PSYCHOMETRIKA
DOI: 10.1007/s11336-014-9406-0

HETEROSCEDASTIC LATENT TRAIT MODELS FOR DICHOTOMOUS DATA

DYLAN MOLENAAR

UNIVERSITY OF AMSTERDAM

Effort has been devoted to account for heteroscedasticity with respect to observed or latent moderator variables in item or test scores. For instance, in the multi-group generalized linear latent trait model, it could be tested whether the observed (polychoric) covariance matrix differs across the levels of an observed moderator variable. In the case that heteroscedasticity arises across the latent trait itself, existing models commonly distinguish between heteroscedastic residuals and a skewed trait distribution. These models have valuable applications in intelligence, personality and psychopathology research. However, existing approaches are only limited to continuous and polytomous data, while dichotomous data are common in intelligence and psychopathology research. Therefore, in present paper, a heteroscedastic latent trait model is presented for dichotomous data. The model is studied in a simulation study, and applied to data pertaining alcohol use and cognitive ability.

Key words: heteroscedasticity, latent trait models, item response theory, two-parameter model, non-normal latent variables.

Generalized linear models constitute an important class of statistical tools in psychological research. The most obvious examples are MANOVA, the linear regression model, and the logit regression model used to test associations among observed variables. Other examples can be found in psychometrics, where generalized linear latent trait models like the common factor model and the two-parameter model are used for psychological and educational measurement (see e.g. Mellenbergh, 1994a).

As a general rule in parametric statistical models, results can only be trusted to reflect an effect that is actually present in the data if the assumptions underlying the statistical model are met. For the generalized linear models that are commonly used in psychology, a central assumption is that of homoscedasticity (e.g. Dobson, 2010, p. 33; Greene, 2011). Homoscedasticity refers to the requirement that the variances of the random effects in a statistical model are constant across units (e.g. Green, 2011).[1] In generalized linear modelling approaches, this requirement implies that the residual covariance matrix does not differ across levels of the independent variables (Slutsky, 1913; see Green, 2011 for the logit and probit case). If this assumption is violated, then one speaks of heteroscedasticity.

Heteroscedasticity is a well-studied phenomenon within generalized linear models. Methods have been proposed to test the equality of the residual covariance matrix across the levels of the independent variables in MANOVA or t test type of analysis (e.g. Anderson, 2006), and methods have been studied to account for possible violations (e.g. Brunner, Dette, & Munk, 1997). In linear regression models, various approaches exist to assess, test, or model heteroscedasticity, including diagnostic graphical approaches (e.g. Stevens, 2009, p. 90), statistical tests (e.g. Jarque & Bera, 1980), corrections (e.g. Long & Ervin, 2000), and approaches to model heteroscedasticity explicitly (e.g. Harvey, 1976).

With respect to the generalized linear latent trait models in psychometrics, approaches to the study of heteroscedasticity differ from those above as these models commonly contain an

Correspondence should be sent to Dylan Molenaar, Psychological Methods, Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, The Netherlands. E-mail: D.Molenaar@uva.nl

[1] A sensible designation as 'skedasis' is the Greek word for scatter or dispersion.

Published online: 01 August 2014

TABLE 1.
Overview of models for heteroscedasticity within the generalized linear latent trait modelling framework.

Data	Categorical moderator		Continuous moderator	
	Manifest	Latent	Manifest	Latent
Continuous	Multi-group FM	Factor Mixtures	Moderated FM	Heteroscedastic FM
Categorical	Multi-group IRT	Mixture IRT	Moderated IRT	Heteroscedastic GRM

FM factor model, *IRT* item response theory, *GRM* graded response model.

additional random subject effect. That is, besides the variable-specific residual effect in the measurement model, there is a latent trait effect in the structural model that is common to all variables. Thus, for a given variable in a psychometric model, heteroscedasticity can have two sources. In the literature, methods to model heteroscedasticity within the measurement and/or structural model have been studied elaborately for continuous and categorical data.[2] See Table 1 for an overview of the different approaches within the generalized linear latent trait framework. As can be seen, the methods are different in what they assume about the nature of the moderator variable across which the heteroscedasticity arises. In the case of heteroscedasticity with respect to a categorical moderator variable (or grouping variable), multi-group models (e.g. Jöreskog, 1971, Lee, Poon, & Bentler, 1989; Meredith, 1993; Muthén & Christoffersson, 1981) can be used to account for differences in the latent trait variance and the residual variances across the categories of a manifest moderator (e.g. gender, see Dolan et al., 2006). In addition, mixture models (e.g. Dolan & Van der Maas, 1998; Jedidi, Jagpal, & DeSarbo, 1997; Mislevy & Verhelst, 1990; Rost, 1990) can be used to test for differences in the latent trait variance and the residual variances across the categories of a latent moderator (i.e. latent to the data at hand; for instance stages of Piagetian conservation, see Jansen & Van der Maas, 1997).

Models that can be used in the case of heteroscedasticity with respect to a continuous moderator variable have been developed only recently for manifest moderators (e.g. Bauer & Hussong, 2009; Mehta & Neale, 2005; Merkle & Zeileis, 2013; Neale, Aggen, Maes, Kubarych, & Schmitt, 2006; Rabe-Hesketh, Skrondal, & Pickles, 2004). As the moderator is a continuous variable (e.g. age, see Molenaar, Dolan, Wicherts, & Van der Maas, 2010), the latent trait variance and/or the residual variance are not estimated for each level of the moderator separately, but these parameters are made a parametric function of the continuous moderator. This approach can be considered generalizations of the multi-group models as they include these models as special cases (see e.g. Bauer & Hussong, 2009). Note, however, that other approaches do not require the specification of the exact form of the function between the variance parameters and the moderator (see Merkle & Zeileis, 2013; Merkle, Fan, & Zeileis, 2013).

In the case that the moderator is a latent continuous variable, the moderator variable and the latent trait become indistinguishable and coincide. This implies that the heteroscedasticity can only come about by residual variances that differ across the levels of the latent trait (Bollen, 1996; Hessen & Dolan, 2009; Lewin-Koh & Amemiya, 2003; Meijer & Mooijaart, 1996). As a skewed latent trait distribution can also result in unequal observed (polychoric) variances for high and low levels of the trait, this effect needs to be taken into account to disentangle the effect of heteroscedastic residuals from the non-normality of the trait distribution. The resulting model is thus a latent trait model with heteroscedastic residuals and a skewed trait distribution, shortly denoted by heteroscedastic latent trait model.

[2] Truly continuous observed scores are rare in psychological and educational measurement. It has been shown that pragmatically, a scale with 7 or more ordered levels can be treated as continuous (Dolan, 1994). Therefore, in this paper, the term 'continuous variable' is used to denote variables with 7 or more ordered levels.

There are a number of reasons why studying this specific type of model is important (see Molenaar, Dolan, & De Boeck, 2012, for a comprehensive discussion). First, as heteroscedastic residuals result in asymmetric item characteristic curves, parameter bias may arise as has been shown by Bazán, Branco, and Bolfarine (2006) and Molenaar et al. (2012). Second, Samejima (1997, 2000, 2008) demonstrated that symmetrical item characteristic functions imply an inconsistent relationship between the order of the latent trait estimates and the difficulty parameters. As discussed by Samejima, models with asymmetric characteristic functions are not subject to this problem. Third, wrongfully assuming a normal latent trait distribution has been shown to bias item parameter estimates (Azevedo, Bolfarine, & Andrade, 2011; Zwinderman & van der Wollenberg, 1990) and ability estimates (Ree, 1979; Seong, 1990; Swaminathan & Gifford, 1983), although the occurrence of bias may depend on factors like the number of items and the sample size (Kirisci, Hsu, & Yu, 2001; Stone, 1992). Fourth, as will be shown in this paper, neglecting heteroscedasticity may bias the item information function. This has implications in, for instance, computerized adaptive testing. Finally, a reason to study the heteroscedastic latent trait model is that it has shown valuable substantive applications in various research fields. These applications include the ability differentiation hypothesis in intelligence research (e.g. Murray, Dixon, & Johnson, 2013), the schematicity or traitedness hypothesis in personality research (Molenaar et al., 2012), and genotype by environment interactions in behaviour genetics (e.g. Van der Sluis, Dolan, Neale, Boomsma, & Posthuma, 2006). In addition, other possible applications may include psychopathology where it is hypothesized that subjects with a higher level of a psychopathological trait like depression are more consistent in their self-reports of the symptoms (Fokkema, Smits, Kelderman, & Cuijpers, 2013).

