THE HETEROSCEDASTIC GRADED RESPONSE MODEL WITH A SKEWED LATENT TRAIT: TESTING STATISTICAL AND SUBSTANTIVE HYPOTHESES RELATED TO SKEWED ITEM CATEGORY FUNCTIONS

DYLAN MOLENAAR, CONOR V. DOLAN, AND PAUL DE BOECK

UNIVERSITY OF AMSTERDAM

The Graded Response Model (GRM; Samejima, Estimation of ability using a response pattern of graded scores, Psychometric Monograph No. 17, Richmond, VA: The Psychometric Society, 1969) can be derived by assuming a linear regression of a continuous variable, Z, on the trait, θ, to underlie the ordinal item scores (Takane & de Leeuw in Psychometrika, 52:393–408, 1987). Traditionally, a normal distribution is specified for Z implying homoscedastic error variances and a normally distributed θ. In this paper, we present the Heteroscedastic GRM with Skewed Latent Trait, which extends the traditional GRM by incorporation of heteroscedastic error variances and a skew-normal latent trait. An appealing property of the extended GRM is that it includes the traditional GRM as a special case. This enables specific tests on the normality assumption of Z. We show how violations of normality in Z can lead to asymmetrical category response functions. The ability to test this normality assumption is beneficial from both a statistical and substantive perspective. In a simulation study, we show the viability of the model and investigate the specificity of the effects. We apply the model to a dataset on affect and a dataset on alexithymia.

Key words: graded response model, normal distribution, heteroscedasticity, skew-normal distribution, non-normality.

The Graded Response Model (Samejima, 1969) is a psychometric generalized linear mixed model to infer a continuously distributed latent trait from a set of ordinal item scores. In psychology, the model is particularly useful in the field of personality (e.g., Fraley, Waller, & Brennan, 2000; Emons, Meijer, & Denollet, 2007) where questionnaires are administered using items with Likert-type scales.

The GRM can be derived by assuming that the ordinal scores on item i have risen from categorization of an underlying normally distributed variable, Z_i, at increasing thresholds. This variable is modeled as a linear function of the latent trait, θ. The category response functions—which specify the relation between the probability of a response and θ—are then given by integrating over the conditional density function of $Z_{i|\theta}$ between the corresponding thresholds (see Takane & de Leeuw, 1987; Wirth & Edwards, 2007). In the present article, we focus on how violations of the normality assumption affect the model. We will present the Heteroscedastic GRM with Skewed Latent Trait which is an extension of the traditional GRM that enables specific tests on this assumption. These tests are useful both in the statistical and the substantive setting.

1. The Normality Assumption for Z_i

The assumption of a normal distribution for Z_i is pragmatic but testable (see, e.g., Jöreskog, 2002, p. 13, for an unconditional test on the bivariate marginal distribution of Z_i). The normal distribution can be replaced by a logistic distribution (Birnbaum, 1968) as an approximation. The shape of the category response functions in the GRM depends on $Z_{i|\theta}$. Thus, using a normal

distribution for Z_i implies a normal distribution for $Z_{i|\theta}$ which results in the normal-ogive GRM. On the contrary, using a logistic distribution for Z_i implies a logistic distribution for $Z_{i|\theta}$ which results in the logistic GRM (see Samejima, 1969). Note that the distribution of θ does not affect the category response function. For example, if the distribution of θ is logistic and the distribution of $Z_{i|\theta}$ is normal, the category response functions are still characterized by a normal-ogive function while Z_i is non-normally distributed (due to the non-normal θ).

Central to both the logistic and normal distribution is that they are symmetric, implying a symmetric distribution for $Z_{i|\theta}$, which results in symmetric category response functions. As a result, the category response functions increase and decrease at the same rate to their upper and lower limits, respectively, i.e., plotting the probability of answering in category c as a function of the underlying trait, θ, will result in symmetric functions for all c. This property of generalized (non-)linear (mixed) models has been questioned in general with respect to binary data (e.g., Czado & Santner, 1992; Chen, Dey, & Shao, 1999), and specifically in item response theory (IRT). Notably, Samejima (1997, 2000, 2008) pointed out that the assumption of a symmetrical distribution for $Z_{i|\theta}$ can lead to an inconsistent relation between the category location parameters, β_{ic} and estimated θ (see Samejima, 1997, 2000). As a solution, Samejima (2000, 2008) proposed the logistic positive exponent family that incorporates an acceleration parameter within the logistic GRM, which allows the category response functions to be asymmetric. In addition, Bazán, Branco, and Bolfarine (2006) (see also Bazán, Bolfarine, & Branco, 2004) obtained asymmetric category response functions by using a skewed normal distribution function for $Z_{i|\theta}$. In applying the model to a dataset on mathematics ability and on weight perception, Bazán et al. found that a model with a skewed distribution function fitted better than a model with the traditional normal distribution function. Finally, Ramsay and Abrahamowicz (1989) proposed using a spline function for the category response functions.

Past research as described above focused on asymmetric category response functions due to skewness in the conditional distribution of $Z_{i|\theta}$. We argue that asymmetry in the category response functions can also arise if the distribution of $Z_{i|\theta}$ is perfectly symmetric. Specifically, we show that when the errors of the underlying regression of Z_i on θ are heteroscedastic, the distribution of $Z_{i|\theta}$ is symmetric, but the category response functions are not. As this effect will make the marginal distribution of Z_i skewed, we also take into account that the marginal distribution could be asymmetric due to a skewed distribution of θ. However, this will not result in asymmetric category response functions.

The heteroscedastic GRM with a skewed latent trait—as presented in this paper—can be used to test for specific basis for the departure from normality in the marginal distribution of Z_i; i.e., heteroscedasticity or a skewed latent trait. The basis of the departure may be of interest for both statistical and substantive reasons.

1.1. Statistical Point of View

1.1.1. Parameter Bias   In the IRT literature, it has been shown that the presence of a non-normal latent trait will bias the parameter estimates of the discrimination parameters (Azevedo, Bolfarine, & Andrade, 2011), the item category parameters (Zwinderman & van der Wollenberg, 1990), and the ability estimates (Seong, 1990; Ree, 1979; Swaminathan & Gifford, 1983). However, some authors found no bias (Stone, 1992) which Kirisci, Hsu, and Yu (2001) attributed to the different simulation conditions used in the different studies. It was concluded that for decreasing test length and decreasing sample size, bias on the item and person parameters increase (see Kirisci et al., 2001).

With respect to heteroscedasticity in IRT, little is known as—interestingly—the subject of heteroscedasticity has not yet been investigated within this field. Heteroscedasticity has been investigated in depth in other fields including (generalized) linear regression (e.g., Agresti, 2002, p. 151; Long & Ervin, 2000), (repeated measures) analysis of (co)variance (e.g., Keselman &

Lix, 1997; Rochon, 1992), and factor analysis (e.g., Bollen, 1996; Meijer & Mooijaart, 1996; Hessen & Dolan, 2009).

As heteroscedasticity has not yet been investigated in the GRM, it is unclear whether the presence of heteroscedastic error variances will bias either item and/or person parameter estimates. We suspected that this would be the case given that heteroscedasticity has a similar effect on the distribution of Z_i compared to a non-normal latent trait. We conducted a small simulation study to investigate the effect of heteroscedasticity in the GRM on the parameter estimates. As it is not the main focus of the present paper, results are presented in the appendix. From the results it is clear that for increasing effect sizes item parameters get more biased (in the most extreme case, the root mean squared differences between the parameter estimates and the true values exceeded twice the values under the tradition GRM).

Thus, we think that, from a statistical point of view, testing for heteroscedasticity and non-normality in the latent trait within the GRM is valuable, at least to compare the results to those obtained from the traditional GRM. If the effects of heteroscedasticity and non-normality of the trait are small, and item parameters are highly similar to those obtained using the traditional GRM, one can safely rely on the traditional GRM. However, when differences are large, and effects appear to be large, it might be better to take the effects of heteroscedasticity and/or non-normality in the trait into account, which can be done using the models outlined in this paper.

1.1.2. Consistency of Estimated θ   As Samejima (1997, 2000) pointed out, estimating θ in the normal-ogive model using symmetric category response curves will result in an inconsistent relation between category location parameters, β_{ic}, and estimated θ. For simplicity, consider two items that follow a 1-parameter normal-ogive model with category location parameter β_1 for item 1 and β_2 for item 2. Assume that item 2 is more difficult than item 1, i.e., $\beta_1 < \beta_2$. Within this model Samejima (1997) showed that for subjects with $\theta < \text{Mean}(\theta)$, answering the more difficult item 2 correctly is given more credit, while for subjects with $\theta > \text{Mean}(\theta)$ answering the easier item 1 correctly is given more credit. Stated differently, for $\theta < \text{Mean}(\theta)$ answering a more difficult item correctly will result in a higher estimate of θ compared to answering an easier item correctly. In addition, for $\theta > \text{Mean}(\theta)$ answering an easier item correctly will result in a higher estimate of θ compared to answering a more difficult item correctly. When category response functions are asymmetric, this inconsistency can be overcome. For a detailed discussion, see Samejima (1997) for the 2-parameter normal-ogive model, and Samejima (2000) for the GRM. For the present undertaking, it thus seems reasonable to take the possibility of asymmetric category response functions into account, to address this inconsistency.

1.1.3. Advantage over Existing Tests   When testing for heteroscedasticity and/or for non-normality in θ, an omnibus test is less sensitive than a more specifically focused test. In an omnibus test, such as the standard tests in LISREL, see Jöreskog (2002, p. 13), or the test proposed by Muthén and Hofacker (1988), no distinction is made between non-normality due to heteroscedastic errors and/or due to a non-normal trait. We see three reasons that specific tests (on the trait and the errors) have added value relative to the omnibus tests on normality. First, omnibus tests suggest no solution when a violation is found, whereas the models in this paper can be used to take the non-normality into account. Second, it is important to distinguish between departures from non-normal Z_i due to heteroscedasticity, and departures due to a non-normal trait. Heteroscedasticity will make the category response functions asymmetric similarly to Samejima (1997, 2000, 2008), but a non-normal trait will not influence the category response functions. Third, the effects of heteroscedasticity and the non-normal trait could be in opposite directions, causing the marginal distribution of Z_i to be reasonably normal while homoscedasticity and normality of the trait are violated.

1.1.4. Non-linear Moderated Factor Analysis and Differential Item Functioning   In non-linear moderated factor analysis (Bauer & Hussong, 2009; see also Neale, 1998; Neale, Aggen, Maes, Kubarych, & Schmitt, 2006; Molenaar, Dolan, Wicherts, & van der Maas, 2010b), the parameters in a given latent variable model are a function of an external variable (e.g., age), which may moderate these parameters. These models have important applications as models for differential item functioning (DIF; Mellenbergh, 1989) with respect to a continuous background variable (see Neale et al., 2006; and Bauer & Hussong, 2009). We note that heteroscedasticity in the GRM, as formulated below, may be attributable to DIF. For instance, suppose that the error variance in the regression of a given item on the latent trait varies with the continuous background variable that is correlated with the latent trait. This heteroscedasticity may be detected and interpreted as a manifestation of DIF by conditioning on the background variable (as in Neale et al., 2006). In the unconditional application of our GRM, such DIF would give rise to heteroscedasticity in the regression of the item on the latent trait.

