, as heteroscedastic errors increased, specifically with skew of 0.5, the rate of choosing the HSGRM over more constrained models decreased notably. As a note, for those cases where the HSGRM was chosen despite not truly being the most appropriate model, as seen previously, the recovery of item parameters was not dramatically impacted. In fact, any added mean bias was generally negligible compared to the correct model. So although incorrectly chosen, this does not necessarily mean the parameters obtained are more biased.

Despite the utility of the HSGRM, it should be noted that there are a few limitations to this research. Under the skew-normal distribution, skew is bound from -1 to 1. Therefore, conclusions from this research must be considered in light of the skew limits of the skew-normal

distribution. As such, future research should consider the performance of heteroscedastic item parameter recovery in the presence of skew outside this range. Also, given that currently the only program capable of estimating the HSGRM is Mx, this approach may necessitate acquiring requisite knowledge in a new software. Furthermore, estimating the HSGRM in Mx is rather time intensive and this process is drastically lengthened as the number of items increases. Finally, given that this model is specific to the skew-normal distribution, future research should consider incorporating alternative non-normal distributional forms that allow for greater skew.

These results nonetheless provide a functional solution for the non-normality predicament. In the presence of non-normality, this alternative graded response model offers clear benefits with regard to accurate estimation of item parameters. As demonstrated, the effects of model misspecification were undeniably atrocious as graphically emphasized in Figures 5 to 13 and most notably for small samples sizes with 5-category response option. Unfortunately, what is considered a small sample size in this study tends to be somewhat the norm, if lucky, in psychological research. For this reason it is even more imperative that sources of non-normality be identified and modeled. The added attractiveness of the HSGRM is that constraints can be imposed on skew and/or heteroscedastic errors. Thus, this model is flexible and can be implemented for data without any sort of non-normality, for data where it is suspected than non-normality is specific to the trait, data with possible heteroscedastic errors, and in other cases, both. Even more attractive is that all four models can be estimated with one command script and likelihood ratio tests can be conducted to identify the best fitting model. Of course, a variety of other methods currently exist (i.e., RC-IRT, log-logistic, etc.) or are being developed to handle non-normality. It is not the opinion of the author to recommend any particular method over the other but rather provide evidence supporting to utility of this model even under the most extreme non-normal conditions.

TABLES

Table 1. Means and Standard Deviations for Skew and Heteroscedastic Errors from Each Model in 5-Category Conditions

	Skew								Heteroscedastic Errors							
N=2000	B		S		H		HS		B		S		H		HS	
h0s0	-	-	0.01	(0.27)	-	-	-0.01	(0.40)	-	-	-	-	0.00	(0.16)	0.00	(0.16)
h0s5	-	-	2.17	(0.26)	-	-	2.17	(0.28)	-	-	-	-	-0.10	(0.16)	0.00	(0.16)
h0s75	-	-	3.62	(0.44)	-	-	3.63	(0.45)	-	-	-	-	-0.15	(0.16)	0.00	(0.16)
h0s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	-0.19	(0.15)	-0.01	(0.16)
h4s0	-	-	-0.29	(0.50)	-	-	-0.01	(0.45)	-	-	-	-	0.40	(0.15)	0.41	(0.15)
h4s5	-	-	1.75	(0.22)	-	-	2.17	(0.25)	-	-	-	-	0.30	(0.16)	0.40	(0.17)
h4s75	-	-	3.16	(0.39)	-	-	3.60	(0.45)	-	-	-	-	0.23	(0.17)	0.38	(0.17)
h4s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.19	(0.18)	0.39	(0.19)
h8s0	-	-	-0.25	(0.57)	-	-	-0.13	(0.31)	-	-	-	-	0.81	(0.15)	0.81	(0.15)
h8s5	-	-	1.41	(0.23)	-	-	2.11	(0.29)	-	-	-	-	0.68	(0.16)	0.78	(0.16)
h8s75	-	-	2.65	(0.34)	-	-	3.41	(0.52)	-	-	-	-	0.62	(0.17)	0.76	(0.17)
h8s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.53	(0.18)	0.79	(0.22)
h1s0	-	-	-0.20	(0.54)	-	-	-0.09	(0.29)	-	-	-	-	1.00	(0.15)	1.00	(0.15)
h1s5	-	-	1.16	(0.27)	-	-	1.95	(0.36)	-	-	-	-	0.89	(0.17)	0.96	(0.17)
h1s75	-	-	2.49	(0.34)	-	-	3.32	(0.53)	-	-	-	-	0.82	(0.18)	0.94	(0.17)
h1s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.69	(0.19)	1.01	(0.24)
N=500																
h0s0	-	-	0.05	(0.46)	-	-	0.05	(0.59)	-	-	-	-	0.01	(0.35)	0.01	(0.35)
h0s5	-	-	2.22	(0.62)	-	-	2.25	(0.69)	-	-	-	-	-0.10	(0.33)	0.01	(0.34)
h0s75	-	-	3.89	(0.99)	-	-	3.95	(1.12)	-	-	-	-	-0.16	(0.33)	0.01	(0.35)
h0s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	-0.20	(0.33)	0.00	(0.35)
h4s0	-	-	-0.22	(0.50)	-	-	-0.11	(0.51)	-	-	-	-	0.40	(0.34)	0.40	(0.34)
h4s5	-	-	1.75	(0.55)	-	-	2.14	(0.66)	-	-	-	-	0.29	(0.35)	0.40	(0.36)
h4s75	-	-	3.19	(0.79)	-	-	3.64	(0.97)	-	-	-	-	0.24	(0.35)	0.39	(0.37)
h4s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.17	(0.36)	0.41	(0.41)
h8s0	-	-	-0.22	(0.54)	-	-	-0.23	(0.48)	-	-	-	-	0.80	(0.33)	0.80	(0.33)
h8s5	-	-	1.34	(0.57)	-	-	1.96	(0.65)	-	-	-	-	0.68	(0.35)	0.78	(0.36)
h8s75	-	-	2.70	(0.70)	-	-	3.48	(0.97)	-	-	-	-	0.62	(0.36)	0.78	(0.37)
h8s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.52	(0.38)	0.82	(0.48)
h1s0	-	-	-0.22	(0.57)	-	-	-0.36	(0.41)	-	-	-	-	1.01	(0.31)	1.00	(0.32)
h1s5	-	-	1.10	(0.73)	-	-	1.68	(0.88)	-	-	-	-	0.89	(0.35)	0.98	(0.36)
h1s75	-	-	2.58	(0.67)	-	-	3.35	(0.96)	-	-	-	-	0.82	(0.38)	0.96	(0.39)
h1s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.68	(0.38)	1.06	(0.53)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Tabled mean values correspond to skew parameter used to compute actual skew value (i.e., skew of 1.0 = 28) Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 2. Means and Standard Deviations for Baseline Residual from Each Model in 5-Category Conditions

N=2000	B		S		H		HS	
h0s0	1.00	(0.09)	1.00	(0.09)	1.00	(0.12)	1.00	(0.12)
h0s5	0.97	(0.09)	1.00	(0.09)	0.93	(0.11)	1.01	(0.12)
h0s75	0.96	(0.09)	1.00	(0.09)	0.90	(0.10)	1.01	(0.11)
h0s1	0.94	(0.09)	1.02	(0.09)	0.86	(0.10)	1.02	(0.11)
h4s0	0.80	(0.08)	0.80	(0.08)	1.01	(0.13)	1.01	(0.13)
h4s5	0.78	(0.07)	0.80	(0.08)	0.92	(0.12)	1.00	(0.13)
h4s75	0.78	(0.07)	0.82	(0.08)	0.89	(0.11)	1.00	(0.13)
h4s1	0.77	(0.07)	0.84	(0.08)	0.85	(0.11)	1.01	(0.13)
h8s0	0.62	(0.06)	0.62	(0.06)	1.01	(0.13)	1.01	(0.13)
h8s5	0.61	(0.06)	0.62	(0.06)	0.91	(0.12)	0.99	(0.13)
h8s75	0.61	(0.06)	0.63	(0.06)	0.87	(0.12)	0.98	(0.14)
h8s1	0.62	(0.06)	0.68	(0.07)	0.83	(0.12)	0.98	(0.14)
h1s0	0.54	(0.06)	0.54	(0.06)	1.00	(0.13)	1.00	(0.13)
h1s5	0.53	(0.06)	0.54	(0.06)	0.91	(0.13)	0.98	(0.13)
h1s75	0.53	(0.06)	0.56	(0.06)	0.87	(0.13)	0.96	(0.13)
h1s1	0.56	(0.06)	0.61	(0.06)	0.82	(0.13)	0.98	(0.15)
N=500								
h0s0	1.01	(0.20)	1.01	(0.20)	1.04	(0.27)	1.04	(0.27)
h0s5	0.97	(0.19)	1.00	(0.19)	0.95	(0.23)	1.03	(0.26)
h0s75	0.96	(0.19)	1.01	(0.19)	0.91	(0.22)	1.03	(0.26)
h0s1	0.95	(0.18)	1.03	(0.19)	0.88	(0.21)	1.05	(0.25)
h4s0	0.80	(0.16)	0.80	(0.16)	1.02	(0.27)	1.02	(0.27)
h4s5	0.78	(0.15)	0.80	(0.16)	0.94	(0.25)	1.02	(0.28)
h4s75	0.78	(0.16)	0.82	(0.16)	0.90	(0.25)	1.02	(0.28)
h4s1	0.77	(0.15)	0.84	(0.16)	0.86	(0.24)	1.03	(0.29)
h8s0	0.61	(0.13)	0.61	(0.13)	1.01	(0.26)	1.01	(0.26)
h8s5	0.60	(0.13)	0.61	(0.13)	0.91	(0.26)	0.99	(0.28)
h8s75	0.61	(0.13)	0.64	(0.14)	0.90	(0.27)	1.01	(0.30)
h8s1	0.61	(0.13)	0.67	(0.14)	0.83	(0.25)	1.01	(0.33)
h1s0	0.54	(0.12)	0.54	(0.12)	1.01	(0.26)	1.00	(0.26)
h1s5	0.53	(0.12)	0.54	(0.12)	0.92	(0.27)	0.99	(0.29)
h1s75	0.53	(0.12)	0.55	(0.12)	0.87	(0.27)	0.98	(0.29)
h1s1	0.55	(0.12)	0.61	(0.13)	0.82	(0.26)	1.01	(0.34)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 3. Means and Standard Deviations for Factor Loadings from Each Model in 5-Category Conditions

N=2000	B		S		H		HS	
h0s0	1.00	(0.05)	1.00	(0.05)	1.00	(0.06)	1.00	(0.06)
h0s5	0.98	(0.05)	1.00	(0.06)	0.96	(0.06)	1.00	(0.07)
h0s75	0.97	(0.05)	1.00	(0.06)	0.94	(0.06)	1.00	(0.07)
h0s1	0.96	(0.05)	1.03	(0.06)	0.91	(0.05)	1.03	(0.07)
h4s0	0.90	(0.05)	0.90	(0.05)	1.00	(0.06)	1.00	(0.06)
h4s5	0.87	(0.05)	0.89	(0.05)	0.95	(0.06)	1.00	(0.07)
h4s75	0.86	(0.05)	0.89	(0.05)	0.92	(0.06)	0.99	(0.07)
h4s1	0.85	(0.05)	0.92	(0.05)	0.90	(0.06)	1.02	(0.08)
h8s0	0.80	(0.04)	0.80	(0.04)	1.00	(0.06)	1.00	(0.06)
h8s5	0.77	(0.04)	0.78	(0.04)	0.94	(0.06)	0.99	(0.07)
h8s75	0.76	(0.04)	0.78	(0.04)	0.91	(0.06)	0.97	(0.07)
h8s1	0.76	(0.04)	0.82	(0.05)	0.88	(0.06)	1.02	(0.09)
h1s0	0.75	(0.04)	0.75	(0.04)	1.00	(0.06)	1.00	(0.06)
h1s5	0.72	(0.04)	0.73	(0.04)	0.94	(0.06)	0.98	(0.07)
h1s75	0.71	(0.04)	0.73	(0.04)	0.91	(0.06)	0.96	(0.07)
h1s1	0.71	(0.04)	0.78	(0.05)	0.87	(0.07)	1.02	(0.10)
N=500								
h0s0	1.01	(0.11)	1.01	(0.11)	1.02	(0.13)	1.02	(0.14)
h0s5	0.98	(0.11)	1.00	(0.11)	0.96	(0.13)	1.01	(0.15)
h0s75	0.97	(0.11)	1.00	(0.11)	0.94	(0.12)	1.01	(0.15)
h0s1	0.96	(0.11)	1.04	(0.12)	0.92	(0.12)	1.05	(0.16)
h4s0	0.90	(0.10)	0.90	(0.10)	1.00	(0.13)	1.00	(0.13)
h4s5	0.87	(0.09)	0.88	(0.10)	0.95	(0.13)	1.00	(0.15)
h4s75	0.86	(0.10)	0.89	(0.10)	0.92	(0.13)	1.00	(0.16)
h4s1	0.85	(0.10)	0.92	(0.11)	0.89	(0.12)	1.03	(0.18)
h8s0	0.79	(0.08)	0.79	(0.08)	1.00	(0.12)	1.00	(0.12)
h8s5	0.77	(0.09)	0.77	(0.09)	0.94	(0.13)	0.98	(0.15)
h8s75	0.76	(0.09)	0.78	(0.10)	0.92	(0.14)	0.99	(0.16)
h8s1	0.75	(0.09)	0.81	(0.10)	0.87	(0.13)	1.03	(0.20)
h1s0	0.75	(0.08)	0.74	(0.08)	1.00	(0.12)	0.99	(0.12)
h1s5	0.72	(0.08)	0.73	(0.08)	0.94	(0.13)	0.98	(0.14)
h1s75	0.71	(0.08)	0.73	(0.09)	0.91	(0.13)	0.97	(0.16)
h1s1	0.71	(0.09)	0.77	(0.10)	0.86	(0.14)	1.03	(0.21)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 4. Means and Standard Deviations for Threshold 3 from Each Model in 5-Category Conditions

N=2000	B		S		H		HS	
h0s0	1.00	(0.11)	1.00	(0.11)	1.00	(0.13)	1.00	(0.13)
h0s5	0.97	(0.11)	1.00	(0.11)	0.93	(0.12)	1.01	(0.13)
h0s75	0.96	(0.11)	1.00	(0.11)	0.90	(0.12)	1.01	(0.13)
h0s1	0.94	(0.11)	1.02	(0.11)	0.86	(0.12)	1.02	(0.12)
h4s0	0.79	(0.10)	0.78	(0.10)	1.01	(0.14)	1.01	(0.14)
h4s5	0.78	(0.10)	0.80	(0.10)	0.93	(0.13)	1.01	(0.14)
h4s75	0.78	(0.10)	0.82	(0.10)	0.89	(0.13)	1.00	(0.14)
h4s1	0.77	(0.10)	0.85	(0.10)	0.86	(0.13)	1.01	(0.14)
h8s0	0.56	(0.09)	0.56	(0.09)	1.01	(0.14)	1.00	(0.14)
h8s5	0.57	(0.09)	0.58	(0.09)	0.92	(0.14)	0.99	(0.14)
h8s75	0.57	(0.09)	0.60	(0.09)	0.88	(0.14)	0.97	(0.14)
h8s1	0.60	(0.09)	0.68	(0.09)	0.85	(0.14)	1.00	(0.16)
h1s0	0.46	(0.08)	0.45	(0.08)	1.00	(0.13)	1.00	(0.13)
h1s5	0.46	(0.08)	0.47	(0.08)	0.92	(0.14)	0.98	(0.14)
h1s75	0.47	(0.09)	0.50	(0.09)	0.88	(0.14)	0.96	(0.14)
h1s1	0.52	(0.09)	0.59	(0.09)	0.84	(0.14)	1.00	(0.17)
N=500								
h0s0	1.01	(0.23)	1.01	(0.23)	1.03	(0.29)	1.03	(0.29)
h0s5	0.98	(0.22)	1.01	(0.22)	0.95	(0.26)	1.03	(0.27)
h0s75	0.96	(0.22)	1.01	(0.22)	0.91	(0.25)	1.03	(0.27)
h0s1	0.95	(0.22)	1.03	(0.22)	0.88	(0.24)	1.05	(0.27)
h4s0	0.78	(0.20)	0.78	(0.20)	1.01	(0.29)	1.01	(0.29)
h4s5	0.78	(0.20)	0.80	(0.20)	0.94	(0.27)	1.02	(0.29)
h4s75	0.78	(0.20)	0.82	(0.21)	0.90	(0.27)	1.02	(0.29)
h4s1	0.77	(0.20)	0.85	(0.21)	0.86	(0.26)	1.03	(0.30)
h8s0	0.55	(0.18)	0.55	(0.18)	0.99	(0.27)	0.99	(0.27)
h8s5	0.55	(0.18)	0.57	(0.19)	0.91	(0.29)	0.98	(0.30)
h8s75	0.57	(0.19)	0.61	(0.20)	0.89	(0.29)	1.00	(0.31)
h8s1	0.59	(0.19)	0.67	(0.20)	0.83	(0.28)	1.02	(0.35)
h1s0	0.45	(0.17)	0.45	(0.17)	1.00	(0.27)	0.99	(0.27)
h1s5	0.46	(0.17)	0.47	(0.18)	0.92	(0.29)	0.98	(0.30)
h1s75	0.47	(0.18)	0.50	(0.19)	0.87	(0.30)	0.97	(0.31)
h1s1	0.51	(0.19)	0.59	(0.20)	0.83	(0.29)	1.03	(0.37)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 5. Means and Standard Deviations for Threshold 4 from Each Model in 5-Category Conditions

N=2000	B		S		H		HS	
h0s0	2.50	(0.18)	2.50	(0.18)	2.51	(0.25)	2.51	(0.25)
h0s5	2.43	(0.18)	2.50	(0.18)	2.32	(0.24)	2.51	(0.25)
h0s75	2.38	(0.18)	2.51	(0.18)	2.23	(0.22)	2.51	(0.25)
h0s1	2.34	(0.17)	2.54	(0.18)	2.14	(0.21)	2.54	(0.25)
h4s0	2.02	(0.16)	2.01	(0.16)	2.52	(0.26)	2.52	(0.26)
h4s5	1.97	(0.15)	2.02	(0.16)	2.31	(0.25)	2.51	(0.27)
h4s75	1.95	(0.15)	2.05	(0.16)	2.22	(0.24)	2.49	(0.27)
h4s1	1.93	(0.15)	2.11	(0.16)	2.13	(0.24)	2.51	(0.27)
h8s0	1.58	(0.14)	1.57	(0.14)	2.51	(0.25)	2.51	(0.25)
h8s5	1.55	(0.14)	1.59	(0.14)	2.31	(0.25)	2.48	(0.27)
h8s75	1.54	(0.14)	1.62	(0.14)	2.20	(0.26)	2.44	(0.28)
h8s1	1.57	(0.14)	1.73	(0.15)	2.11	(0.25)	2.49	(0.31)
h1s0	1.38	(0.13)	1.37	(0.13)	2.50	(0.24)	2.51	(0.24)
h1s5	1.35	(0.13)	1.38	(0.13)	2.30	(0.25)	2.45	(0.26)
h1s75	1.36	(0.13)	1.43	(0.14)	2.20	(0.26)	2.42	(0.27)
h1s1	1.41	(0.14)	1.56	(0.14)	2.09	(0.26)	2.49	(0.33)
N=500								
h0s0	2.53	(0.38)	2.53	(0.38)	2.57	(0.55)	2.57	(0.55)
h0s5	2.44	(0.37)	2.52	(0.38)	2.36	(0.50)	2.57	(0.55)
h0s75	2.39	(0.37)	2.52	(0.38)	2.25	(0.48)	2.56	(0.56)
h0s1	2.36	(0.36)	2.57	(0.38)	2.18	(0.46)	2.60	(0.55)
h4s0	2.02	(0.32)	2.01	(0.32)	2.53	(0.54)	2.52	(0.54)
h4s5	1.98	(0.32)	2.03	(0.33)	2.34	(0.52)	2.54	(0.57)
h4s75	1.96	(0.32)	2.05	(0.33)	2.24	(0.51)	2.53	(0.58)
h4s1	1.93	(0.32)	2.11	(0.34)	2.14	(0.48)	2.56	(0.60)
h8s0	1.56	(0.28)	1.56	(0.28)	2.49	(0.50)	2.49	(0.50)
h8s5	1.53	(0.29)	1.57	(0.30)	2.29	(0.53)	2.46	(0.57)
h8s75	1.55	(0.30)	1.63	(0.31)	2.24	(0.54)	2.50	(0.60)
h8s1	1.55	(0.30)	1.71	(0.32)	2.08	(0.52)	2.52	(0.68)
h1s0	1.38	(0.26)	1.37	(0.26)	2.51	(0.49)	2.49	(0.49)
h1s5	1.35	(0.27)	1.38	(0.28)	2.30	(0.53)	2.46	(0.56)
h1s75	1.36	(0.28)	1.43	(0.29)	2.20	(0.54)	2.45	(0.59)
h1s1	1.40	(0.29)	1.56	(0.31)	2.07	(0.53)	2.55	(0.71)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 6. Means and Standard Deviations for Intercept from Each Model in 5-Category Conditions

N=2000	B		S		H		HS	
h0s0	0.00	(0.07)	0.00	(0.07)	0.00	(0.08)	0.00	(0.08)
h0s5	-0.02	(0.07)	0.00	(0.07)	-0.05	(0.08)	0.00	(0.08)
h0s75	-0.03	(0.07)	0.00	(0.07)	-0.07	(0.07)	0.00	(0.08)
h0s1	-.04	(0.07)	-0.02	(0.07)	-0.09	(0.07)	-0.02	(0.08)
h4s0	-0.13	(0.06)	-0.14	(0.06)	0.00	(0.09)	0.00	(0.09)
h4s5	-0.14	(0.06)	-0.13	(0.06)	-0.05	(0.08)	0.00	(0.09)
h4s75	-0.14	(0.06)	-0.12	(0.06)	-0.07	(0.08)	-0.01	(0.09)
h4s1	-0.14	(0.06)	-0.13	(0.06)	-0.09	(0.08)	-0.03	(0.08)
h8s0	-0.26	(0.05)	-0.26	(0.05)	0.00	(0.08)	0.00	(0.08)
h8s5	-0.26	(0.05)	-0.25	(0.05)	-0.05	(0.08)	-0.01	(0.08)
h8s75	-0.26	(0.05)	-0.24	(0.05)	-0.07	(0.08)	-0.02	(0.09)
h8s1	-0.24	(0.05)	-0.23	(0.06)	-0.10	(0.08)	-0.05	(0.09)
h1s0	-0.32	(0.05)	-0.32	(0.05)	0.00	(0.08)	0.00	(0.08)
h1s5	-0.32	(0.05)	-0.31	(0.05)	-0.05	(0.08)	-0.01	(0.08)
h1s75	-0.31	(0.05)	-0.29	(0.05)	-0.07	(0.08)	-0.03	(0.08)
h1s1	-0.29	(0.05)	-0.27	(0.05)	-0.10	(0.08)	-0.05	(0.09)
N=500								
h0s0	0.01	(0.15)	0.01	(0.15)	0.03	(0.18)	0.03	(0.18)
h0s5	-0.01	(0.14)	0.01	(0.14)	-0.03	(0.16)	0.02	(0.17)
h0s75	-0.03	(0.14)	0.00	(0.14)	-0.06	(0.16)	0.01	(0.17)
h0s1	-0.03	(0.13)	-0.01	(0.13)	-0.07	(0.15)	0.00	(0.16)
h4s0	-0.13	(0.13)	-0.13	(0.13)	0.01	(0.18)	0.01	(0.18)
h4s5	-0.13	(0.13)	-0.12	(0.13)	-0.04	(0.17)	0.02	(0.18)
h4s75	-0.14	(0.12)	-0.11	(0.12)	-0.06	(0.17)	0.01	(0.18)
h4s1	-0.15	(0.12)	-0.13	(0.12)	-0.09	(0.16)	-0.02	(0.18)
h8s0	-0.26	(0.11)	-0.26	(0.11)	0.00	(0.16)	0.00	(0.17)
h8s5	-0.26	(0.11)	-0.25	(0.11)	-0.05	(0.17)	0.00	(0.18)
h8s75	-0.25	(0.12)	-0.23	(0.12)	-0.06	(0.17)	0.01	(0.19)
h8s1	-0.25	(0.11)	-0.23	(0.12)	-0.10	(0.16)	-0.04	(0.19)
h1s0	-0.32	(0.10)	-0.32	(0.10)	0.00	(0.16)	0.00	(0.16)
h1s5	-0.32	(0.10)	-0.32	(0.10)	-0.06	(0.17)	-0.01	(0.17)
h1s75	-0.31	(0.11)	-0.29	(0.11)	-0.08	(0.17)	-0.01	(0.18)
h1s1	-0.29	(0.11)	-0.28	(0.11)	-0.11	(0.17)	-0.04	(0.20)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 7. Means and Standard Deviations for Skew and Heteroscedastic Errors for Each model in 3-Category Conditions

	Skew								Heteroscedastic Error							
N=2000	B		S		H		HS		B		S		H		HS	
h0s0	-	-	-0.02	(0.31)	-	-	-0.01	(0.49)	-	-	-	-	0.00	(0.20)	0.00	(0.20)
h0s5	-	-	2.17	(0.28)	-	-	2.16	(0.31)	-	-	-	-	-0.19	(0.20)	-0.01	(0.20)
h0s75	-	-	3.63	(0.47)	-	-	3.63	(0.51)	-	-	-	-	-0.27	(0.19)	-0.01	(0.20)
h0s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	-0.35	(0.18)	-0.01	(0.20)
h4s0	-	-	-0.48	(0.71)	-	-	-0.20	(0.56)	-	-	-	-	0.41	(0.20)	0.40	(0.20)
h4s5	-	-	1.44	(0.25)	-	-	2.15	(0.28)	-	-	-	-	0.21	(0.20)	0.40	(0.21)
h4s75	-	-	2.68	(0.41)	-	-	3.54	(0.51)	-	-	-	-	0.11	(0.20)	0.39	(0.21)
h4s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.01	(0.19)	0.40	(0.23)
h8s0	-	-	-0.98	(1.00)	-	-	-0.23	(0.41)	-	-	-	-	0.81	(0.21)	0.80	(0.21)
h8s5	-	-	0.56	(0.53)	-	-	1.89	(0.76)	-	-	-	-	0.60	(0.21)	0.78	(0.22)
h8s75	-	-	1.91	(0.29)	-	-	3.41	(0.51)	-	-	-	-	0.47	(0.21)	0.75	(0.22)
h8s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.32	(0.19)	0.83	(0.32)
h1s0	-	-	-1.05	(1.14)	-	-	-0.25	(0.40)	-	-	-	-	1.02	(0.22)	1.02	(0.23)
h1s5	-	-	-0.15	(0.68)	-	-	1.09	(1.05)	-	-	-	-	0.79	(0.23)	0.89	(0.24)
h1s75	-	-	1.63	(0.28)	-	-	3.34	(0.59)	-	-	-	-	0.63	(0.21)	0.91	(0.22)
h1s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.46	(0.20)	1.09	(0.43)
N=500																
h0s0	-	-	0.00	(0.61)	-	-	-0.02	(0.94)	-	-	-	-	0.01	(0.43)	0.01	(0.43)
h0s5	-	-	2.35	(0.66)	-	-	2.45	(0.79)	-	-	-	-	-0.18	(0.43)	0.02	(0.46)
h0s75	-	-	3.73	(1.05)	-	-	3.88	(1.31)	-	-	-	-	-0.29	(0.41)	-0.01	(0.47)
h0s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	-0.36	(0.38)	0.01	(0.48)
h4s0	-	-	-0.59	(0.78)	-	-	-0.31	(0.71)	-	-	-	-	0.43	(0.44)	0.41	(0.45)
h4s5	-	-	1.47	(0.58)	-	-	2.28	(0.85)	-	-	-	-	0.23	(0.44)	0.43	(0.47)
h4s75	-	-	2.76	(0.83)	-	-	3.81	(1.31)	-	-	-	-	0.11	(0.44)	0.41	(0.51)
h4s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.01	(0.40)	0.50	(0.67)
h8s0	-	-	-1.08	(1.09)	-	-	-0.45	(0.57)	-	-	-	-	0.86	(0.47)	0.85	(0.48)
h8s5	-	-	0.44	(0.82)	-	-	1.41	(1.18)	-	-	-	-	0.64	(0.49)	0.78	(0.51)
h8s75	-	-	1.92	(0.58)	-	-	3.46	(1.16)	-	-	-	-	0.50	(0.51)	0.82	(0.60)
h8s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.33	(0.42)	1.02	(0.88)
h1s0	-	-	-1.08	(1.19)	-	-	-0.39	(0.52)	-	-	-	-	1.06	(0.47)	1.05	(0.48)
h1s5	-	-	-0.05	(0.88)	-	-	0.91	(1.17)	-	-	-	-	0.84	(0.51)	0.94	(0.52)
h1s75	-	-	1.65	(0.71)	-	-	3.34	(1.31)	-	-	-	-	0.67	(0.48)	0.97	(0.55)
h1s1	-	-	28.00	(0.00)	-	-	28.00	(0.00)	-	-	-	-	0.48	(0.47)	1.33	(1.02)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Tabled mean values correspond to skew parameter used to compute actual skew value (i.e., skew of 1.0 = 28) Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 8. Means and Standard Deviations for Baseline Residual for Each Model in 3-Category Conditions

N=2000	B		S		H		HS	
h0s0	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)
h0s5	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)
h0s75	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)
h0s1	1.01	(0.06)	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)
h4s0	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)	1.00	(0.06)
h4s5	0.99	(0.06)	0.99	(0.06)	0.99	(0.06)	1.00	(0.06)
h4s75	0.99	(0.06)	0.99	(0.06)	0.99	(0.06)	1.00	(0.06)
h4s1	0.98	(0.06)	0.98	(0.06)	0.98	(0.06)	0.99	(0.06)
h8s0	1.01	(0.06)	1.01	(0.06)	1.00	(0.06)	1.00	(0.06)
h8s5	0.99	(0.06)	0.99	(0.06)	0.98	(0.06)	1.00	(0.06)
h8s75	0.98	(0.06)	0.97	(0.06)	0.97	(0.06)	1.00	(0.06)
h8s1	0.97	(0.05)	0.97	(0.06)	0.97	(0.06)	0.98	(0.06)
h1s0	1.01	(0.06)	1.01	(0.06)	1.00	(0.06)	1.00	(0.06)
h1s5	0.98	(0.06)	0.98	(0.06)	0.98	(0.06)	0.99	(0.06)
h1s75	0.97	(0.06)	0.97	(0.06)	0.97	(0.06)	1.00	(0.06)
h1s1	0.96	(0.06)	0.96	(0.06)	0.96	(0.06)	0.98	(0.06)
N=500								
h0s0	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)
h0s5	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)
h0s75	1.00	(0.11)	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)
h0s1	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)
h4s0	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)	1.00	(0.12)
h4s5	0.99	(0.12)	0.99	(0.12)	0.99	(0.12)	1.00	(0.12)
h4s75	0.99	(0.11)	0.99	(0.11)	0.99	(0.11)	1.00	(0.12)
h4s1	0.98	(0.12)	0.97	(0.12)	0.98	(0.12)	0.99	(0.13)
h8s0	1.01	(0.12)	1.01	(0.12)	1.00	(0.12)	1.00	(0.12)
h8s5	0.98	(0.11)	0.98	(0.11)	0.98	(0.12)	0.99	(0.12)
h8s75	0.97	(0.11)	0.97	(0.11)	0.98	(0.12)	1.00	(0.13)
h8s1	0.97	(0.11)	0.96	(0.11)	0.97	(0.11)	0.99	(0.13)
h1s0	1.01	(0.12)	1.01	(0.12)	1.00	(0.13)	1.00	(0.13)
h1s5	0.98	(0.12)	0.98	(0.12)	0.98	(0.12)	0.99	(0.13)
h1s75	0.97	(0.12)	0.97	(0.12)	0.97	(0.12)	1.00	(0.13)
h1s1	0.96	(0.11)	0.96	(0.12)	0.96	(0.12)	0.98	(0.13)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1).