Heteroscedastic latent trait models have been proposed for continuous data (Molenaar, Dolan, & Van der Maas, 2011; Molenaar, Dolan, & Verhelst, 2010a), and for polytomous data (Molenaar et al., 2012). However, for dichotomous data, no suitable procedure has yet been proposed, while these kind of data are common in intelligence and psychopathology research. Therefore, in present paper, a generalized linear latent trait modelling approach is presented to model heteroscedastic residuals and a skewed latent trait in dichotomous data. The outline of this paper is as follows: First, the homoscedastic and heteroscedastic case of the generalized linear latent trait model are presented for continuous and polytomous data. Then, this approach is extended to enable modelling of dichotomous data. Next, the new model is studied in a simulation study to investigate the parameter recovery, the required sample size, the resolvability of the different effects, and the power to detect the effects. Subsequently, the model is applied to two real datasets pertaining to alcohol use and cognitive ability. Finally, some limitations and future directions are discussed.

1. The Generalized Linear Latent Trait Model

1.1. *The Homoscedastic Case*

The models covered in this paper concern unidimensional generalized linear latent trait models (Mellenbergh, 1994a) for continuous and ordered categorical data. These models are part of the more general class of models commonly referred to as the generalized latent variable modelling framework (Bartholomew, Knott, & Moustaki, 2011; Moustaki & Knott, 2000; Skrondal & Rabe-Hesketh, 2004). It is thus assumed that the data consist of either sum scores, responses to continuous items, responses to Likert scale items, item scores that are recoded into correct and false, or items with a dichotomous answer scale (e.g. yes/no questions). Poisson models (e.g. for counts) and gamma models (e.g. for time taken to complete a task) which are part of the generalized framework, are not considered in this paper as these commonly do not have a separate dispersion parameter.

In the unidimensional generalized linear latent trait model, given local independence, the marginal likelihood of a response vector, **y**, is given by (Bock & Aitkin, 1981; Moustaki & Knott, 2000)

$$L(\mathbf{y};\boldsymbol{\tau}) = \int\limits_{-\infty}^{\infty} \prod_{i=1}^{n} h(y_i|\theta)g(\theta)\mathrm{d}\theta, \tag{1}$$

where $\boldsymbol{\tau}$ is a vector of parameters, n is the number of observed variables, $h(.)$ is the distribution of the observed variables under the measurement model, and $g(.)$ denotes the distribution of the latent trait, θ, under the structural model. The general measurement model considered in this paper is given by

$$y_i^* = \nu_i + \alpha_i\theta + \varepsilon_i, \tag{2}$$

where y_i^* is an unobserved continuously distributed variable underlying item i, ν_i is the fixed intercept parameter, α_i is the fixed discrimination parameter and ε_{pi} is a random residual effect with variance $\sigma_{\varepsilon i}^2$. If the observed data, y_i, are continuous, then in Eq. 2, $y_i = y_i^*$. If y_i is ordered categorical or dichotomous, then

$$y_i = c \quad if \quad y_i^* \in (\beta_{ci}, \beta_{(c+1)i}) \quad c = 0, \ldots, C-1. \tag{3}$$

That is, the continuous y_i^* is categorized at increasing thresholds, β_{ic}, where $\beta_{0i} = -\infty$ and $\beta_{Ci} = \infty$. This approach is sometimes referred to as item factor analysis (Christofferson, 1975; Muthén, 1978; Olssen, 1979) or the underlying variable approach (Jöreskog & Moustaki, 2001), and is based on Thurston's model for categorical judgement (1920; see Master, 1982; Skrondal, 1996, Chapt. 10). As shown by Takane and de Leeuw (1987), this approach is equivalent to the item response theory approach for categorical data.

By assuming a normal distributions for ε_i, the distribution of y_i under the measurement model in the homoscedastic case is

$$h(y_i|\theta) = f(y_i^*|\theta) \text{ for continuous } y_i \tag{4}$$

$$h(y_i|\theta) = \int\limits_{\beta_{ci}}^{\beta_{(c+1)i}} f(y_i^*|\theta)\mathrm{d}y_i^* \text{ for ordinal } y_i \tag{5}$$

with

$$f(y_i^*|\theta) = \frac{1}{\sigma_{\varepsilon i}}\varphi\left(\frac{y_i^* - \nu_i - \alpha_i\theta}{\sigma_{\varepsilon i}}\right), \tag{6}$$

where $\varphi(.)$ denotes the standard normal density function. By further assuming a normal distribution for θ, the density function in the structural model for the homoscedastic case is given by

$$g(\theta) = \frac{1}{\sigma_\theta}\varphi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right). \tag{7}$$

In the case of ordinal data, Eq. 5 gives the item category response functions (for $C \geq 3$) or the item characteristic function (for $C = 2$).

The formulation above has the advantage that it includes a number of important and commonly used measurement models as special cases. That is, if y_i has a normal distribution and $\mathrm{cov}(\theta, \varepsilon_i) =$

0, then the common factor model arises (Mellenbergh, 1994b). If $C \geq 3$, then the model is equivalent to a graded response model (Samejima, 1969) with $\nu_i = \nu = 0$ and $\sigma^2_{\varepsilon i} = \sigma^2_{\varepsilon} = 1$. Subsequently, if $C = 2$, then the model is a two-parameter normal ogive item response theory model (Lord, 1952). Other models that can be formulated within the present approach are the linear logistic model (Fischer, 1983) and the nominal response model (Bock, 1972).

Some well-known models cannot be formulated within the present approach because an underlying variable formulation is not possible. These models include the rating scale model (Andrich, 1978), the model for guessing by Thissen and Steinberg (1984), the partial credit model (Masters, 1982), and the three-parameter model (Birnbaum, 1968). However, in the case of the rating scale model and the partial credit model, the graded response model might be considered as an alternative because this model is highly similar (but not equivalent, see Masters, 1982; Thissen & Steinberg, 1986).

1.2. The Heteroscedastic Case

For each variable i in the general model in Eq. 2, there are two random effects, ε_i and θ, that may be subject to heteroscedasticity. Heteroscedasticity is formalized by considering the variance of the random effects conditional on a so-called moderator variable, M. By assuming that $\mathrm{E}(\varepsilon_i) = \mathrm{cor}(\varepsilon_i, \theta) = 0$, the variance of y^*_i is then given by

$$\begin{aligned} \sigma^2_{y*i|M} &= \alpha^2_i \sigma_{\theta|} M^2 + \sigma_{\varepsilon i|} M^2 \\ \mu^2_{y*i|M} &= \nu_i + \alpha_i \mu_{\theta|M}, \end{aligned} \tag{8}$$

where $\sigma^2_{\theta|M}$ and $\sigma^2_{\varepsilon i|M}$ denote the conditional variance of θ and ε_i, respectively, and $\mu_{\theta|M}$ is the conditional mean of θ. Introducing the idea in Eq. 8 into the general model in Eqs. 6 and 7, the general model for heteroscedasticity is given by

$$\begin{aligned} f\left(y^*_i|\theta, M\right) &= \frac{1}{\sigma_{\varepsilon i|M}} \varphi\left(\frac{y^*_i - \nu_i - \alpha_i\theta}{\sigma_{\varepsilon i|M}}\right) \\ g(\theta) &= \frac{1}{\sigma_{\theta|M}} \varphi\left(\frac{\theta - \mu_{\theta|M}}{\sigma_{\theta|M}}\right) \end{aligned} \tag{9}$$

Note that if M is manifest and categorical, then Eq. 9 is equivalent to the strong measurement invariance model proposed by Meredith (1993); if M is latent and categorical, then the model is equivalent to a mixture of two latent trait models; and if M is manifest and continuous, then the model in Eq. 9 is a moderated latent trait model where the residual variances, the trait variance, and the trait mean are some parametric functions of the moderator.