1.2. Substantive Point of View

Finding that a set of items that purport to measure the same underlying latent trait is associated with heteroscedasticity across the trait can have interesting implications for the theory of that trait. This applies equally well to the finding that a given latent trait has a skewed distribution. We consider three examples, which we discuss shortly below.

1.2.1. Schematicity   In personality research, schematicity refers to the hypothesis that people differ in the certainty and accuracy with which they answer to self-report measures because of differences in their cognitive structures concerned with processing information about the self (Markus, 1977; Rogers, Kuiper & Kirker, 1977; Tellegen, 1988). Research indicates that being highly schematic (i.e., having strong cognitive structures about the self) is associated with an extreme position on the trait (Markus, 1977). Note that this phenomenon implies heteroscedastic errors in the GRM, i.e., conditional variances decrease across the trait. We are not aware of any latent variable approaches to modeling schematicity. However, using the present model this is possible, as we illustrate later.

1.2.2. Gene-by-Environment Interaction   In behavior genetics, a gene-by-environment interaction (GxE) will result in heteroscedasticity of the environment factors with respect to the genetic factors (Jinks & Fulker, 1970). Within the linear factor model, van der Sluis, Dolan, Neale, Boomsma, and Posthuma (2006) propose testing for heteroscedastic errors as a test of GxE, i.e., the variance conditional on the genetic factor increases or decreases across that factor (see also Molenaar, van der Sluis, Boomsma, & Dolan, 2012). This approach is, however, limited to continuous factor indicators (e.g., subtest scores). The GRM proposed in this paper can be used to investigate GxE at the item level.

1.2.3. Ability Differentiation   This hypothesis from intelligence research refers to the claim that the general intelligence factor is not equally strong across its range (Spearman, 1927; Tucker-Drob, 2009). This phenomenon could be attributed to heteroscedastic errors, i.e., increasing conditional variances across the intelligence factor (Hessen & Dolan, 2009), non-linear factor loadings (Tucker-Drob, 2009) or a skewed ability factor (Molenaar, Dolan, & van der Maas, 2011). All previous studies have focused on the subtest level; however, using the heteroscedastic GRM with non-normal trait enables the investigation of the phenomenon at the item level.

In conclusion, the heteroscedastic GRM with non-normal trait can be used to test statistical and substantive hypotheses. The outline of the present paper is as follows: First, we present the derivation of the traditional GRM using the underlying Z_i. Next we discuss how category

response functions can be skewed. Then, we extend the GRM to incorporate heteroscedastic errors and a skewed θ. Next, we present a simulation study to show the viability of the model and to investigate the power to detect both effects. Finally, we apply the model to a questionnaire measuring affect, and to a dataset on alexithymia to test the schematicity hypothesis.

1.3. The Graded Response Model

In the GRM, it is assumed that a normally distributed variable Z_i underlies Y_i, the observed scores on item i. To enable inferences about a latent trait θ, Z_i is regressed on θ resulting in

$$Z_i = \upsilon_i + \lambda_i \theta + \varepsilon_i, \tag{1}$$

where υ_i is the intercept, λ_i is the regression weight or factor loading, and ε_i is the error term.[1] The marginal distribution of Z_i, $g(.)$, is given by

$$g(Z_i) = \int_{-\infty}^{\infty} f(Z_{i|\theta}) h(\theta) \, d\theta, \tag{2}$$

where $h(.)$ is the density function for θ, and $Z_{i|\theta}$ is the score on Z_i conditional on θ with conditional density function $f(.)$ and

$$E(Z_{i|\theta}) = \upsilon_i + \lambda_i \theta \tag{3}$$

and

$$\text{Var}(Z_{i|\theta}) = \sigma_{\varepsilon i}^2. \tag{4}$$

The probability of a subject with a given θ answering in category c on item i is given by

$$P(Y_{i|\theta} = c) = \int_{\tau_{ic}}^{\tau_{i(c+1)}} f(Z_{i|\theta}) \, dZ_{i|\theta} \quad \text{for } c = 0, \ldots, C_i - 1, \tag{5}$$

where $Y_{i|\theta}$ is the observed score on item i conditional on θ, C_i is the number of answer categories of item i, $\tau_{i0} = -\infty$, and $\tau_{i(Ci-1)} = \infty$. Equations (3), (4), and (5) give

$$P(Y_{i|\theta} = c) = F\left(\frac{\upsilon_i + \lambda_i \theta - \tau_{ic}}{\sigma_{\varepsilon i}}\right) - F\left(\frac{\upsilon_i + \lambda_i \theta - \tau_{i(c+1)}}{\sigma_{\varepsilon i}}\right), \tag{6}$$

where $F(.)$ is the conditional distribution function of $Z_{i|\theta}$. Substituting $\upsilon_i = 0, \alpha_i = \frac{\lambda_i}{\sigma_{\varepsilon i}}$, and $\beta_{ic} = -\frac{\tau_{ic}}{\sigma_{\varepsilon i}}$, we obtain

$$P(Y_{i|\theta} = c) = F(\alpha_i \theta + \beta_{ic}) - F(\alpha_i \theta + \beta_{i(c+1)}) \tag{7}$$

which is the normal-ogive GRM with item scale parameter α_i and category location parameter β_{ic}; see Takane and de Leeuw (1987), McDonald (1999).

2. Asymmetric Z_i

In Equations (1) and (2), it holds that if θ and ε_i are both normally distributed, Z_i is normally distributed by convolution. Conversely, Cramér's theorem (Cramér, 1937) ensures that when Z_i is normally distributed, θ is normally distributed and $Z_{i|\theta}$ (i.e., ε_i) is normally distributed.[2] Additionally, for a normal distribution for Z_i to hold, $Z_{i|\theta}$ should be homoscedastic (i.e., constant across θ; see Meijer & Mooijaart, 1996; Hessen & Dolan, 2009). See Figure 1 for a graphical representation for the case where $C = 3$.

---

[1]Note that the error term may include a systematic component due to misfit.

[2]Cramér's theorem states that if X_1 and X_2 are independent random variables and $X_1 + X_2$ is normally distributed, it follows that both X_1 and X_2 are normally distributed.
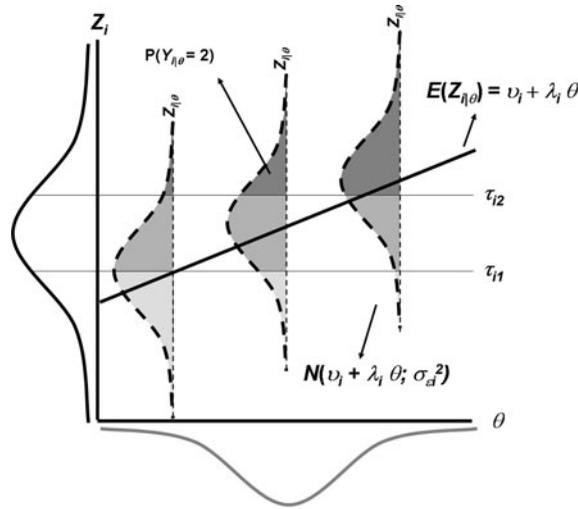
FIGURE 1.
Graphical representation of the relation between $P(Y_{i|\theta} = c)$ and $Z_{i|\theta}$ for $C = 3$. In the figure, only three of the infinite conditional distributions are depicted. $N(.)$ refers to a normal density.

Thus, if we accept the linear relation between Z_i and θ in Equation (1), then any violation of symmetry in the marginal distribution of Z_i is caused by the following not mutually exclusive causes:

1. $\sigma_{\varepsilon i}^2$ is heteroscedastic (i.e., variable across θ), see Figure 2(b).
2. The distribution of θ is skewed, see Figure 2(c).
3. The distribution of $Z_{i|\theta}$ is skewed, see Figure 2(d).

Existing approaches by Samejima (1997, 2000, 2008), Bazán et al. (2006), and Ramsay and Abrahamowicz (1989), as discussed above, focus exclusively on cause 3. However, cause 1 will also result in asymmetric category response functions. Thus, if normality of $Z_{i|\theta}$ is not rejected by one of the existing approaches, category response functions can still be asymmetric due to heteroscedastic $\sigma_{\varepsilon i}^2$. In addition, non-normality can arise in Z_i due to a skewed θ.

Specific tests on the violations as presented above are developed within the framework of factor analysis for continuous variables; Hessen and Dolan (2009) proposed a one-factor model to test for heteroscedastic errors, and Molenaar, Dolan, and Verhelst (2010a) proposed a one-factor model to test for heteroscedastic errors, non-linear factor loadings, and/or a non-normal factor distribution. In the framework of factor analysis for ordinal variables, including the GRM, such specific tests have received relatively little attention. Effort has been invested in approximating the distribution of θ using mixture distributions (Muthén & Muthén, 2007; Vermunt, 2004; Vermunt & Hagenaars, 2004; Schmitt, Mehta, Aggen, Kubarych, & Neale, 2006), extensions of the normal distribution (van den Oord, 2005; Verhelst, 2009; Azevedo et al., 2011), and splines (Woods, 2007). However, we are not aware of any work on heteroscedastic errors in the GRM.