Table 9. Means and Standard Deviations for Factor Loadings from each Model in 3-Category Conditions

N=2000	B		S		H		HS	
h0s0	1.00	(0.04)	1.00	(0.04)	1.00	(0.04)	1.00	(0.04)
h0s5	0.99	(0.04)	1.00	(0.04)	0.98	(0.04)	1.00	(0.04)
h0s75	0.98	(0.04)	1.00	(0.04)	0.98	(0.04)	1.00	(0.04)
h0s1	0.97	(0.04)	1.02	(0.04)	0.96	(0.04)	1.02	(0.04)
h4s0	1.01	(0.04)	1.01	(0.04)	1.00	(0.04)	1.00	(0.04)
h4s5	0.99	(0.04)	0.99	(0.04)	0.98	(0.04)	1.00	(0.04)
h4s75	0.97	(0.04)	0.98	(0.04)	0.97	(0.04)	1.00	(0.04)
h4s1	0.95	(0.04)	1.00	(0.04)	0.95	(0.04)	1.02	(0.04)
h8s0	1.03	(0.04)	1.04	(0.04)	1.00	(0.04)	1.00	(0.04)
h8s5	1.00	(0.04)	1.00	(0.04)	0.98	(0.04)	1.00	(0.04)
h8s75	0.97	(0.04)	0.98	(0.04)	0.97	(0.04)	0.99	(0.04)
h8s1	0.94	(0.04)	0.99	(0.04)	0.94	(0.04)	1.02	(0.04)
h1s0	1.05	(0.04)	1.05	(0.04)	1.00	(0.04)	1.00	(0.04)
h1s5	1.01	(0.04)	1.01	(0.04)	0.98	(0.04)	0.99	(0.04)
h1s75	0.98	(0.04)	0.98	(0.04)	0.97	(0.04)	0.99	(0.04)
h1s1	0.94	(0.04)	0.98	(0.04)	0.93	(0.04)	1.02	(0.04)
N=500								
h0s0	1.00	(0.07)	1.00	(0.08)	1.00	(0.07)	1.00	(0.08)
h0s5	0.99	(0.08)	1.00	(0.08)	0.98	(0.08)	1.00	(0.08)
h0s75	0.98	(0.07)	1.00	(0.08)	0.97	(0.07)	1.00	(0.08)
h0s1	0.97	(0.07)	1.02	(0.08)	0.96	(0.07)	1.01	(0.08)
h4s0	1.01	(0.08)	1.01	(0.08)	1.00	(0.08)	1.00	(0.08)
h4s5	0.99	(0.07)	0.99	(0.07)	0.98	(0.07)	1.00	(0.08)
h4s75	0.97	(0.07)	0.98	(0.08)	0.97	(0.07)	1.00	(0.08)
h4s1	0.95	(0.07)	1.00	(0.08)	0.95	(0.07)	1.02	(0.08)
h8s0	1.04	(0.08)	1.05	(0.08)	1.00	(0.08)	1.01	(0.08)
h8s5	1.00	(0.08)	1.00	(0.08)	0.99	(0.08)	0.99	(0.08)
h8s75	0.97	(0.08)	0.98	(0.08)	0.96	(0.08)	0.99	(0.08)
h8s1	0.94	(0.07)	0.99	(0.08)	0.94	(0.07)	1.02	(0.08)
h1s0	1.04	(0.08)	1.05	(0.08)	1.00	(0.08)	1.00	(0.08)
h1s5	1.01	(0.07)	1.01	(0.07)	0.98	(0.08)	0.99	(0.07)
h1s75	0.98	(0.07)	0.98	(0.07)	0.96	(0.08)	0.98	(0.08)
h1s1	0.94	(0.07)	0.99	(0.08)	0.94	(0.07)	1.02	(0.08)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 10. Means and Standard Deviations for Intercepts for Each Model in 3-Category Conditions

N=2000	B		S		H		HS	
h0s0	0.00	(0.04)	0.00	(0.04)	0.00	(0.04)	0.00	(0.04)
h0s5	-0.01	(0.04)	0.00	(0.04)	-0.02	(0.04)	0.00	(0.04)
h0s75	-0.02	(0.03)	0.00	(0.04)	-0.04	(0.04)	0.00	(0.04)
h0s1	-0.02	(0.03)	-0.02	(0.03)	-0.05	(0.04)	-0.03	(0.04)
h4s0	-0.03	(0.04)	-0.03	(0.04)	0.00	(0.04)	0.00	(0.04)
h4s5	-0.04	(0.03)	-0.03	(0.03)	-0.02	(0.04)	0.00	(0.04)
h4s75	-0.04	(0.03)	-0.03	(0.03)	-0.03	(0.04)	0.00	(0.04)
h4s1	-0.04	(0.03)	-0.06	(0.03)	-0.04	(0.04)	-0.03	(0.04)
h8s0	-0.06	(0.04)	-0.07	(0.04)	0.00	(0.04)	0.00	(0.04)
h8s5	-0.06	(0.03)	-0.06	(0.03)	-0.02	(0.04)	0.00	(0.04)
h8s75	-0.07	(0.03)	-0.06	(0.03)	-0.03	(0.03)	-0.01	(0.04)
h8s1	-0.06	(0.03)	-0.09	(0.03)	-0.04	(0.04)	-0.04	(0.04)
h1s0	-0.07	(0.04)	-0.08	(0.04)	0.00	(0.04)	0.00	(0.04)
h1s5	-0.08	(0.04)	-0.08	(0.04)	-0.02	(0.04)	-0.01	(0.04)
h1s75	-0.07	(0.03)	-0.07	(0.03)	-0.03	(0.04)	-0.01	(0.04)
h1s1	-0.07	(0.03)	-0.10	(0.03)	-0.04	(0.04)	-0.05	(0.04)
N=500								
h0s0	0.00	(0.07)	0.00	(0.07)	0.00	(0.07)	0.00	(0.07)
h0s5	0.00	(0.07)	0.01	(0.07)	-0.02	(0.07)	0.01	(0.08)
h0s75	-0.02	(0.07)	0.00	(0.07)	-0.04	(0.08)	-0.01	(0.08)
h0s1	-0.02	(0.07)	-0.03	(0.07)	-0.05	(0.07)	-0.03	(0.07)
h4s0	-0.03	(0.07)	-0.04	(0.07)	0.00	(0.07)	-0.01	(0.07)
h4s5	-0.03	(0.07)	-0.02	(0.07)	-0.02	(0.07)	0.01	(0.07)
h4s75	-0.04	(0.07)	-0.03	(0.07)	-0.04	(0.07)	-0.01	(0.07)
h4s1	-0.04	(0.07)	-0.06	(0.07)	-0.04	(0.07)	-0.03	(0.07)
h8s0	-0.06	(0.07)	-0.07	(0.07)	0.00	(0.07)	-0.01	(0.07)
h8s5	-0.06	(0.07)	-0.06	(0.07)	-0.02	(0.08)	0.00	(0.08)
h8s75	-0.06	(0.07)	-0.05	(0.07)	-0.03	(0.07)	0.00	(0.08)
h8s1	-0.06	(0.07)	-0.08	(0.07)	-0.04	(0.07)	-0.04	(0.07)
h1s0	-0.07	(0.07)	-0.08	(0.07)	0.00	(0.07)	0.00	(0.07)
h1s5	-0.08	(0.07)	-0.08	(0.07)	-0.02	(0.08)	-0.01	(0.08)
h1s75	-0.07	(0.07)	-0.07	(0.07)	-0.03	(0.08)	-0.01	(0.08)
h1s1	-0.07	(0.07)	-0.09	(0.07)	-0.04	(0.07)	-0.05	(0.07)

Note. Standard deviations are presented in parentheses. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). The standard normal-ogive model or baseline model is denoted “B”, skew-only is denoted “S”, het-only model is denoted “H”, and full HSGRM is denoted “HS”.

Table 11. Average Bias and RMSD Due to Model Misspecification in 5-Category Response Conditions

	Bias					RMSD				
N=2000	Res	Load	Thresh3	Thresh4	Int	Res	Load	Thresh3	Thresh4	Int
h0s0	0.00	0.00	0.00	0.00	0.00	0.08	0.05	0.10	0.16	0.06
h0s5	-0.03	-0.02	-0.03	-0.07	-0.02	0.09	0.05	0.10	0.17	0.06
h0s75	-0.04	-0.03	-0.04	-0.12	-0.03	0.09	0.05	0.10	0.19	0.07
h0s1	-0.06	-0.04	-0.06	-0.16	-0.04	0.09	0.06	0.11	0.21	0.07
h4s0	-0.20	-0.10	-0.21	-0.48	-0.13	0.19	0.10	0.21	0.45	0.13
h4s5	-0.22	-0.13	-0.22	-0.53	-0.14	0.21	0.12	0.22	0.50	0.14
h4s75	-0.22	-0.14	-0.22	-0.55	-0.14	0.21	0.13	0.22	0.51	0.14
h4s1	-0.23	-0.15	-0.23	-0.57	-0.14	0.21	0.14	0.22	0.53	0.14
h8s0	-0.38	-0.20	-0.44	-0.92	-0.26	0.35	0.18	0.40	0.83	0.24
h8s5	-0.39	-0.23	-0.43	-0.95	-0.26	0.35	0.21	0.40	0.86	0.24
h8s75	-0.39	-0.24	-0.43	-0.96	-0.26	0.36	0.22	0.39	0.87	0.23
h8s1	-0.38	-0.24	-0.40	-0.93	-0.24	0.34	0.22	0.37	0.84	0.22
h1s0	-0.46	-0.25	-0.54	-1.12	-0.32	0.41	0.23	0.49	1.01	0.29
h1s5	-0.47	-0.28	-0.54	-1.15	-0.32	0.42	0.25	0.49	1.03	0.29
h1s75	-0.47	-0.29	-0.53	-1.14	-0.31	0.42	0.26	0.48	1.03	0.28
h1s1	-0.44	-0.29	-0.48	-1.09	-0.29	0.40	0.26	0.44	0.98	0.26
N=500										
h0s0	0.01	0.01	0.01	0.03	0.01	0.18	0.10	0.21	0.34	0.13
h0s5	-0.03	-0.02	-0.02	-0.06	-0.01	0.17	0.10	0.20	0.33	0.13
h0s75	-0.04	-0.03	-0.04	-0.11	-0.03	0.17	0.10	0.20	0.34	0.13
h0s1	-0.05	-0.04	-0.05	-0.14	-0.03	0.17	0.10	0.20	0.34	0.12
h4s0	-0.20	-0.10	-0.22	-0.48	-0.13	0.23	0.13	0.27	0.52	0.16
h4s5	-0.22	-0.13	-0.22	-0.52	-0.13	0.24	0.14	0.27	0.55	0.16
h4s75	-0.22	-0.14	-0.22	-0.54	-0.14	0.24	0.15	0.27	0.57	0.16
h4s1	-0.23	-0.15	-0.23	-0.57	-0.15	0.25	0.16	0.27	0.58	0.17
h8s0	-0.39	-0.21	-0.45	-0.94	-0.26	0.36	0.20	0.43	0.87	0.25
h8s5	-0.40	-0.23	-0.45	-0.97	-0.26	0.38	0.22	0.43	0.90	0.25
h8s75	-0.39	-0.24	-0.43	-0.95	-0.25	0.37	0.23	0.42	0.89	0.24
h8s1	-0.39	-0.25	-0.41	-0.95	-0.25	0.37	0.24	0.41	0.89	0.24
h1s0	-0.46	-0.25	-0.55	-1.12	-0.32	0.43	0.24	0.51	1.03	0.30
h1s5	-0.47	-0.28	-0.54	-1.15	-0.32	0.44	0.26	0.51	1.06	0.30
h1s75	-0.47	-0.29	-0.53	-1.14	-0.31	0.43	0.27	0.50	1.05	0.29
h1s1	-0.45	-0.29	-0.49	-1.10	-0.29	0.42	0.27	0.47	1.02	0.28

Note. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). Baseline residuals denoted = Res, factor loadings = Load, threshold 3 = Thresh3, threshold 4 = Thresh4, intercept = Int.

Table 12. Average bias and RMSD due to model misspecification in 3-Category Response Conditions

	Bias			RMSD		
N=2000	Res	Load	Int	Res	Load	Int
h0s0	0.00	0.00	0.00	0.05	0.03	0.03
h0s5	0.00	-0.01	-0.01	0.05	0.03	0.03
h0s75	0.00	-0.02	-0.02	0.05	0.04	0.03
h0s1	0.01	-0.03	-0.02	0.05	0.04	0.04
h4s0	0.00	0.01	-0.03	0.05	0.03	0.04
h4s5	-0.01	-0.01	-0.04	0.05	0.04	0.05
h4s75	-0.01	-0.03	-0.04	0.05	0.04	0.05
h4s1	-0.02	-0.05	-0.04	0.05	0.05	0.05
h8s0	0.01	0.03	-0.06	0.05	0.04	0.06
h8s5	-0.01	0.00	-0.06	0.05	0.03	0.07
h8s75	-0.02	-0.03	-0.07	0.06	0.04	0.07
h8s1	-0.03	-0.06	-0.06	0.06	0.06	0.06
h1s0	0.01	0.05	-0.07	0.06	0.05	0.07
h1s5	-0.02	0.01	-0.08	0.05	0.03	0.08
h1s75	-0.03	-0.02	-0.07	0.06	0.04	0.07
h1s1	-0.04	-0.06	-0.07	0.06	0.07	0.07
N=500						
h0s0	0.00	0.00	0.00	0.11	0.07	0.06
h0s5	0.00	-0.01	0.00	0.10	0.07	0.06
h0s75	0.00	-0.02	-0.02	0.10	0.07	0.07
h0s1	0.00	-0.03	-0.02	0.10	0.07	0.06
h4s0	0.00	0.01	-0.03	0.10	0.07	0.07
h4s5	-0.01	-0.01	-0.03	0.10	0.07	0.07
h4s75	-0.01	-0.03	-0.04	0.10	0.07	0.07
h4s1	-0.02	-0.05	-0.04	0.11	0.08	0.07
h8s0	0.01	0.04	-0.06	0.11	0.08	0.09
h8s5	-0.02	0.00	-0.06	0.10	0.07	0.08
h8s75	-0.03	-0.03	-0.06	0.10	0.07	0.08
h8s1	-0.03	-0.06	-0.06	0.10	0.08	0.08
h1s0	0.01	0.04	-0.07	0.11	0.08	0.09
h1s5	-0.02	0.01	-0.08	0.11	0.07	0.09
h1s75	-0.03	-0.02	-0.07	0.11	0.07	0.09
h1s1	-0.04	-0.06	-0.07	0.11	0.08	0.09

Note. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). Baseline residuals denoted = Res, factor loadings = Load, intercept = Int.

Table 13. Bias and RMSD for Item Parameter Estimate from Constrained HSGRM for 5-Category Response Conditions

N=2,000	Bias							RMSD						
Item	Skew	Res	Het	Load	t3	t4	Int	Skew	Res	Het	Load	t3	t4	Int
1	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.08	0.00	0.05	0.09	0.15	0.06
2	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.08	0.00	0.05	0.10	0.17	0.06
3	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00	0.08	0.00	0.05	0.10	0.16	0.06
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.05	0.09	0.16	0.06
5	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.08	0.00	0.05	0.09	0.16	0.06
6	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.08	0.00	0.05	0.10	0.16	0.06
7	0.00	-0.01	0.00	-0.01	-0.02	-0.01	-0.01	0.00	0.08	0.00	0.05	0.10	0.16	0.06
8	0.00	-0.01	0.00	0.00	-0.01	-0.02	-0.01	0.00	0.08	0.00	0.05	0.10	0.16	0.06
9	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.08	0.00	0.05	0.09	0.16	0.06
10	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.05	0.10	0.16	0.06
Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.05	0.10	0.16	0.06
SD	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N=500														
1	0.00	-0.01	0.00	-0.01	-0.03	-0.03	-0.01	0.00	0.18	0.00	0.09	0.20	0.37	0.13
2	0.00	0.01	0.00	0.01	0.02	0.02	0.02	0.00	0.16	0.00	0.08	0.18	0.30	0.12
3	0.00	0.02	0.00	0.02	0.04	0.07	0.02	0.00	0.18	0.00	0.11	0.21	0.32	0.13
4	0.00	0.00	0.00	0.01	0.01	0.05	0.01	0.00	0.17	0.00	0.10	0.19	0.34	0.13
5	0.00	0.00	0.00	0.00	0.00	-0.01	0.01	0.00	0.17	0.00	0.10	0.21	0.34	0.13
6	0.00	-0.01	0.00	-0.01	-0.04	-0.06	-0.02	0.00	0.18	0.00	0.10	0.21	0.35	0.14
7	0.00	0.03	0.00	0.01	0.03	0.06	0.02	0.00	0.17	0.00	0.09	0.21	0.34	0.13
8	0.00	-0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.17	0.00	0.10	0.21	0.33	0.13
9	0.00	0.03	0.00	0.01	0.02	0.04	0.01	0.00	0.18	0.00	0.11	0.20	0.32	0.12
10	0.00	0.03	0.00	0.02	0.05	0.09	0.04	0.00	0.20	0.00	0.12	0.25	0.40	0.16
Mean	0.00	0.01	0.00	0.01	0.01	0.03	0.01	0.00	0.18	0.00	0.10	0.21	0.34	0.13
SD	0.00	0.02	0.00	0.01	0.02	0.04	0.01	0.00	0.01	0.00	0.01	0.02	0.03	0.01

Note. Res = Baseline residual, Het = heteroscedastic errors, Load = factor loadings, t3= threshold 3, t4 = threshold 4, Int = intercept

Table 14. Bias and RMSD for Item Parameter Estimates from Constrained HSGRM for 3-Category Response Conditions

N=2000			Bias					RMSD		
Item	Skew	Res	Het	Load	Int	Skew	Res	Het	Load	Int
1	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.03
2	0.00	0.01	0.00	0.00	0.00	0.00	0.06	0.00	0.03	0.03
3	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.04	0.03
4	0.00	0.00	0.00	0.01	0.00	0.00	0.06	0.00	0.04	0.03
5	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.03
6	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.03
7	0.00	0.00	0.00	0.00	-0.01	0.00	0.05	0.00	0.03	0.03
8	0.00	0.01	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.03
9	0.00	0.01	0.00	0.01	0.00	0.00	0.05	0.00	0.04	0.03
10	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.03
Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.03
SD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N=500										
1	0.00	0.00	0.00	0.01	0.00	0.00	0.11	0.00	0.07	0.06
2	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.07	0.06
3	0.00	-0.01	0.00	-0.01	0.00	0.00	0.10	0.00	0.07	0.06
4	0.00	0.01	0.00	0.00	0.00	0.00	0.11	0.00	0.07	0.06
5	0.00	0.00	0.00	0.00	0.01	0.00	0.11	0.00	0.06	0.06
6	0.00	0.01	0.00	0.01	0.00	0.00	0.11	0.00	0.07	0.06
7	0.00	-0.02	0.00	0.00	0.00	0.00	0.10	0.00	0.06	0.06
8	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.07	0.06
9	0.00	0	0.00	0.00	0.00	0.00	0.11	0.00	0.07	0.06
10	0.00	-0.01	0.00	0.00	0.00	0.00	0.10	0.00	0.07	0.06
Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.07	0.06
SD	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note. Res = Baseline residual, Het = heteroscedastic errors, Load = factor loadings, Int = intercept

Table 15. Bias and RMSD for Item Parameter Estimates from Full HSGRM for 5-Category Response Conditions

	Bias							RMSD						
N=2000	Skew	Res	Het	Load	T3	T4	Int	Skew	Res	Het	Load	T3	T4	Int
1	-0.01	0.00	0.01	0.00	0.01	0.02	0.00	0.36	0.10	0.14	0.06	0.11	0.22	0.07
2	-0.01	0.01	0.01	0.00	0.01	0.03	0.01	0.36	0.11	0.13	0.06	0.12	0.23	0.08
3	-0.01	0.00	-0.02	-0.01	-0.01	-0.02	-0.01	0.36	0.10	0.12	0.06	0.12	0.22	0.07
4	-0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.36	0.11	0.15	0.05	0.11	0.22	0.07
5	-0.01	0.02	0.02	0.00	0.02	0.04	0.01	0.36	0.10	0.14	0.05	0.11	0.22	0.08
6	-0.01	0.00	-0.01	0.00	0.01	0.01	0.00	0.36	0.11	0.15	0.06	0.12	0.24	0.08
7	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	0.36	0.10	0.15	0.06	0.11	0.23	0.07
8	-0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.36	0.10	0.14	0.06	0.12	0.22	0.07
9	-0.01	0.01	0.01	0.00	0.01	0.02	0.01	0.36	0.10	0.13	0.06	0.12	0.23	0.07
10	-0.01	0.00	-0.02	0.00	0.00	-0.01	0.00	0.36	0.11	0.15	0.06	0.12	0.23	0.08
Mean	-0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.36	0.10	0.14	0.06	0.12	0.23	0.07
SD	0.00	0.01	0.01	0.00	0.01	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00
N=500														
1	0.05	0.00	-0.01	-0.01	-0.01	-0.01	0.00	0.53	0.22	0.29	0.11	0.24	0.49	0.15
2	0.05	0.04	0.01	0.02	0.05	0.08	0.04	0.53	0.21	0.33	0.10	0.22	0.44	0.15
3	0.05	0.06	0.02	0.04	0.07	0.14	0.04	0.53	0.24	0.35	0.12	0.26	0.48	0.17
4	0.05	0.02	-0.04	0.02	0.02	0.05	0.02	0.53	0.23	0.32	0.12	0.25	0.51	0.17
5	0.05	0.02	-0.02	0.00	0.02	0.01	0.01	0.53	0.24	0.29	0.13	0.27	0.50	0.17
6	0.05	0.02	0.02	0.01	-0.01	-0.01	0.00	0.53	0.24	0.32	0.12	0.24	0.46	0.16
7	0.05	0.05	-0.01	0.02	0.05	0.09	0.03	0.53	0.23	0.31	0.11	0.25	0.49	0.16
8	0.05	0.04	0.04	0.03	0.04	0.11	0.03	0.53	0.25	0.28	0.12	0.26	0.51	0.17
9	0.05	0.08	0.05	0.03	0.06	0.13	0.03	0.53	0.27	0.30	0.13	0.27	0.51	0.17
10	0.05	0.07	0.01	0.03	0.08	0.15	0.05	0.53	0.28	0.33	0.15	0.31	0.59	0.19
Mean	0.05	0.04	0.01	0.02	0.03	0.07	0.03	0.53	0.24	0.31	0.12	0.26	0.50	0.17
SD	0.00	0.02	0.03	0.01	0.03	0.06	0.02	0.00	0.02	0.02	0.01	0.02	0.04	0.01

Note. Res = Baseline residual, Het = heteroscedastic errors, Load = factor loadings, T3= threshold 3, T4 = threshold 4, Int = intercept

Table 16. Bias and RMSD for Item Parameter Estimates from Full HSGRM for 3-Category Response Conditions

	Bias					RMSD				
N=2000	Skew	Res	Het	Load	Int	Skew	Res	Het	Load	Int
1	-0.01	0.00	0.01	0.00	0.00	0.44	0.05	0.19	0.03	0.03
2	-0.01	0.01	0.00	0.00	0.00	0.44	0.06	0.17	0.03	0.03
3	-0.01	0.00	0.00	0.00	0.00	0.44	0.06	0.17	0.04	0.03
4	-0.01	0.00	0.00	0.00	0.00	0.44	0.06	0.18	0.04	0.04
5	-0.01	0.00	0.01	0.00	0.00	0.44	0.05	0.18	0.03	0.03
6	-0.01	0.00	0.00	0.00	0.00	0.44	0.05	0.18	0.03	0.04
7	-0.01	0.00	-0.02	0.00	-0.01	0.44	0.05	0.18	0.03	0.03
8	-0.01	0.01	0.02	0.00	0.00	0.44	0.05	0.17	0.03	0.03
9	-0.01	0.01	-0.01	0.00	0.00	0.44	0.05	0.17	0.04	0.03
10	-0.01	0.00	0.01	0.00	0.00	0.44	0.05	0.17	0.03	0.03
Mean	-0.01	0.00	0.00	0.00	0.00	0.44	0.05	0.18	0.03	0.03
SD	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
N=500										
1	-0.02	0.00	0.03	0.00	0.00	0.84	0.11	0.40	0.07	0.06
2	-0.02	0.00	-0.06	0.00	0.00	0.84	0.11	0.37	0.07	0.07
3	-0.02	-0.01	0.00	-0.01	0.00	0.84	0.10	0.41	0.07	0.07
4	-0.02	0.01	0.01	0.00	0.00	0.84	0.11	0.38	0.07	0.07
5	-0.02	0.00	-0.03	0.00	0.00	0.84	0.11	0.39	0.07	0.07
6	-0.02	0.00	0.01	0.01	0.00	0.84	0.11	0.37	0.07	0.07
7	-0.02	-0.02	0.01	0.00	0.00	0.84	0.10	0.37	0.06	0.06
8	-0.02	0.00	0.03	0.00	0.00	0.84	0.11	0.42	0.07	0.07
9	-0.02	0.01	0.00	0.00	0.00	0.84	0.11	0.40	0.07	0.07
10	-0.02	-0.01	0.05	0.00	0.01	0.84	0.10	0.36	0.07	0.07
Mean	-0.02	0.00	0.01	0.00	0.00	0.84	0.11	0.39	0.07	0.07
SD	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.02	0.00	0.00

Note. Res = Baseline residual, Het = heteroscedastic errors, Load = factor loadings, T3= threshold 3, T4 = threshold 4, Int = intercept