If the continuous moderator is a latent variable, then M and θ are indistinguishable and coincide, which implies that—conditional on θ (i.e. M)—the only source of (polychoric) variance is the residual variance, $\sigma^2_{\varepsilon i}$. Thus, heteroscedasticity can be modelled by making $\sigma^2_{\varepsilon i}$ a function of θ itself, that is $\sigma^2_{\varepsilon i|M} = \sigma^2_{\varepsilon i|\theta}$. For the residuals, a suitable function needs to be specified for $\sigma^2_{\varepsilon i|\theta} = \mathrm{w}(\theta; \boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a parameter vector. In the case of continuous data, an exponential function is commonly used (see e.g. Bauer & Hussong, 2009; Harvey, 1976; Hessen & Dolan, 2009). However, as pointed out by Molenaar et al. (2012), in models for categorical data, this function causes undesirable behaviour of the item category response functions (Eq. 5). Therefore, for polytomous item scores, the following function is proposed (Molenaar et al., 2012):

$$\sigma^2_{\varepsilon i|\theta} = w(\theta; \boldsymbol{\delta}) = 2\delta_0[1 + \exp(-\delta_1\theta)]^{-1}, \tag{10}$$

where δ_0 is a baseline parameter, $\delta_0 \in (-\infty, \infty)$, and δ_1 is a heteroscedasticity parameter, $\delta_1 \in (-\infty, \infty)$. Note that if δ_1=0, then the residual variances are homoscedastic with $\sigma^2_{\varepsilon i|\theta} = \delta_0$; if $\delta_1 > 0$, then the residual variances are increasing across θ; and if $\delta_1 < 0$, residual variances are decreasing across θ. For polytomous items, the resulting model can be identified by fixing two adjacent thresholds, β_{ic}, in Eq. 5 (see Mehta, Neale, & Flay, 2004). Due to this identification constraint, the model is only suitable in the case of $C \geq 3$, and not in the case of dichotomous data where $C = 2$.

As a skewed latent trait distribution causes unequal observed (polychoric) variances at high and low levels of the trait, this effect needs to be taken into account to disentangle the effect of heteroscedastic residuals from the effect of non-normality in the trait distribution. Within latent trait modelling, a number of authors (Azzevado, Bolfarine, & Andrade, 2011; Molenaar, Dolan, & De Boeck, 2012; Molenaar, Dolan et al., 2010a) have proposed the skew-normal distribution (Azzalini, 1985, 1986; Azzalini & Capatanio, 1999). This distribution has the convenient property that it includes the normal distribution as a special case. However, any other distribution that allows for skew can be considered for pragmatic or theoretical reasons.

1.3. An Approach for Dichotomous Data

In this section, a suitable approach is presented to account for heteroscedasticity in the case of dichotomous data. For dichotomous data, the function in Eq. 10 can be used, but different identification constraints are necessary. The required constraints could be inferred from results in Millsap and Yun-Tein (2004) for dichotomous data and a categorical moderator, M (Eq. 9). That is, in addition to the standard scale and location constraints (e.g. Bollen, 1989, p. 238), the following restrictions are required in Eq. 9 to identify both $\sigma^2_{\varepsilon i|M}$ and $\sigma^2_{\theta|M}$ across groups (i.e. across the levels of M). First, $\nu_i = 0$ for all i and $\sigma^2_{\varepsilon i|M} = 1$ for all i in an arbitrary reference group. Next, $\sigma^2_{\varepsilon i|M} = 1$ for all groups for some i (the anchor item).[3] Now, as α_i in Eq. 9 and β_{ic} in Eq. 5 do not depend on M (i.e. they are invariant across levels of M), the multi-group model for dichotomous data is identified.

For the heteroscedastic latent trait model for dichotomous data, the results above imply that (1) δ_0 should be constrained to equal 1 for all i, such that for $\theta = 0 \rightarrow \sigma^2_{\varepsilon i|\theta} = 1$ (the reference point); and (2) δ_1 should be constrained to equal 0 for some i (the anchor item). The parameter ζ will then pick up the effect of heteroscedasticity that is common to all items, where δ_{1i} models item-specific departures from this main effect. As in the multi-group case, these two sets of parameter(s) capture the same effect of heteroscedasticity. Thus, two models are possible that are just identified in their heteroscedasticity parameters:

#1: A model with δ_{1i} free for all i and ζ fixed.
#2: A model with δ_{1i} free for all i except for the anchor item and ζ free.

This is opposed to the models for polytomous and continuous data, where the two effects can be combined for all items without further restrictions.

[3] In fact, Millsap and Yun-Tein (2004) use the restriction of VAR(y^*_{pi}) $= 1$ instead of fixing $\sigma^2_{\varepsilon i} = 1$ as is done here. Because, VAR(y^*_{pi}) $= \alpha^2_i \times \sigma^2_\theta + \sigma^2_{\varepsilon i}$ in which σ^2_θ is already identified by fixing $\sigma^2_\theta = 1$ (or $\alpha_i = 1$), fixing VAR(y_{pi}) $= 1$ will result in $\sigma^2_{\varepsilon i} = 1 - \alpha^2_i$ (or $\sigma^2_{\varepsilon i} = 1 - \sigma^2_\theta$) which thus implies a fixed $\sigma^2_{\varepsilon i}$. The opposite holds as well, that is, fixing $\sigma^2_{\varepsilon i} = 1$ implies a fixed VAR(y^*_{pi}).

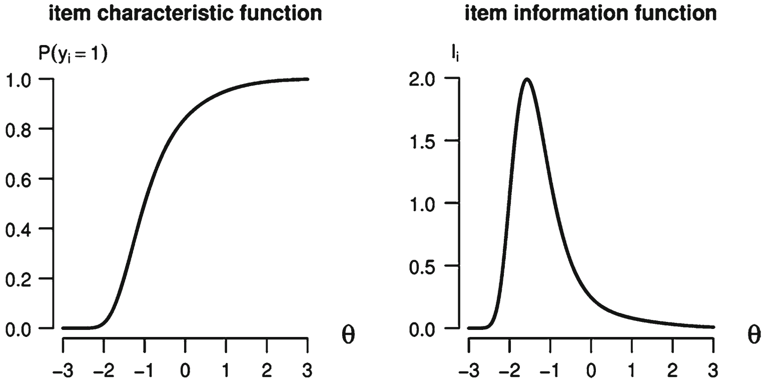

FIGURE 1.
Example of an item characteristic function (*left*) and the corresponding item information function (*right*) in the heteroscedastic latent trait model for $\alpha_i = 1$, $\beta_i = 1$, and $\delta_1 = 0.8$.

The resulting heteroscedastic latent trait model for dichotomous data is given by

$$\begin{aligned} f(y_i^*|\theta, M) &= \frac{1}{\sqrt{2}\left[1 + \exp\left(-\delta_{1\mathrm{i}}\theta\right)\right]^{-\frac{1}{2}}}\varphi\left(\frac{y_i^* - \alpha_i\theta}{\sqrt{2}\left[1 + \exp\left(-\delta_{1\mathrm{i}}\theta\right)\right]^{-\frac{1}{2}}}\right) \\ g(\theta) &= \frac{2}{\sigma_\theta}\Phi\left(\zeta\frac{\theta - \mu_\theta}{\sigma_\theta}\right)\varphi\left(\frac{\theta - \mu_\theta}{\sigma_\theta}\right) \end{aligned} \tag{11}$$

where $g(.)$ is a skew-normal density function with location parameter μ_θ and scale parameter σ_θ for which it holds that

$$E(\theta) = \mu_\theta + \sigma_\theta\sqrt{\frac{2}{\pi}}\frac{\zeta}{\sqrt{1+\zeta^2}}, \tag{12}$$

and

$$\mathrm{SD}(\theta) = \sigma_\theta\sqrt{1 - \frac{2}{\pi}\left(\frac{\zeta^2}{1+\zeta^2}\right)}, \tag{13}$$

where ζ is a shape parameter for which it holds that if $\zeta = 0$, then $g(.)$ in Eq. 11 will simplify to a normal distribution with $\mathrm{E}(\theta) = \mu_\theta$ and $\mathrm{SD}(\theta) = \sigma_\theta$. The item characteristic function of the heteroscedastic latent trait model can be derived by substituting $f(.)$ from Eq. 11 into Eq. 5 for $C = 2$ which gives

$$P\left(y_i = 1|\theta\right) = \Phi\left(\frac{\alpha_i\theta_p + \beta_i}{\sqrt{2}\left[1 + \exp\left(-\delta_{1i}\theta\right)\right]^{-\frac{1}{2}}}\right),$$

where β_i is the item difficulty parameter. See Figure 1 for an example of the item characteristic function (left) and the corresponding item information function (right) for $\alpha_i = 1$, $\beta_i = 1$, and $\delta_1 = 0.8$.