3. Extending the GRM

We extended the GRM to include heteroscedastic $\sigma_{\varepsilon i}^2$ and a skewed θ. We present the model development in factor analysis notation, as conceptualization of the GRM in terms of an underlying Z_i variable originates in this field (Wirth & Edwards, 2007). In addition, within factor
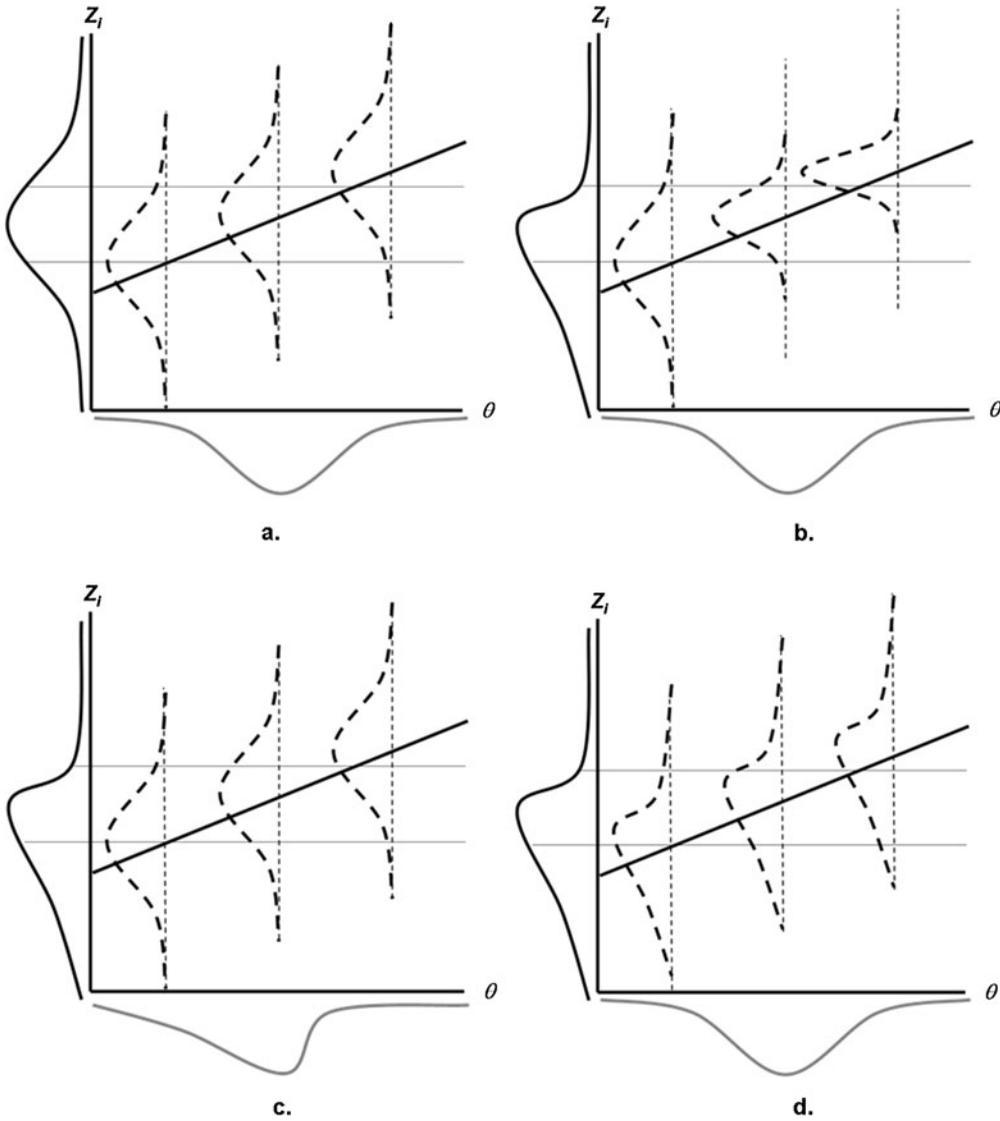
FIGURE 2.
Graphical representation of the causes of non-normality in Z_i: (**a**) traditional model (see also Figure 1), (**b**) heteroscedastic $\sigma^2_{\varepsilon i}$, (**c**) skewed θ (**d**) skewed $Z_{i|\theta}$. For clarity, effect sizes are exaggerated.

analysis, heteroscedasticity has already been studied to some degree, so that our presentation connects better with the existing literature. Still, translating the final result to IRT notation is straightforward (using the results in Takane & de Leeuw, 1987, as motivated above).

3.1. Heteroscedastic Errors

In Equations (2), (3), and (4), $\sigma^2_{\varepsilon i}$ are assumed to be homoscedastic. Thus, we propose to account for heteroscedasticity by making $\sigma^2_{\varepsilon i}$ a function of θ, i.e.,
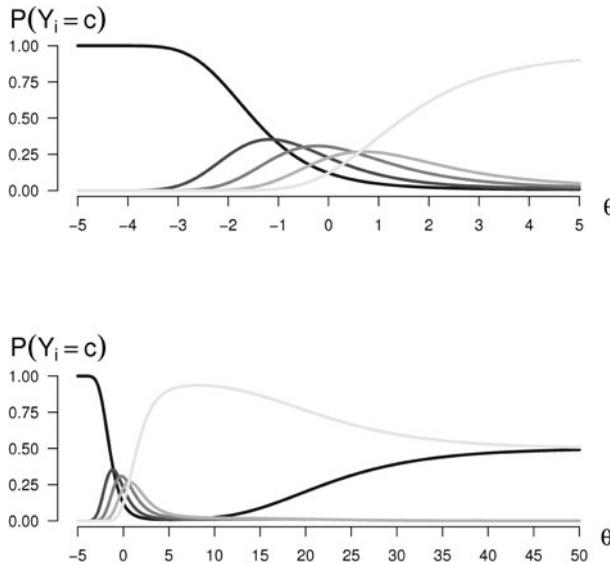
$$\sigma^2_{\varepsilon i|\theta} = k(\theta; \boldsymbol{\delta}_i),$$

FIGURE 3.
Heteroscedastic item category curves for $-5 \leq \theta \leq 5$ (*top*) and $-5 \leq \theta \leq 50$ (*bottom*).

where $\boldsymbol{\delta}_i$ is a vector of parameters with elements $\delta_{i0}, \ldots, \delta_{ir}$, and $k(.)$ is any function that is strictly positive. Within the framework of linear factor analysis for continuous variables, Hessen and Dolan (2009) used an exponential function for $k(.)$, i.e.,

$$\sigma^2_{\varepsilon i|\theta} = \exp(\delta_{i0} + \delta_{i1}\theta + \delta_{i2}\theta^2 + \cdots + \delta_{ir}\theta_r).$$

Thus, $\log(\sigma^2_{\varepsilon i})$ is modeled as an rth degree polynomial function of θ. Bauer and Hussong (2009) used a similar function to model the relation between $\sigma^2_{\varepsilon i}$ and a moderator variable. Generally, as a simple test on heteroscedasticity it will suffice to consider the minimal heteroscedasticity model, i.e., $r = 1$,

$$\sigma^2_{\varepsilon i|\theta} = \exp(\delta_{i0} + \delta_{i1}\theta), \tag{8}$$

with baseline parameter $\delta_{i0} \in (-\infty, \infty)$ and heteroscedasticity parameter $\delta_{i1} \in (-\infty, \infty)$. When $\delta_{i1} = 0$, error variances are homoscedastic. When $\delta_{i1} > 0$ error variances are increasing for increasing levels of θ, and when $\delta_{i1} < 0$ error variances are decreasing for increasing levels if θ. Note that, in principle, in the homoscedastic case of the GRM (i.e., $\delta_{i1} = 0$) and the heteroscedastic case of the GRM (i.e., $\delta_{i1} \neq 0$) all parameters are interpreted the same (the thresholds, factor loadings and intercepts), except for δ_{i0}. In the homoscedastic model, δ_{i0} represents $\log(\sigma^2_{\varepsilon i})$ for all θ's, while in the heteroscedastic model it represents $\log(\sigma^2_{\varepsilon i|\theta})$ for $\theta = 0$.

The function in Equation (8) is not generally suitable within the GRM. To see this, consider the top graph in Figure 3 in which $P(Y_i = c|\theta)$ from Equation (6) is plotted for $C_i = 5$, $\lambda_i = 1$, $\upsilon_i = 0$, $\tau_{ic} = \{-3, -1, 1, 3\}$, $-5 \leq \theta \leq 5$, and $\sigma^2_{\varepsilon i|\theta} = \exp(0.5 + 0.5 \times \theta)$. It appears that the heteroscedastic errors underlying the distribution of Z_i make the category response curves skewed. In the bottom graph of Figure 3, the behavior of the same category response functions is shown for $-5 \leq \theta \leq 50$. Clearly, the category response functions for $c = 0$ and $c = 4$ (first and last category) behave undesirably as both have an upper limit of 0.5. How this happens is clear from the category response functions in Equation (6), e.g., for $c = 0$

$$P(Y_{i|\theta} = 0) = 1 - \Phi\left(\frac{\upsilon_i + \lambda_i\theta - \tau_{i(c+1)}}{\exp(\delta_{i0} + \delta_{i1}\theta)^{\frac{1}{2}}}\right)$$

because $\tau_{i0} = -\infty$. From this equation it can be seen that when θ increases, the denominator of the fraction in the normal distribution function, $\Phi(.)$, accelerates more than the numerator (as the denominator involves an exponential function of θ and the numerator a linear function of θ). Therefore, for increasing θ the fraction approaches 0 causing $P(Y_{i|\theta} = c)$ to approach $1 - \Phi(0) = 0.5$ in the limit. Key of the problem is that category 0 has only one effective threshold (τ_{i1}) as the other threshold (τ_{i0}) equals minus infinity. Therefore, a similar problem occurs for $c = 4$ (i.e., the upper category), but not for the intermediate categories as these categories have thresholds that do not involve (minus) infinity. The problem outlined in Figure 3 appears to be occurring on a range of θ that will generally fall outside the observed data. However, when the degree of heteroscedasticity increases, the category response curves of category 0 and 4 will approach 0.5 in the more reasonable ranges for θ as well.

We therefore propose an alternative function for $k(.)$ to link $\sigma_{\varepsilon i}^2$ and θ, i.e.,

$$\sigma_{\varepsilon i|\theta}^2 = \frac{2\delta_{i0}}{1 + \exp(-\delta_{i1}\frac{\theta - E(\theta)}{SD(\theta)})} \tag{9}$$

with baseline parameter $\delta_{i0} \in [0, \infty)$ and heteroscedasticity parameter $\delta_{i1} \in (-\infty, \infty)$.

Note that when $\delta_{i1} > 0$ and $\theta \to \infty \Rightarrow \sigma_{\varepsilon i|\theta}^2 \to \delta_0$, i.e., the function has an upper bound which prevents the category functions of $c = 0$ and $c = C - 1$ to approach 0.5. In addition, when $\theta \to -\infty \Rightarrow \sigma_{\varepsilon i|\theta}^2 \to 0$, i.e., the function has a lower bound of 0. Similarly when $\delta_{i1} < 0$ and $\theta \to -\infty \Rightarrow \sigma_{\varepsilon i|\theta}^2 \to \delta_{i0}$, and when $\theta \to -\infty \Rightarrow \sigma_{\varepsilon i|\theta}^2 \to 0$. In addition, δ_{i1} is a heteroscedasticity parameter, i.e., when $\delta_{i1} = 0 \Rightarrow \sigma_{\varepsilon i|\theta}^2 = \delta_{i0}$.

3.2. Skewed Latent Trait

Within the 2-parameter logistic IRT model, Azevedo et al. (2011) proposed the use of the skew-normal density function (Azzalini, 1985, 1986; Azzalini & Capatanio, 1999; Arnold, Beaver, Groeneveld & Meeker, 1993).[3] Note that within the field of latent variable modeling, this distribution is also used by Bazán et al. (2006) to model skewness in the category response function of a 2-parameter normal-ogive model (as discussed above) and the distribution is also used by Molenaar et al. (2010a) and Molenaar et al. (2011) to approximate the latent variable distribution in the linear factor model.