Table 17. Bias and RMSD Between True Population Value and Correct Model in 5-Category Conditions

	Bias							RMSE						
	Skew	Res	Het	Load	T3	T4	Int	Skew	Res	Het	Load	T3	T4	Int
h0s0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.05	0.10	0.16	0.06
h0s5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.08	0.00	0.05	0.10	0.16	0.06
h0s75	-0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.39	0.08	0.00	0.05	0.10	0.17	0.06
h0s1	0.00	0.02	0.00	0.03	0.02	0.04	-0.02	0.00	0.08	0.00	0.06	0.10	0.17	0.06
h4s0	0.00	0.01	0.00	0.00	0.01	0.02	0.00	0.00	0.11	0.14	0.06	0.12	0.23	0.08
h4s5	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.22	0.11	0.15	0.06	0.12	0.24	0.08
h4s75	-0.04	0.00	-0.02	-0.01	0.00	-0.01	-0.01	0.40	0.11	0.16	0.07	0.12	0.24	0.08
h4s1	0.00	0.01	-0.01	0.02	0.01	0.01	-0.03	0.00	0.11	0.17	0.08	0.12	0.25	0.08
h8s0	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.11	0.13	0.05	0.12	0.23	0.07
h8s5	-0.06	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	0.26	0.12	0.15	0.06	0.13	0.24	0.08
h8s75	-0.23	-0.02	-0.04	-0.03	-0.03	-0.06	-0.02	0.51	0.12	0.16	0.07	0.13	0.25	0.08
h8s1	0.00	-0.02	-0.01	0.02	0.00	-0.01	-0.05	0.00	0.13	0.20	0.08	0.14	0.27	0.09
h1s0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.13	0.05	0.12	0.22	0.07
h1s5	-0.22	-0.02	-0.04	-0.02	-0.02	-0.05	-0.01	0.38	0.12	0.15	0.06	0.13	0.24	0.07
h1s75	-0.32	-0.04	-0.06	-0.04	-0.04	-0.08	-0.03	0.55	0.13	0.16	0.07	0.13	0.25	0.08
h1s1	0.00	-0.02	0.01	0.02	0.00	-0.01	-0.05	0.00	0.14	0.22	0.09	0.15	0.29	0.09
N=500														
h0s0	0.00	0.01	0.00	0.01	0.01	0.03	0.01	0.00	0.18	0.00	0.10	0.21	0.34	0.13
h0s5	0.05	0.00	0.00	0.00	0.01	0.02	0.01	0.55	0.17	0.00	0.10	0.20	0.34	0.13
h0s75	0.24	0.01	0.00	0.00	0.01	0.02	0.00	0.91	0.17	0.00	0.10	0.20	0.34	0.13
h0s1	0.00	0.03	0.00	0.04	0.03	0.07	-0.01	0.00	0.17	0.00	0.11	0.20	0.35	0.12
h4s0	0.00	0.02	0.00	0.00	0.01	0.03	0.01	0.00	0.25	0.30	0.12	0.26	0.49	0.16
h4s5	-0.03	0.02	0.00	0.00	0.02	0.04	0.02	0.59	0.25	0.32	0.13	0.26	0.51	0.16
h4s75	0.00	0.02	-0.01	0.00	0.02	0.03	0.01	0.86	0.25	0.33	0.14	0.26	0.52	0.16
h4s1	0.00	0.03	0.01	0.03	0.03	0.06	-0.02	0.00	0.26	0.37	0.16	0.27	0.54	0.16
h8s0	0.00	0.01	0.00	0.00	-0.01	-0.01	0.00	0.00	0.23	0.29	0.11	0.24	0.45	0.15
h8s5	-0.21	-0.01	-0.02	-0.02	-0.02	-0.04	0.00	0.61	0.25	0.32	0.13	0.27	0.51	0.16
h8s75	-0.16	0.01	-0.02	-0.01	0.00	0.00	0.01	0.87	0.27	0.33	0.15	0.28	0.53	0.17
h8s1	0.00	0.01	0.02	0.03	0.02	0.02	-0.04	0.00	0.30	0.43	0.18	0.32	0.61	0.18
h1s0	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.23	0.28	0.11	0.24	0.44	0.14
h1s5	-0.49	-0.01	-0.02	-0.02	-0.02	-0.04	-0.01	0.90	0.26	0.32	0.13	0.27	0.50	0.16
h1s75	-0.29	-0.02	-0.04	-0.03	-0.03	-0.05	-0.01	0.89	0.26	0.35	0.14	0.28	0.53	0.16
h1s1	0.00	0.01	0.06	0.03	0.03	0.05	-0.04	0.00	0.31	0.48	0.19	0.34	0.63	0.18

Note. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). Baseline residuals denoted = Res, Heteroscedastic error = het, factor loadings = Load, threshold 3 = T3, threshold 4 = T4, intercept = Int.

Table 18. Bias and RMSD for True Population Values against Correct Model Parameters in 3-Category Conditions

	Bias					RMSD				
N=2000	Skew	Res	Het	Load	Int	Skew	Res	Het	Load	Int
h0s0	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.03	0.03
h0s5	0.00	0.00	0.00	0.00	0.00	0.25	0.05	0.00	0.03	0.03
h0s75	-0.02	0.00	0.00	0.00	0.00	0.42	0.05	0.00	0.04	0.03
h0s1	0.00	0.00	0.00	0.02	-0.02	0.00	0.05	0.00	0.04	0.04
h4s0	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.18	0.03	0.03
h4s5	-0.02	0.00	0.00	0.00	0.00	0.25	0.05	0.19	0.03	0.03
h4s75	-0.10	0.00	-0.01	0.00	0.00	0.47	0.05	0.19	0.03	0.03
h4s1	0.00	-0.01	0.00	0.02	-0.03	0.00	0.06	0.21	0.04	0.04
h8s0	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.19	0.03	0.03
h8s5	-0.28	0.00	-0.02	0.00	0.00	0.72	0.06	0.20	0.03	0.03
h8s75	-0.23	0.00	-0.05	-0.01	-0.01	0.50	0.05	0.20	0.04	0.03
h8s1	0.00	-0.02	0.03	0.02	-0.04	0.00	0.06	0.29	0.04	0.05
h1s0	0.00	0.00	0.02	0.00	0.00	0.00	0.06	0.20	0.03	0.03
h1s5	-1.08	-0.01	-0.11	-0.01	-0.01	1.35	0.06	0.23	0.04	0.04
h1s75	-0.30	0.00	-0.09	-0.01	-0.01	0.59	0.06	0.21	0.04	0.03
h1s1	0.00	-0.02	0.09	0.02	-0.05	0.00	0.06	0.39	0.04	0.06
N=500										
h0s0	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.07	0.06
h0s5	0.18	0.00	0.00	0.00	0.01	0.62	0.10	0.00	0.07	0.06
h0s75	0.09	0.00	0.00	0.00	0.00	0.94	0.10	0.00	0.07	0.06
h0s1	0.00	0.00	0.00	0.02	-0.03	0.00	0.11	0.00	0.07	0.07
h4s0	0.00	0.00	0.03	0.00	0.00	0.00	0.11	0.40	0.07	0.07
h4s5	0.11	0.00	0.03	0.00	0.01	0.76	0.11	0.42	0.07	0.07
h4s75	0.17	0.00	0.01	0.00	-0.01	1.18	0.11	0.45	0.07	0.07
h4s1	0.00	-0.01	0.10	0.02	-0.03	0.00	0.11	0.60	0.08	0.07
h8s0	0.00	0.00	0.06	0.00	0.00	0.00	0.11	0.42	0.07	0.07
h8s5	-0.76	-0.01	-0.02	-0.01	0.00	1.25	0.11	0.46	0.07	0.07
h8s75	-0.18	0.00	0.02	-0.01	0.00	1.05	0.11	0.53	0.07	0.07
h8s1	0.00	-0.01	0.22	0.02	-0.04	0.00	0.11	0.81	0.07	0.07
h1s0	0.00	0.00	0.06	0.00	0.00	0.00	0.11	0.43	0.07	0.07
h1s5	-1.26	-0.01	-0.06	-0.01	-0.01	1.54	0.11	0.47	0.07	0.07
h1s75	-0.30	0.00	-0.03	-0.02	-0.01	1.20	0.12	0.49	0.07	0.07
h1s1	0.00	-0.02	0.33	0.02	-0.05	0.00	0.12	0.95	0.07	0.08

Note. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). Baseline residuals denoted = Res, Heteroscedastic error = het, factor loadings = Load, intercept = Int.

Table 19. Bias and RMSD When Fitting HSGRM to Models with Only Skew, Heteroscedastic Errors, or Neither in 5-Category Conditions

	Bias							RMSD						
N=2000	Skew	Res	Het	Load	T3	T4	Int	Skew	Res	Het	Load	T3	T4	Int
h0s0	-0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.36	0.10	0.14	0.06	0.12	0.23	0.07
h0s5	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.25	0.11	0.14	0.06	0.12	0.23	0.07
h0s75	-0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.40	0.10	0.14	0.06	0.11	0.23	0.07
h0s1	0.00	0.02	-0.01	0.03	0.02	0.04	-0.02	0.00	0.10	0.14	0.07	0.11	0.22	0.07
h4s0	-0.01	0.01	0.01	0.00	0.01	0.02	0.00	0.41	0.11	0.14	0.06	0.12	0.23	0.08
h8s0	-0.13	0.01	0.01	0.00	0.00	0.01	0.00	0.30	0.11	0.13	0.05	0.12	0.23	0.08
h1s0	-0.09	0.00	0.00	0.00	0.00	0.01	0.00	0.27	0.11	0.13	0.05	0.12	0.22	0.07
N=500														
h0s0	0.05	0.04	0.01	0.02	0.03	0.07	0.03	0.53	0.24	0.31	0.12	0.26	0.50	0.17
h0s5	0.08	0.03	0.01	0.01	0.03	0.07	0.02	0.62	0.23	0.31	0.13	0.25	0.49	0.16
h0s75	0.31	0.03	0.01	0.01	0.03	0.06	0.01	1.04	0.24	0.32	0.14	0.25	0.50	0.15
h0s1	0.00	0.05	0.00	0.05	0.05	0.10	0.00	0.00	0.23	0.31	0.15	0.24	0.50	0.15
h4s0	-0.11	0.02	0.00	0.00	0.01	0.02	0.01	0.46	0.25	0.30	0.12	0.26	0.49	0.16
h8s0	-0.23	0.01	0.00	0.00	-0.01	-0.01	0.00	0.48	0.24	0.30	0.11	0.24	0.45	0.15
h1s0	-0.36	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.49	0.23	0.28	0.11	0.24	0.44	0.14

Note. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). Baseline residuals denoted = Res, Heteroscedastic error = het, factor loadings = Load, threshold 3 = T3, threshold 4 = T4, intercept = Int.

Table 20. Bias and RMSD When Fitting HSGRM to Models with Only Skew, Het, or Neither in 3-Category Conditions

	Bias					RMSD				
N=2000	Skew	Res	Het	Load	Int	Skew	Res	Het	Load	Int
h0s0	-0.01	0.00	0.00	0.00	0.00	0.44	0.05	0.18	0.03	0.03
h0s5	-0.01	0.00	-0.01	0.00	0.00	0.28	0.05	0.18	0.03	0.03
h0s75	-0.02	0.00	-0.01	0.00	0.00	0.45	0.05	0.18	0.04	0.03
h0s1	0.00	0.00	-0.01	0.02	-0.03	0.00	0.05	0.18	0.04	0.04
h4s0	-0.20	0.00	0.00	0.00	0.00	0.53	0.05	0.18	0.03	0.03
h8s0	-0.23	0.00	0.00	0.00	0.00	0.42	0.05	0.19	0.03	0.03
h1s0	-0.25	0.00	0.02	0.00	0.00	0.42	0.06	0.20	0.03	0.03
N=500										
h0s0	-0.02	0.00	0.01	0.00	0.00	0.84	0.11	0.39	0.07	0.07
h0s5	0.28	0.00	0.02	0.00	0.01	0.75	0.11	0.41	0.07	0.07
h0s75	0.24	0.00	-0.01	0.00	-0.01	1.19	0.10	0.42	0.07	0.07
h0s1	0.00	0.00	0.01	0.01	-0.03	0.00	0.11	0.43	0.08	0.07
h4s0	-0.31	0.00	0.01	0.00	-0.01	0.69	0.11	0.40	0.07	0.07
h8s0	-0.45	0.00	0.05	0.01	-0.01	0.65	0.11	0.44	0.07	0.07
h1s0	-0.39	0.00	0.05	0.00	0.00	0.58	0.11	0.43	0.07	0.06

Note. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). Baseline residuals denoted = Res, Heteroscedastic error = het, factor loadings = Load, intercept = Int.

Table 21. Percentage of Time More Parameterized Model Chosen

	5 Categories						3 Categories					
N=2,000	BvS	BvH	BvHS	SvH	SvHS	HvHS	BvS	BvH	BvHS	SvH	SvHS	HvHS
h0s0	0.02	0.07	0.06	0.08	0.07	0.01	0.03	0.04	0.04	0.08	0.05	0.02
h0s5	1.00	0.21	1.00	0.00	0.07	1.00	1.00	0.53	1.00	0.01	0.06	1.00
h0s75	1.00	0.51	1.00	0.00	0.07	1.00	1.00	0.91	1.00	0.00	0.04	1.00
h0s1	1.00	0.78	1.00	0.00	0.03	1.00	1.00	0.99	1.00	0.00	0.05	1.00
h4s0	0.19	1.00	1.00	1.00	1.00	0.03	0.32	1.00	1.00	1.00	1.00	0.03
h8s0	0.18	1.00	1.00	1.00	1.00	0.01	0.50	1.00	1.00	1.00	1.00	0.01
h1s0	0.14	1.00	1.00	1.00	1.00	0.11	0.47	1.00	1.00	1.00	1.00	0.03
h4s5	1.00	0.98	1.00	0.45	1.00	1.00	0.94	0.59	1.00	0.25	0.99	1.00
h4s75	1.00	0.91	1.00	0.00	1.00	1.00	1.00	0.19	1.00	0.00	1.00	1.00
h4s1	1.00	0.70	1.00	0.00	1.00	1.00	1.00	0.04	1.00	0.00	1.00	1.00
h8s5	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89
h8s75	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00
h8s1	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.97	1.00	0.00	1.00	1.00
h1s5	0.75	1.00	1.00	1.00	1.00	0.99	0.10	1.00	1.00	1.00	1.00	0.95
h1s75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
h1s1	1.00	1.00	1.00	0.23	1.00	1.00	1.00	1.00	1.00	0.11	1.00	1.00
N=500												
h0s0	0.00	0.13	0.08	0.16	0.12	0.00	0.04	0.08	0.08	0.09	0.09	0.05
h0s5	0.93	0.09	0.65	0.01	0.05	0.91	0.91	0.13	0.61	0.01	0.09	0.88
h0s75	1.00	0.14	1.00	0.00	0.07	1.00	1.00	0.34	0.95	0.01	0.07	0.99
h0s1	1.00	0.22	1.00	0.00	0.04	1.00	1.00	0.44	1.00	0.00	0.05	1.00
h4s0	0.02	0.80	0.76	0.82	0.80	0.02	0.21	0.65	0.60	0.60	0.58	0.03
h8s0	0.07	1.00	1.00	1.00	1.00	0.05	0.49	1.00	1.00	1.00	0.99	0.01
h1s0	0.10	1.00	1.00	1.00	1.00	0.06	0.46	1.00	1.00	1.00	1.00	0.01
h4s5	0.77	0.41	0.90	0.25	0.71	0.92	0.43	0.18	0.66	0.15	0.52	0.83
h4s75	1.00	0.31	1.00	0.02	0.72	1.00	0.94	0.08	0.94	0.02	0.49	1.00
h4s1	1.00	0.23	1.00	0.00	0.70	1.00	0.99	0.04	1.00	0.01	0.56	1.00
h8s5	0.41	1.00	1.00	0.99	1.00	0.94	0.06	0.90	0.98	0.91	0.98	0.62
h8s75	0.99	0.99	1.00	0.76	1.00	1.00	0.71	0.76	0.99	0.52	0.98	1.00
h8s1	1.00	0.94	1.00	0.09	1.00	1.00	0.93	0.38	0.99	0.08	0.99	1.00
h1s5	0.30	1.00	1.00	1.00	1.00	0.86	0.07	1.00	1.00	1.00	1.00	0.43
h1s75	0.98	1.00	1.00	0.97	1.00	1.00	0.58	0.94	1.00	0.79	1.00	0.97
h1s1	1.00	0.99	1.00	0.22	1.00	1.00	0.92	0.67	1.00	0.17	1.00	1.00

Note. Heteroscedasticity denoted “h” and skew is denoted “s” (i.e., h0s0), Skew values are 0.0, 0.5, 0.75, and 1.0 (noted as 0, 5, 75, 1). Heteroscedastic values include 0, 0.4, 0.8, and 1.0 (noted as 0, 4, 8, 1). Model comparisons are Baseline versus Skew-only (BvS), Baseline versus Het-only (BvH), Baseline versus HSGRM (BvHS), Skew-only versus Het-only (SvH), Skew-only versus HSGRM (SvHS), and Het-only versus HSGRM (HvHS).

Table 22. Best fitting model as determined by AIC for 5- and 3-category conditions

5 Categories	Best Fitting			
N=2000	B	S	H	HS
B	0.920	0.030	0.045	0.005
S	0.000	0.970	0.000	0.030
H	0.000	0.000	0.830	0.170
HS	0.000	0.000	0.001	0.999
N=500	B	S	H	HS
B	0.955	0.000	0.045	0.000
S	0.005	0.968	0.000	0.027
H	0.077	0.005	0.805	0.105
HS	0.005	0.122	0.017	0.856
3 Categories				
N=2000	B	S	H	HS
B	0.935	0.035	0.025	0.005
S	0.000	0.970	0.000	0.030
H	0.000	0.002	0.907	0.092
HS	0.000	0.003	0.063	0.934
N=500	B	S	H	HS
B	0.870	0.080	0.035	0.015
S	0.013	0.950	0.002	0.035
H	0.123	0.052	0.765	0.060
HS	0.035	0.170	0.105	0.690

Note. True models are in the far-left column. Baseline GRM is denoted (B), Skew-only denoted (S), het-only denoted (H), and HSGRM denoted (HS)

FIGURES

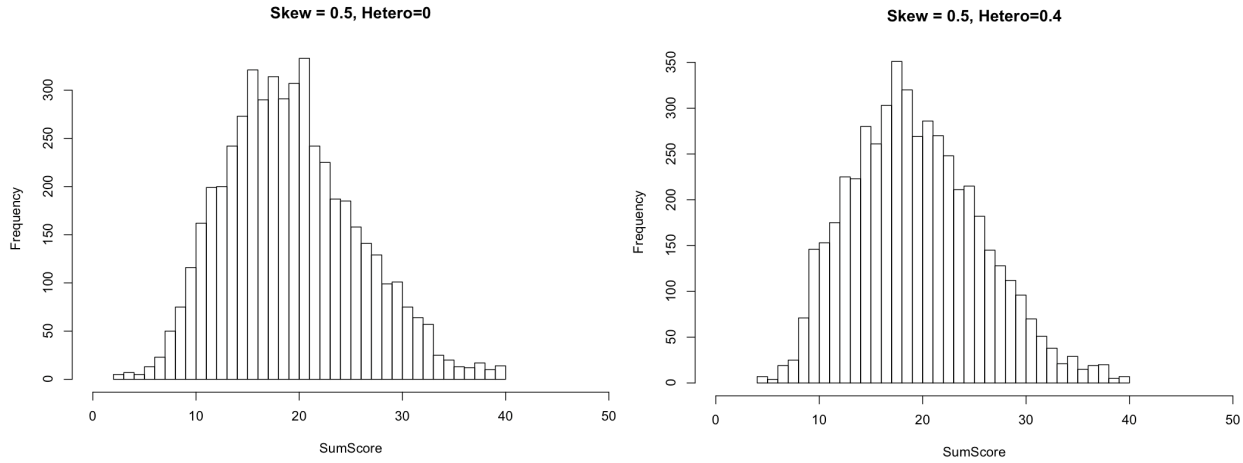
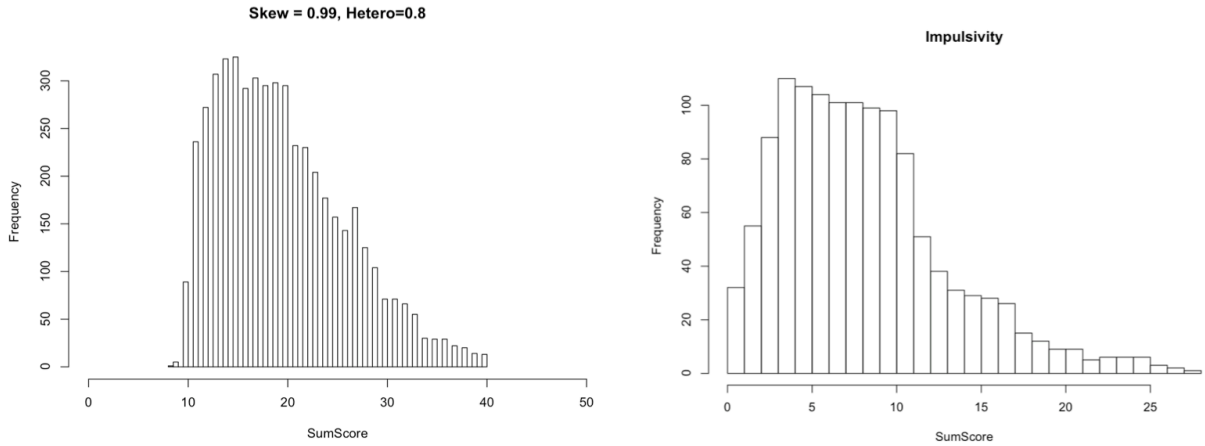


Figure 1a and b. Skew-normal data generated with skew=0 or 0.5 and/or heteroscedasticity of 0 or 0.4.



(a) (b)

Figure 2a and 2b. Simulated data with skew of ~1.0 and heteroscedasticity of 0.80 and real impulsivity data.

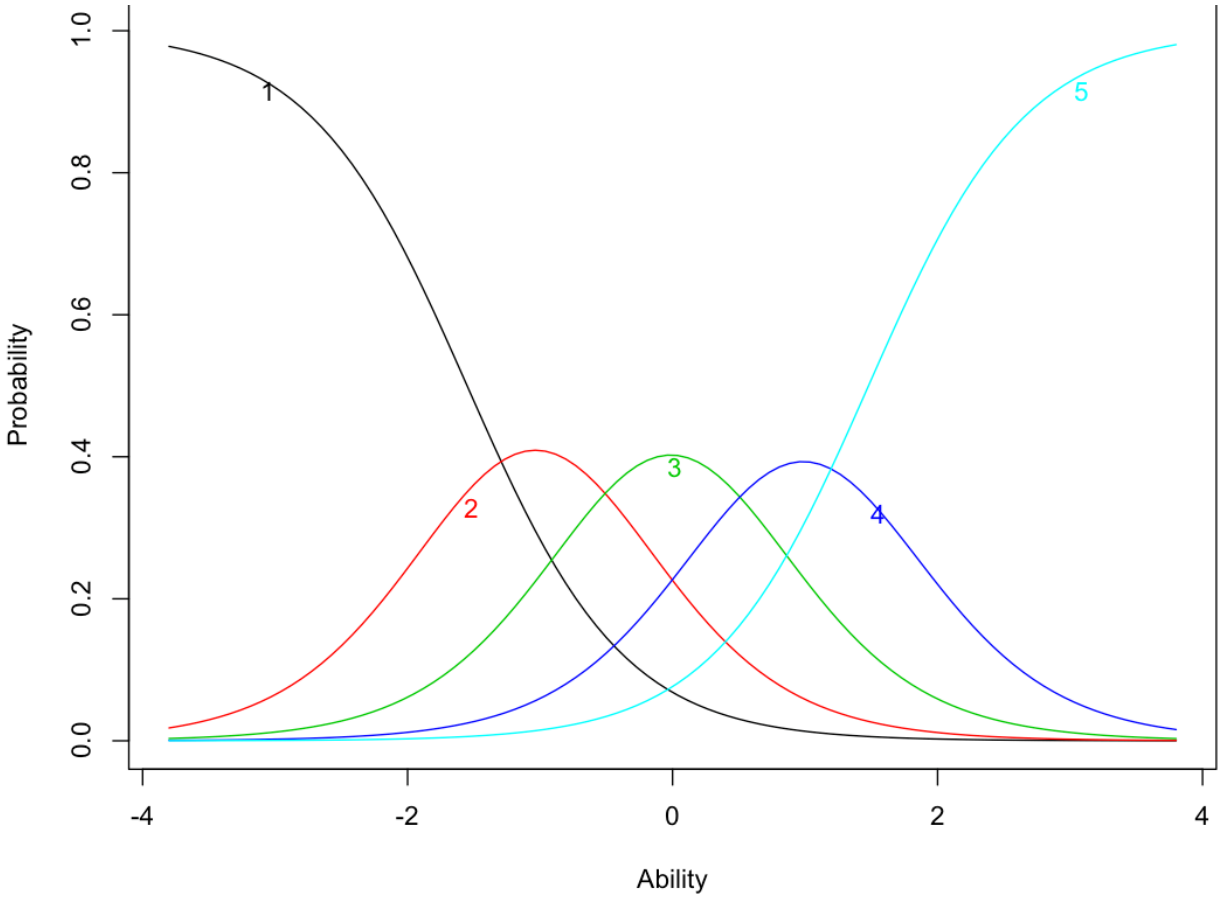


Figure 3. Category Response Curves (CRCs) for a polytomous item with five response category options.

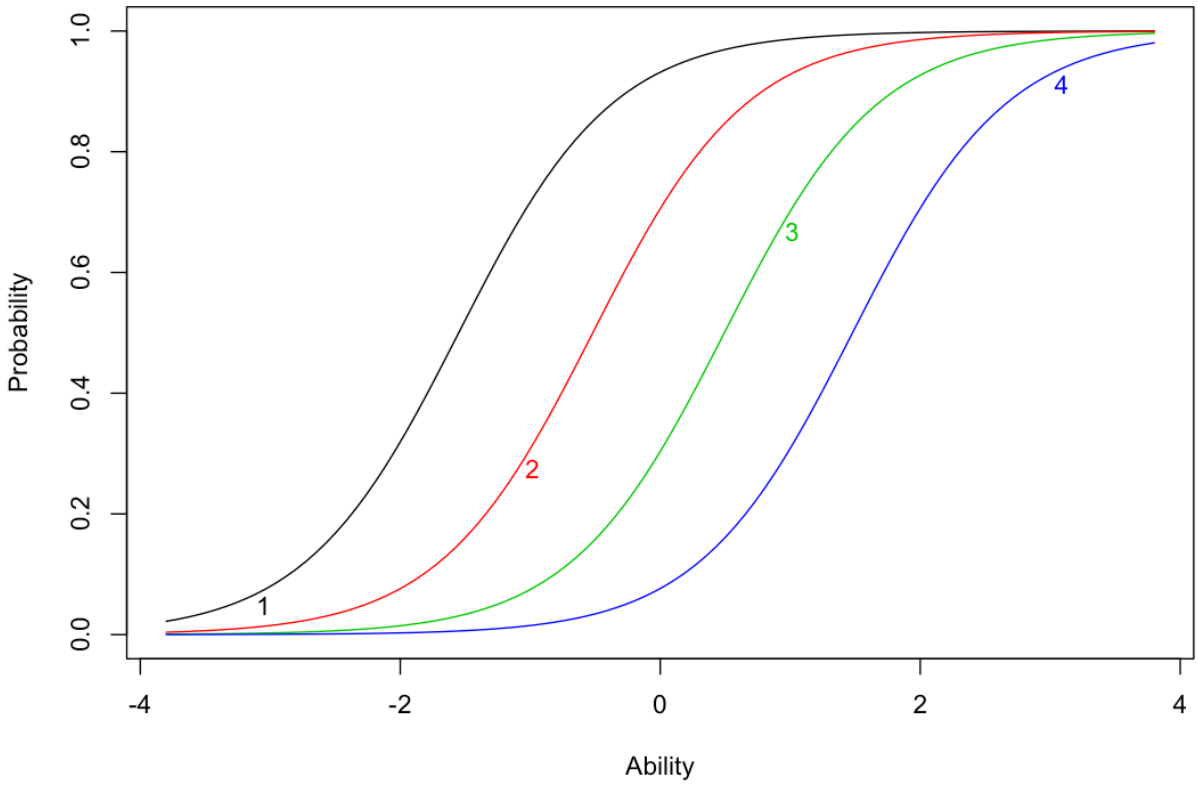


Figure 4. Operating Characteristic Curves (OCCs) for a 5-category item.