1.4. Parameter Estimation and Implementation

The heteroscedastic latent trait models discussed in this paper can be fit to data using marginal maximum likelihood estimation (MML; Bock & Aitkin, 1981). For the new model, an MML estimation procedure is implemented in the statistical software package R (R Core Team, 2012). Specifically, -2 times the log of the marginal likelihood function (Eq. 1 with Eqs. 5 and 11) is minimized using the R built-in function 'optim'. This optimizer uses a Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS; see e.g. Nocedal & Wright, 2006; p. 194) which is a quasi-Newton algorithm that uses first-order derivatives. The likelihood function is approximated using 50 Gauss-Hermite quadrature points (see Molenaar et al., 2012). It is also possible to fit the model using existing software like Mx (Neale, Aggen, Maes, Kubarych & Schmitt, 2006) and SAS (SAS Institute, 2011). The R scripts used here are available on the personal webpage of the author, together with an Mx script to fit the model.

2. Simulation Study

In this simulation study, the viability of the model for dichotomous data is investigated. That is, this simulation study is conducted to (1) investigate whether true parameters are adequately recovered, (2) to assess what sample sizes are at least needed to apply the model, (3) to examine whether δ_{1i} and ζ can be disentangled satisfactorily, and (4) to see whether the statistical power to detect the effects is acceptable given a reasonable effect size.

2.1. Design

Data are generated according to either (1) the full model with heteroscedastic residuals and a skew-normal trait where both effects are in the same direction, denoted 'het(+)' and 'skw(−)' (i.e. a negatively skewed trait and increasing residuals for increasing trait levels); (2) the full model with heteroscedastic residuals and a skew-normal trait where both effects are in the opposite direction, denoted 'het(+)' and 'skw(+)' (i.e. a positively skewed trait and increasing residuals for increasing trait levels); (3) a model with het(+) only; and 4) a model with a skw(−) only. Sample sizes equalled 500, 1,000, 3,000 and 5,000; the number of items equalled 10 and 500 replications were conducted for each condition. Parameter values in the case of item-specific heteroscedasticity effects (i.e. the effect of 'het(+)') are $\delta_{1i} = 0.4$ ('small effect'), $\delta_{1i} = 0.6$ ('medium effect'), and $\delta_{1i} = 0.8$ ('large effect') for all i except for one item ($\delta_{1i} = 0$ for this anchor item). For skw(+) and skw(−), $\zeta = 2.17$ and $\zeta = -2.17$ are used, respectively ('medium effect'). The above effect sizes are based on Molenaar et al. (2010a, 2012).[4] The anchor item was the 5th item. The remaining parameters were chosen to be $\alpha_i = 1$ for all i and the β_i parameters were chosen equally spaced in the interval $[-1.5, 1.5]$. In addition, $\mathrm{E}(\theta)$ and $\mathrm{SD}(\theta)$ were restricted to equal 0 and 1 respectively.

2.2. Models and Power

Four models are fit to the generated datasets:

1. **M0:2PM**. A homoscedastic two-parameter model with a normal distribution for the trait (i.e. a traditional two-parameter model (i.e. $\delta_{1i} = 0$ for all i and $\zeta = 0$);
2. **M1:het**. A model with heteroscedastic residuals and a normal distribution for the trait (i.e. $\zeta = 0$, and δ_{1i} is free for all i except the anchor item);

[4] Note that Molenaar et al. (2012) used $\delta_{0i} = 1.5$ instead of $\delta_{0i} = 1,$; therefore, in the present case, δ_{1i} is rescaled to correspond to the effect size in Molenaar et al (i.e. due to the difference in δ_{0i}, there is no one-to-one correspondence).

3. **M1:skw**. A model with a skew-normal trait distribution and with homoscedastic residual variances (i.e. ζ is free and $\delta_{1i} = 0$ for all i);
4. **M2:full.** A full model with both heteroscedastic residuals and a skewed trait (i.e. ζ is free and δ_{1i} is free for all i except the anchor item).

In all models, α_i and β_i are estimated for all i, and identification is accomplished by fixing $\text{E}(\theta) = 0$, $\text{SD}(\theta) = 1$, $\delta_{1i} = 0$ for the anchor item, and $\delta_{0i} = 1$, for all i.

For each model except M0:2PM (as this model serves as the baseline model), the power of the likelihood ratio test to detect the effect(s) in the model is determined for a 0.05 level of significance (e.g. Satorra & Saris, 1985). In the likelihood ratio test, H_A is the model under consideration (M1:het, M1:skw or M2:full) and H_0 is the model without the effect of interest that is nested in H_A (i.e. M0:2PM, M1:het, or M1:skw). The likelihood ratio statistic is then given by

$$T = -2 \times \left[\ln L\left(\boldsymbol{y_p}; \hat{\boldsymbol{\tau}}_{\boldsymbol{0}}\right) - \ln L\left(\boldsymbol{y_p}; \hat{\boldsymbol{\tau}}_{\boldsymbol{A}}\right)\right],$$

where $L(.)$ is given by Eq. 1, $\hat{\boldsymbol{\tau}}_{\boldsymbol{0}}$ is the vector of parameter estimates of H_0 and $\hat{\boldsymbol{\tau}}_{\boldsymbol{A}}$ is the vector of parameter estimates of H_A. As H_0 is nested in H_A, the parameter vector $\hat{\boldsymbol{\tau}}_{\boldsymbol{0}}$ is a restricted version of $\hat{\boldsymbol{\tau}}_{\boldsymbol{A}}$. If the parameter constraints in H_0 are not on the boundary of the parameter space, then 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, T has a non-central χ^2-distribution with non-centrality parameter λ and with *df* equal to the number of restrictions in H_0. To calculate power, the non-centrality parameter needs to be estimated. As it holds that $\text{E(T)} = df + \lambda$ (see Fisher, 1928), an estimate of λ could be obtained by averaging T-*df* over the replications in the simulation study. Next, power is obtained by integrating the non-central χ^2 distribution from the critical value to ∞.

3. Results

3.1. *Parameter Recovery and Sample Size*

In Table 2, results are depicted for the conditions in which the true model includes heteroscedastic residuals with a small effect size. Items are sorted according to their item difficulty parameter, where item 1 is located at the lower θ range and item 10 is located at the upper θ range. As can be seen, for $N = 5,000$ and $N = 3,000$, parameters are acceptably recovered. For item 10, variability is large due to, respectively, 1 and 9 diverged estimates in the case of $N = 5,000$ and $N = 3,000$. For $N = 1,000$ and $N = 500$, estimates of δ_{1i} are diverging in an increasing number of cases causing large parameter variability for the items at the upper and lower end of the θ range. See Figure 2 for a histogram of the parameter estimates of δ_{1i} for item 10 (an item at the upper range of θ) for $N = 1,000$ and $N = 500$. As can be seen, the estimates centre approximately around the true value (0.4), but a number of estimates diverged. Apparently, for the cases that diverged, there was not enough information in the data concerning the δ_{1i} parameter.

Results for the model with a skewed latent trait only ($\delta_{1i} = 0$ and $\zeta = -2.17$) are not tabulated but were generally good. Specifically, parameter estimates (SD) for ζ in this condition are $-2.19(0.30)$, $-2.22(0.39)$, $-2.27(0.81)$ and $-2.54(1.48)$ for samples sizes of, respectively, 5,000, 3,000, 1,000 and 500. From these results and the results concerning ζ in Table 2, it could be concluded that the true value is well recovered for all sample sizes. Notably, in the case of $N = 500$, the parameter estimate variability tends to be quite large.

3.2. *Power and Resolvability*

Table 3 depicts the power to detect heteroscedasticity in the items (i.e. the power to detect that $\delta_{1i} \neq 0$) for a model with heteroscedastic residuals only (*M1:het*) and a model with both

TABLE 2.
Mean (standard deviation) of the estimates for δ_{1i} and ζ in the correct model for each condition in the case of a small effect size.