An appealing feature of the class of skew-normal distributions is that it includes the normal distribution as a special case, which makes it straightforward to test a given variable for normality. Specifically, if random variable η has a skew-normal distribution, its probability density function is given by

$$h(\eta) = \frac{2}{\omega} \times \Phi\left(\zeta \frac{\eta - \kappa}{\omega}\right) \times \varphi\left(\frac{\eta - \kappa}{\omega}\right) \tag{10}$$

where κ is a location parameter, $\kappa \in (-\infty, \infty)$, ω is a scale parameter, $\omega \in [0, \infty)$, and ζ is a shape parameter, $\zeta \in (-\infty, \infty)$. From these parameters the expected value, variance, and skewness of η can be calculated using the so-called centered parameterization of the skew-normal distribution (see Azzalini, 1985)

$$E(\eta) = \kappa + \omega\sqrt{\frac{2}{\pi}}\rho, \tag{11}$$

$$\text{Var}(\eta) = \omega^2\left(1 - \frac{2\rho^2}{\pi}\right), \tag{12}$$

---

[3]Specifically, Azevedo et al. (2011) used the centered skew-normal distribution, see below.

and

$$\mathrm{Skew}(\eta) = \frac{4 - \pi}{2} \frac{(\rho\sqrt{2/\pi})^3}{(1 - 2\rho^2/\pi)^{\frac{3}{2}}},$$

with

$$\rho = \frac{\zeta}{\sqrt{1 + \zeta^2}}.$$

Note that in Equation (10), for $\zeta = 0$, $h(.)$ reduces to the normal density function. Within the heteroscedastic GRM with skewed θ, the latent trait, θ, can still be interpreted as in the traditional GRM. That is, a high position on θ in the traditional GRM corresponds with a high value on θ in the extended GRM.

3.3. Identification

The present form of the model in Equation (6), is unidentified as both θ and Z_i are unobserved variables. Identification of θ traditionally proceeds by restricting $E(\theta) = 0$, $\mathrm{Var}(\theta) = 1$, and identification of Z_i proceeds by restricting $E(Z_i) = 0$ and $\mathrm{Var}(Z_i) = 1$. Because of Equation (13), these restrictions have the following implications:

$$\sigma_{\varepsilon i|\theta}^2 = \mathrm{Var}(Z_i) - \lambda_i^2 \times \mathrm{Var}(\theta) = 1 - \lambda_i^2$$

and

$$\upsilon_i = E(Z_i) - \lambda_i \times E(\theta) = 0.$$

That is, intercepts are not in the model and the error variances are a deterministic function of the factor loadings. Thus, in the present model, we could not introduce heteroscedastic errors, as $\sigma_{\varepsilon i}^2$ is not a free parameter. We therefore identify the scale of Z_i by fixing two adjacent thresholds to some arbitrary values (see Mehta, Neale, & Flay, 2004). The scale of Z_i is now identified because the distance between the fixed thresholds define one unit on the Z_i scale. An advantage is that we can estimate $\sigma_{\varepsilon i}^2$ and υ_i for all i. We are still free to fix the scale of θ by either fixing $\mathrm{Var}(\theta)$ to equal 1, or by fixing λ_i to equal 1 for a given i. In the latter case we can estimate ω from Equation (10). In addition, in multi-group applications we can fix $E(\theta)$ to equal 0 in the first group and estimate κ from Equation (10) in the second group. However, in single group applications $E(\theta)$ (and thus κ) is fixed.

The restrictions presented above result in an identified model, as we verify in our simulation study (see below). In addition, in all applications of the model (among which are the simulations and illustrations that are presented below), using (marginal) maximum likelihood, the Hessian was positive definite, parameter recovery was satisfactory, and altering starting values did not result in a different solution with the same likelihood. We also note that within the linear one-factor model, we already established that heteroscedastic errors and a skew-normal distribution are simultaneously identified (see Molenaar et al., 2010a). Taking these results together, we are confident that the present model is identified.

3.4. Estimation

Equation (6), with $\sigma_{\varepsilon i}^2$ given by Equation (9) and θ distributed with density function given by Equation (10), constitutes the heteroscedastic GRM with a skewed trait distribution. To fit this model to data, we propose a marginal maximum likelihood procedure (MML; Bock & Aitkin, 1981). That is, we maximize the log-marginal likelihood function, i.e.,

$$\ell(\boldsymbol{\gamma}|\mathbf{X}) = \sum_{p=1}^{N} \log \int_{-\infty}^{\infty} \prod_{i=1}^{q} P(Y_{i|\theta} = x_{(pi)}) h(\theta) \, d\theta, \tag{13}$$

where $\mathbf{X}$ denotes the $N \times q$ matrix with item scores of N subjects on q items, $x_{(pi)}$ is element (p, i) from this matrix, and $\boldsymbol{\gamma}$ is the vector of parameters in the model. This vector includes: δ_{i0}, δ_{i1}, λ_i, υ_i, τ_{i2}, τ_{i3}, and ζ for all $i = 1, \ldots, q$. Note that κ and ω are not free parameters as these are fixed, such that $E(\theta) = 0$ and $\text{Var}(\theta) = 1$ according to Equations (11) and (12) as discussed above.

4. Simulation Study

In the present section, we present results of a small simulation study. The simulation study served three aims. First, we wanted to establish the viability of the model. Second, we wanted to investigate the statistical power to detect the different effects simultaneously and in isolation. Finally, we wanted to investigate the specificity of both effects, i.e., whether the models can distinguish between a dataset which is generated according to a skew-normal θ and a dataset which is generated according to heteroscedastic $\sigma_{\varepsilon i}^2$.

4.1. Design

We simulated data according to the model and varied three aspects in the data: (i) The number of respondents: $N = 400$ or $N = 800$; (ii) the number of answer categories, $C = 3$ or $C = 5$; and (iii) the nature of the effect(s) in the data: no effects in the data, heteroscedastic $\sigma_{\varepsilon i}^2$, skew-normal θ, or both. We simulated data for 10 items (i.e., $q = 10$). In the case that one or both of the effects were present in the data, we chose $\zeta = 2.17$ [Skew$(\theta) = 0.5$] and/or $\delta_{i1} = 0.4$. All other parameters were fixed to $\lambda_i = 1$, $\upsilon_i = 0$, $\delta_{i0} = 1.5$, $E(\theta) = 0$, $\text{Var}(\theta) = 1$, and $\tau_{ic} = \{-2, -0.75, 0.75, 2\}$ when $C = 5$, and $\tau_{ic} = \{-0.75, 0.75\}$ when $C = 3$. The models were identified by fixing $E(\theta) = 0$, $\text{Var}(\theta) = 1$, and $\tau_{i1} = -2$, $\tau_{i2} = -0.75$ (if $C = 5$) and $\tau_{i1} = -0.75$, $\tau_{i2} = 0.75$ (if $C = 3$).

For each of the 16 conditions ($2 \times 2 \times 4$) we simulated 40 data sets and fitted four different models: the baseline model with no effects (i.e., $\zeta = 0$ and $\delta_{i1} = 0$ for all i), a heteroscedasticity model (i.e., δ_{i1} is free for all i but $\zeta = 0$), a skewed-trait model (i.e., ζ is free but $\delta_{i1} = 0$ for all i), and a full model with both effects simultaneously (i.e., ζ is free and δ_{i1} is free for all i).

4.2. Likelihood Ratio and Power Calculation

In testing for skewness in the distribution of θ and/or heteroscedasticity of the errors, we use the likelihood ratio test (LRT). In this test, H$_0$ states that $\zeta = 0$ and/or $\delta_{i1} = 0$ for some or all of the items, and H$_A$ states that $\zeta \neq 0$ and/or $\delta_{i1} \neq 0$. The LRT statistic, T, is calculated as

$$T = -2 \times \left[\ell(\hat{\boldsymbol{\gamma}}_0 | \mathbf{X}) - \ell(\hat{\boldsymbol{\gamma}}_4 | \mathbf{X}) \right], \tag{14}$$

where $\hat{\boldsymbol{\gamma}}_0$ is the estimated vector of parameters from Equation (13) subjected to the parameter restrictions of H$_0$, and $\hat{\boldsymbol{\gamma}}_A$ is the vector of parameter estimates under H$_A$. Under H$_0$, T has a central-χ^2 distribution with degrees of freedom (df) equal to the number of restrictions in H$_0$. Under H$_A$, this statistic has a non-central χ^2 distribution with a non-centrality parameter that depends on the effect size and sample size. Two important conditions that need to be satisfied for the LRT statistic to approach these theoretical distributions are (i) that H$_0$ is nested under H$_A$ and (ii) that the restrictions in H$_0$ should not be on a boundary of the parameter space (Cramér, 1946). As our likelihood ratio tests concern either fixing ζ with parameter space $(-\infty, \infty)$ to 0, and/or fixing δ_{i1} with parameter space $(-\infty, \infty)$ to 0, these regularity conditions are satisfied.

In the present simulations, we study the empirical power of the LRT to reject H$_0$ of a normal θ and/or homoscedastic ε_i. To this end, we calculated T from Equation (14) for each replication in the design. As T has a non-central χ^2 distribution under H$_A$, it holds that (Fisher, 1928):

$$E(T) = \text{df} + \text{ncp}.$$

Thus, the non-centrality parameter (ncp) of the LRT statistic, T, could empirically be approximated by subtracting the df from the average T across the replications in a given cell in the simulation study. As now the ncp of the distribution of T is known under H_A, we can choose a nominal alpha level (α) and calculate the power to reject H_0 by integrating over the non-central χ^2 distribution. Within covariance structure modeling, Satorra and Saris (1985), (see Molenaar, Dolan, & Wicherts, 2009, for an illustrative example) developed an analytical approach to calculate the ncp given that N is large, and that the misspecification under H_0 is not severe. We use essentially the same approach as that of Satorra and Saris; however, because we have no analytical expression for the ncp, we therefore establish the ncp empirically from the data, as described above.

We used Gauss–Hermite quadratures with 50 quadrature points to approximate the integral in the log marginal likelihood function in Equation (13). We fitted the model using the freely available software package Mx (Neale, Boker, Xie, & Maes, 2002). A script to fit the heteroscedastic GRM with skewed θ is available from the website of the first author.

4.3. Results

In Figures 4 and 5, box plots of the parameter estimates are shown for the heteroscedasticity parameters, δ_{i1}, and the skewness parameter, ζ, for the case that both effects were in the data and the full model was fitted to that data. As it appears, true parameter values are recovered quite well. Variability in the parameter estimates is highest when $N = 400$ and $C = 3$ but decreases with increasing N and increasing C.

In Table 1, power coefficients ($\alpha = 0.05$) are shown for the 16 conditions in the simulation study and for three classes of models. The first class of models is the full model, i.e., the heteroscedastic GRM with a skewed trait (abbreviated het-GRM-skew in Table 1), second is the heteroscedastic GRM (het-GRM; i.e., a GRM with heteroscedastic errors only), and third is the GRM with a skewed trait only (GRM-skew). Ideally, when a given effect is in the data (e.g., heteroscedastic $\sigma_{\varepsilon i}^2$) it is not detected by a misspecified model (i.e., the GRM-skew). Thus, power of the misspecified model should ideally be equal to 0.05 (i.e., equal to the Type I error probability specified a priori). If this is the case, the model is said to be highly specific, that is, it detects only the effects that are in the model and it does not pick up any additional effects. In Table 1, the power coefficients of the misspecified models are underlined and are desired to be close to the nominal alpha level. All other coefficients should be reasonably large.