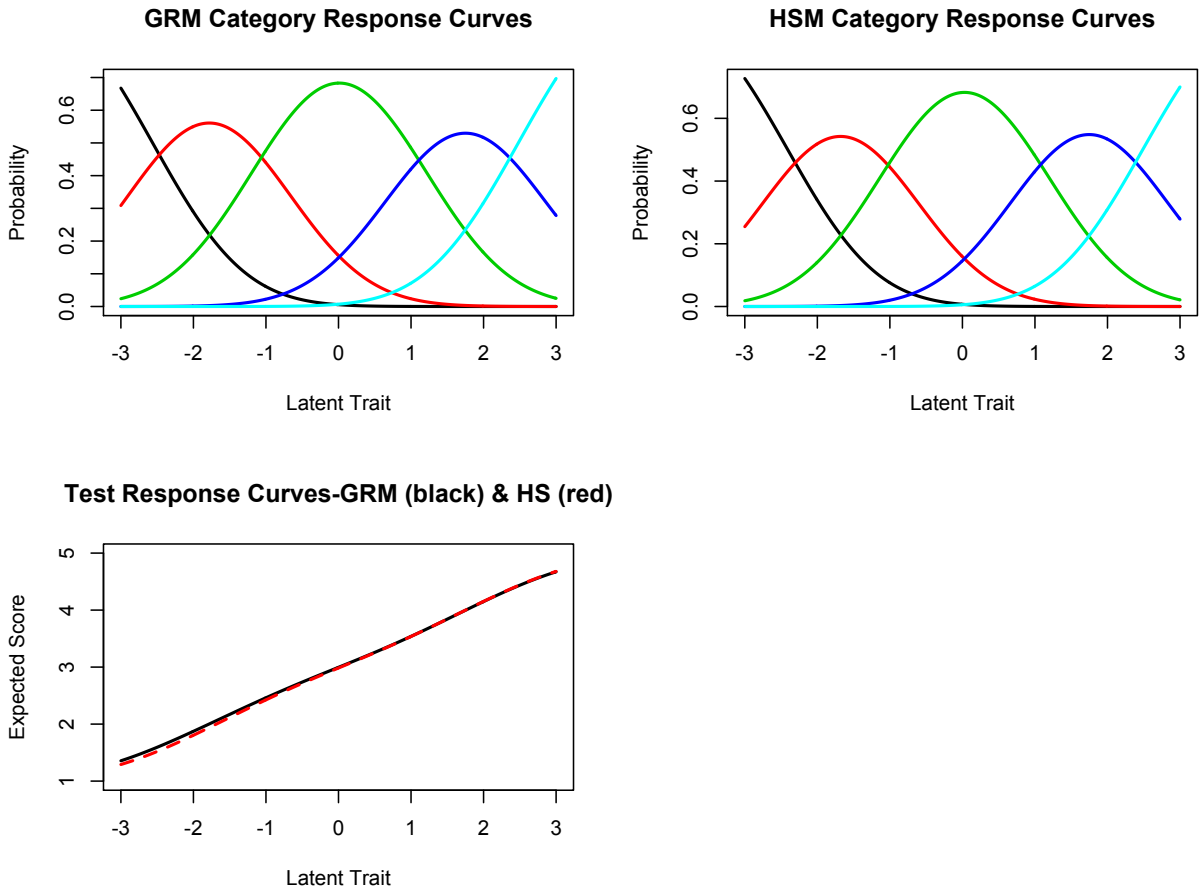


Figure 5. CRCs and TRCs for Baseline GRM and HSGRM for large sample 5-category condition with no heteroscedastic errors and only skew of 1.0

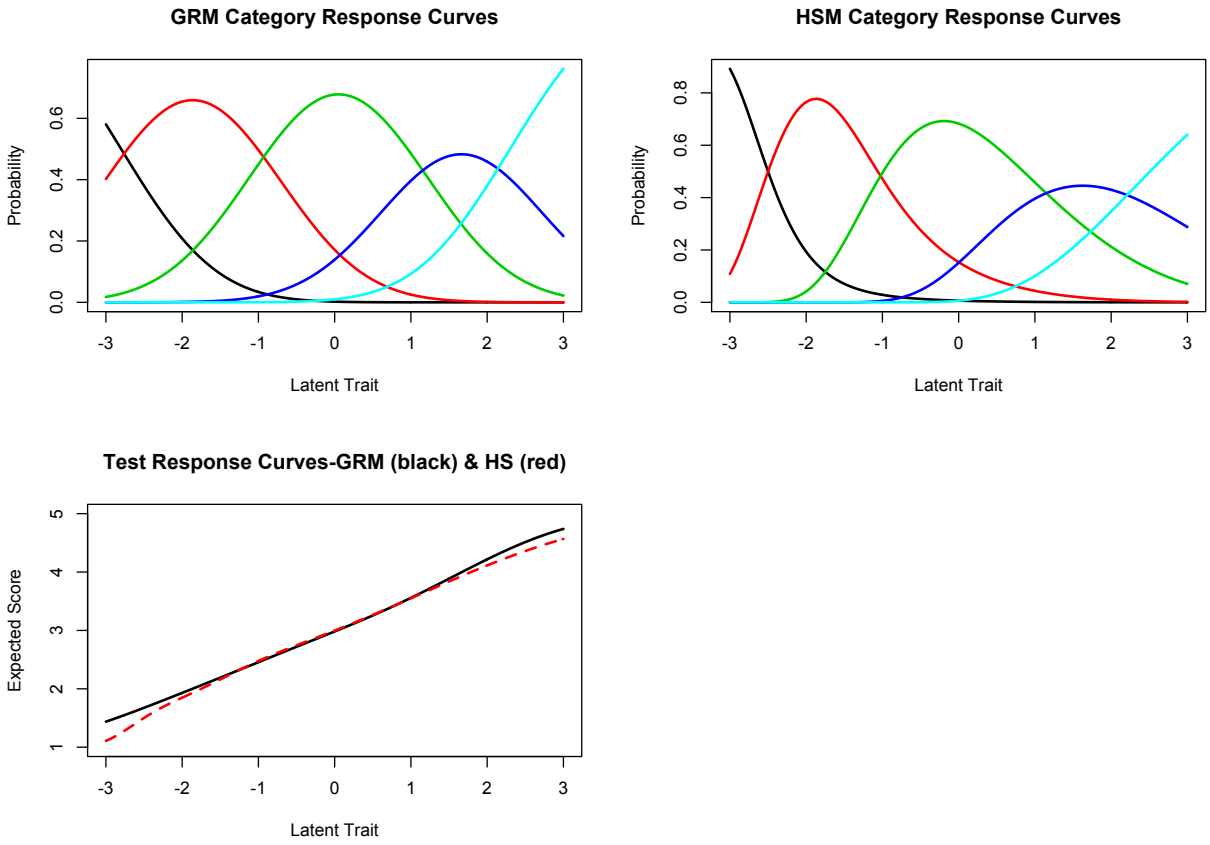


Figure 6. CRCs and TRC for Baseline GRM and HSGRM for large sample with 5 categories with heteroscedasticity of 0.8 and skew of 0

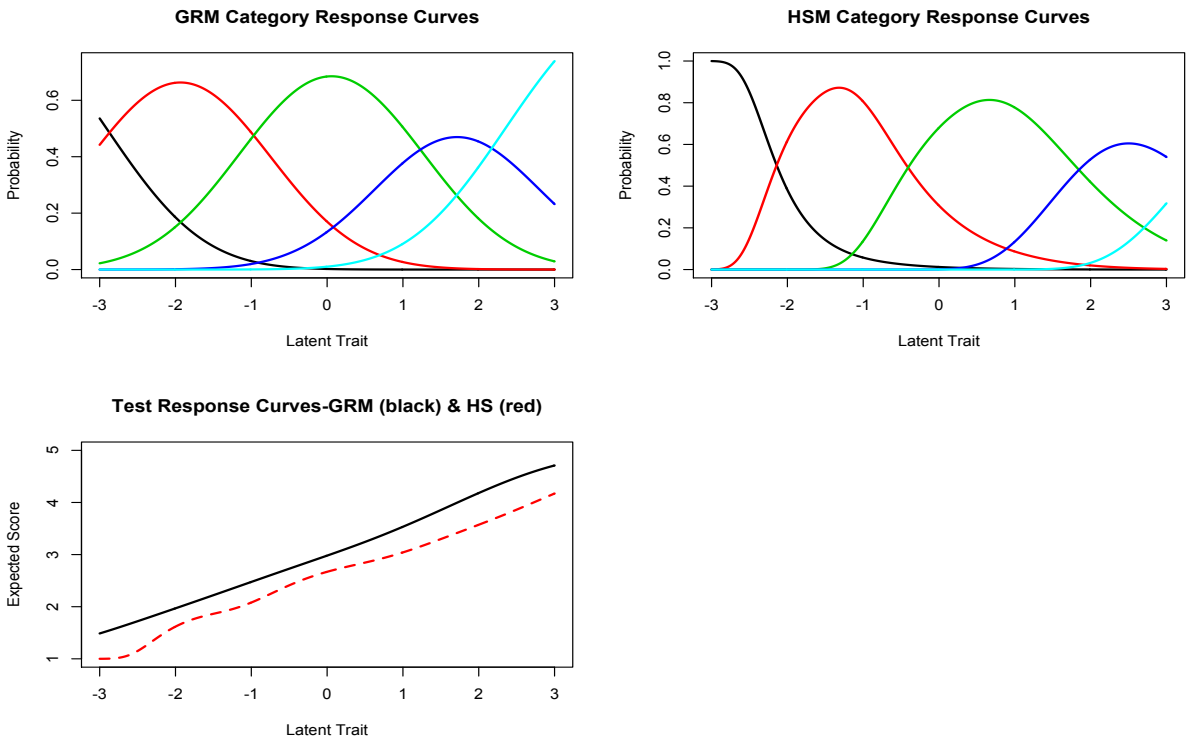


Figure 7. CRCs and TRC for both baseline GRM and HSGRM for large sample with 5 categories with heteroscedastic errors of 0.8 and skew of 0.5

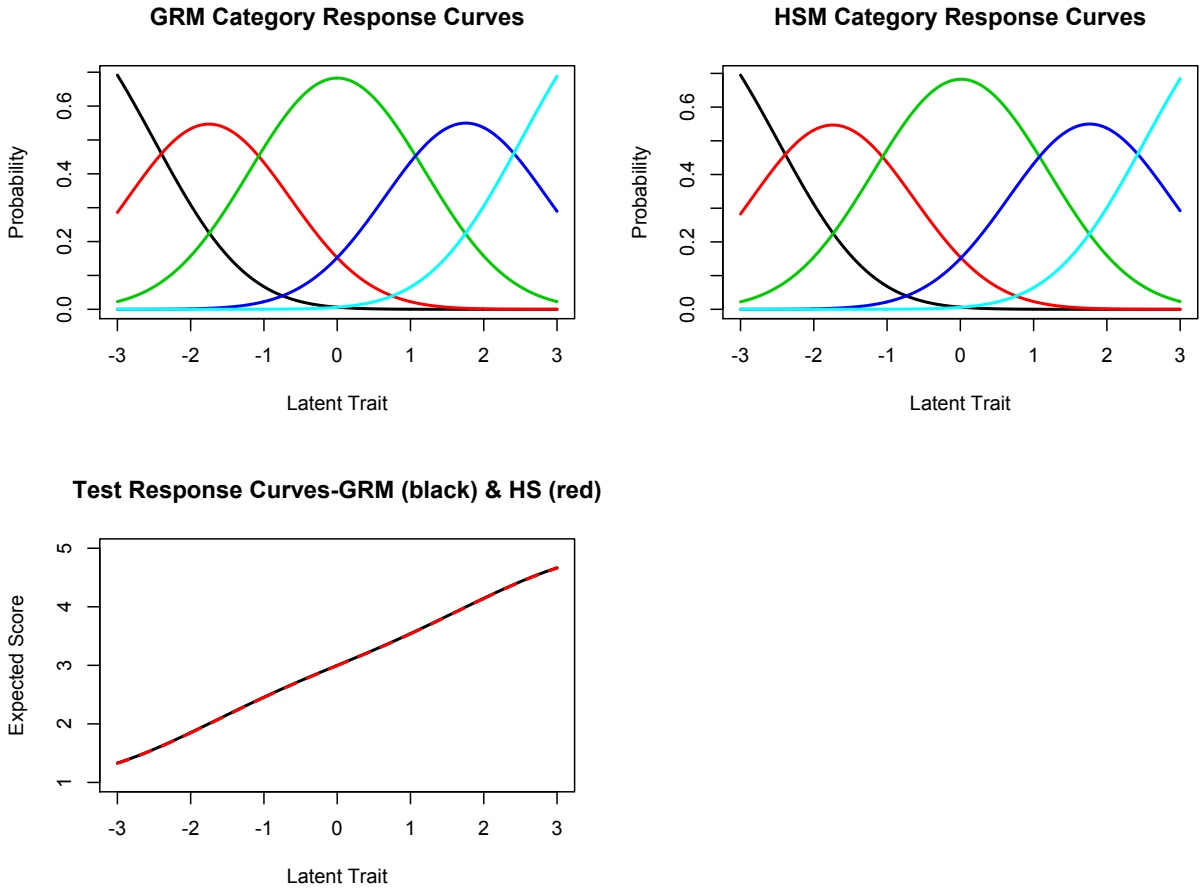


Figure 8. CRCs and TRC for Baseline GRM and full HSGRM for control condition with no skew or heteroscedastic errors in large sample with 5 categories

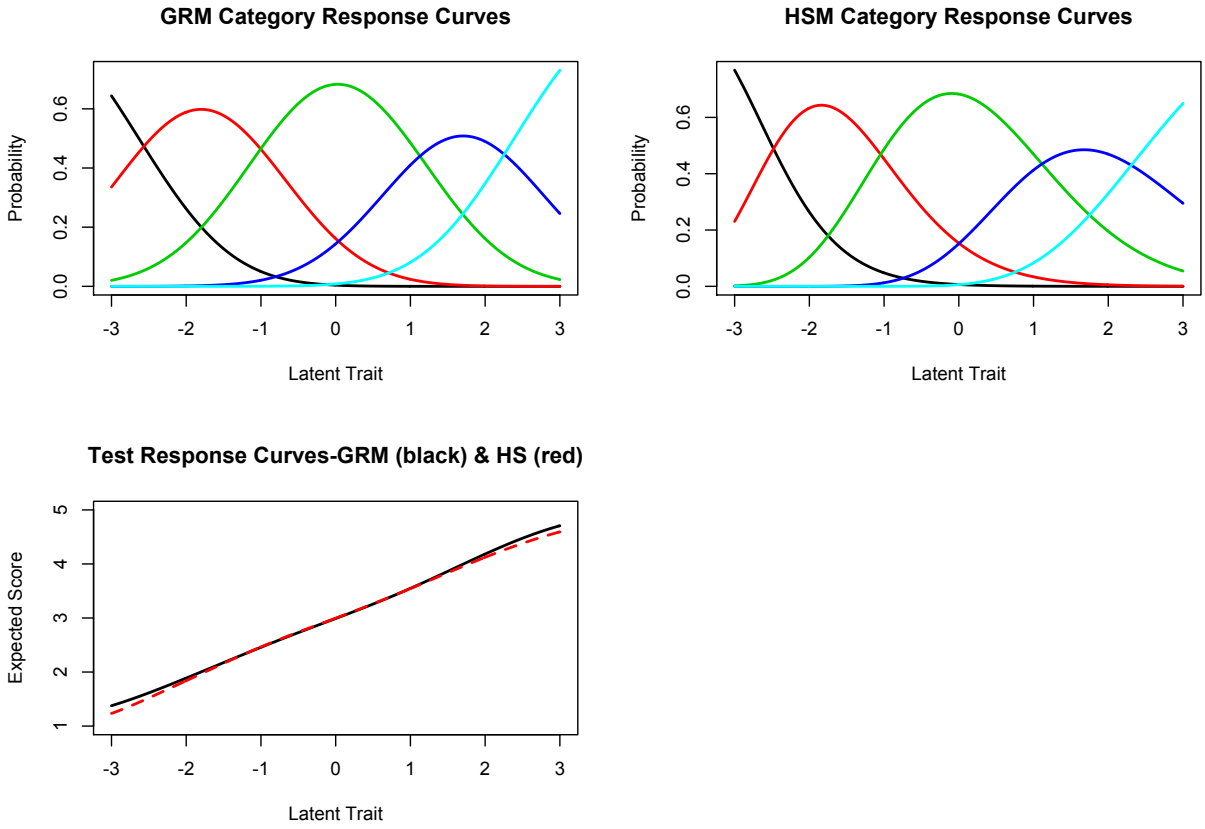


Figure 9. CRCs and TRC for Baseline GRM and HSGRM for large sample with 5 categories with heteroscedasticity of 0.4 and skew of 0.

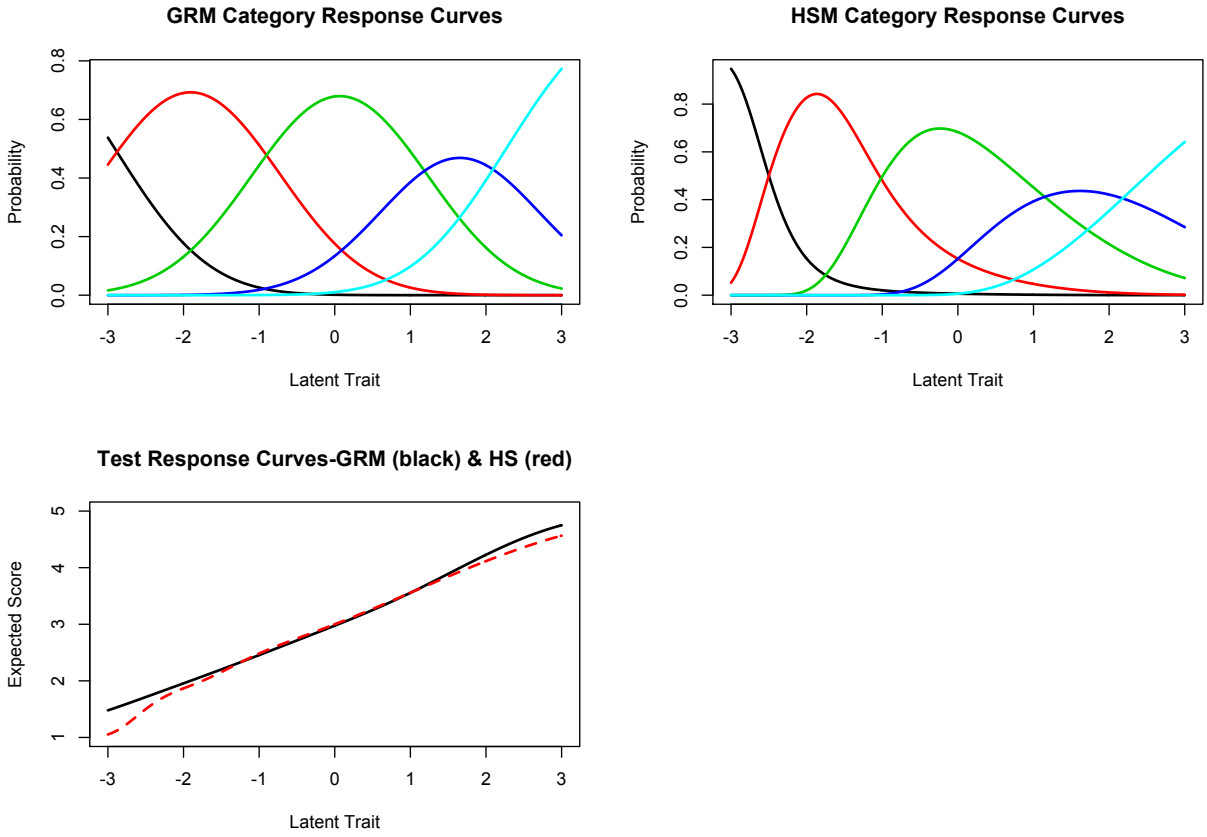


Figure 10. CRCs and TRC for Baseline GRM and HSGRM for large sample with 5 categories with heteroscedasticity of 1 and skew of 0

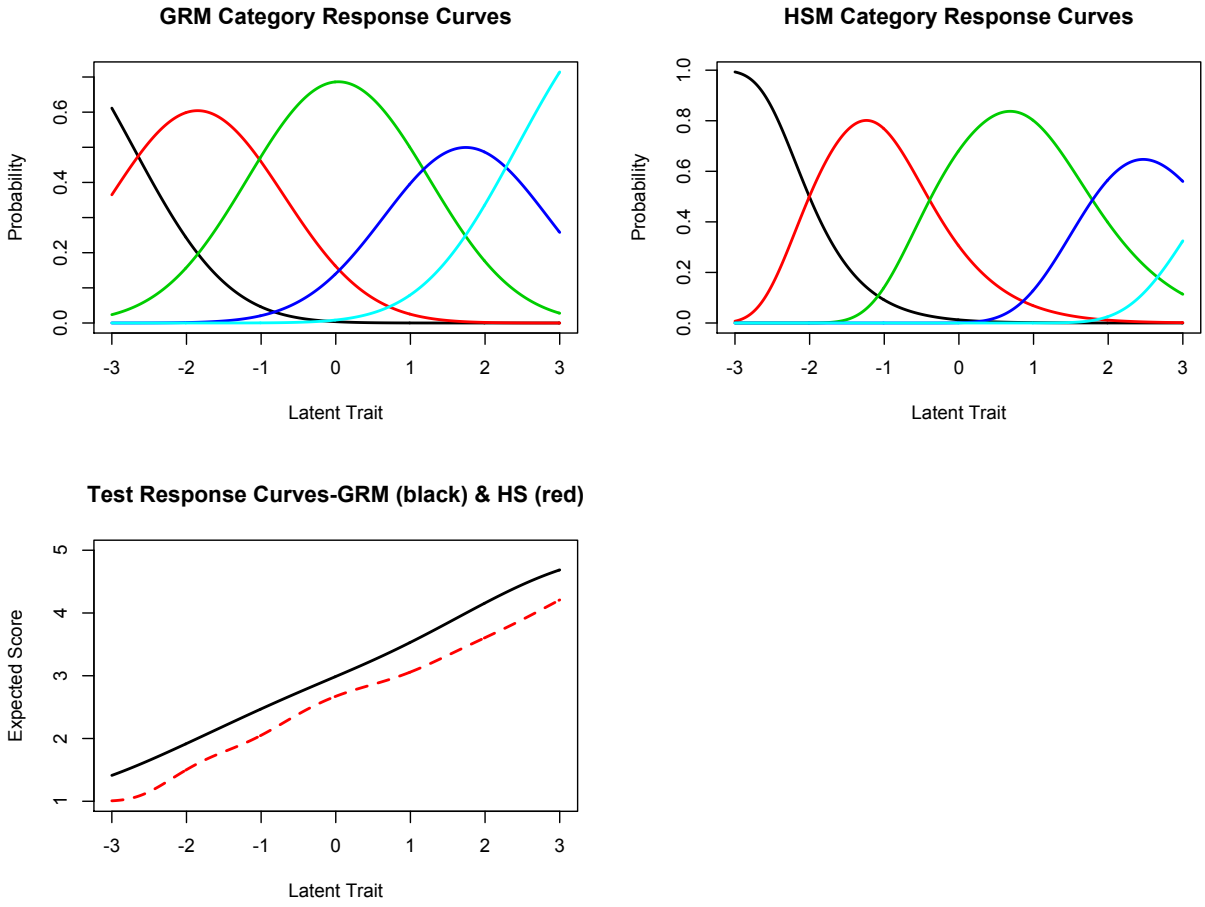


Figure 11. CRCs and TRC for Baseline GRM and HSGRM for large sample with 5 categories with heteroscedasticity of 0.4 and skew of 5

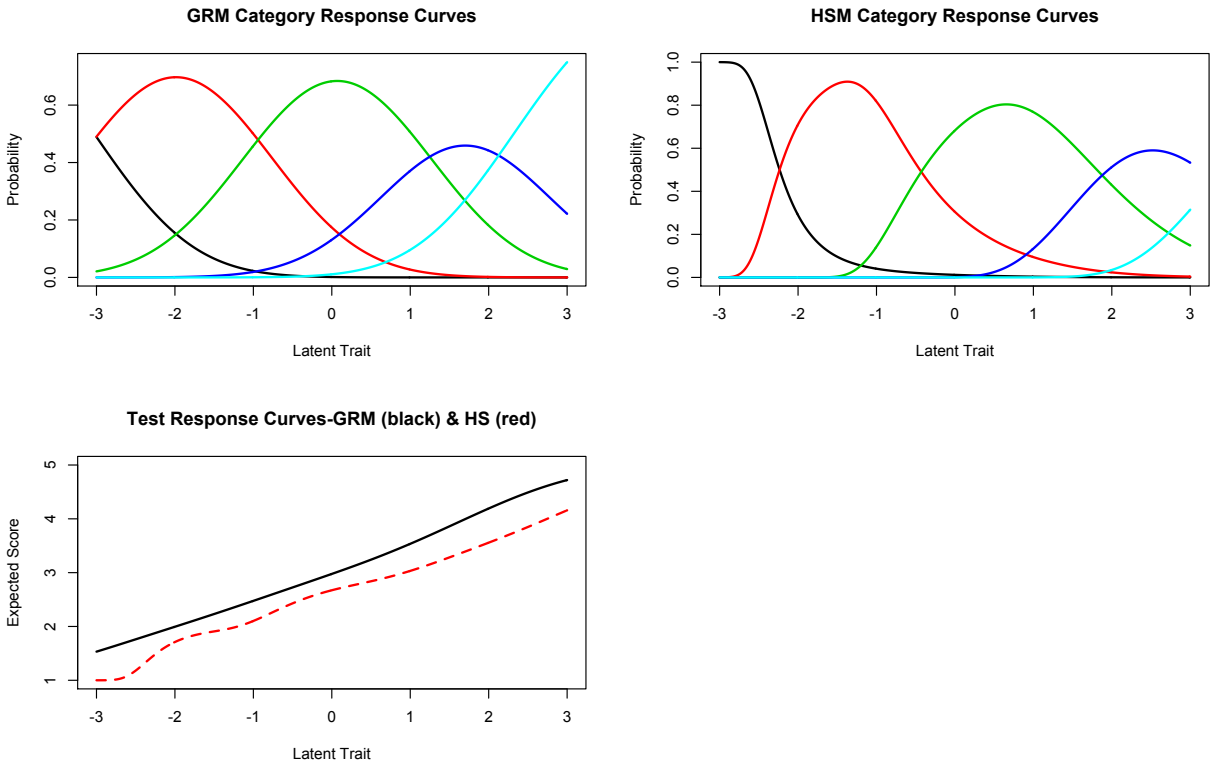


Figure 12. CRCs and TRC for Baseline GRM and HSGRM for large sample with 5 categories with heteroscedasticity of 1 and skew of 5

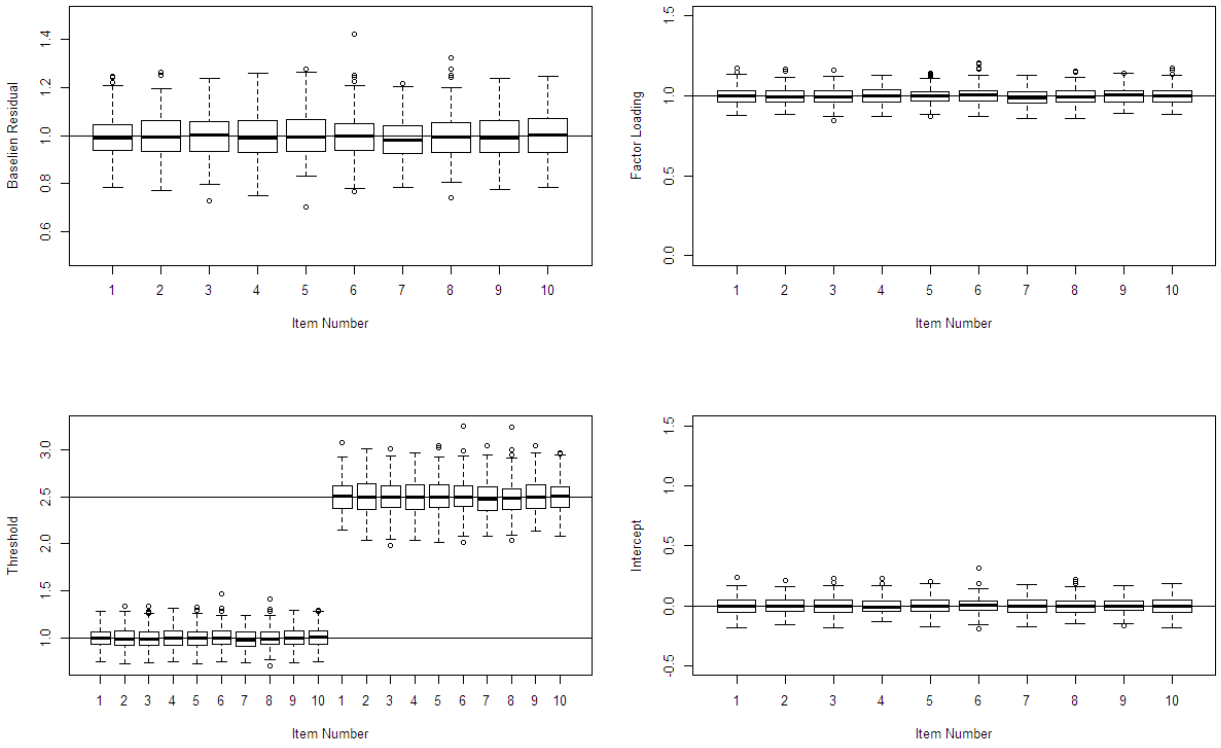


Figure 13. Baseline model boxplots for item parameters in large sample with 5-category response options with no heteroscedasticity or skew. Each plot contains boxplots for each item.