True values		$i=1$	$i=2$	$i=3$	$i=4$	$i=6$	$i=7$	$i=8$	$i=9$	$i=10$	
		δ_{1i}									ζ
$\delta_{1i} = 0.4; \zeta = 0$	$N=5{,}000$	0.42 (0.22)	0.39 (0.21)	0.39 (0.20)	0.40 (0.21)	0.42 (0.22)	0.41 (0.25)	0.45 (0.28)	0.46 (0.35)	0.50 (0.51)	–
	$N=3{,}000$	0.39 (0.30)	0.41 (0.27)	0.41 (0.27)	0.41 (0.26)	0.41(0.29)	0.44 (0.37)	0.46 (0.44)	0.48 (0.46)	0.80 (3.04)	–
	$N=1{,}000$	0.07 (3.21)	0.41 (0.60)	0.40 (0.54)	0.46 (0.57)	0.46 (0.66)	0.54 (0.71)	0.61 (1.03)	0.98 (3.53)	1.93 (6.45)	–
	$N=500$	−1.36 (8.81)	0.08 (3.09)	0.40 (1.19)	0.47 (0.97)	0.63 (2.65)	1.06 (3.22)	1.46 (4.78)	2.47 (7.51)	4.59 (11.01)	–
$\delta_{1i} = 0.4; \zeta =$-2.17	$N=5{,}000$	0.41 (0.22)	0.43 (0.22)	0.43 (0.25)	0.43 (0.25)	0.44 (0.31)	0.43 (0.36)	0.43 (0.38)	0.45 (0.45)	0.58 (2.12)	−2.22 (0.90)
	$N=3{,}000$	0.41 (0.30)	0.44 (0.30)	0.45 (0.31)	0.43 (0.33)	0.47 (0.41)	0.44 (0.45)	0.46 (0.51)	0.64 (2.49)	1.21 (4.98)	−2.29 (1.14)
	$N=1{,}000$	0.22 (3.02)	0.47 (0.72)	0.46 (0.59)	0.45 (0.60)	0.65 (1.19)	0.56 (0.99)	0.71 (2.86)	2.24 (7.97)	3.68 (10.98)	−2.44 (1.97)
	$N=500$	−0.54 (6.19)	0.28 (3.87)	0.60 (1.31)	0.81 (2.78)	1.11 (3.62)	1.31 (4.93)	3.06 (9.48)	5.01 (12.84)	7.57 (15.34)	−2.53 (2.57)
$\delta_{1i} = 0.4; \zeta = 2.17$	$N=5{,}000$	0.39 (0.37)	0.42 (0.35)	0.44 (0.36)	0.44 (0.34)	0.45 (0.30)	0.41 (0.30)	0.40 (0.30)	0.41 (0.34)	0.50 (1.53)	2.25 (0.95)
	$N=3{,}000$	0.44 (0.46)	0.42 (0.42)	0.46 (0.47)	0.45 (0.46)	0.44 (0.39)	0.42 (0.39)	0.43 (0.35)	0.43 (0.43)	0.55 (2.02)	2.30 (1.20)
	$N=1{,}000$	−0.06 (3.84)	0.40 (1.29)	0.42 (0.85)	0.47 (0.80)	0.45 (0.66)	0.52 (0.76)	0.64 (2.00)	1.11 (4.37)	2.37 (8.06)	2.31 (1.49)
	$N=500$	−2.56 (10.59)	−0.12 (5.03)	0.26 (3.77)	0.80 (2.69)	0.44 (2.60)	0.90 (3.88)	1.75 (6.67)	3.76 (10.58)	6.29 (14.19)	2.54 (1.71)

δ_{1i} is fixed to 0 for $i = 5$ (anchor item). Items are sorted according to their item difficulty. For the results concerning the condition with a skewed latent trait only, see text.

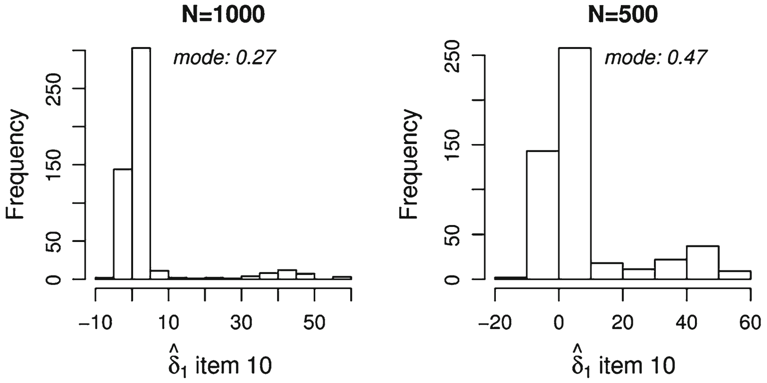

FIGURE 2.
Histograms of the heteroscedasticity parameter estimates for item 10 across replications. The true value equals 0.4.

TABLE 3.
Power of a model with heteroscedastic residuals only (M1:het) and the full model (M2:full) to detect heteroscedasticity in the items for a 0.05 level of significance.

Condition	N	Small effect		Medium effect		Large effect	
		M1:het	*M2:full*	*M1:het*	*M2:full*	*M1:het*	*M2:full*
het(+) & skew(−)	5,000	1.00 (126.15)	0.60 (10.70)	1.00(191.71)	0.98 (27.94)	1.00 (264.19)	1.00 (53.63)
	3,000	1.00 (76.43)	0.35 (6.13)	1.00 (116.03)	0.86 (17.89)	1.00 (157.33)	0.99 (31.67)
	1,000	0.96 (24.58)	0.12 (1.77)	1.00 (38.29)	0.34 (5.95)	1.00 (53.36)	0.61 (10.83)
	500	0.71 (13.14)	0.07 (0.69)	0.89 (19.32)	0.16 (2.84)	0.97 (26.57)	0.19 (3.37)
het(+)	5,000	1.00 (35.25)	0.22 (3.88)	1.00 (74.30)	0.59 (10.53)	1.00 (124.41)	0.94 (22.92)
	3,000	0.92 (20.74)	0.12 (1.88)	1.00 (45.18)	0.35 (6.24)	1.00 (75.33)	0.75 (14.09)
	1,000	0.40 (7.09)	0.07 (0.69)	0.76 (14.33)	0.13 (2.03)	0.96 (25.34)	0.26 (4.62)
	500	0.25 (4.42)	0.07 (0.65)	0.46 (8.16)	0.09 (1.19)	0.70 (12.84)	0.15 (2.55)
skew(−)	5,000	0.99 (30.84)	0.05 (0.14)	0.99 (30.84)	0.05(0.14)	0.99 (30.99)	0.05 (0.04)
	3,000	0.88 (18.86)	0.05 (0.07)	0.89 (19.04)	0.06 (0.29)	0.87 (18.44)	0.06 (0.18)
	1,000	0.38 (6.78)	0.06 (0.39)	0.37 (6.59)	0.05 (0.00)	0.40 (6.98)	0.05 (0.02)
	500	0.21 (3.77)	–	0.21 (3.79)	–	0.24 (4.20)	0.06 (0.32)
het(+) & skew(+)	5000	0.06 (0.27)	0.25 (4.49)	0.51 (9.08)	0.52 (9.19)	0.98 (29.00)	0.77 (14.85)
	3,000	0.05 (0.03)	0.15 (2.45)	0.26 (4.60)	0.29 (5.17)	0.82 (16.29)	0.47 (8.28)
	1,000	0.07 (0.47)	0.07 (0.50)	0.13 (2.04)	0.11 (1.64)	0.35 (6.13)	0.10 (1.38)
	500	0.09 (1.09)	–	0.11 (1.62)	–	0.21 (3.69)	–

het(+): heteroscedastic residuals that increase for increasing levels of θ; skw(−): negatively skewed trait; skw(+): positively skewed trait. '−': in these cases, the non-centrality parameter was slightly negative due to a too small sample size.

heteroscedastic residuals and a skew-normal trait (*M2:full*). First, it can be seen from the table that in the case of the 'het(+) & skew(−)' condition, the pattern of results is generally acceptable. That is, for increasing sample sizes and increasing effect sizes, power coefficients approach at least a 0.80 level. In the case of the 'het(+)' condition, the *M1:het* model has acceptable power to detect the effect for increasing effect size and sample size, while the *M2:full* model has only acceptable power for a large effect. Apparently, incorporation of the additional shape parameter ζ, while the effect is actually not in the data, results in a decrease in power. An interesting condition is the case that there is a skewed trait in the data only (i.e. the 'skew(+)' condition), as conclusions concerning resolvability of the effects could be inferred from the results in this condition. As can be seen, the skewed trait effect is absorbed in the heteroscedasticity parameters δ_{1i} in the *M1:het* model (i.e. the power is large). Thus, unmodelled skewness in the trait will incorrectly be detected as heteroscedasticity (i.e. false positives). However, in the full model *M2:full*, where the skewness in the trait is accounted for by introducing the ζ parameter, the power to detect heteroscedasticity adequately equals the Type I error rate (0.05). This indicates that the effect of skewness in the trait is well separable from heteroscedastic residuals. Finally, it can be seen from the table that in the 'het(+) & skw(+)' condition, the effects may cancel out depending on the effect size. That is, for the small and medium heteroscedasticity effect size, power is much smaller as compared to the 'het(+) & skew(−)' condition. For the large effect size and $N = 5,000$, power is acceptable again.