As can be seen from the table, all underlined values are close to or equal to 0.05 for the het-GRM-skew, which suggests that the two effects (heteroscedastic $\sigma_{\varepsilon i}^2$ and skewed θ) are well separable within this model. For the het-GRM and GRM-skew, the underlined values are departing somewhat more from the 0.05 level (at most 0.31). This indicates that unmodeled skewness in the trait or heteroscedasticity in the errors will increase the false positive rate somewhat.

For $N = 400$ and $C = 3$, power values are around 0.4–0.5 which is judged to be low power (exception is the 0.74 power of the GRM-skew to detect the skewness in θ, which is already acceptable). For $C = 5$, power is acceptable, as power coefficients vary around 0.7–0.8. An exception is the power of the het-GRM-skew to detect the skewed θ in the data. This situation is associated with a power of 0.46 which is still low. However, this is not surprising as the model has 11 extra parameters compared to the standard GRM of which only 1 is concerned with the skewed θ. Power is thus lowered because of the 10 parameters that remain unaffected. For $N = 800$ power is judged to be acceptable-to-high as coefficients are around 0.8–1.0, with the larger power coefficients when $C = 5$.

From Table 1 it also appears that when both effects are in the data, the het-GRM-skew outperforms both the het-GRM and GRM-skew in terms of power. For instance, in the cases that $N = 800$ and $C = 3$, power to detect the heteroscedastic errors is 0.72 and 0.67, respectively, in the het-GRM-skew, compared to 0.53 and 0.29, respectively, in the het-GRM. In addition,
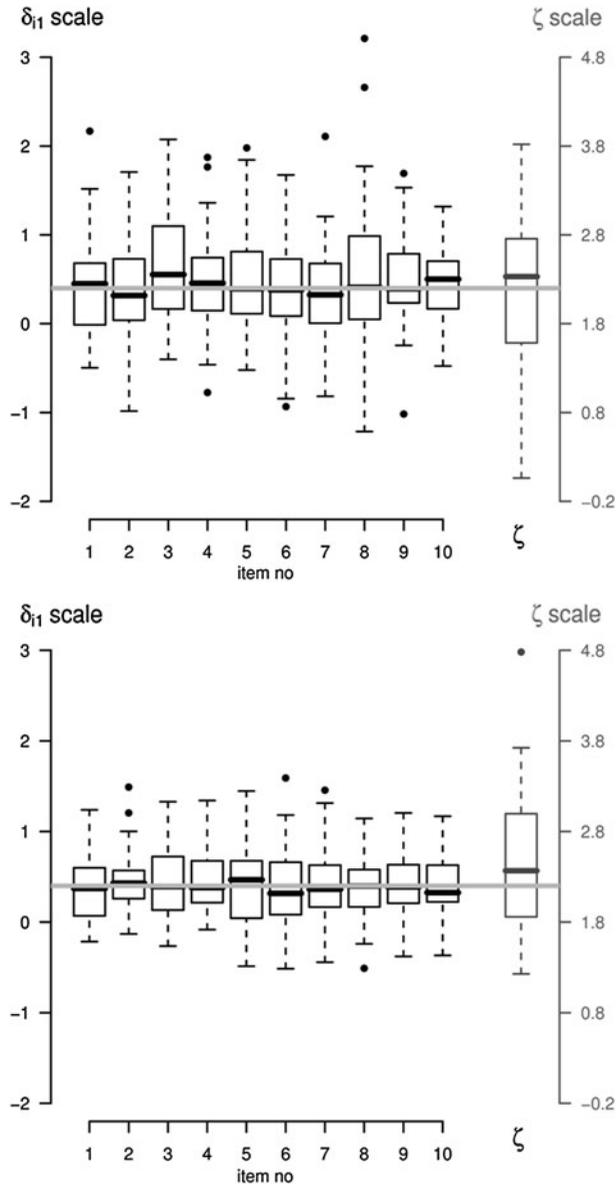
FIGURE 4.

Box plots of the parameter estimates of the heteroscedasticity parameters, δ_{i1} (in *black*; *left axis*) and the skewness parameter, ζ (in *grey*, *right axis*) for $N = 400$ and respectively $C = 3$ (*top figure*) and $C = 5$ (*bottom figure*). The *horizontal line* represents the true value on both the δ_{i1} axis and the ζ axis.

power of the het-GRM-skew to detect the skewness in θ is 0.84 compared to 0.25 in the GRM-skew. When only one effect is in the data but the het-GRM-skew is applied nevertheless (which includes both effects), power is somewhat smaller compared to the GRM-skew and het-GRM. For instance, when $N = 800$ and $C = 3$ and a skewed θ underlies the data, the het-GRM-skew has a power of 0.88 to detect the effect, while the GRM-skew has a power of 0.96. In addition, when heteroscedastic errors underlie the data, the het-GRM skew has power of 0.82 to detect this effect, and the het-GRM has a power of 0.88. In the case of one effect in the data, power is, thus, somewhat increased in the models with a single effect only. However, benefits are not that large
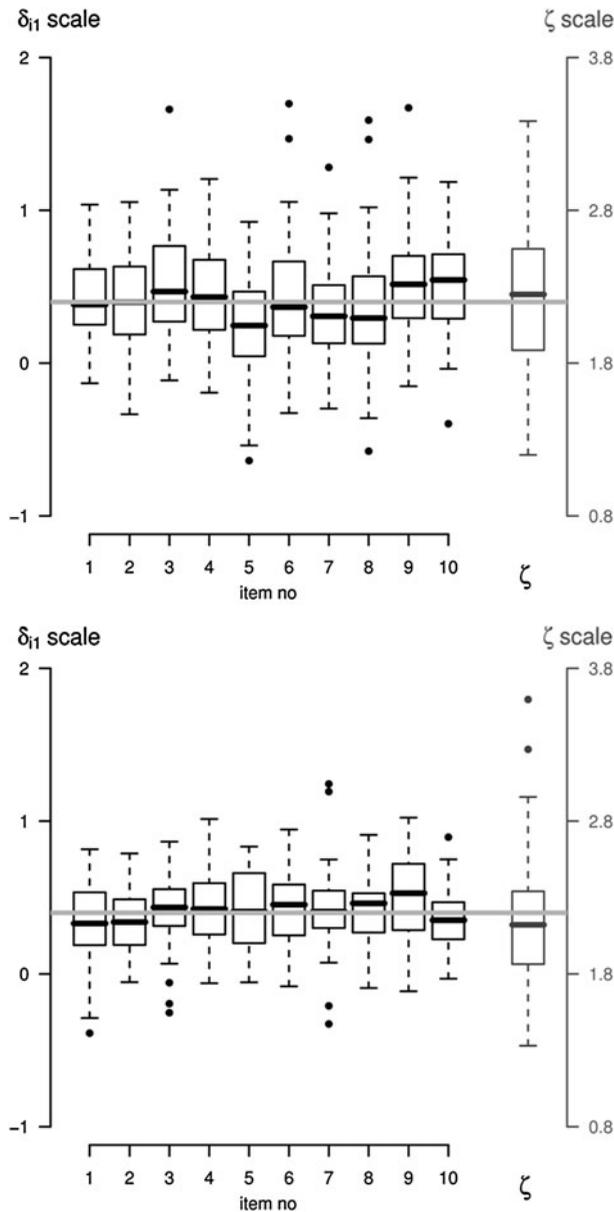
FIGURE 5.
Box plots of the parameter estimates of the heteroscedasticity parameters, δ_{i1} (in *black*; *left axis*) and the skewness parameter, ζ (in *grey*, *right axis*) for $N = 800$ and respectively $C = 3$ (*top figure*) and $C = 5$ (*bottom figure*). The *horizontal line* represents the true value on both the δ_{i1} axis and the ζ axis.

as compared to when both effects are in the data and the het-GRM-skew is applied. It seems, thus, to be advisable that if one is interested in detecting one effect, e.g., heteroscedastic errors, it is safest to take the possibility of the skewed θ into account as well because of substantial power gains.

Results above depend highly on the present choice of effect sizes. However, we showed that given the chosen effect sizes power can be acceptable for reasonable sample sizes. Regardless of

TABLE 1.
Power to detect heteroscedastic errors and skewness in the trait simultaneously and in isolation.

N	C	het-GRM-skew				het-GRM	GRM-skew
		effect	Heterosc. σ	Skew θ	both	Heterosc. σ	Skew θ
400	3	none	0.09 (1.23)	0.05 (0.04)	0.09 (1.19)	0.09 (1.15)	0.05 (0.00)
		heterosc. σ	0.46 (8.40)	0.05 (0.00)	0.47 (8.94)	0.50 (9.23)	0.11 (0.54)
		skew θ	0.06 (0.24)	0.62 (5.14)	0.36 (6.93)	0.11 (1.79)	0.74 (6.70)
		both	0.42 (7.78)	0.56 (4.51)	0.43 (8.27)	0.20 (3.76)	0.11 (0.49)
	5	none	0.05 (0.00)	0.05 (0.00)	0.05 (0.00)	0.05 (0.00)	0.05 (0.00)
		heterosc. σ	0.79 (15.87)	0.07 (0.14)	0.80 (16.76)	0.81 (16.62)	0.16 (0.89)
		skew θ	0.06 (0.42)	0.82 (8.35)	0.46 (8.87)	0.07 (0.52)	0.83 (8.45)
		both	0.72 (14.04)	0.88 (9.77)	0.87 (19.54)	0.53 (9.77)	0.65 (5.50)
800	3	none	0.07 (0.53)	0.05 (0.00)	0.06 (0.34)	0.07 (0.62)	0.05 (0.00)
		heterosc. σ	0.82 (16.94)	0.05 (0.00)	0.86 (19.09)	0.88 (19.26)	0.31 (2.15)
		skew θ	0.06 (0.24)	0.88 (9.78)	0.70 (13.83)	0.22 (4.05)	0.96 (13.59)
		both	0.67 (12.59)	0.84 (8.84)	0.71 (14.22)	0.29 (5.38)	0.25 (1.63)
	5	none	0.05 (0.12)	0.05 (0.00)	0.05 (0.00)	0.05 (0.01)	0.05 (0.00)
		heterosc. σ	0.95 (24.90)	0.05 (0.00)	0.96 (26.19)	0.97 (26.44)	0.21 (1.30)
		skew θ	0.08 (1.03)	0.98 (16.60)	0.84 (18.36)	0.11 (1.76)	0.99 (17.33)
		both	0.99 (32.14)	0.98 (16.31)	1.00 (40.38)	0.95 (24.08)	0.82 (8.24)

Note. Non-centrality parameters are in parentheses. 'het-GRM-skew' denotes 'heteroscedastic GRM with a skewed trait'. 'heterosc. σ' denotes the power to detect the heteroscedastic $\sigma_{\varepsilon i}^2$, 'skew θ' denotes the power to detect skweness in the trait, and 'both' denotes the power to detect both effects simultaneously. The power coefficients that are underlined should be close to the Type I error probability specified a priori, i.e., $\alpha = 0.05$, see the text for explanation.

effect size, we can conclude that the model is viable and that the effects are highly specific and, therefore, well separable.