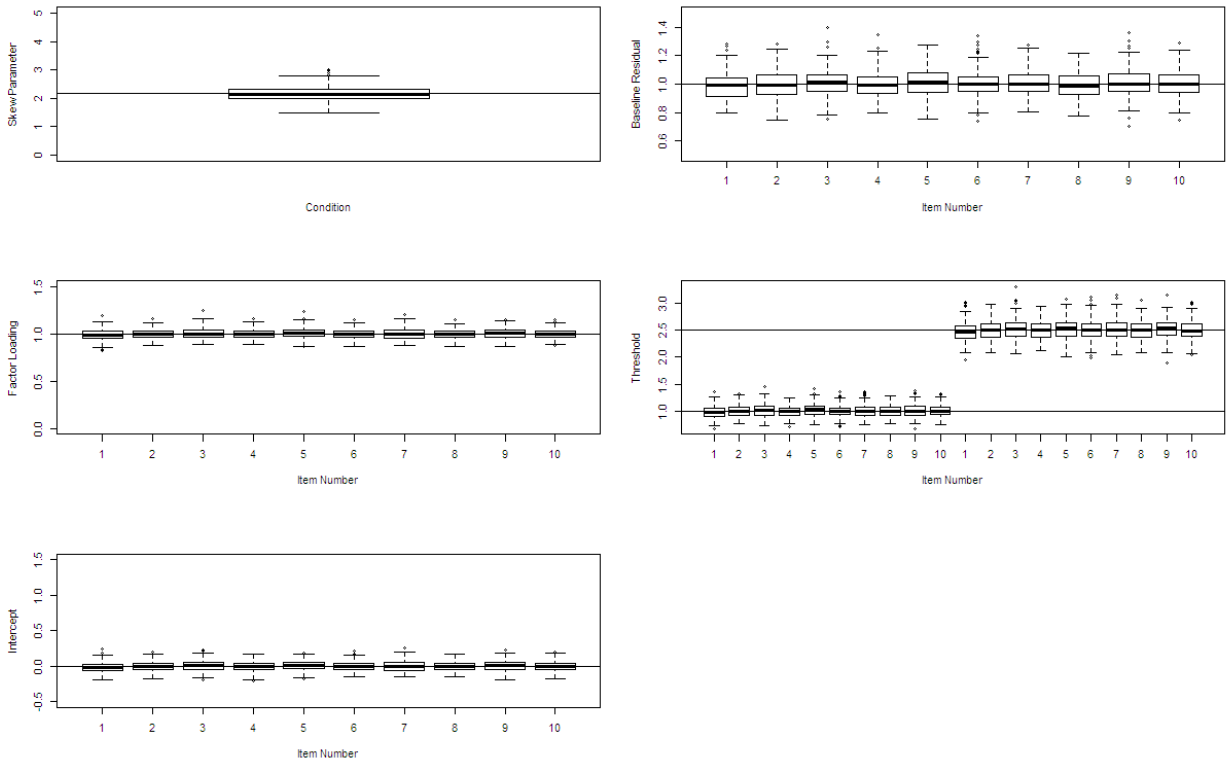


Figure 14. Baseline model boxplots for item parameters in large sample with 5-category response options with no heteroscedasticity and skew of 0.5. Each plot contains boxplots for each item.

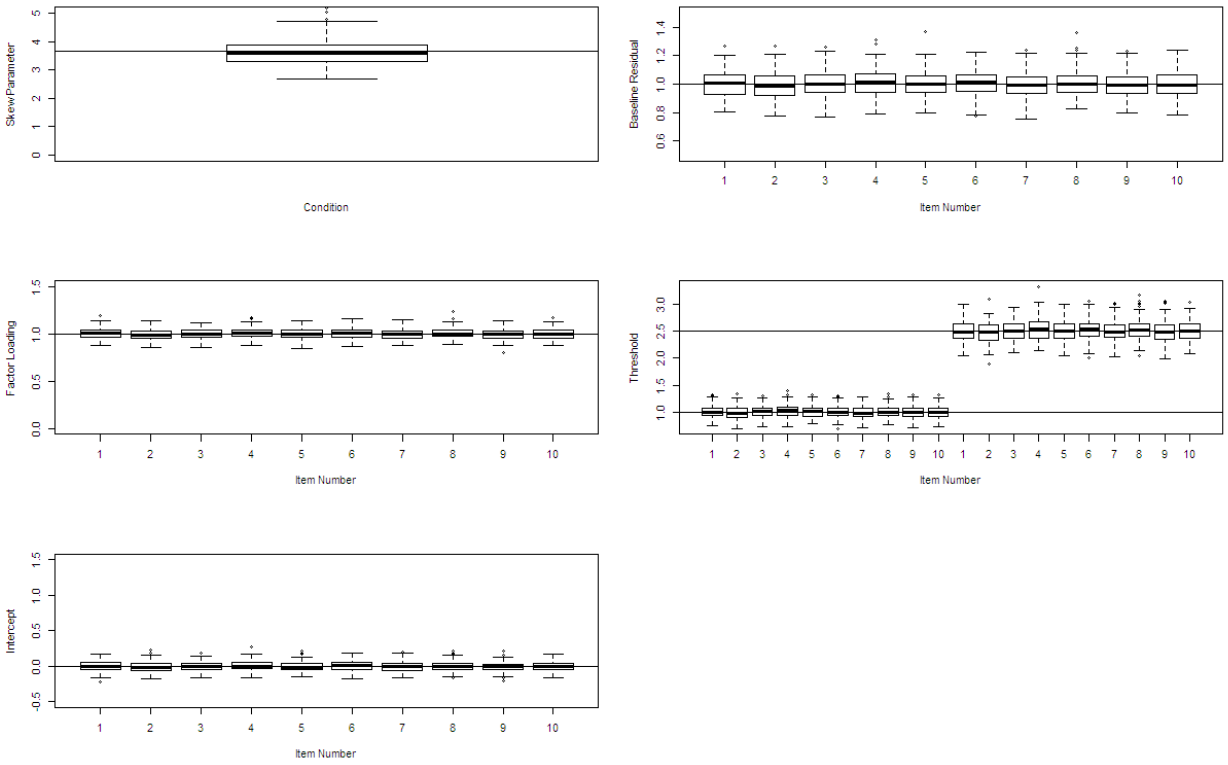


Figure 15. Baseline model boxplots for item parameters in large sample with 5-category response options with no heteroscedasticity and skew of 0.75. Each plot contains boxplots for each item.

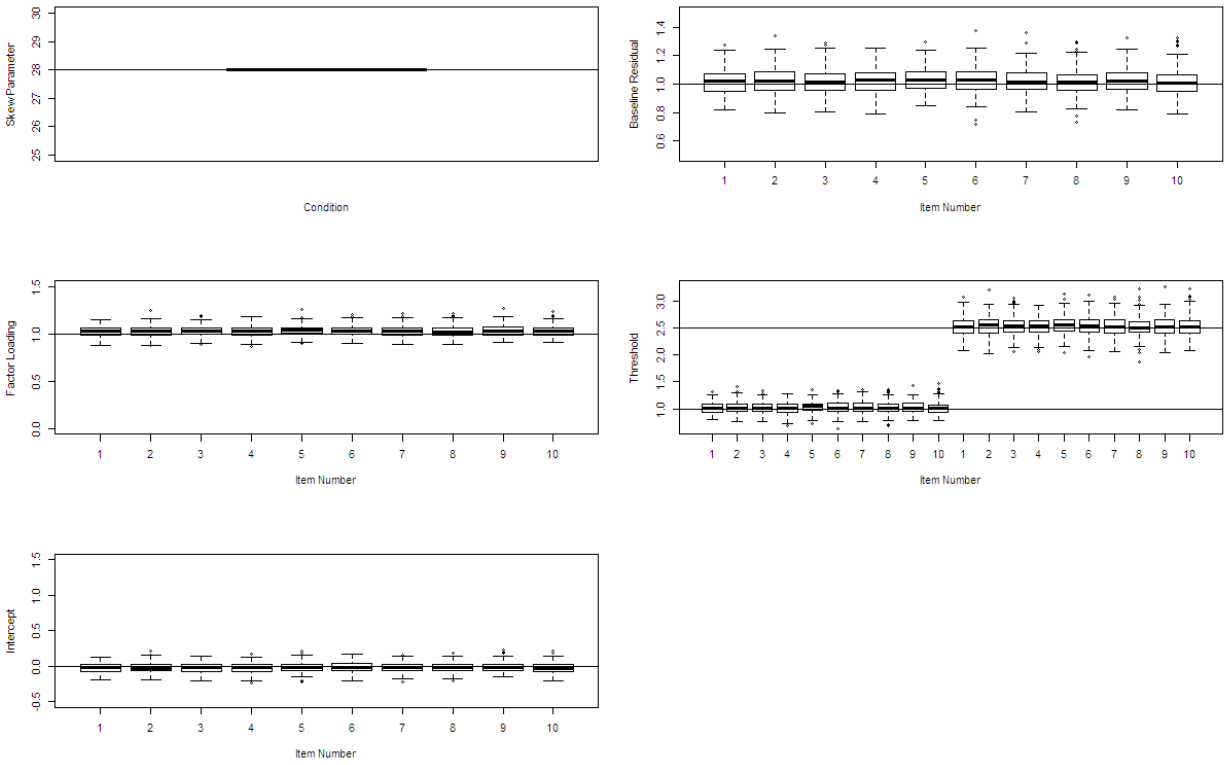


Figure 16. Baseline model boxplots for item parameters in large sample with 5-category response options with no heteroscedasticity and skew of 1.0. Each plot contains boxplots for each item.

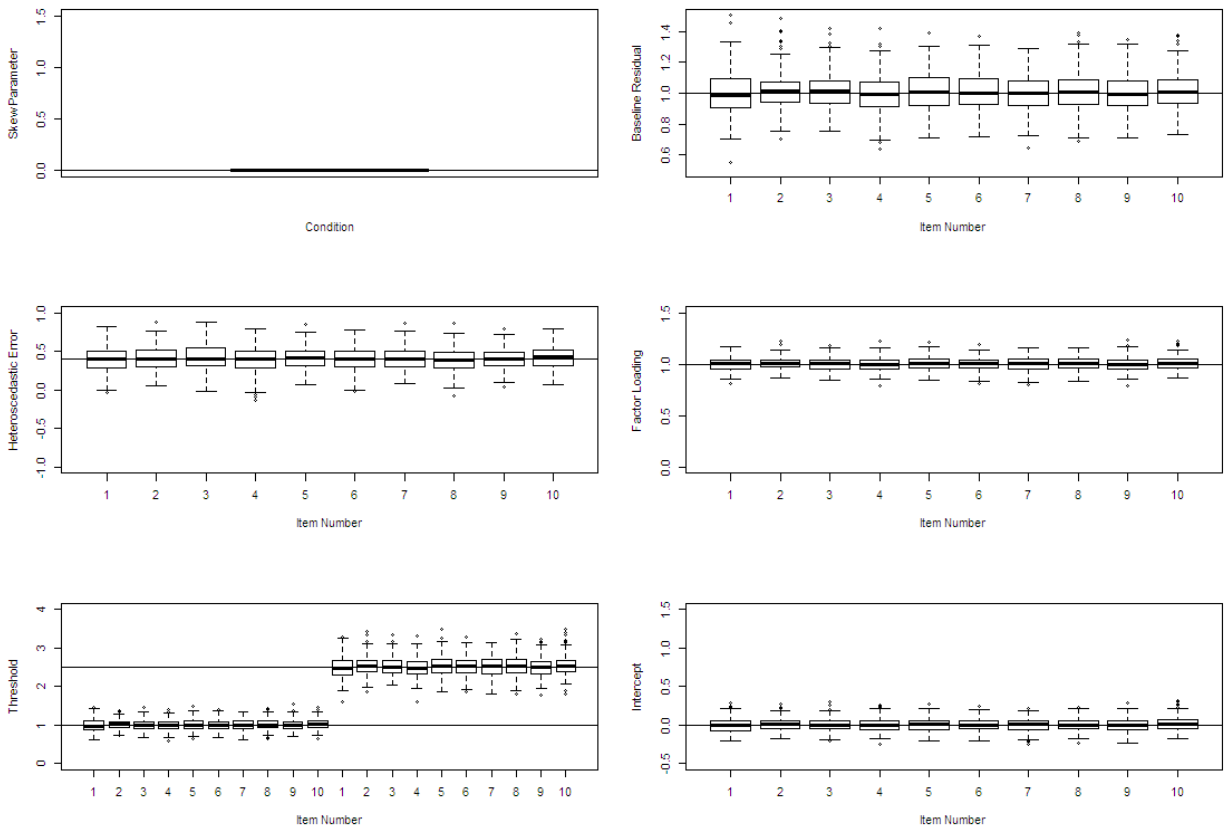


Figure 17. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 0. Each plot contains boxplots for each item.

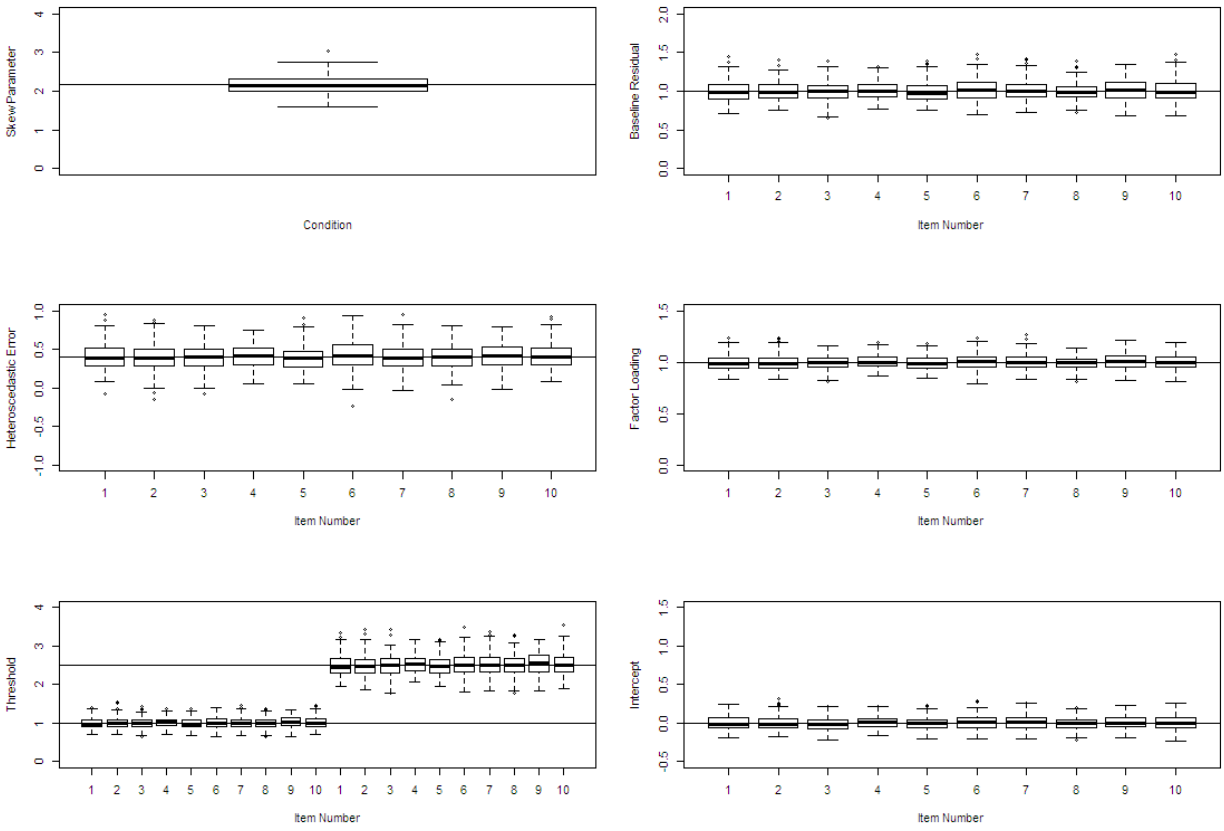


Figure 18. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 0.5. Each plot contains boxplots for each item.

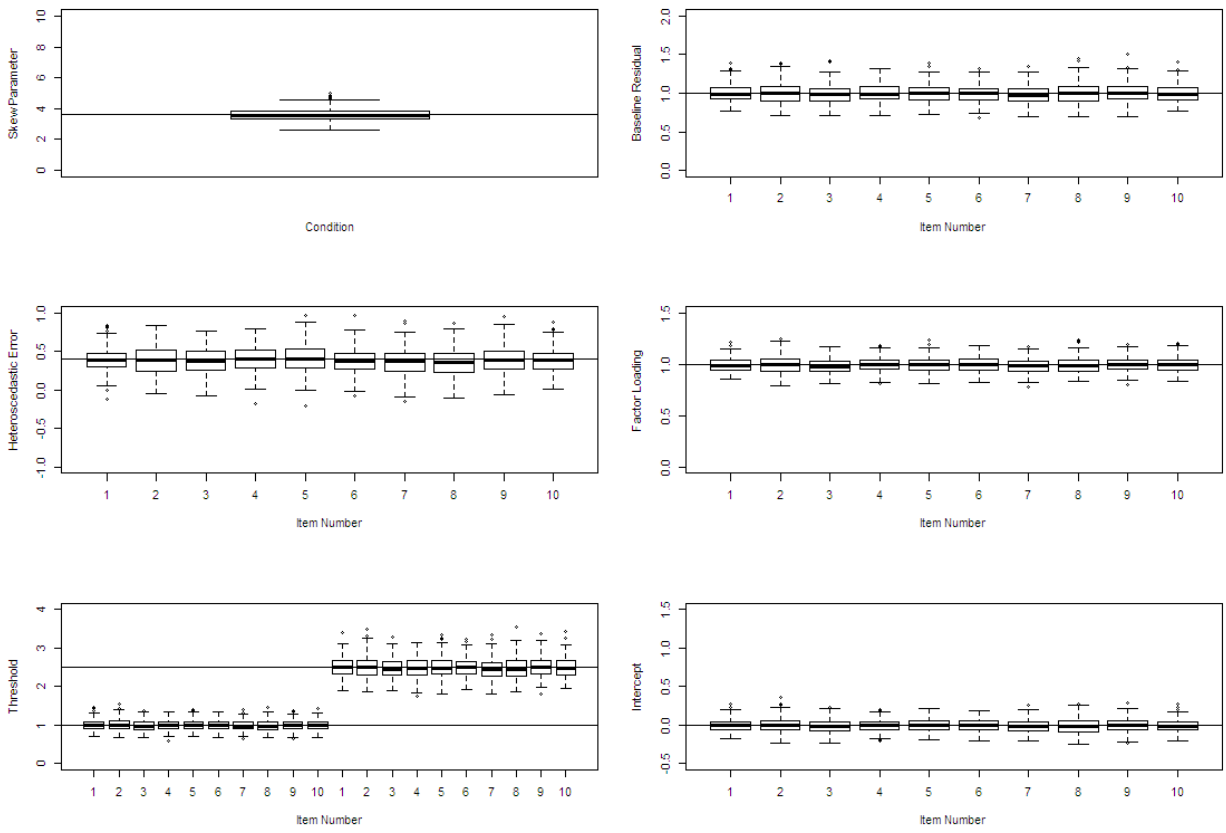


Figure 19. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 0.75. Each plot contains boxplots for each item.

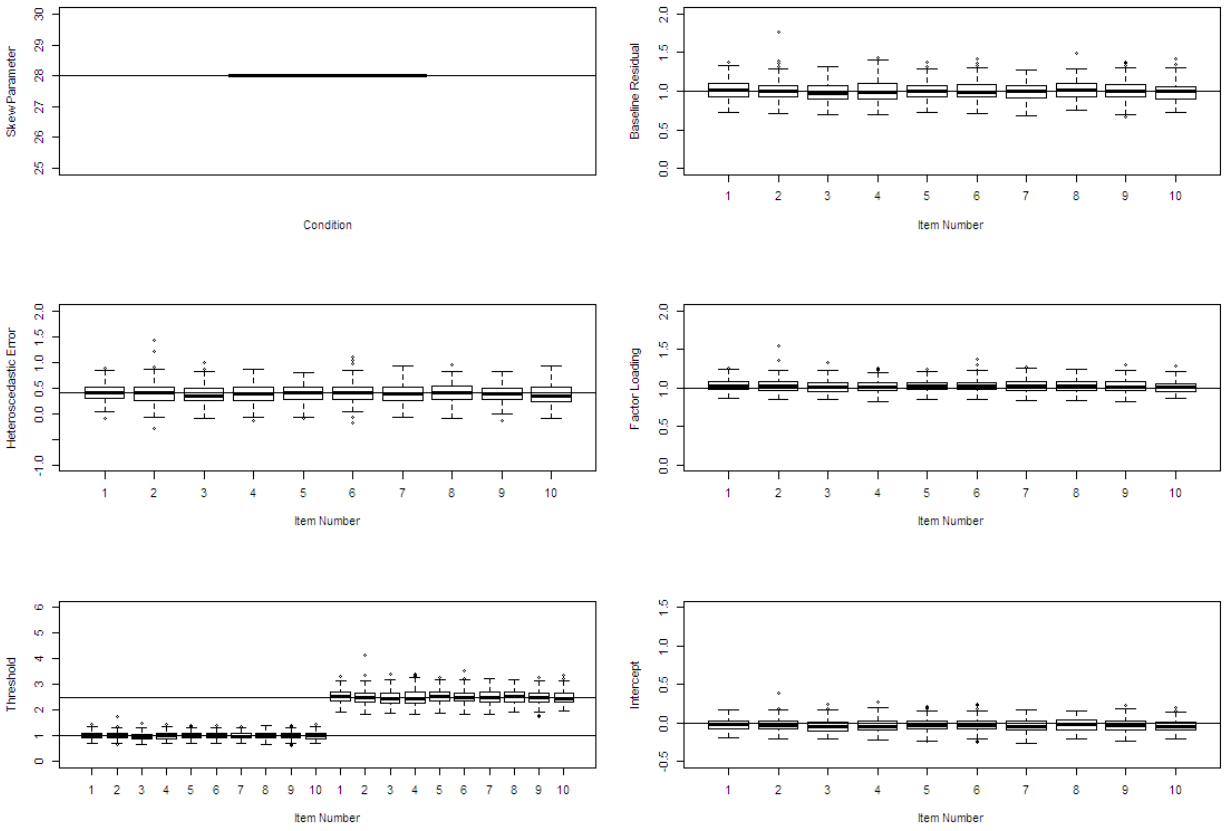


Figure 20. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 1.0. Each plot contains boxplots for each item.

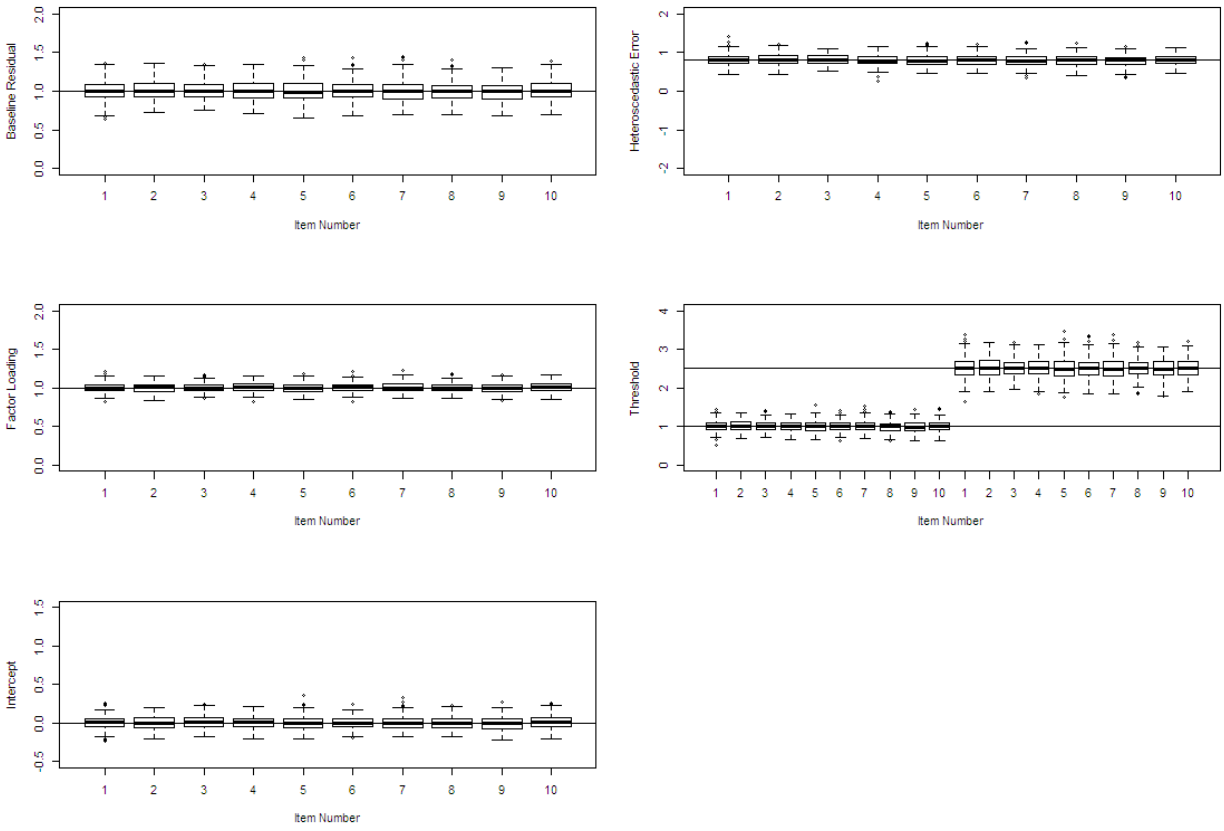


Figure 21. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 0. Each plot contains boxplots for each item.

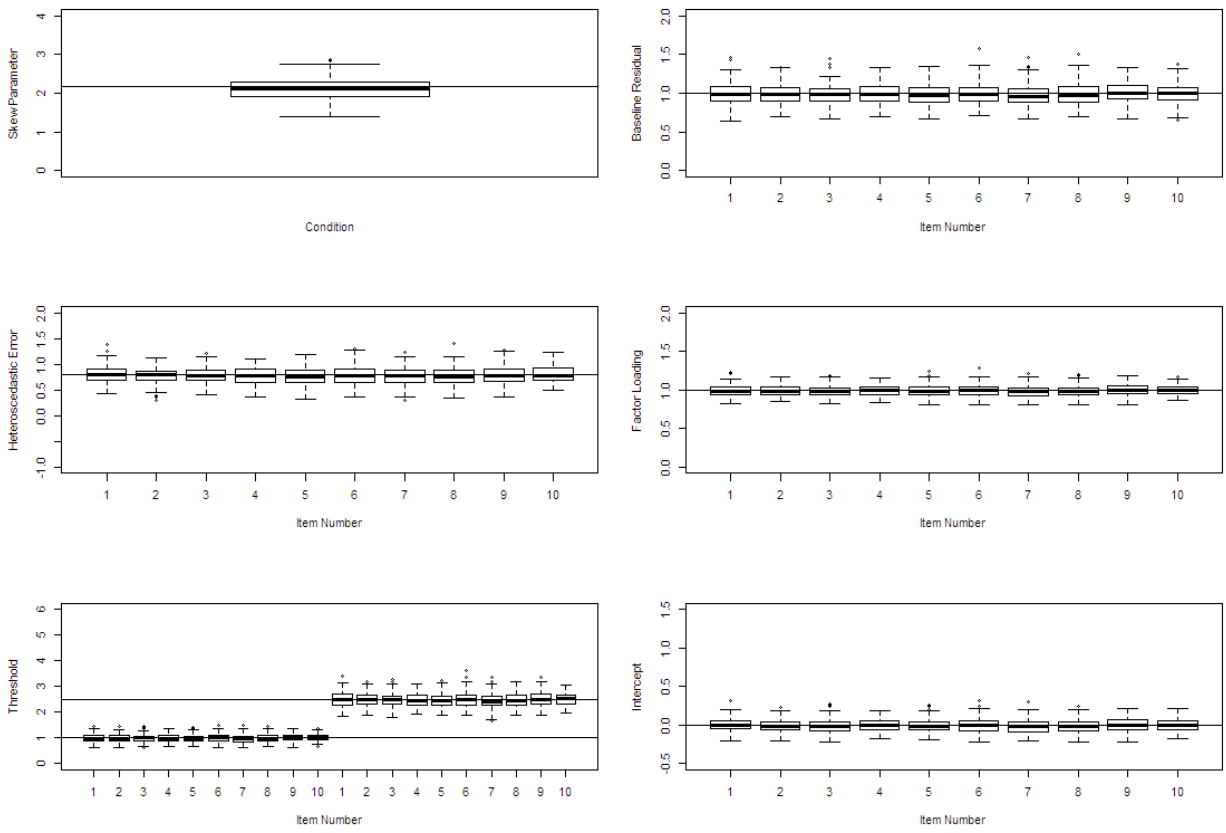


Figure 22. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 0.5. Each plot contains boxplots for each item.