4. Conclusion

From the simulation study, it appears that parameter recovery is generally acceptable. However, in smaller sample sizes, estimates for the heteroscedasticity parameter δ_{1i} are much more likely to diverge from the population value for items that have difficulty parameters at the extremes of the trait. The power results indicated that by the identification constraint $\delta_{1i} = 0$ for the anchor item, the effects on the trait distribution and the residual variances are resolvable. That is, one can distinguish between a common effect of heteroscedasticity (formalized as a skewed trait) and/or item-specific effects (operationalized as heteroscedastic residual variances), given that the effects are not in the same direction. In the case that the effects are in the same direction (e.g. a positively skewed trait and increasing residual variance across the trait), they may cancel each other out depending on the effect size. With respect to the sample size needed to apply the model, it could be concluded that relatively large sample sizes are needed. Power is only acceptable for sample sizes of at least 1,000 subjects when a single effect is considered, or sample sizes of at least 3,000 subjects when both effects are combined. This will be further discussed in the final section.

While the simulation study shows that the anchor item is useful to distinguish item common effects and item-specific effects of heteroscedasticity, the question arises of how an anchor item should be identified. In application 2 below, the use of Lagrange Multipliers (LM) is illustrated. The multivariate LM statistic signifies the approximate increase in loglikelihood that will result from freeing each constraint parameter. This statistic can be calculated by $\mathbf{G}_0'\mathbf{H}_0\mathbf{G}_0$, where $\mathbf{G}_0$ is the vector of first-order derivatives of the loglikelihood function and $\mathbf{H}_0$ is the matrix of second-order derivatives to the model parameters in the baseline model. Using this formula, univariate LM statistics could be calculated for each constrained parameter separately. As will be shown in application 2, univariate LM statistics can be calculated for the δ_{1i} parameter of each item in a model with a skewed trait. Then, these LM statistics will indicate whether —after taking the trait skewness into account—it is beneficial to incorporate additional item-specific heteroscedasticity effects. In the present paper, LM statistics are obtained by the exact first-order derivatives, $\mathbf{G}_0$. The second-order derivatives in $\mathbf{H}_0$ are obtained by a finite difference approximation. Simulations showed that this works well (see also Dolan & Molenaar, 1991).

5. Illustrations

5.1. Illustration 1: Alcohol Use

The data comprises scores of 4,627 subjects on 5 items of the Michigan Alcohol Screening Test (Selzer, 1971) which was administered in the National Survey of Midlife Development in the United States (MIDUS) in 1995–1996 under the auspices of the Inter-university Consortium for

TABLE 4.
Model fit results for application 1.

	−2LL	LRT	df	AIC	BIC	sBIC	DIC
1. Baseline	6,370.02			6,390	6,454	6,423	6,436
2. Heteroscedastic residuals	6,363.39	6.63	5	6,393	6,490	6,442	6,462
3. Skew trait distribution	6,354.51	8.88	1	**6,377**	**6,447**	**6,412**	**6,427**

−2LL denotes −2 times the loglikelihood. The likelihood ratio test (LRT) is conducted between the corresponding model and the baseline model. For the fit indices, the best values are in boldface.

Political and Social Research (ICPSR), see (Grzywacz & Marks, 1999).[5] The items are yes/no questions about possible consequences of alcohol use, for instance

> *'Did you ever, during the past 12 months, have any emotional or psychological problems from using alcohol – such as feeling depressed, being suspicious of people, or having strange ideas?'*

Unidimensionality of the data was assessed to ensure that possible heteroscedasticity effects are not due to misfit (e.g. multidimensionality). Using Mplus (Muthén & Muthén, 2007), a two-parameter probit model was fit to the tetrachoric correlation matrix using weighted least-squares estimation. The model fit was considered good according to the RMSEA which equals 0.03, the CFI which equals 0.993, and the TLI which equals 0.986.[6]

6. Results

Different models were fit to investigate whether heteroscedastic residuals and/or a skewed trait distribution underlie the data. See Table 4 for the results. First, a model with homoscedastic residuals and a normal trait distribution was considered. This model is a traditional two-parameter model and served as a baseline model. Then, all heteroscedasticity parameters δ_{1i} were freed resulting in a model with heteroscedastic residual variances in all items. As indicated by the likelihood ratio test, AIC, BIC, sBIC, and DIC, the fit of this model was not an improvement over the baseline model, that is, the restriction $\delta_{1i} = 0$ were tenable. Next, a model with a skew-normal trait distribution was fit to test whether a general effect can be detected that is not picked up by the individual items. According to the fit indices, this model was the best-fitting model. As can be seen from the table, the estimate of ζ equals 2.739, see Figure 3 for the implied distribution of θ. As can be seen from the figure and the parameter estimate, the alcohol use trait is positively skewed, indicating that more subjects are relatively unproblematic alcohol users. This is in line with the Lucke's (2012) argument that a normal distribution for traits like addiction is theoretically suboptimal. An even better theoretically motivated distribution might be a distribution with support $[0,\infty)$, see Lucke (2012).

6.1. *Illustration 2: Scores on the Raven test*

The data consist of the responses of 2,301 first-year psychology students to the Raven's Progressive Matrices (Raven, 2000) collected between 2001 and 2009 at the University of Amsterdam (Bakker & Wicherts, 2013; see also Wicherts & Bakker, 2012). From the 36 items, 10

[5] The opinions expressed in this article are those of the author and do not necessarily reflect the views of the ICPSR.

[6] RMSEA values smaller than 0.05 are generally considered to indicate good model fit. CFI and TLI values larger than 0.95 are considered to indicate good model fit, see Hu and Bentler (1999)

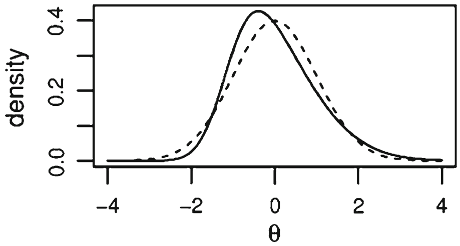

FIGURE 3.
Solid line The estimated distribution of the alcohol trait in application 1. *Dashed line* a normal distribution as specified under the baseline model.

items were selected. Proportions correct of these items are 0.82, 0.71, 0.70, 0.71, 0.65, 0.70, 0.55, 0.59, 0.36 and 0.48. Models were again fit in R as described in application 1. As before, to assess unidimensionality, a one factor model was fit to the data. Results indicated that unidimensionality is tenable as judged by the RMSEA (0.040), the CFI (0.952) and the TLI (0.957).

7. Results

First, a traditional two-parameter model and a model with heteroscedastic residuals only were fit. See Table 5 for the modelling results. As can be seen, it was unclear whether the model with heteroscedastic residuals (model 2) is an improvement over the baseline model (model 1). That is, the LRT, AIC and sBIC indicate that model 2 is the better fitting model, while the BIC and DIC indicate that model 1 is the better fitting model. However, the large likelihood ratio might indicate that at least some items are associated with heteroscedastic residuals. This was studied next. First, a model with a skewed trait distribution was fit (model 3a). All fit indices indicated that this model fits better than the traditional model. As the likelihood ratio already indicated that heteroscedastic residuals characterize these data, the question arose whether this heteroscedasticity is due to the skewed trait, or whether some additional item-specific heteroscedasticity is present. To investigate this question, univariate LM statistics were calculated within model 3a for the δ_{1i} parameters of all items as discussed above.