In the present simulation study, we only used effects on θ and $\sigma_{\varepsilon i}^2$ that are in the same direction, i.e., they both make the distribution of Z_i positively skewed. We did consider the case in which both effects were in opposite directions (specifically, $\zeta = -2.17$ and $\delta_{i1} = 0.4$), but this did not affect the results as presented above. This finding is not surprising as both effects are highly specific and, therefore, do not interfere with each other.

5. Illustration

In this section, we present two applications purported to test the schematicity hypothesis from personality research (Markus, 1977; Rogers et al., 1977; Tellegen, 1988). As discussed above, this hypothesis supposes that people who are low on a given personality trait have a less clear self-schemata, i.e., they are more ambiguous about their trait level compared to people high on that trait. In the literature, a similar phenomenon is the hypothesis that not all personality traits apply equally well to everybody (Allport, 1937; Baumeister & Tice, 1988). As discussed previously, the schematicity hypothesis predicts that people who are low on a given personality trait are less accurate in reporting their exact position on the answer scale of a given item from a questionnaire. On the item level, this prediction implies heteroscedasticity of the error variances in the GRM. In the applications below, we investigate the schematicity hypothesis by testing for heteroscedastic errors. We do also take the possibility of skewness in the trait into account as this might benefit the power to detect the heteroscedastic errors (as we showed in the simulation).

The first application involves an investigation of the schematicity hypothesis in an affect questionnaire, and the second application involves an investigation of the schematicity hypothesis in an alexithymia questionnaire.

6. Application 1

6.1. Description of the Data

A total of 557 psychology freshman completed a Dutch version of the Positive Affect and Negative Affect scale (Guadagnoli & Mor, 1989). All items in these scales consist of words that describe a particular affect (e.g., wistful, desperate, happy, cheerful). Respondents are asked to report on a 5-point Likert scale to what degree they experienced each affect during past week. The questionnaire contains 40 items in total of which 20 purport to measure positive affect and 20 items purport to measure negative affect. Here we limit the analysis to the Negative Affect scale only.[4] The negative affect scale was judged to have acceptable fit to a discrete one-factor model (RMSEA = 0.066).[5] We identified the model by fixing $E(\theta) = 0$, $\text{Var}(\theta) = 1$, $\tau_{i0} = -2$, and $\tau_{i1} = -0.75$.

In the data analysis, we proceeded as follows: First, we fitted the full model (including the skew-normal distribution for θ and the heteroscedastic $\sigma^2_{\varepsilon i}$). From the full model, we dropped the skewness parameter, ζ, resulting in a model with heteroscedastic errors only. Next, from the full model, we dropped the heteroscedastic errors, resulting in a model with a skewed θ only. Finally, we fitted a model with no effects, i.e., the traditional GRM. To see which of these models fitted the data best, we considered various fit indices. First, we conducted the LRT between the model of interest and the full model as described above. Next, we consulted Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both the AIC and BIC are calculated on basis of the marginal likelihood, i.e., the incidental parameters, θ, are integrated out and are not included in these indices. For the AIC and BIC, a lower value suggests a better model fit. Note that these indices are not tests in a statistical sense (e.g., Konishi & Kitagawa, 2008), i.e., they are used to decide which model predicts future data best among competing models. For present purposes we adopt this model selection based strategy using the AIC, BIC and LRT. However, we note that a more thorough approach would be to test whether data are a true realization of the stochastic model under investigation (Schmueli, 2010).

6.2. Results

Table 2 contains the parameter estimates of υ_i, τ_{i2}, τ_{i3}, δ_{0i}, δ_{1i}, and ζ in the full model. From Table 3, it can be seen that dropping the skewness parameter, ζ, from the full model did not result in a significant LRT. However, dropping the heteroscedasticity parameters, δ_{i1}, did result in a significant LRT. Thus, judged by the LRT we would choose a model with heteroscedastic errors only. The AIC provides the same conclusion; however, judged by the BIC the model with no effects should be preferred. Thus, results concerning the Negative Affect scale are mixed. The observation that the AIC and the BIC do not agree can be attributed to differences in the penalty that is used in these indices. BIC is known to favor more parsimonious models, while the AIC tends to favor less parsimonious models. We tend to favor the model without effects as this model is more parsimonious, i.e., we put more weight on the BIC. We can thus conclude that we did not find any support for the schematicity hypothesis with respect to self-reported negative affect.

---

[4]We note however that for the Positive Affect scale, results are similar.

[5]Because of the illustrational purposes, we judge an RMSEA smaller than 0.08 to be an indication of acceptable model fit (see Schermelleh-Engel, Moosbrugger & Müller, 2003).

TABLE 2.
Parameter estimates of the full model in Application 1.

Item	v_i	τ_{i2}	τ_{i3}	λ_i	δ_{0i}	δ_{1i}
1	−1.59	−0.09	1.05	1.23	0.65	0.27
2	−1.91	−0.34	0.65	1.18	0.58	−0.45
3	−2.16	−0.05	1.06	1.33	0.94	−0.13
4	−1.77	−0.07	1.16	1.33	0.40	1.39
5	−1.80	−0.06	0.92	1.28	0.52	−0.48
6	−1.84	−0.19	0.70	0.84	1.22	−0.27
7	−2.17	0.13	1.11	0.83	1.03	−0.05
8	−0.96	−0.21	1.38	0.92	0.74	1.17
9	−1.45	−0.02	1.81	0.79	0.60	1.62
10	−1.49	−0.08	1.31	0.83	0.67	1.49
11	−1.41	−0.27	1.05	0.70	0.49	0.61
12	−1.89	−0.12	1.69	0.89	0.70	0.78
13	−1.20	−0.25	0.95	0.81	0.69	0.12
14	−1.58	−0.04	0.86	1.17	0.72	−0.09
15	−1.54	−0.09	1.02	0.95	0.36	2.37
16	−1.04	−0.32	0.96	0.71	0.48	0.35
17	−1.26	−0.31	0.80	0.94	1.34	−0.20
18	−2.13	0.42	1.42	1.41	0.99	0.24
19	−1.55	−0.44	0.42	1.03	0.30	0.09
20	−1.34	−0.33	0.69	0.68	0.92	0.02
ζ	−0.70					

TABLE 3.
Modeling results of Application 1.

Model	-2ℓ	df	LRT	AIC	BIC
Full model	23518.31	10901	–	1716.31	−22701.99
Heterosc. σ	23520.14	10902	$\chi^2(1) = \mathbf{1.84}$	**1716.14**	−22704.23
Skew θ	23606.28	10921	$\chi^2(20) = 87.97$	1764.28	−22721.23
No effects	23606.29	10922	$\chi^2(21) = 87.99$	1762.29	**−22724.38**

Note. LRT for a given model concerns a likelihood ratio test between that model and the 'full model'. For each fit index, the best value is boldfaced.

7. Application 2

7.1. Description of the Data

A sample purported to be representative of the Dutch population ($N = 816$) completed the Bermond–Vorst alexithymia questionnaire (Vorst & Bermond, 2001) at home on a computer.[6] Alexithymia is a personality trait that reflects the inability to understand, process, or describe emotions. The Bermond–Vorst alexithymia questionnaire consists of 6 subscales: Emotionalizing, Fantasizing, Identifying, Verbalizing, Cognitive Analyzing, and Affective Analyzing. Each

---

[6]This sample was selected from a data set that is much larger ($N = 5780$). The original data included various manipulations (15 in total). We selected subjects from 2 conditions ($N_1 = 410$ and $N_2 = 406$) that appeared to be homogeneous with respect to the manipulation (specifically, the subjects we selected completed the questionnaire with respectively a 'scroll' answer scale and a 'static' answer scale).

TABLE 4.
Parameter estimates of the full model in Application 2.

Scale	ζ		Item no.							
			1	2	3	4	5	6	7	8
Fantasizing	−1.22	υ_i	−0.86	−0.20	−0.53	−0.47	−0.76	−0.37	−0.67	−0.57
		τ_{i2}	−0.10	0.12	0.47	−0.14	0.18	−0.07	0.23	−0.02
		τ_{i3}	1.13	1.15	2.25	0.91	1.60	1.68	1.83	1.04
		λ_i	−0.50	−0.74	1.17	0.50	−0.85	−0.67	1.07	0.62
		δ_{0i}	0.99	0.40	0.61	0.76	0.73	0.73	0.45	0.49
		δ_{1i}	−0.49	1.20	3.85	−0.11	−0.66	−0.13	3.19	−0.10
Cognitive analyzing	−3.47	υ_i	1.84	0.72	−0.71	−0.94	−0.38	1.47	−1.01	−0.71
		τ_{i2}	0.68	0.56	−0.26	−0.07	0.15	0.52	−0.05	−0.24
		τ_{i3}	6.10	3.94	0.35	1.14	1.79	4.65	0.89	0.62
		λ_i	−1.77	−1.24	0.43	0.52	−0.56	−1.63	0.66	0.36
		δ_{0i}	2.31	1.45	0.32	0.36	0.67	1.47	0.40	0.32
		δ_{1i}	−2.42	−1.39	−1.56	0.55	−0.62	−1.32	−0.17	−0.81
Affective analyzing	−2.05	υ_i	−1.00	1.13	−1.15	0.82	−1.03	2.99	−1.27	1.54
		τ_{i2}	−0.28	0.60	−0.31	0.60	−0.21	1.23	−0.33	0.60
		τ_{i3}	0.30	3.81	0.10	3.27	0.38	6.55	0.11	2.97
		λ_i	0.47	−1.12	0.55	−0.81	0.52	−1.96	0.53	−1.00
		δ_{0i}	0.35	1.60	0.24	1.23	0.28	3.00	0.34	0.67
		δ_{1i}	−1.43	−1.45	−1.72	−0.87	−0.91	−1.47	−1.68	0.42
Identifying	−2.09	υ_i	−0.88	−0.88	1.69	0.98	−1.03	−0.85	0.44	1.90
		τ_{i2}	−0.34	−0.27	0.54	0.33	−0.28	−0.26	0.26	0.68
		τ_{i3}	0.08	0.27	5.28	4.09	0.29	0.18	2.42	4.16
		λ_i	0.39	0.36	−1.35	−0.93	0.51	0.37	−0.50	−0.79
		δ_{0i}	0.20	0.27	3.56	1.68	0.38	0.23	1.10	2.38
		δ_{1i}	−1.92	−1.14	−1.84	−1.57	−1.41	−1.58	−0.30	−0.33

subscale consists of 8 items measured on a 5-point Likert scale. An example of an item in the Identifying subscale is: "If I am nervous, it is unclear to me which emotion caused this feeling."