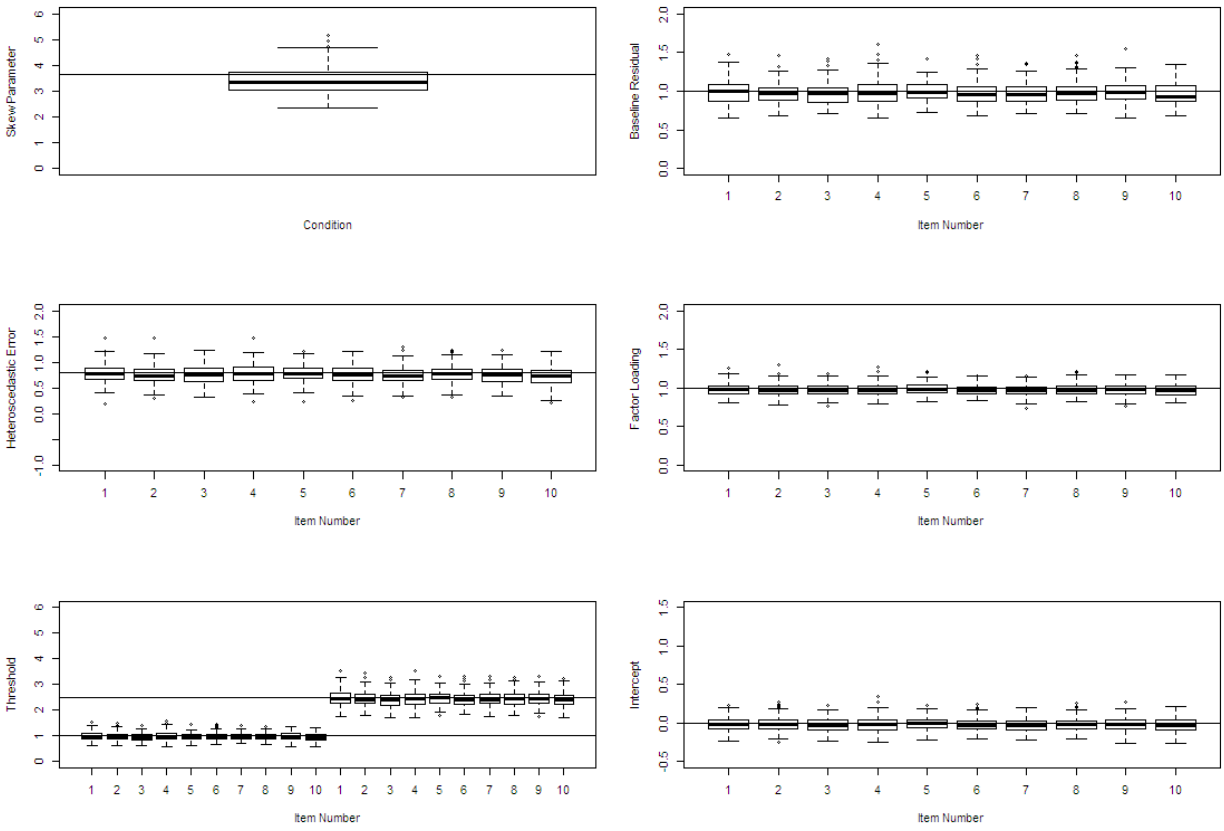


Figure 23. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 0.75. Each plot contains boxplots for each item.

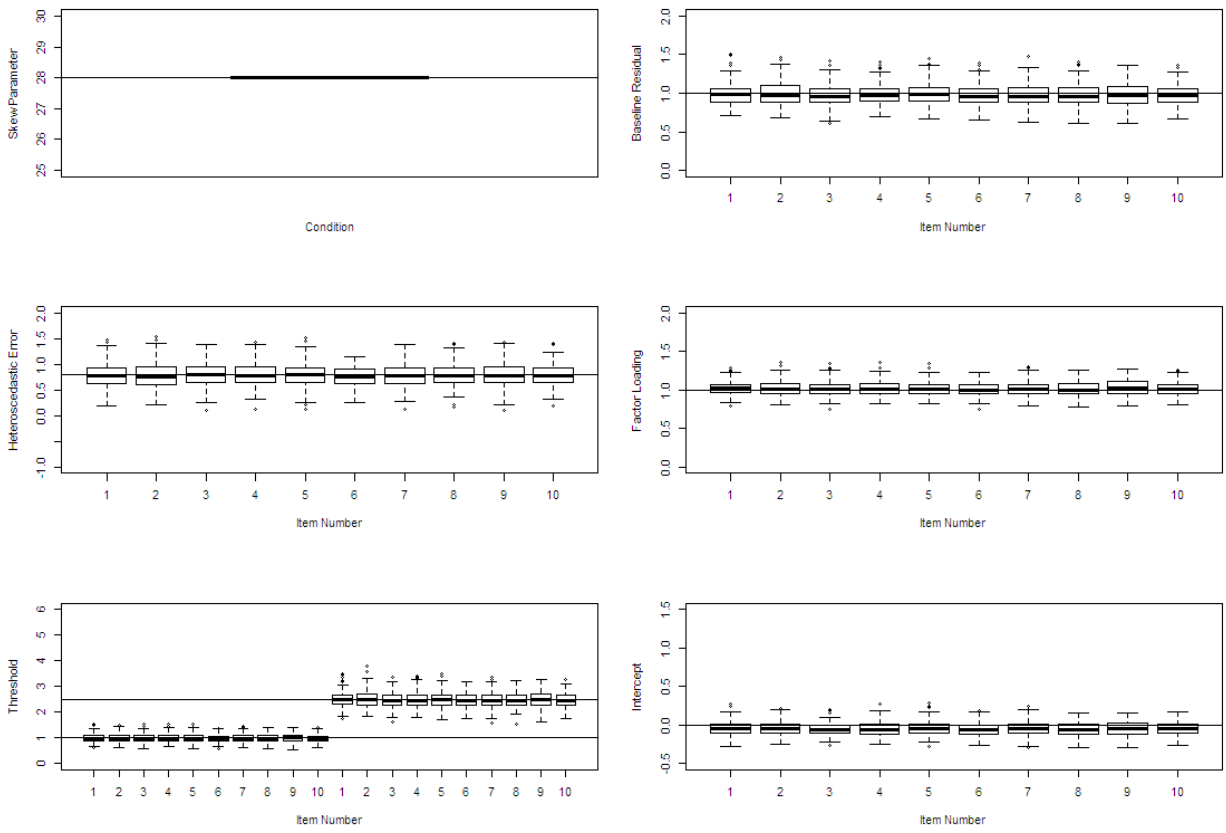


Figure 24. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 1.0. Each plot contains boxplots for each item.

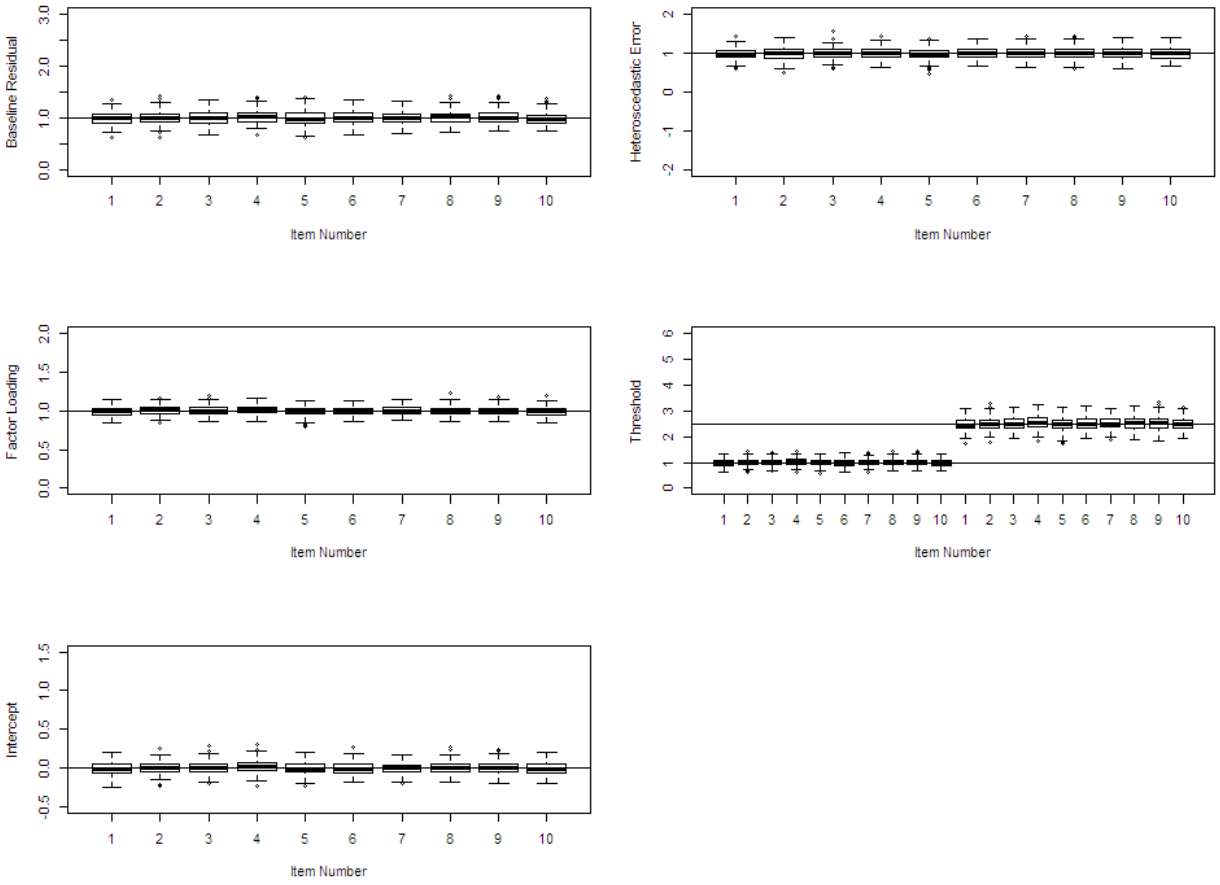


Figure 25. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 0. Each plot contains boxplots for each item.

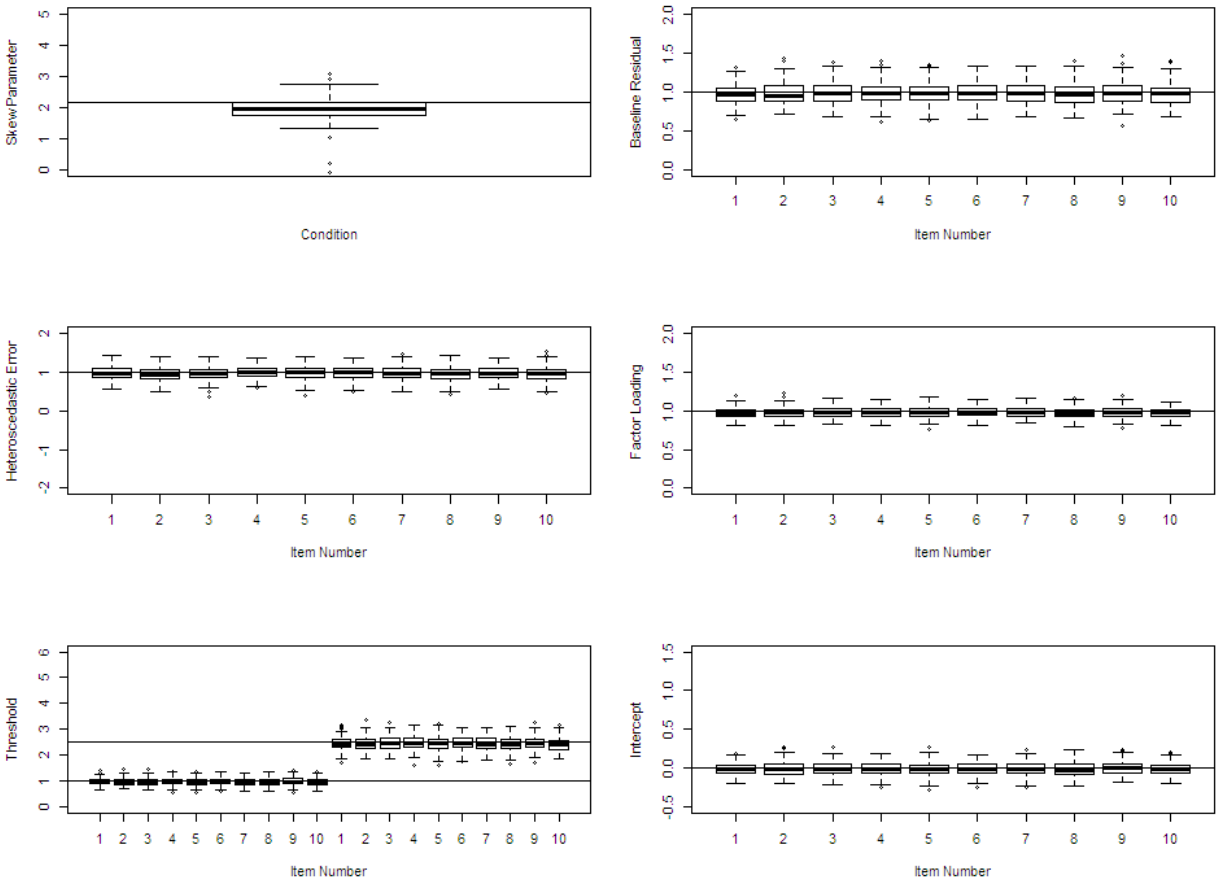


Figure 26. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 0.5. Each plot contains boxplots for each item.

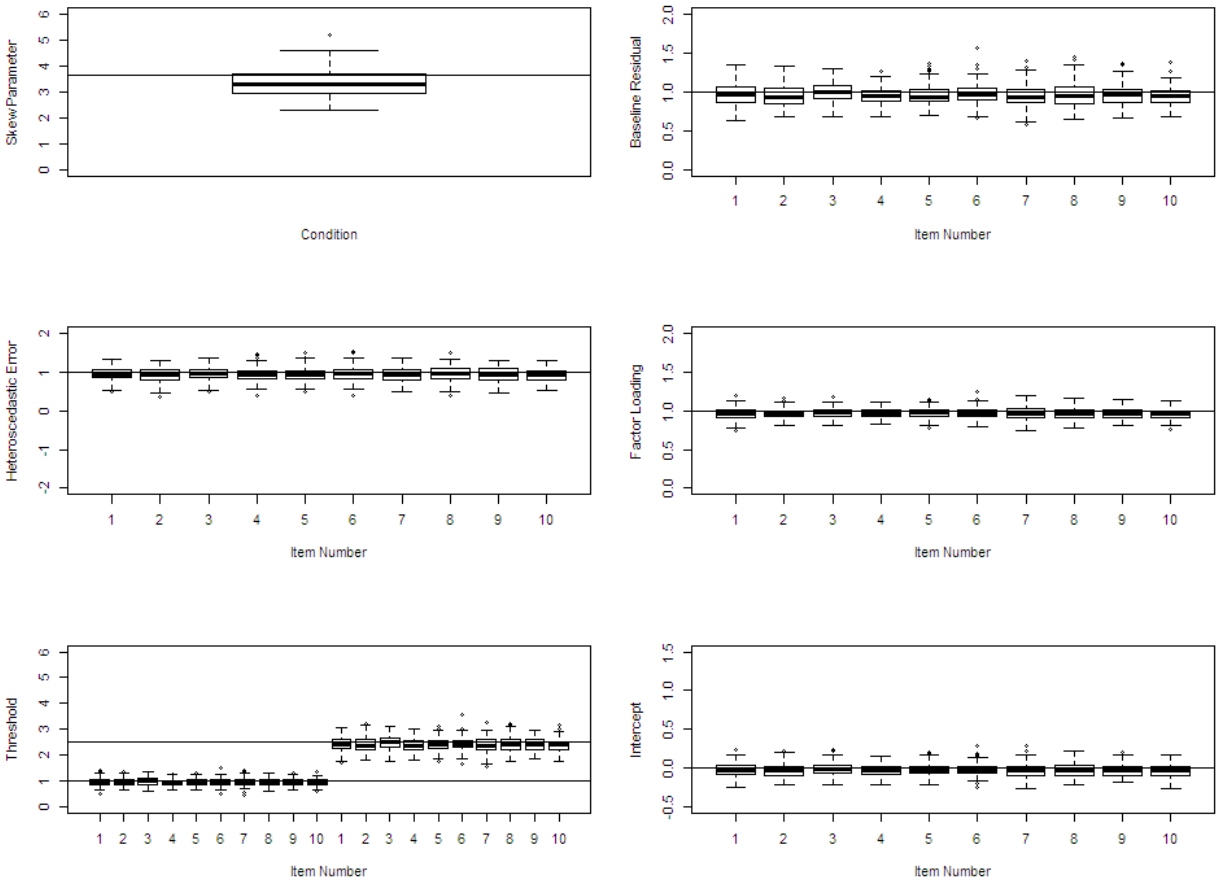


Figure 27. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 0.75. Each plot contains boxplots for each item.

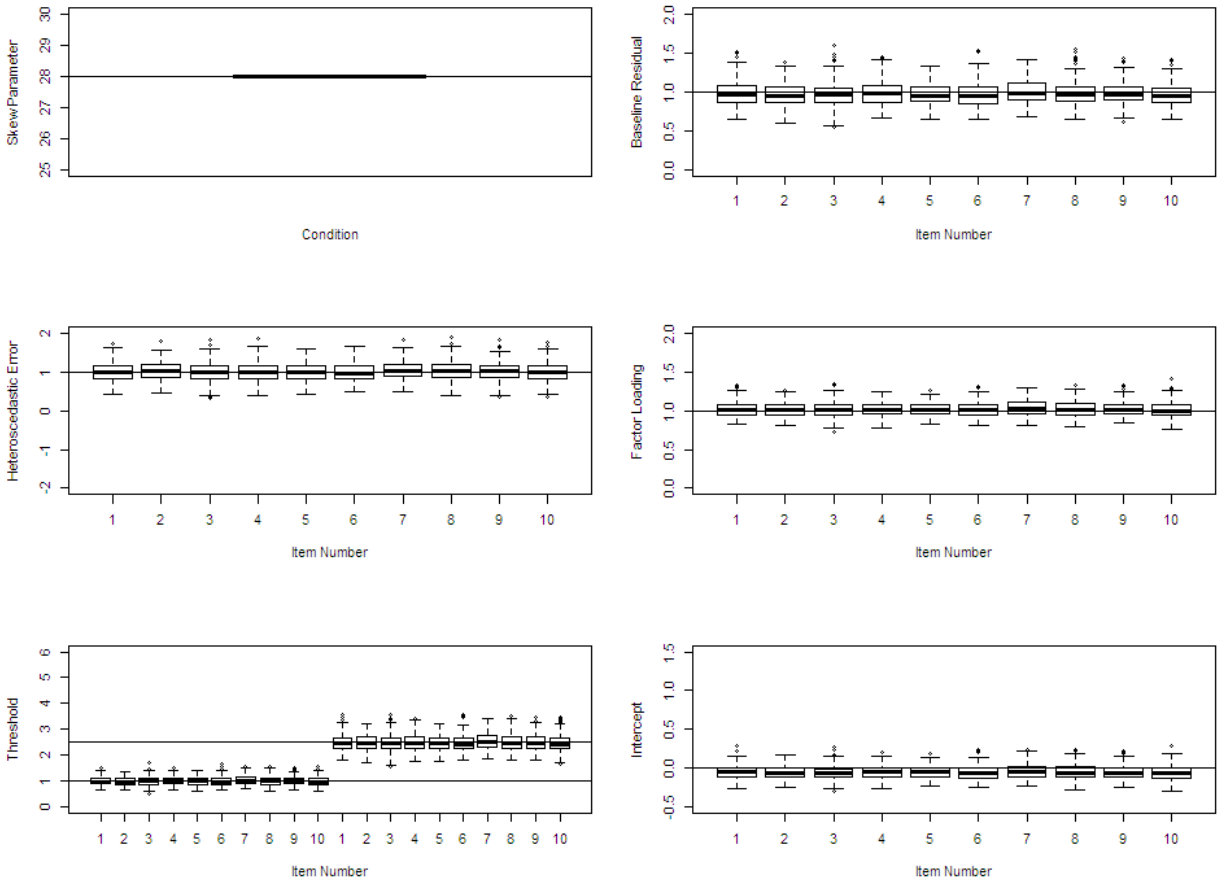


Figure 28. Baseline model boxplots for item parameters in large sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 1.0. Each plot contains boxplots for each item.

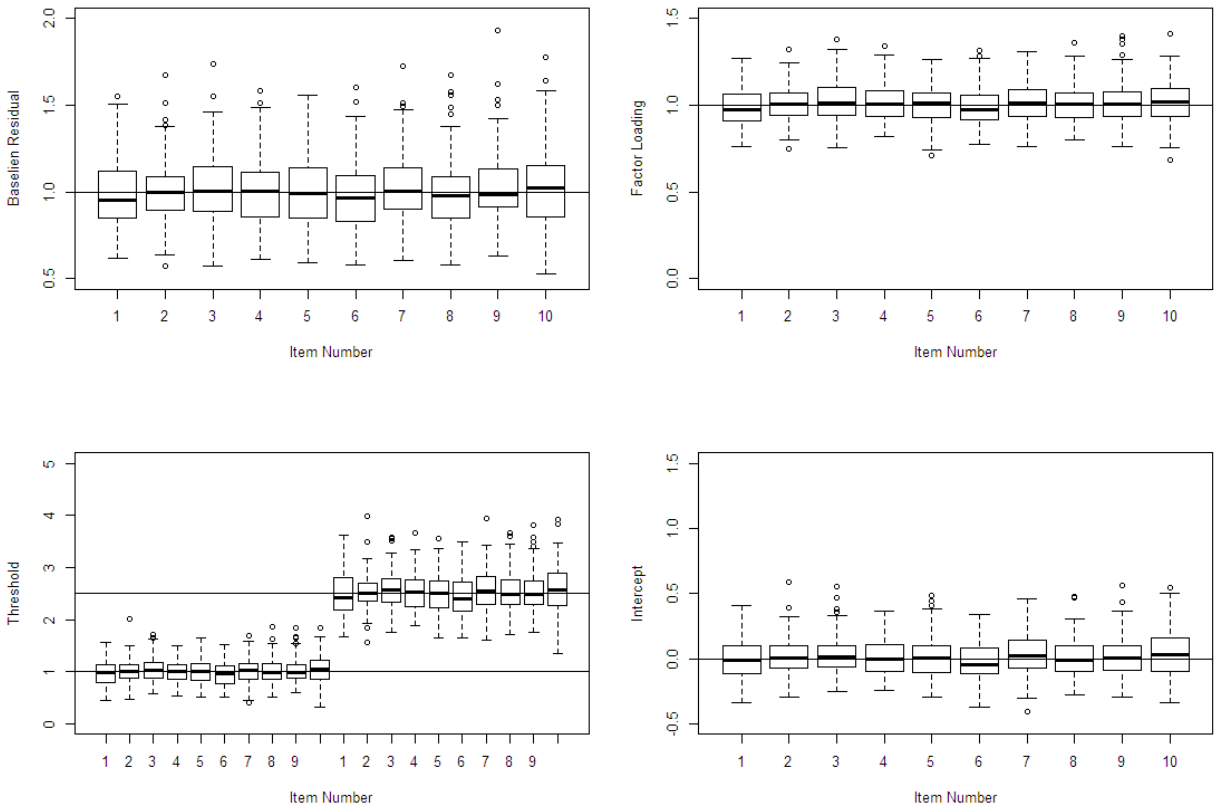


Figure 29. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0 and skew of 0. Each plot contains boxplots for each item.

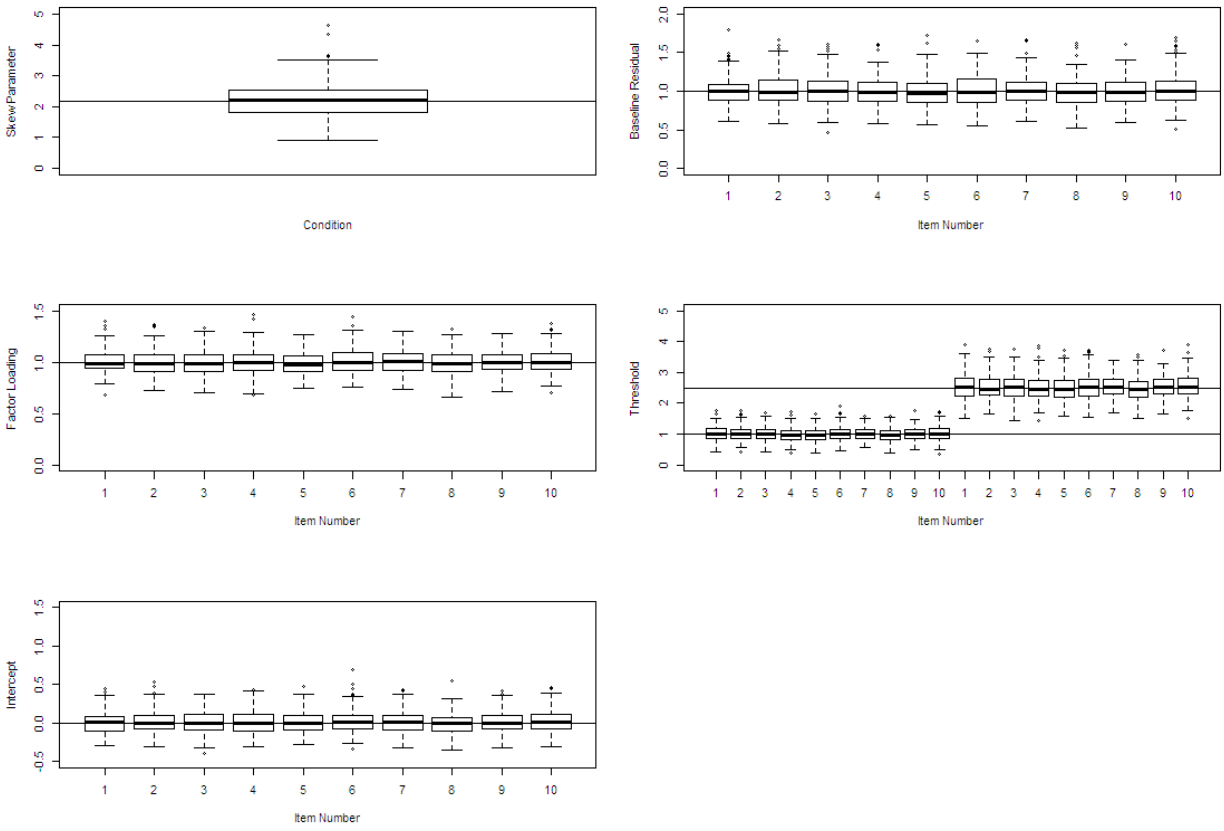


Figure 30. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0 and skew of 0.5. Each plot contains boxplots for each item.