The univariate LM statistics for δ_{1i} equaled 1.05, 0.17, 16.14, 13.07, 6.52, 8.51, 0.48, 28.01, 0.71 and 0.22 for items 1–10, respectively. As these statistics are $\chi^2(1)$ distributed, δ_{1i} was freed for items 3, 4, 5, 6 and 8. In the resulting model 3b, Lagrange multipliers for the fixed δ_{1i} equaled, respectively, 0.11, 1.73, 1.73, 0.01 and 2.98 for item 1, 2, 7, 9 and 10. That is, there was no indication of additional item-specific heteroscedasticity. As can be seen in Table 5, the LRT, AIC and sBIC indicate that this model fits best among the models considered. The BIC favours the

TABLE 5.
Model fit results for application 2.

Model	−2LL	LRT	*df*	AIC	BIC	sBIC	DIC
1. Baseline	26,854.90			26,895	27,010	26,946	26,973
2. Heteroscedastic residuals	26,801.26	2 vs. 1: 53.64	10	26,861	27,034	26,938	26,978
3a. Skew trait distribution	26,835.15	3 vs. 1: 19.75	1	26,877	**26,998**	26,931	**26,959**
3b. Skew and het. on 3,4,5,6, 8	26,805.48	**3b vs. 3a: 29.67**	5	**26,857**	27,007	**26,924**	**26,959**

−2LL denotes −2 times the loglikelihood. For the fit indices, best values are in boldface.

TABLE 6.
Parameter estimates for the final model (model 3b) in application 2.

Item no	Baseline		Model 3b		
	α_i	β_i	α_i	β_i	δ_{1i}
1	0.509	1.037	0.492	1.029	–
2	0.370	0.584	0.367	0.582	–
3	0.468	0.581	0.484	0.635	0.797
4	0.591	0.642	0.593	0.713	1.046
5	0.423	0.430	0.427	0.484	0.773
6	0.633	0.619	0.633	0.678	0.751
7	0.915	0.179	0.926	0.164	–
8	1.158	0.337	1.001	0.009	−3.651
9	0.670	−0.437	0.668	−0.443	–
10	0.700	−0.062	0.693	−0.071	–
ζ	–		−1.570		

model without item-specific heteroscedasticity, and the DIC is undecided. As most indices favour model 3b, this model was accepted as the final model. See Table 6 for the parameter estimates in this model and the parameter estimates in the baseline model (model 1, i.e. a traditional two-parameter model). The item characteristic functions of items 3, 4, 5, 6 and 8 that are associated with item-specific heteroscedasticity are in Figure 4 both under the homoscedastic model (striped line) and under the heteroscedastic model (solid line). The item information functions for these items are in Figure 5 again both under the homoscedastic model (dashed line) and under the heteroscedastic model (solid line). As can be seen from Figure 4, the item characteristic function in the homoscedastic case approximates the function in the heteroscedastic case, with systematic under- or over-prediction. Most importantly, as the heteroscedastic ICFs are asymmetric, the point at which the slope of the ICF has it maximum can differ noticeably between the homoscedastic and heteroscedastic cases. This causes the estimated amount of item information and the location on the θ-scale at which the maximum amount of information is obtained to differ considerably between the homoscedastic and heteroscedastic models, as can be seen from Figure 5.

8. Discussion

From the results in this paper, it can be concluded that disentangling heteroscedastic residuals from skewness in the latent trait is a possible but demanding endeavour. As dichotomous data contain less information concerning individual differences, larger sample sizes were needed as compared to the polytomous case. In the polytomous case, Molenaar et al. (2012) obtained satisfactory modelling results (acceptable power and parameter recovery) for $N = 400$, while in the present paper, at least 1,000 to 3,000 subjects were needed. Therefore, the model seems most useful in cases where large sample sizes are involved (e.g. computerized adaptive testing or national surveys). That is, in these cases, effects of heteroscedasticity could be discovered. However, for smaller sample sizes, it could be sufficient to only consider a skew-normal distribution for the trait, as from the simulation study and the applications, it appears that most of the effects of heteroscedasticity could be captured in this way.

In the present paper, a distinction was made between latent and observed moderators. In the case of an observed moderator, the moderator is external to the measurement model (i.e. the

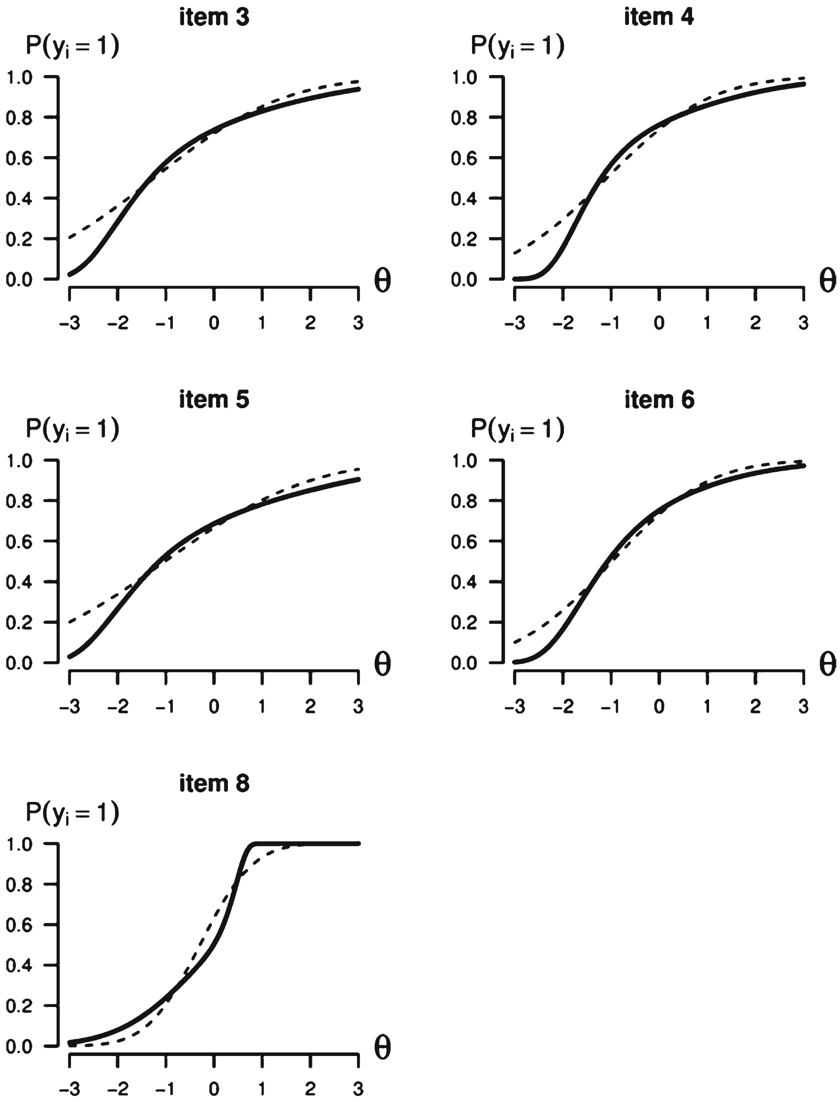

FIGURE 4.
Solid line The model implied item characteristic functions for the items that display heteroscedasticity in application 2. *Dashed line* the item characteristic functions for the corresponding item under the baseline model (a traditional two-parameter model).

moderator is not an indicator from the measurement model). In the case of a latent moderator variable, it was assumed that the moderator is latent to the data at hand. An interesting possibility to consider in future research is the case in which the latent moderator will be an external latent trait with its own measurement model. This will address the latent variable interaction literature (Kenny & Judd, 1984; Klein & Moosbrugger, 2000). The models could be interesting in for instance testing the personality differentiation hypothesis (Austin, Deary, & Gibson, 1997), which postulates that personality and intelligence interact. In addition, in the field of behaviour genetics, genotype by environment interactions are investigated by either the latent continuous moderator approach (Molenaar, van der Sluis, Boomsma, & Dolan, 2012; Van der Sluis et al., 2006) or the observed continuous moderator approach (Purcell, 2002; van der Sluis et al., 2012). For instance, it was