We analyzed all subscales separately. A discrete one-factor model was judged to fit in an acceptable way to the Fantasizing (RMSEA = 0.061), Identifying (RMSEA = 0.073), Affective Analyzing (RMSEA = 0.064), and Cognitive Analyzing (RMSEA = 0.060) subscales. The subscales Emotionalizing (RMSEA = 0.110) and Verbalizing (RMSEA = 0.091) were judged to fit poor to a one-factor model and are, therefore, not included in the analysis. We identified the model by fixing $E(\theta) = 0$, $\text{Var}(\theta) = 1$, $\tau_{i0} = -2$, and $\tau_{i1} = -0.75$.

7.2. Results

Table 4 contains the parameter estimates of υ_i, τ_{i2}, τ_{i3}, δ_{0i}, δ_{1i}, and ζ in the full model for the six subscales. In Table 5, modeling results are given. We followed the same procedure as in Application 1. As shown in Table 5, for all subscales, the full model appears to fit best according to the AIC and BIC. See Figure 6 for the model implied category response functions for the Affective Analyzing subscale under the full model.

Taken together, in contrast to Application 1, conclusions are relatively clear in the sense that we can conclude that for all 4 subscales analyzed, the distribution of Z_i is characterized by both heteroscedastic errors and a skewed θ distribution. Additionally, we tested for equality of δ_{1i} to see whether we could model the heteroscedasticity with a single parameter for all items instead of an individual parameter for each item. We found that this equality restriction held

TABLE 5.
Modeling results of Application 2.

	Full model		Heterosc. σ		Skew θ		No effects	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Fantasizing	**3053.8**	**−13713.1**	3065.36	−13709.62	3132.48	−13692.53	3145.79	−13688.23
Cognitive analyzing	**1269.75**	**−14605.1**	1296.76	−14593.92	1400.78	−14558.38	1403.32	−14559.46
Affective analyzing	**49.15**	**−15215.4**	105.91	−15189.35	260.91	−15128.31	279.17	−15121.54
Identifying	**1272.38**	**−14603.76**	1296.21	−14594.20	1384.96	−14566.29	1387.15	−14567.55

Note. For each fit index, the best value for each of the four subscales is boldfaced.
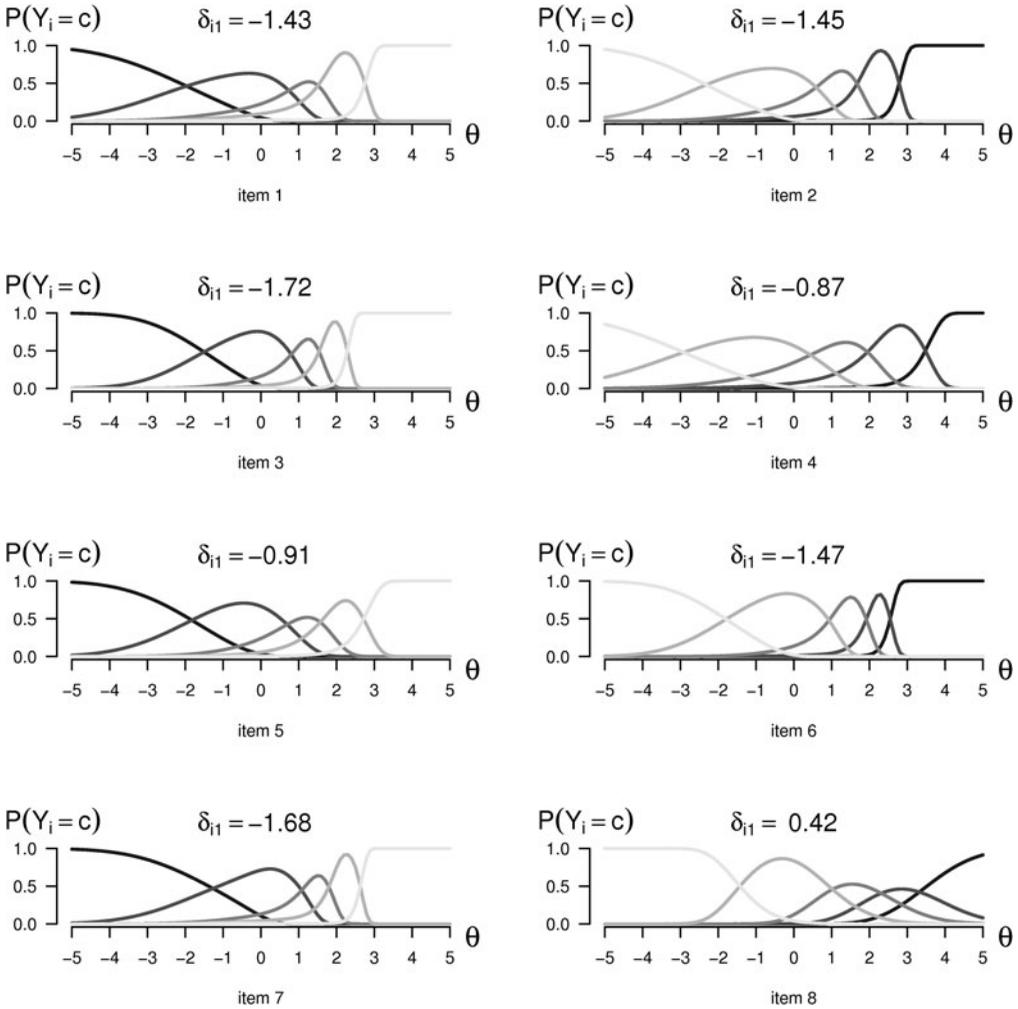

FIGURE 6.
Model implied item category functions for the Affective Analyzing subscale of Application 2. Estimated heteroscedasticity parameters, δ_{i1}, are given in the figure. Estimated baseline parameters, δ_{i0}, are 0.35, 1.60, 0.24, 1.23, 0.28, 3.00, 0.34, and 0.67 for, respectively, items 1 to 8. See also Table 4. Category 0 is *colored darkest*. Items 2, 4, 6, and 8 are contra indicative items (i.e., $\lambda_i < 0$), as they are worded as such in the questionnaire.

for the Cognitive Analyzing subscale only (AIC reduced to 1267.99 and the BIC reduced to -14622.42). For the other subscales, δ_{1i} differed between items.

It can be seen in Table 4 that for the Affective Analyzing, Cognitive Analyzing, and Identifying subscales the heteroscedasticity is systematically in the same direction for almost all items.[7] That is, lower levels of θ are associated with higher error variance. Thus individuals with high levels of alexithymia are more consistent in their answers compared to individuals with lower levels of alexithymia. This finding is in line with the schematicity hypothesis, suggesting that high alexithymia individuals are less ambiguous about their psychological functioning in this regard because they are in little doubt concerning their lack of insight and knowledge of the emotions that they experience. Allport (1937) and Baumeister and Tice (1988) questioned whether personality traits do apply equally well to everybody. In perspective of this line of work, the present results suggest that for people that are well able to describe their emotions, the alexithymia trait applies less well (i.e., is a weaker source of individuals differences) compared to people that are less able to describe their emotion.

8. Discussion

In the present paper, we presented the Heteroscedastic Graded Response Model with a skewed latent trait—a unified model that extends the traditional Graded Response Model by adding heteroscedastic errors and a non-normal latent trait. In introducing heteroscedastic errors, we specified a parametric function between the trait and the error variances. As we showed, this choice is not straightforward, as the commonly used exponential function (Hessen & Dolan, 2009; Bauer & Hussong, 2009) gives undesirable results in the limits of the trait. We therefore adopted a logistic function, which was shown to be adequate. However, the choice for this function was pragmatic. The function could be replaced with any other function that behaves reasonably given the purposes of the researcher.

Although the logistic function performs well in the limits of the trait, it has some drawbacks. When effect sizes are very large (i.e., the logistic function tends to a step function), category response functions of some categories will show patterns that are not realistic from a psychological or measurement point of view. That is, category response curves will have more than one stationary point; for an example, see Figure 7. However, we stress that this occurs only given very large effect sizes, which we consider unrealistic. In addition, in applications of the model, this behavior of the category response function can easily be detected by inspecting relevant plots of the results. When a given item shows these patterns, a different function can be used for this item or all items. Or a non-parametric approach could be considered, as explained in the following.

An alternative to specifying a parametric function between the trait and the error variances is to consider a non-parametric approach. In this approach, heteroscedasticity is detected using a step function. An advantage is that the method is more flexible, i.e., highly specific forms of heteroscedasticity can be tested depending on the substantive theory and/or the aims of the researcher. However, the method will likely be associated with smaller power as compared to the parametric function approach, particularly as the number of steps in the step function increases. This remains to be investigated in depth.

With respect to the distribution of the trait, we considered the skew-normal distribution because, first, this distribution enables straightforward tests on normality; second, its statistical properties are well documented (e.g., Azzalini & Dalla Valle, 1996; Azzalini & Capatanio, 1999;

---

[7]For the Cognitive Analyzing and Affective Analyzing subscales, only one item is associated with heteroscedasticity in the opposing direction.
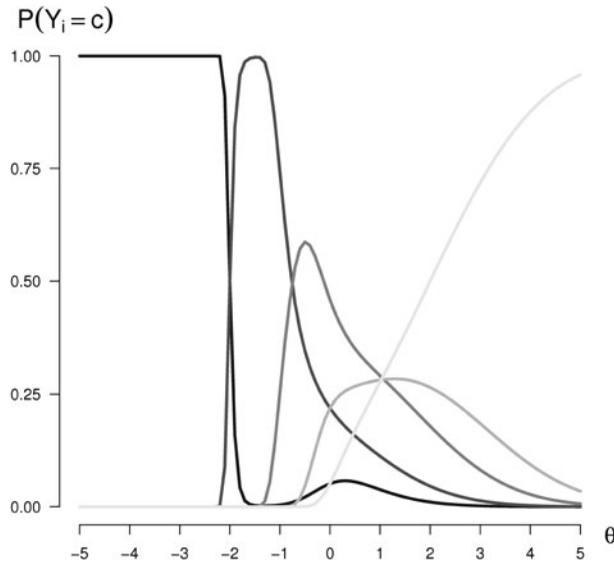
FIGURE 7.

Example of category response functions in case of an extreme effect size for the heteroscedasticity in the error variances. Parameters used to create the picture are: $\delta_{i0} = 1.5$, $\delta_{i1} = 3$, $\lambda_i = 1$, $\upsilon_i = 0$, and $\tau_{ic} = \{-2, -0.75, 0, 0.75, 2\}$.