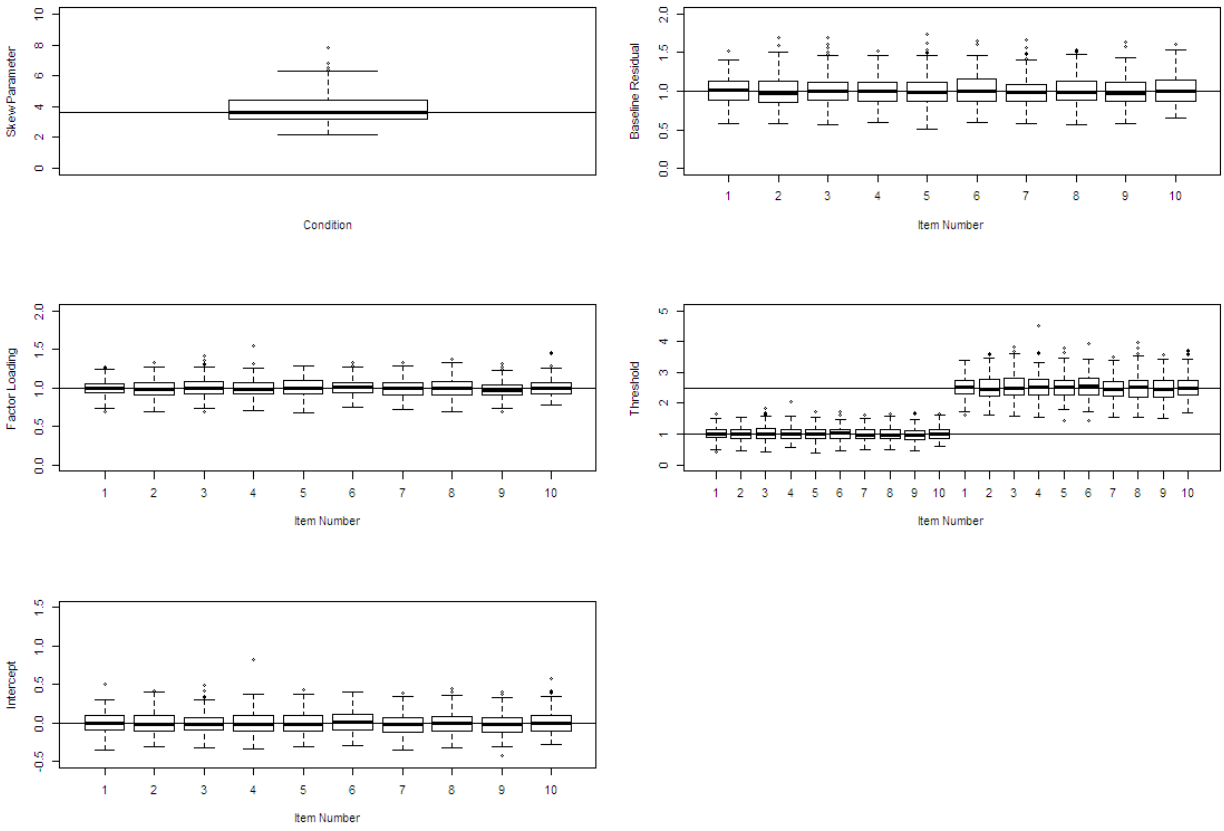


Figure 31. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0 and skew of 0.75. Each plot contains boxplots for each item.

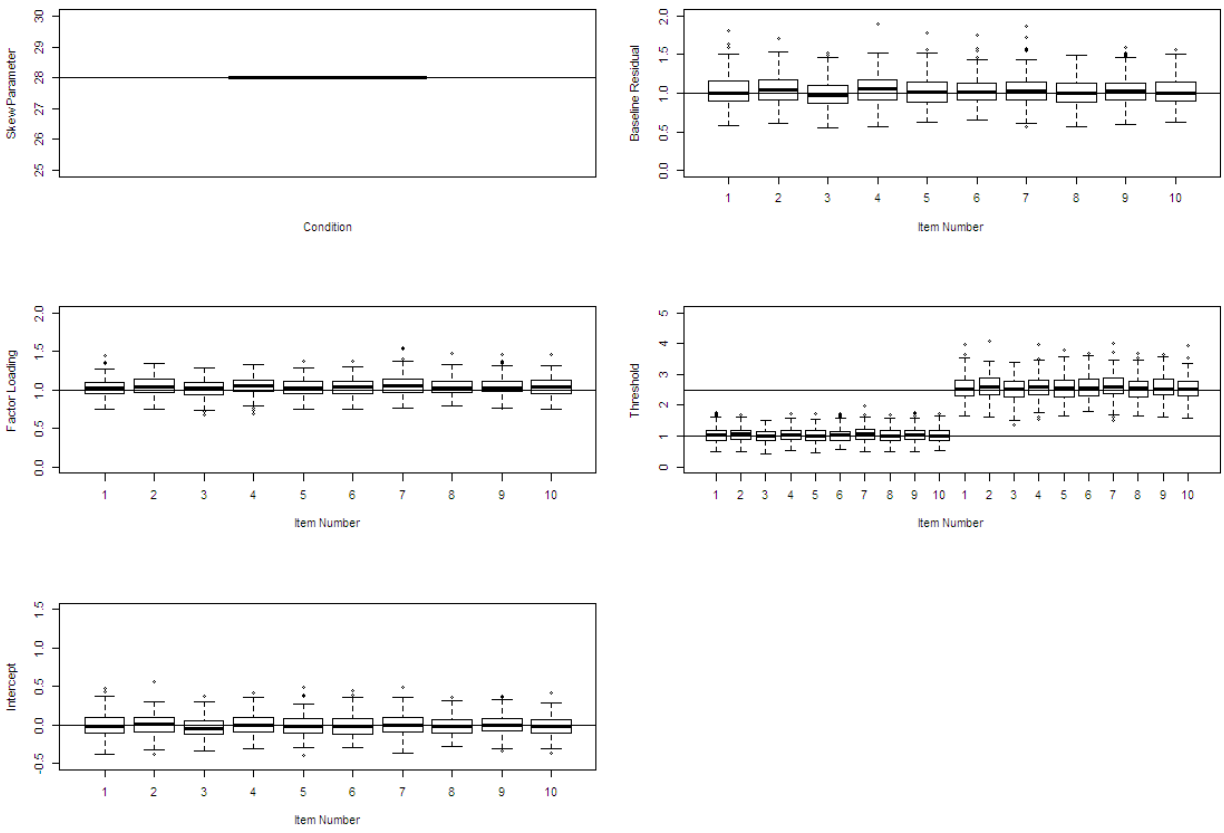


Figure 32. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0 and skew of 1.0. Each plot contains boxplots for each item.

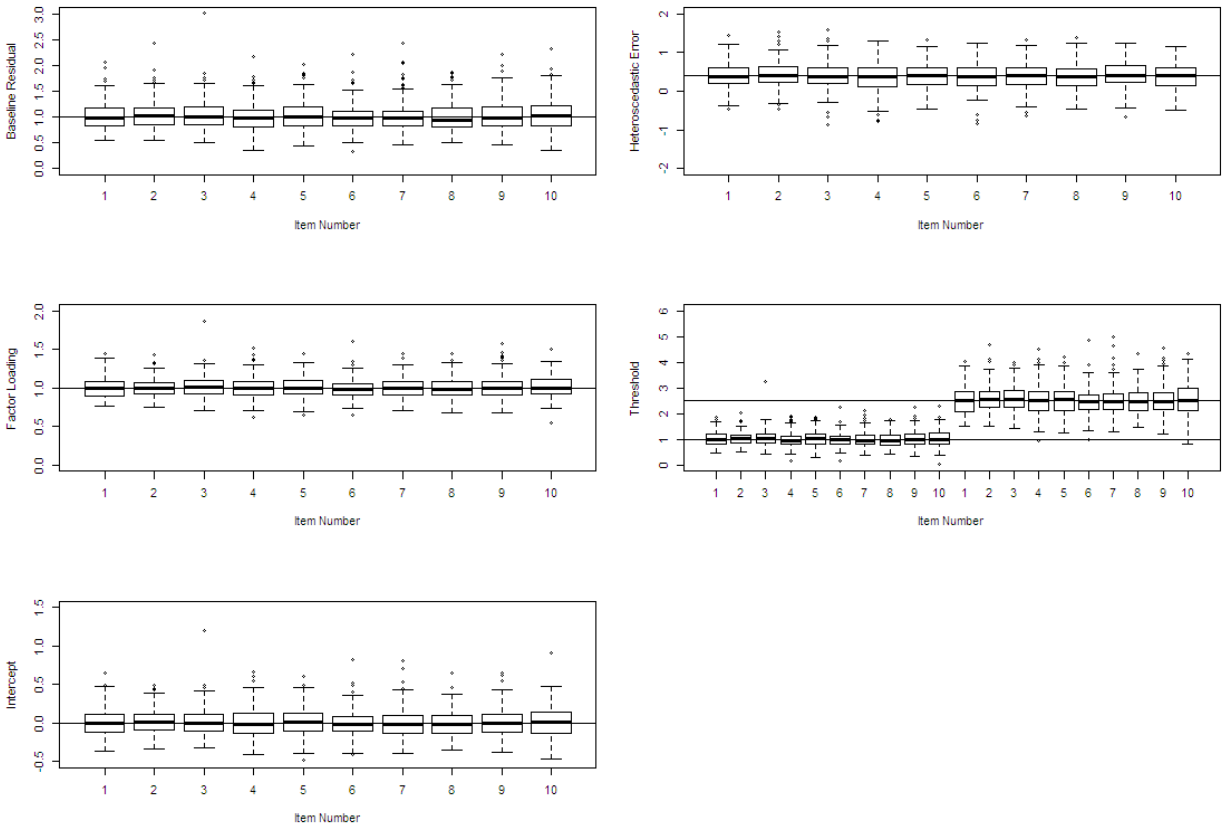


Figure 33. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 0. Each plot contains boxplots for each item.

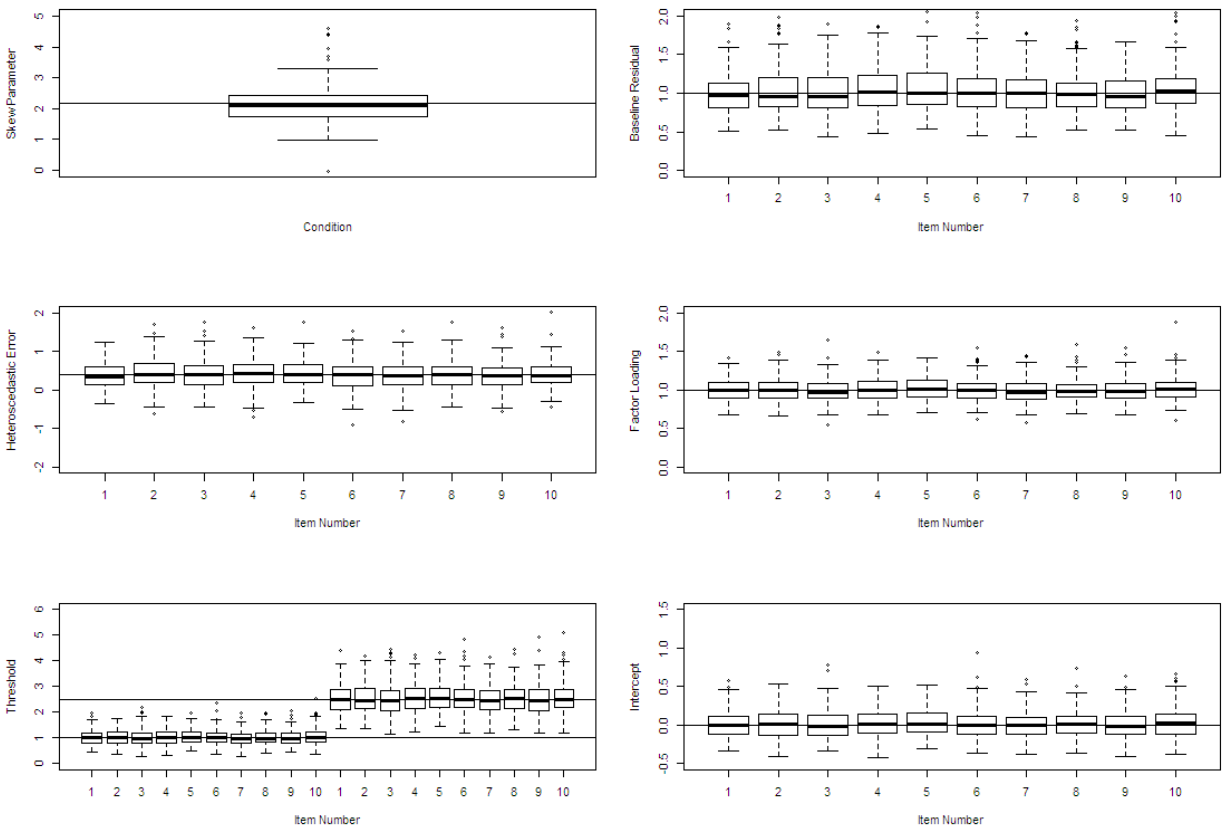


Figure 34. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 0.5. Each plot contains boxplots for each item.

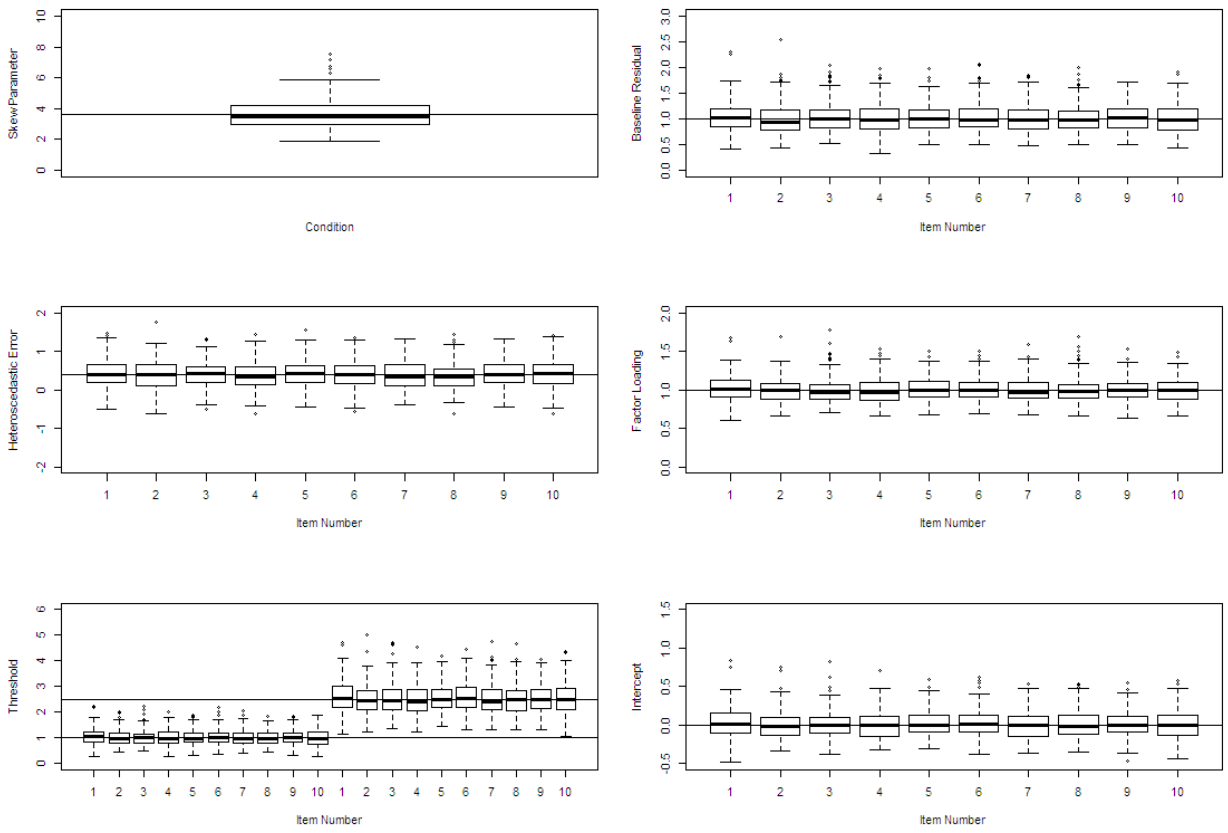


Figure 35. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 0.75. Each plot contains boxplots for each item.

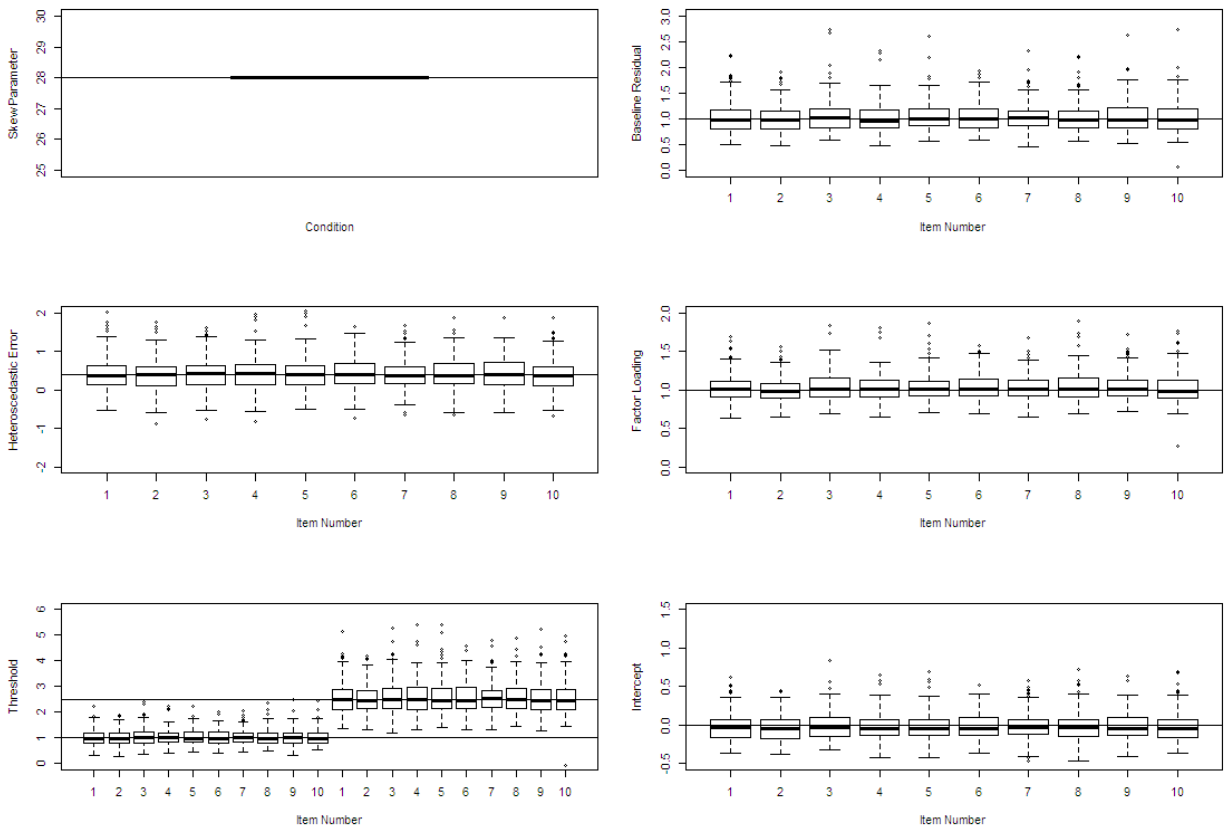


Figure 36. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.4 and skew of 1.0. Each plot contains boxplots for each item.

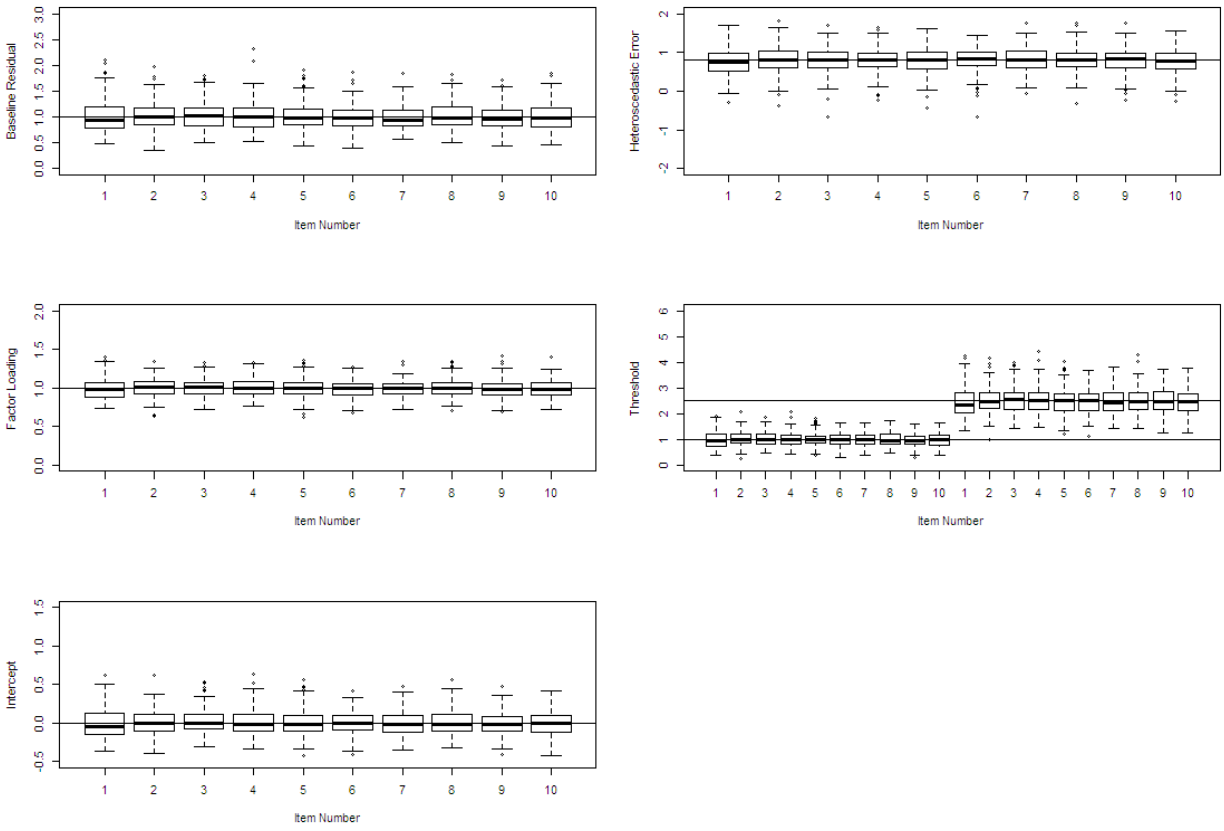


Figure 37. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 0. Each plot contains boxplots for each item.

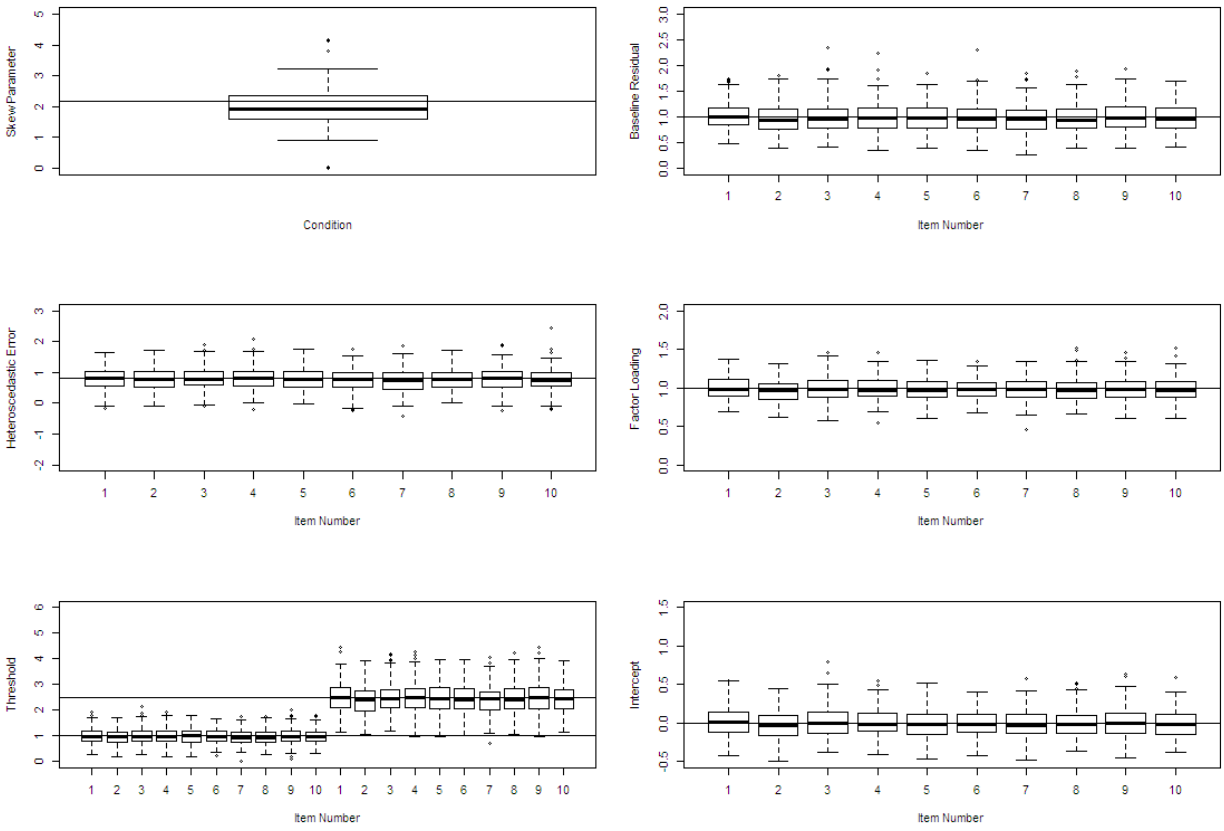


Figure 38. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 0.5. Each plot contains boxplots for each item.

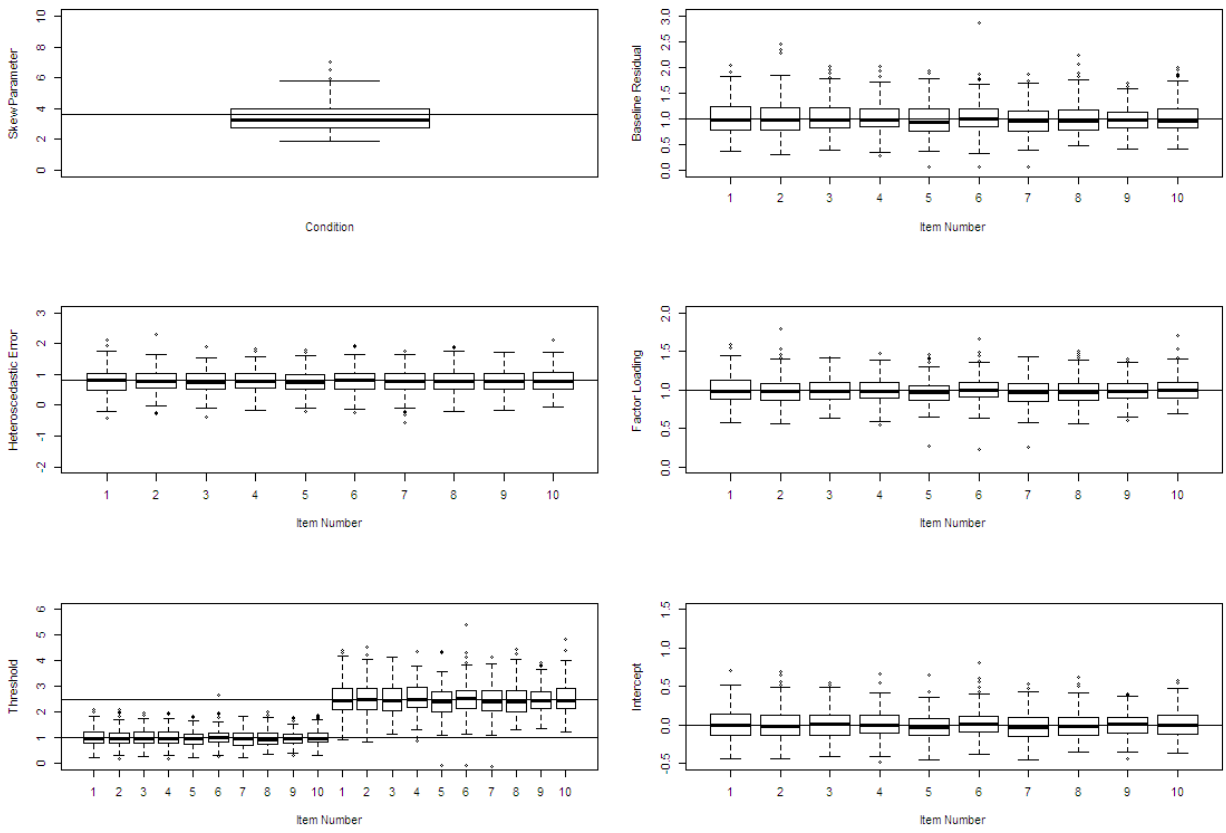


Figure 39. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 0.75. Each plot contains boxplots for each item.