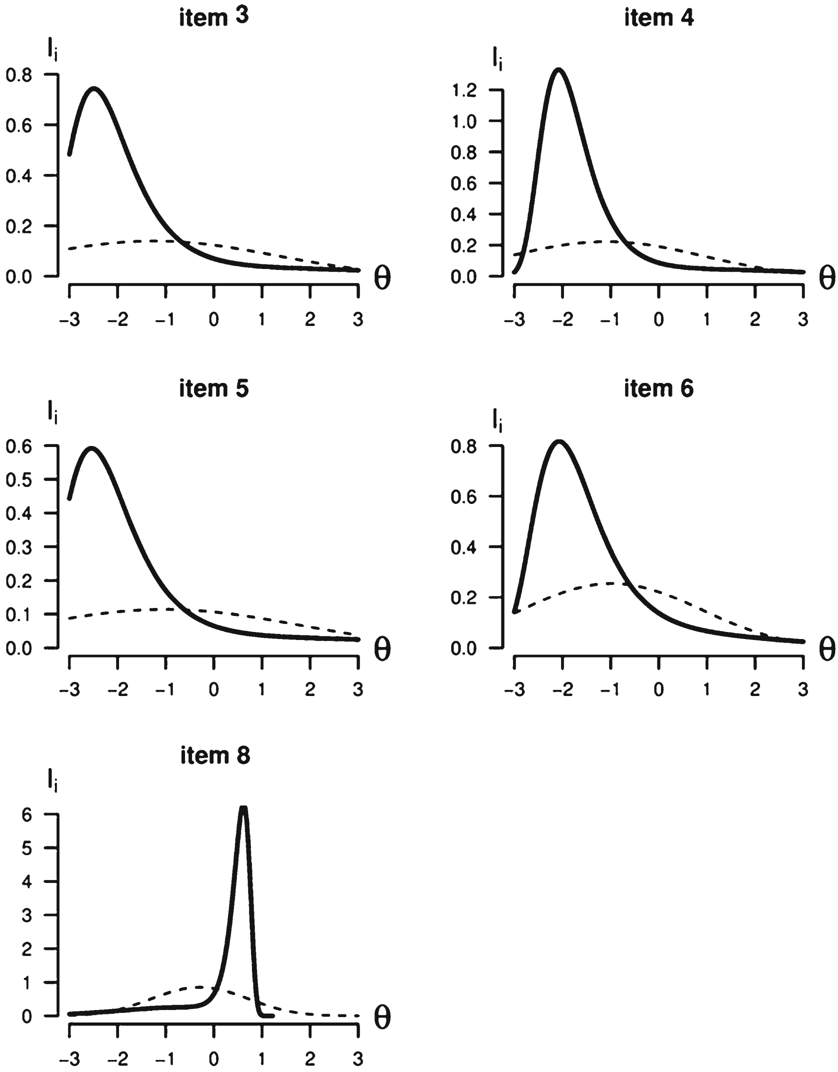

FIGURE 5.

Solid line The model implied item information functions for items that display heteroscedasticity in application 2. *Dashed line* the item information function for the corresponding item under the baseline model (a traditional two-parameter model). The x-scales of the plots are deliberately chosen to be the same as in Figure 5.

found that for cognitive ability tests, the genetic factor is heteroscedastic across socioeconomic status (Turkheimer, Haley, Waldron, D'Onofrio, & Gottesman, 2003) indicating a genotype by environment interaction on cognitive ability. If heteroscedasticity of the genetic factor across a latent trait is of interest (e.g. depression), then an external latent moderator approach might be useful.

Acknowledgments

In memory of Roger Millsap whose comments and guidance were of great help in preparing the final version of this paper. I am also indebted to Mijke Rhemtulla, Conor Dolan, and two anonymous reviewers for their valuable comments on a previous draft of the manuscript.

References

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, *62*, 245–253.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Austin, E. J., Deary, I. J., & Gibson, G. J. (1997). Relationships between ability and personality: Three hypotheses tested. *Intelligence*, *25*(1), 49–70.

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.

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

Bakker, M., & Wicherts, J. M. (2013). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*. doi:10.1037/met0000014.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. UK: Wiley.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125.

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 E. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores (chap* (pp. 17–20). Reading, MA: Addison Wesley.

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

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. (1989). *Structural equations with latent variables*. New York: Wiley.

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, NJ: Erlbaum.

Brunner, E., Dette, H., & Munk, A. (1997). Box-Type Approximations in Nonparametric Factorial Designs. *Journal of the American Statistical Association*, *92*, 1494–1502.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32.

Dobson, A. J. (2010). *An introduction to generalized linear models*. Boca Raton, FL: CRC Press.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309–326.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & van de Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence*, *34*, 193–210.

Dolan, C. V., & van der Maas, H. L. (1998). Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*, 227–253.

Dolan, C. V., & Molenaar, P. (1991). A comparison of four methods of calculating standard errors of maximum-likelihood estimates in the analysis of covariance structure. *British Journal of Mathematical and Statistical Psychology*, *44*, 359–368.

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.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*, 3–26.

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*, 520–531.

Greene, W. (2011). *Econometric analysis* (7th ed.). New York: Prentice Hall.

Grzywacz, J. G., & Marks, N. F. (1999). Family solidarity and health behaviors: Evidence from the National Survey of Midlife Development in the United States. *Journal of Family Issues, 20*(2), 243–268.

Harvey, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. Econometrica: Journal of the Econometric Society, 44, 461–465.

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

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Jansen, B. R., & van der Maas, H. L. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*, 321–357.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*, 255–259.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite mixture structural equation models for response based segmentation and unobserved heterogeneity. *Marketing Science*, *16*, 39–59.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347–387.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201–210.

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.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474.

Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1989). Simultaneous analysis of multivariate polytomous variates in several groups. *Psychometrika*, *54*, 63–73.

Lewin-Koh, S., & Amemiya, Y. (2003). Heteroscedastic factor analysis. *Biometrika*, *90*, 85–97.

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

Lord, F. M. (1952). *A theory of test scores*. New York: Psychometric Society.

Lucke, J.F. (2012). Positive Trait Item Response Models. Proceedings of the 2012 Joint Statistical Meetings of the American Statistical Association, the International Biometric Society, the Institute of Mathematical Statistics, and Statistica Canada. San Diego, CA. 2012. Retrieved from http://works.bepress.com/joseph_lucke/35

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284.

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.

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

Merkle, E. C., Fan, J., & Zeileis, A. (2013). Testing for measurement invariance with respect to an ordinal variable. Psychometrika, 1–16. doi:10.1007/s11336-013-9376-7.

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*, 59–82.

Mellenbergh, G. J. (1994a). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223–236.

Mellenbergh, G. J. (1994b). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195–215.

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

Molenaar, D., Dolan, C. V., & van der Maas, H. L. (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 and Statistical Psychology*, *63*, 293–317.

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

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.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*, 391–411.

Murray, A. L., Dixon, H., & Johnson, W. (2013). Spearman's law of diminishing returns: A statistical artifact? *Intelligence*, *41*, 439–451.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551–560.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407–419.

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

Neale, M. C., Aggen, S. H., Maes, H. H., Kubarych, T. S., & Schmitt, J. E. (2006a). Methodological issues in the assessment of substance use phenotypes. *Addicitive Behavior*, *31*, 1010–34.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2006b). *Mx: Statistical modeling* (7th ed.). VCU, Richmond, VA: Author.

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). New York: Springer.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research*, *5*, 554–571.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1–48.

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

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores (Psychometric Monograph No. 17). Richmond, VA: 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.

SAS Institute. (2011). SAS/STAT 9.3 user's guide. SAS Institute.

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

Selzer, M. L. (1971). The michigan alcohol screening test: The quest for a new diagnostic instrument. *American Journal of Psychiatry*, *127*, 89–94.

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.

Skrondal, A. (1996). *Latent trait, multilevel and repeated measurement modeling with incomplete data of mixed measurement levels*. Oslo: UiO .

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC.

Slutsky, E. E. (1913). On the criterion of goodness of fit of the regression lines and on the best method of fitting them to the data. *Journal of the Royal Statistical Society*, *77*, 78–84.

Stevens, J. (2009). *Applied multivariate statistics for the social sciences*. USA: Taylor & Francis.

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.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501–519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological science*, *14*(6), 623–628.

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.

van der Sluis, S., Posthuma, D., & Dolan, C. V. (2012). A note on false positives and power in G× E modelling of twin data. *Behavior Genetics*, *42*, 170–186.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*, *40*, 73–76.

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.

Manuscript Received: 7 NOV 2013