Arnold et al., 1993; Arnold & Beaver, 2002; Chiogna, 2005; Monti, 2003); and, third, the distribution is applied in general (for examples, see Azzalini, 2005) and specifically with respect to psychometrics (e.g., Bazán et al., 2006; Molenaar et al., 2010b, 2011; Azevedo et al., 2011). However, our choice is based on pragmatic considerations; other distributions can be considered as well, e.g., mixture distributions (Muthén & Muthén, 2007; Vermunt, 2004; Vermunt & Hagenaars, 2004; Schmitt et al., 2006), the shifted-log normal (Verhelst, 2009), or distributions based on Johnson curves (van den Oord, 2005) or splines (Woods, 2007).

Acknowledgements

Appendix

We conducted a small simulation study to see how the presence of heteroscedasticity influences the parameter estimates of the traditional GRM. In this simulation study, we simulated 50 data sets according to the extended GRM. We simulated either heteroscedastic residuals or homoscedastic residuals. In case of heteroscedastic residuals, we simulated either a small ($\delta_1 = 0.5$), medium ($\delta_1 = 1$), or large effect size ($\delta_1 = 1.5$). Please note that the large effect size is still realistic, as we found such effect sizes in Application 2 of the manuscript (for instance, for the Cognitive Analyzing subscale 5 of the 7 items showed effects around $\delta_1 = -1.5$, see

TABLE A.1.
Root mean squared difference between estimated discrimination parameters (using the traditional GRM) and true discrimination parameters (within the het-GRM-skew).

N	Effect	Discrimination parameters of item no.:									
		1	2	3	4	5	6	7	8	9	10
1000	no	0.13	0.14	0.14	0.13	0.12	0.12	0.13	0.12	0.13	0.15
	small	0.15	0.14	0.13	0.13	0.14	0.13	0.15	0.14	0.13	0.14
	medium	0.18	0.19	0.17	0.17	0.18	0.19	0.18	0.15	0.16	0.18
	large	0.22	0.21	0.22	0.22	0.21	0.21	0.20	0.20	0.21	0.25
3000	no	0.07	0.08	0.08	0.08	0.08	0.08	0.07	0.08	0.07	0.07
	small	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.10
	medium	0.13	0.12	0.14	0.13	0.12	0.11	0.12	0.11	0.13	0.12
	large	0.18	0.19	0.17	0.18	0.18	0.18	0.19	0.18	0.19	0.17

Table 4). We carried out this procedure for $N = 1000$ and $N = 3000$. Next, we fitted the traditional GRM to see how parameter estimates differed from the true values. We found effects on the discrimination parameters and the thresholds, but not on the trait estimates.

From Table A.1 it can be seen that a small degree of heteroscedasticity does not have much effect, but when the heteroscedasticity increases, the root mean squared difference (RMSD) increases when it is not recognized. In the case of a large degree of heteroscedasticity and $N = 3000$, the RMSD is more than two times larger (0.17 to 0.19) compared to the case in which they are homoscedastic (0.07 to 0.08). We found negligible effects on the standard errors (both the empirical standard errors and the theoretical standard errors). It is clear that the discrimination parameters of the traditional GRM are biased if the true model is heteroscedastic. For the threshold, we found even larger effects but this is to be expected as the non-normality in the data is partly captured by the threshold. As already mentioned, for the trait estimates we found no (clear) effects.

References

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

Allport, G.W. (1937). *Personality. A psychological interpretation*. New York: Henry Holt.

Arnold, B., & Beaver, R. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test*, *11*, 7–54.

Arnold, B.C., Beaver, R.J., Groeneveld, R.A., & Meeker, W.Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, *58*, 471–488.

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*, 199–208.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, *32*, 159–188.

Azzalini, A., & Capatanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society. Series B*, *61*, 579–602.

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*, 715–726.

Azevedo, C.L.N., Bolfarine, H., & Andrade, D.F. (2011). Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics & Data Analysis*, *55*, 353–365.

Bauer, D.J., & Hussong, A.M. (2009). Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychological Methods*, *14*, 101–125.

Baumeister, R.E., & Tice, T.M. (1988). Metatraits. *Journal of Personality*, *56*, 571–598.

Bazán, J.L., Bolfarine, H., & Branco, D.M. (2004). *A new family of asymmetric models for item response theory: a skew-normal IRT family* (Technical Report No. RT-MAE-2004-17). Department of Statistics, University of São Paulo.

Bazán, J.L., Branco, M.D., & Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis*, *1*, 861–892.

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., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bollen, K.A. (1996). A limited-information estimator for LISREL models with or without heteroscedastic errors. In G.A. Marcoulides & R.E. Schumacker (Eds.), *Advanced structural equation modeling: issues and techniques* (pp. 227–241). Mahwah: Erlbaum.

Chen, M.-H., Dey, D.K., & Shao, Q.M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, *94*, 1172–1186.

Chiogna, M. (2005). A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. *Statistical Methods & Applications*, *14*, 331–334.

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

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton: Princeton University Press.

Czado, C., & Santner, T.J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, *33*, 213–231.

Emons, W.H., Meijer, R.R., & Denollet, J. (2007). Negative affectivity and social inhibition in cardiovascular disease: evaluating type-D personality and its assessment using item response theory. *Journal of Psychosomatic Research*, *63*, 27–39.

Fisher, R.A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London. Series A*, *121*, 654–673.

Fraley, R.C., Waller, N.G., & Brennan, K.A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*, 350–365.

Guadagnoli, E., & Mor, V. (1989). Measuring cancer patients' affect: revision and psychometric properties of the Profile of Mood States (POMS). *Psychological Assessment*, *1*, 150–154.

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

Jinks, J.L., & Fulker, D.W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin*, *73*, 311–349.

Jöreskog, K.J. (2002). *Structural equation modeling with ordinal variables using LISREL*. Scientific Software International Inc. Retrieved November 3, 2010, from: http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf.

Keselman, H.J., & Lix, L.M. (1997). Analyzing multivariate repeated measures designs when covariance matrices are heterogeneous. *British Journal of Mathematical & Statistical Psychology*, *50*, 319–338.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146–162.

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York: Springer.

Long, J.S., & Ervin, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, *54*, 217–224.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, *35*, 63–78.

McDonald, R.P. (1999). *Test theory: a unified treatment*. Mahwah: Lawrence Erlbaum.

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*, 301–333.

Meijer, E., & Mooijaart, A. (1996). Factor analysis with heteroscedastic errors. *British Journal of Mathematical & Statistical Psychology*, *49*, 189–202.

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Molenaar, D., Dolan, C.V., & van der Maas, H.L.J. (2011). Modeling ability differentiation in the second-order factor model. *Structural Equation Modeling*, *18*, 578–594.

Molenaar, D., Dolan, C.V., & Verhelst, N.D. (2010a). Testing and modeling non-normality within the one factor model. *British Journal of Mathematical & Statistical Psychology*, *63*, 293–317.

Molenaar, D., Dolan, C.V., & Wicherts, J.M. (2009). The power to detect sex differences in IQ test scores using multigroup covariance and mean structure analysis. *Intelligence*, *37*, 396–404.

Molenaar, D., Dolan, C.V., Wicherts, J.M., & van der Maas, H.L.J. (2010b). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence*, *38*, 611–624.

Molenaar, D., van der Sluis, S., Boomsma, D.I., & Dolan, C.V. (2012). Detecting specific genotype by environment interaction using marginal maximum likelihood estimation in the classical twin design. *Behavior Genetics*, *42*, 483–499.

Monti, A.C. (2003). A note on the estimation of the skew normal and the skew exponential power distributions. *Metron*, *LXI*, 205–219.

Muthén, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika*, *53*, 563–578.

Muthén, L.K., & Muthén, B.O. (2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthén & Muthén.

Neale, M.C. (1998). Modeling interaction and nonlinear effects with Mx: a general approach. In G. Marcoulides & R. Schumacker (Eds.), *Interaction and non-linear effects in structural equation modeling* (pp. 43–61). New York: Lawrence Erlbaum Associates.

Neale, M.C., Aggen, S.H., Maes, H.H., Kubarych, T.S., & Schmitt, J.E. (2006). Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors*, *31*, 1010–1034.

Neale, M.C., Boker, S.M., Xie, G., & Maes, H.H. (2002). *Mx: statistical modeling* (6th ed.). Richmond: VCU.

Ramsay, J.O., & Abrahamowicz, M. (1989). Binomial regression with monotone splines: a psychometric application. *Journal of the American Statistical Association*, *84*, 906–915.

Ree, M.J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement*, *3*, 371–385.

Rochon, J. (1992). ARMA covariance structures with time heteroscedasticity for repeated measures experiments. *Journal of the American Statistical Association*, *87*, 777–784.

Rogers, T.B., Kuiper, N.A., & Kirker, W.S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*, 677–688.

Samejima, F. (1969). *Psychometric monograph: Vol. 17. Estimation of ability using a response pattern of graded scores*. Richmond: The Psychometric Society.

Samejima, F. (1997). Departure from normal assumptions: a promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, *62*, 471–493.

Samejima, F. (2000). Logistic positive exponent family of models: virtue of asymmetric item characteristic curves. *Psychometrika*, *65*, 319–335.

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. *Psychometrika*, *73*, 561–578.

Satorra, A., & Saris, W.E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*, 23–74.

Schmitt, J.E., Mehta, P.D., Aggen, S.H., Kubarych, T.S., & Neale, M.C. (2006). Semi-nonparametric methods for detecting latent non-normality: a fusion of latent trait and ordered latent class modeling. *Multivariate Behavioral Research*, *41*, 427–443.

Schmueli, G. (2010). To explain or to predict. *Statistical Science*, *25*, 289–310.

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*, 299–311.

Spearman, C.E. (1927). *The abilities of man: their nature and measurement*. New York: Macmillan.

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*, 1–16.

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.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*, 621–663.

Tucker-Drob, E.M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology*, *45*, 1097–1118.

van den Oord, E.J. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement*, *29*, 45–64.

van der Sluis, S., Dolan, C.V., Neale, M.C., Boomsma, D.I., & Posthuma, D. (2006). Detecting genotype-environment interaction in monozygotic twin data: comparing the Jinks & Fulker test and a new test based on marginal maximum likelihood estimation. *Twin Research and Human Genetics*, *9*, 377–392.

Verhelst, N.D. (2009). *Latent variable analysis with skew distributions*. Manuscript in preparation.

Vermunt, J.K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, *58*, 220–233.

Vermunt, J.K., & Hagenaars, J.A. (2004). Ordinal longitudinal data analysis. In R.C. Hauspie, N. Cameron, & L. Molinari (Eds.), *Methods in human growth research* (pp. 374–393). Cambridge: Cambridge University Press.

Vorst, H.C.M., & Bermond, B. (2001). Validity and reliability of the Bermond–Vorst alexithymia questionnaire. *Personality and Individual Differences*, *30*, 413–434.

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*, 58–79.

Woods, C.M. (2007). Ramsay-curve IRT for Likert type data. *Applied Psychological Measurement*, *31*, 195–212.

Zwinderman, A.H., & van den Wollenberg, A.L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, *14*, 73–81.