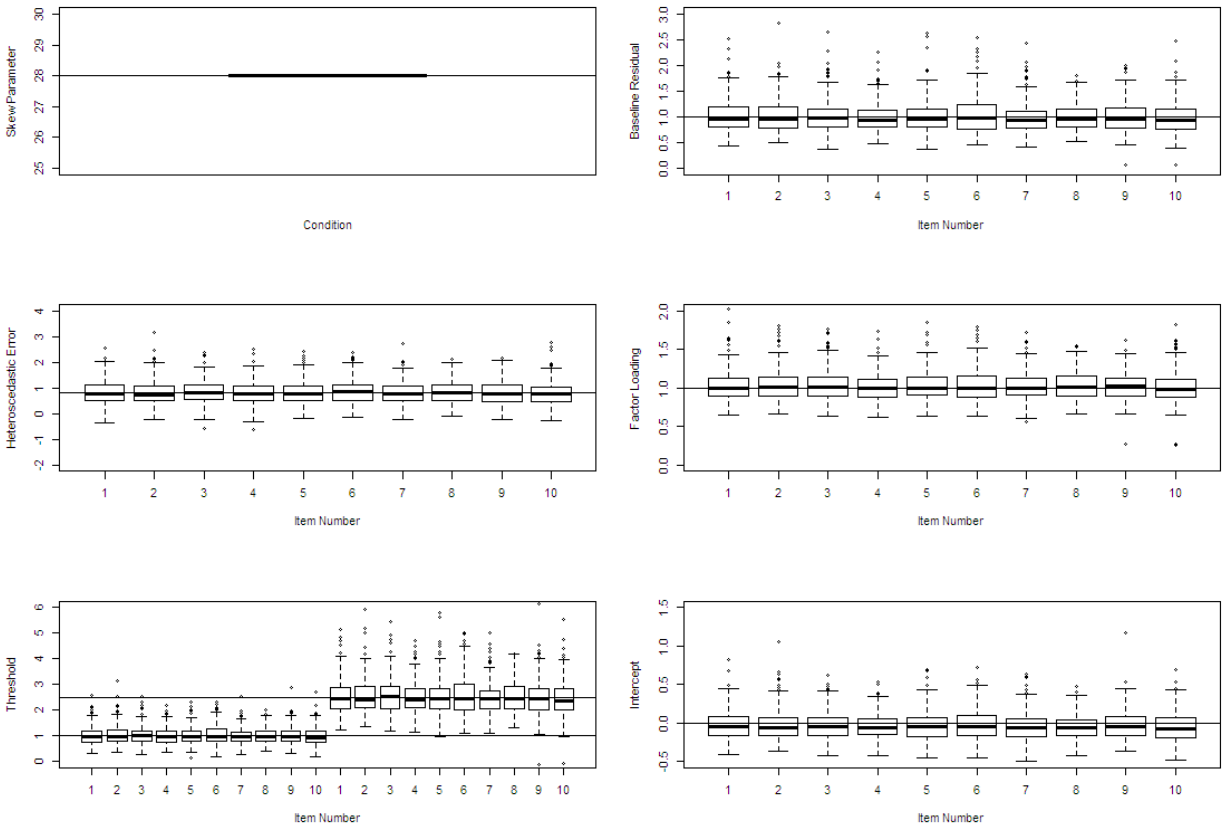


Figure 40. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 0.8 and skew of 1.0. Each plot contains boxplots for each item.

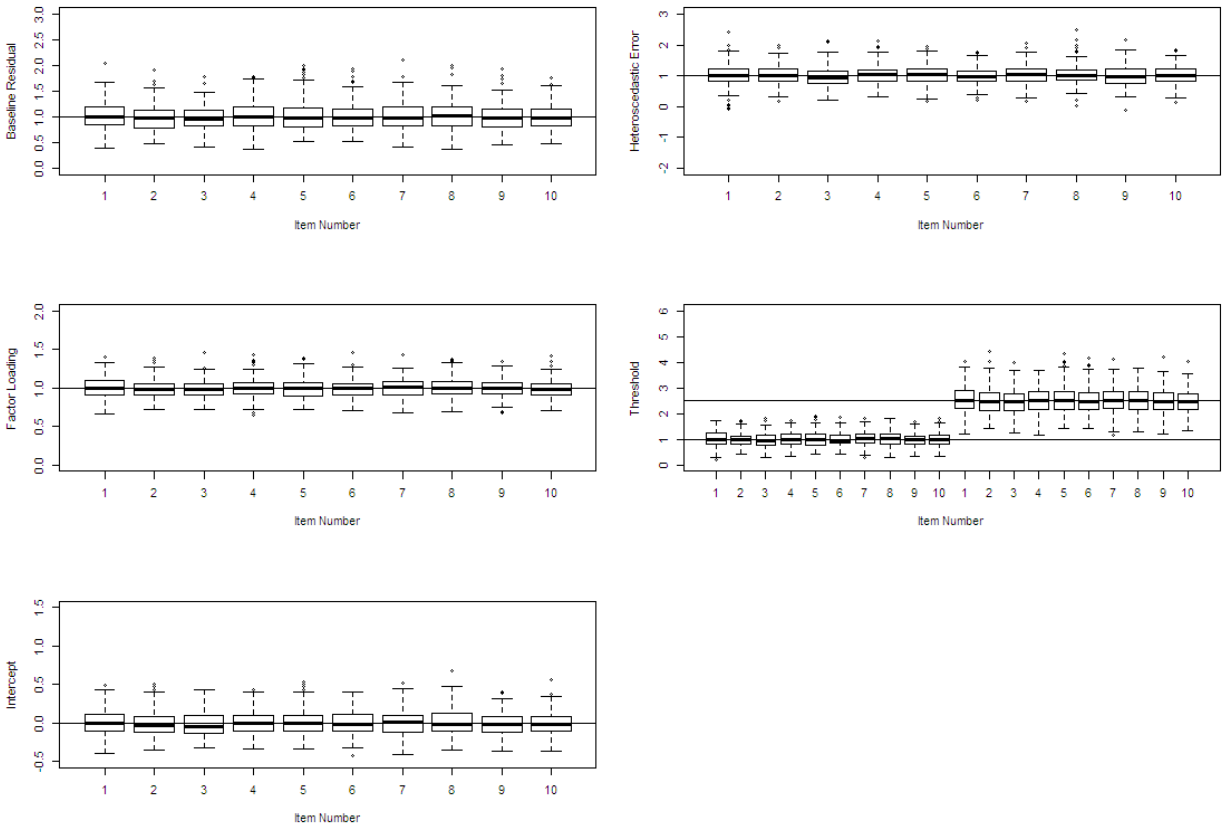


Figure 41. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 0. Each plot contains boxplots for each item.

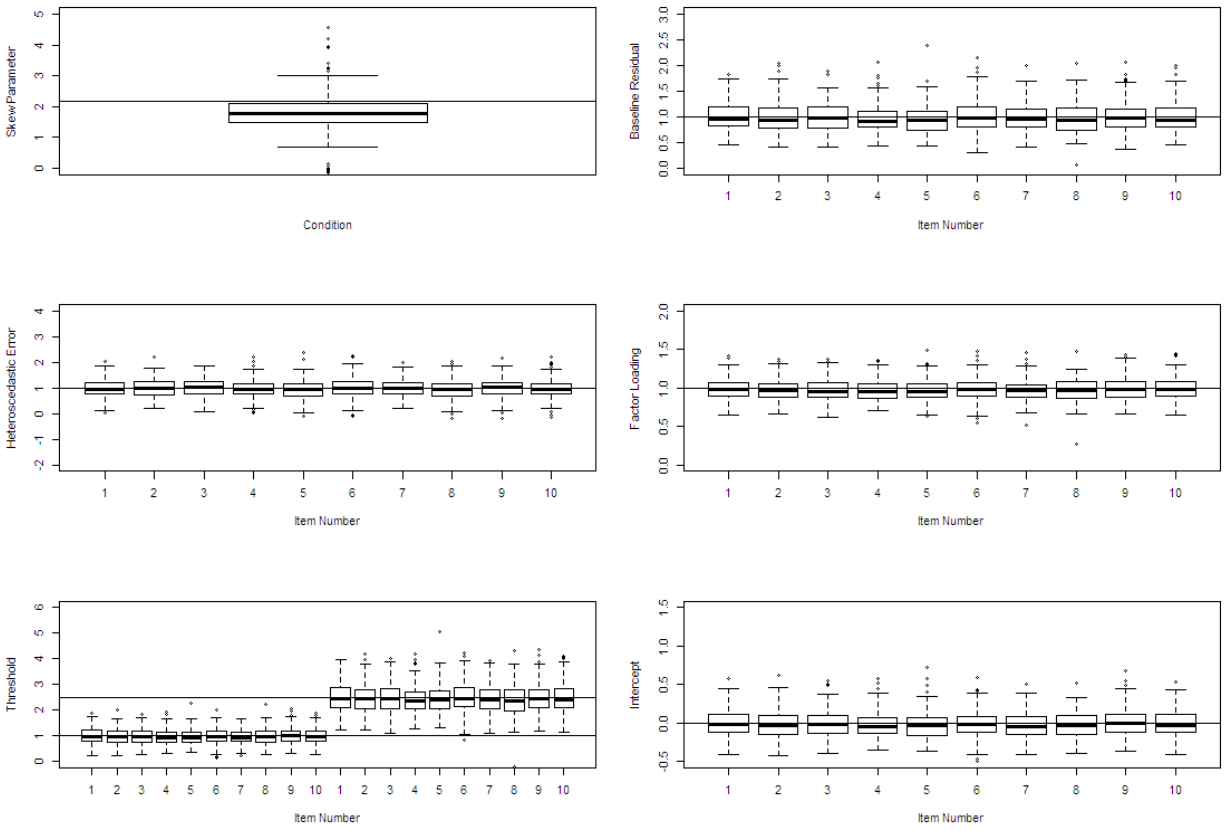


Figure 42. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 0.5. Each plot contains boxplots for each item.

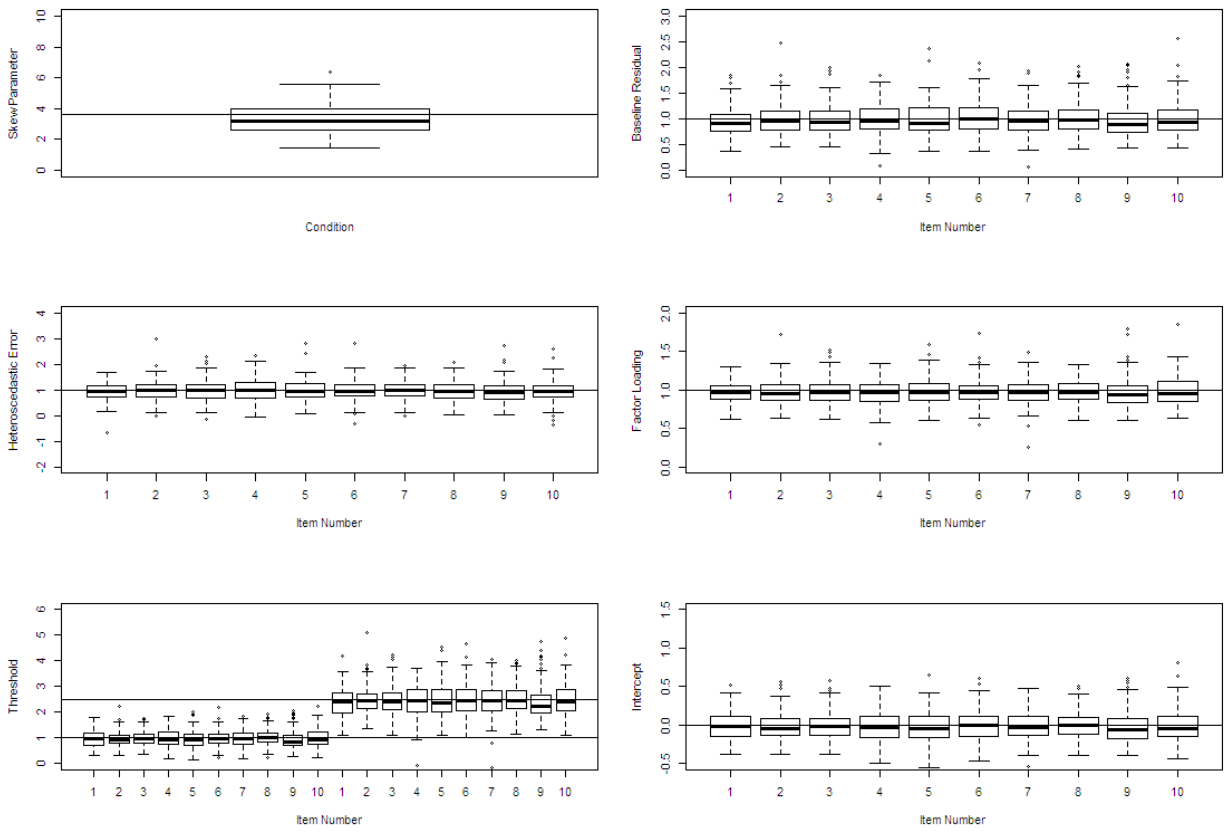


Figure 43. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 0.75. Each plot contains boxplots for each item.

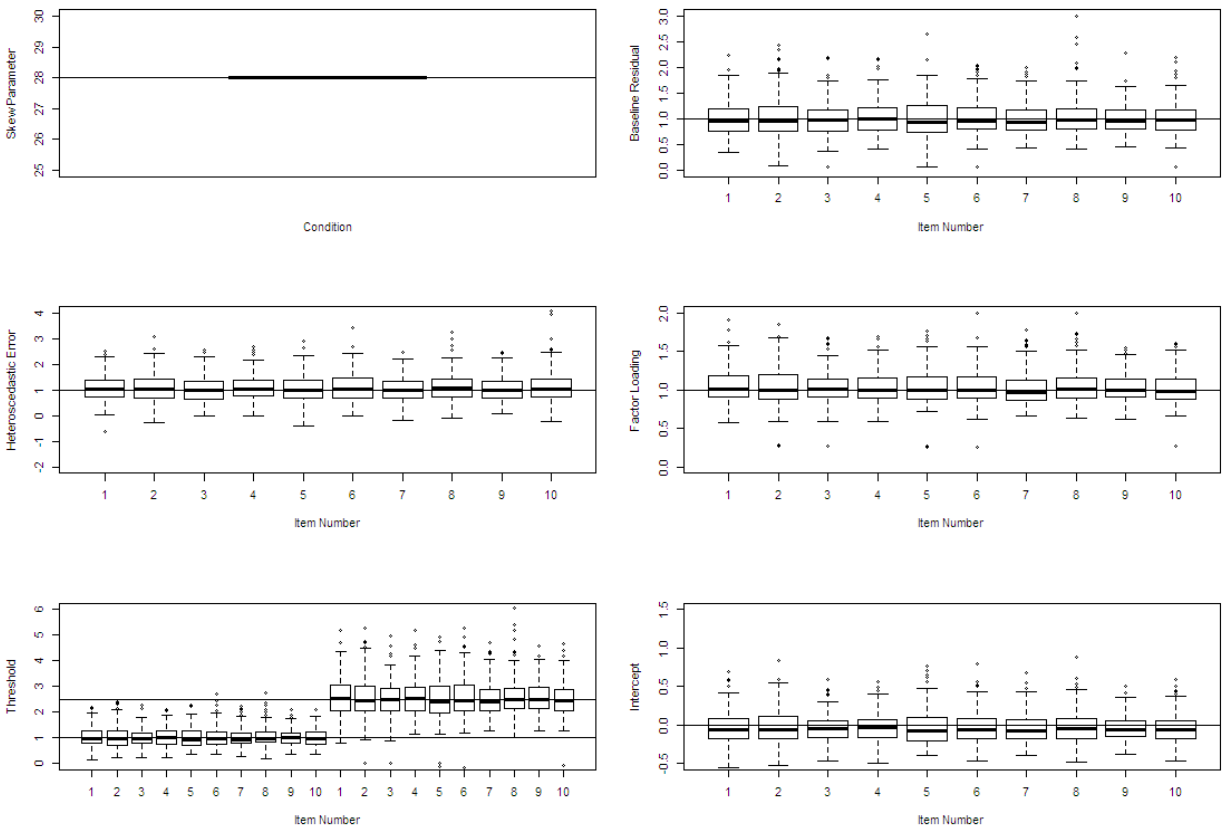


Figure 44. Baseline model boxplots for item parameters in small sample with 5-category response options with heteroscedastic errors of 1.0 and skew of 1.0. Each plot contains boxplots for each item.

REFERENCES

Abdel-fattah, A. A. (1994, April). *Comparing BILOG and LOGIST estimates for normal, truncated normal, and beta ability distributions*. Paper presented at the annual meeting of the American Education Research Association, New Orleans, G.A.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81. doi:10.1007/BF02293746

Andersen, E. B. (1997). The rating scale model. In W.J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 67-84). New York, NY: Springer.

Andrich, D. (1978a) Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4)*,* 581-594. doi:10.1177/014662167800200413

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi:10.1007/BF02293814

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement*, *19*(1), 101-119. doi:10.1177/014662169501900111

Azevedo, C. L., Bolfarine, H., & Andrade, D. F. (2011). Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics & Data Analysis*, *55*(1), 353-365. doi:10.1016/j.csda.2010.05.003

Azevedo, C. L., Bolfarine, H., & Andrade, D. F. (2012). Parameter recovery for a skew-normal IRT model under a Bayesian approach: hierarchical framework, prior and kernel sensitivity and sample size. *Journal of Statistical Computation and Simulation*, *82*(11), 1679-1699. doi:10.1080/00949655.2011.591798

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*, 171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal

ones. *Statistica*, *46*(2), 199-208. doi:10.6092/issn.1973-2201/711

Azzalini, A. (2015). The R package 'sn': The Skew-Normal and Skew-t distributions (version 1.3-0). URL http://azzalini.stat.unipd.it/SN

Azzalini, A., & Capatanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B*, *61*(3), 579-602. doi:10.1111/1467-9868.00194

Bazán, J. L., Branco, M. D., & Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis*, *1*(4), 861-892. doi:10.1214/06-BA128

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores*. Reading: Addison Wesley (Chapters 17–20).

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(3)*,* 29-51. doi:10.1007/BF02291411

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored item. *Psychometrika, 35*(2), 179-197. doi:10.1007/BF02291262

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. doi:10.1007/BF02293801

Boulet, J. R. (1996). *The effect of nonnormal ability distributions on IRT parameter estimation using full-information and limited-information methods (item response theory, nonlinear factor analysis).* Dissertation abstracts online; University of Ottawa (Canada).

Cella, D., & Stone, A. A. (2015). Health-related quality of life measurement in oncology: Advances and opportunities. *American Psychologist*, *70*(2), 175-185. doi:10.1037/a0037821

Cooper, L. D., Balsis, S., & Zimmerman, M. (2010). Challenges associated with a polythetic diagnostic system: Criteria combinations in the personality disorders. *Journal of*

Abnormal Psychology, 119(4), 886-895. doi:10.1037/a0021078

Cramér, H. (1937). *Random variables and probability distributions*. Cambridge: Cambridge University Press.

Davis, P. J., & Rabinovitz, P. (1975). *Methods of numerical integration.* New York: Academic Press.

de Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*(1), 3-19. doi:10.1177/01466219922031130

DeMars, C. E. (2003). Sample size and the recovery of the nominal response model item parameters. *Applied Psychological Measurement, 27*(4)*,* 275-288. doi:10.1177/0146621603027004003

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1-38.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77-90. doi:10.1177/014662168901300108

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

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists,* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Garcia, E. A., Aryal, S., & Walters, S. T. (2015). Using item response theory (IRT) to describe at-risk patients participating in health coaching programs. http://digitalcommons.hsc.unt.edu/thdc/thdc15/CommunityHealthandPrevention/2/.

Gorter, R., Fox, J. P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research*

Methodology, *15*(1), 1-12. doi:10.1186/s12874-015-0050-x

Greco, L. A., Lambert, W., & Baer, R. A. (2008). Psychological inflexibility in childhood and adolescence: Development and evaluation of the avoidance and fusion questionnaire for youth. *Psychological Assessment, 20*(2), 93-102. doi:10.1037/1040-3590.20.2.93

Hemker, B. T., Andries van der Ark, L., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*(4), 487-506. Doi:10.1007/BF02296191

Hessen, D. J., & Dolan, C. V. (2009). Heteroscedastic one-factor models and marginal maximum likelihood estimation. *British Journal of Mathematical & Statistical Psychology*, *62*(1), 57-77. doi:10.1348/000711007X248884

Keeley, J. W., Webb, C., Peterson, D., Roussin, L., & Flanagan, E. H. (2016). Development of a Response Inconsistency Scale for the Personality Inventory for DSM–5. *Journal of Personality Assessment*, *98*(4), 351-359. doi:10.1080/00223891.2016.1158719

Kim, S., Kim, S., & Kamphaus, R. W. (2010). Is aggression the same for boys and girls? Assessing measurement invariance with confirmatory factor analysis and item response theory. *School Psychology Quarterly, 25*(1)*,* 45-61. doi:10.1037/a0018768

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality, *Applied Psychological Measurement, 25*(2), 146-162. doi:10.1177/01466210122031975

Krueger, R. F., & Finger, M. S. (2010). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment, 13*(1), 140-151. doi:10.1037/1040-3590.13.1.140

Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistical Probability Letter, 9*(1), 91-97. doi:10.1016/0167-7152(90)90100-L

Lucke, J. F. (2014). Unipolar Item Response Models. In S.P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical*

Performance Assessment (pp. 272-284). New York, NY: Routledge/Taylor & Francis Group.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2)*,* 149-174. doi:10.1007/BF02296272

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W.J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-121). New York, NY: Springer.

McCracken, L. M., Chilcot, J., & Norton, S. (2015). Further development in the assessment of psychological flexibility: A shortened Committed Action Questionnaire (CAQ-8). *European Journal of Pain*, *19*(5), 677-685. doi:10.1002/ejp.589

McGlinchey, J. B., & Zimmerman, M. (2007). Examining a dimensional representation of depression and anxiety disorders' comorbidity in psychiatric outpatients with item response modeling. *Journal of Abnormal Psychology, 116*(3) 464-474. doi:10.1037/0021-843X.116.3.464

Meade, A. M. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728-743. doi:10.1037/a0018966

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: latent growth curves with ordinal outcomes. *Psychological Methods*, *9*(3), 301-333. doi:10.1037/1082-989X.9.3.301

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*(1), 91-100. doi:10.1177/014662169501900110

Micceri, T. (1989). The unicorn, the normal distribution, and other mythical creatures. *Psychological Bulletin, 105*(1)*,* 156-166. doi:10.1037/0033-2909.105.1.156.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika, 49*(3)*,* 359-381. doi:10.1007/BF02306026

Mislevy, R., & Bock, D. (1990). BILOG 3: Item analysis and test scoring with binary

logistic models [Computer software]. Chicago: Scientific Software International, Inc.

Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika, 80*(3), 625-644. doi:10.1007/s11336-014-9406-0

Molenaar, D., Dolan, C. V., & de Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, *77*(3), 455-478. doi:10.1007/s11336-012-9273-5

Molenaar, I. W. (1983). Item steps (Heymans Bulletin 83-630-EX). Groningen, The Netherlands: University of Groningen, Department of Statistics and Measurement Theory.

Monroe, S. & Cai, L. (2014). Estimation of a Ramsay-Curve item response theory model by Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement, 74*(2), 343-369. doi:10.1177/0013164413499344

Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement, 14*(1)*,* 59-71. doi:10.1177/014662169001400106

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2)*, 159-176.* doi:10.1177/014662169201600206

Muraki, E. (1993). Information functions of the generalized partial credit model. *ETS Research Report Series*, *1993*(1), i-12.

Murray, A. L., Molenaar, D., Johnson, W., & Krueger, R. F. (2016). Dependence of Gene-by-Environment Interactions (GxE) on Scaling: Comparing the Use of Sum Scores, Transformed Sum Scores and IRT Scores for the Phenotype in Tests of GxE. *Behavior Genetics*, *46*(4), 552-572. doi:10.1007/s10519-016-9783-5

Neale, M. C., Boker, S. M., Xie, G. & Maes, H. H. (2003). *Mx: Statistical Modeling*. VCU Box 900126, Richmond,VA 23298: Department of Psychiatry.

Olino, T. M. (2016). Future research directions in the positive valence systems: Measurement, development, and implications for youth unipolar depression. *Journal of Clinical Child &*

Adolescent Psychology, *45*(5), 681-705. doi:10.1080/15374416.2015.1118694

Parent, J., McKee, L. G., Rough, J. N., & Forehand, R. (2016). The association of parent mindfulness with parenting and youth psychopathology across three developmental stages. *Journal of Abnormal Child Psychology*, *44*(1), 191-202. doi:10.1007/s10802-015-9978-x

Patton, J. H., & Stanford, M. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*(6), 768-774. doi:10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Preston, K. S. J., & Reise, S. P. (2014). Estimating the Nominal Response Model under nonnormal conditions. *Educational and Psychological Measurement*, *74*(3), 377-399. doi:10.1177/0013164413507063

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2015). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453-469. doi:10.1002/bdm.1883

Purpura, D.J., Wilson, S.B., & Lonigan, C.J. (2010). Attention-deficit/hyperactivity disorder symptoms in preschool children: Examining psychometric properties using item response theory. *Psychological Assessment, 22*(3)*,* 546-558. doi:10.1037/a0019581

R Development Core Team. (2014). R: A language and environment for statistical computing [Software]. Vienna, Austria: R Foundation for Statistical Computing, Retrieved from http://www.R-project.org

Ramsay, J. O., & Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, *84*(408), 906-915. doi:10.1080/01621459.1989.10478854

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement*, *3*(3), 371-385. doi:10.1177/014662167900300309

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005) Item response theory: Fundamentals,

applications, and promise in psychological research. *Current Directions in Psychological Science, 14*(2), 95-101. doi:10.1111/j.0963-7214.2005.00342.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 35*(1), 139-139.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement.*

Samejima F. (1996). The graded response model. In W.J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Samejima, F. (1997). Departure from normal assumptions: a promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, *62*(4), 471-493. doi:10.1007/BF02294639

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, *65*(3), 319-335. doi:10.1007/BF02296149

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. *Psychometrika*, *73*(4), 561-578. doi:10.1007/s11336-008-9071-2

Samuel, D. B., Simms, L. J., Clark, L. A., Livesley, W. J., & Wigider, T. A. (2010). An item response theory integration of normal and abnormal personality scales. *Personality Disorders: Theory, Research, and Treatment, 1*(1), 5-21. doi:10.1037/a0018136

Schmitt, J. E., Mehta, P. D., Aggen, S. H., Kubarych, T. S., & Neale, M. C. (2006). Semi-nonparametric methods for detecting latent nonnormality: A fusion of latent trait and ordered latent class modeling. *Multivariate Behavioral Research, 41*(4), 427-443. doi:10.1207/s15327906mbr4104_1

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological*

Measurement, *14*(3), 299-311. doi:10.1177/014662169001400307

Shi, J. Q., & Lee, S. Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society, 62*(1), 77-87. doi:10.1111/1467-9868.00220

Shono, Y., Ames, S. L., & Stacy, A. W. (2015). Evaluation of internal validity using modern test theory: Application to word association. *Psychological Assessment, 28*(2)*,* 194-204. doi:10.1037/pas0000175

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1)*,* 1-16. doi:10.1177/014662169201600101

Swaminathan, H., & Gifford, J. (1983). Estimation of parameters in the three- parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, *13*(3), 201-214. doi:10.1111/j.1745-3984.1976.tb00011.x

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*(4), 501-519. doi:10.1007/BF02302588

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika,51*(4), 567-577. doi:10.1007/BF02295596

Thomas, M. L., & Locke, D. E. C. (2010). Psychometric properties of the MMPI-2-RF somatic complaints (RC1) scale. *Psychological Assessment, 22*(3), 492-503. doi:10.1037/a0019229

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39-55. doi:10.1111/j.2044-8317.1990.tb00925.x

van den Oord, E. J. C. G. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement, 29*(1), 23-30. doi:10.1177/0146621604269791

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. New York: Springer.

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement*, *39*(8), 583-597. doi:10.1177/0146621615588184

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.S., (2002). Recovery of item parameters in the nominal response model: a comparison of marginal maximum likelihood estimation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement, 26*(3)*,* 339-352. doi:10.1177/0146621602026003007

Woods, C. M. (2006). Ramsay-curve item response theory to detect and correct for nonnormal latent variables. *Psychological Methods, 11*(3), 253-270. doi:10.1037/1082-989X.11.3.253

Woods, C. M. (2007). Ramsay-curve IRT for Likert- type data. *Applied Psychological Measurement, 31*(3)*,* 195-212. doi:10.1177/0146621606291567

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102-117. doi:10.1177/0146621608319512

Woods, C. M., & Thissen, D. (2004). RCLOG v. 1: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities. C*hapel Hill, NC: LL Thurstone Psychometric Laboratory*.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*(2)*,* 281-301. doi:10.1007/s11336-004-1175-8

Zwinderman, A. H., & van den Wollenberg, A. L. (1990) Robustness of marginal maximum

likelihood estimation in the Rasch model. *Applied Psychological Measurement, 14*(1), 73-81. doi:10.1177/014662169001400107