British Journal of Mathematical and Statistical Psychology (2012), 65.1: 32-55

DOI: http://dx.doi.org/ 10.1111/j.2044-8317.2010.02009.x

LOGISTIC RESPONSE MODELS WITH ITEM INTERACTIONS

Javier Revuelta

Address:

Javier Revuelta
Department of Social Psychology and Methodology
Autonoma University of Madrid
28049 Madrid
SPAIN

e.mail: javier.revuelta@uam.es

Note:

I would like to express my gratitude to the editor and two anonymous reviewers, whose comments contributed to improve earlier versions of this paper.

Published in: Revuelta, J. (2011). Logistic response models with item interactions. *British Journal of Mathematical & Statistical Psychology, 65(1)*, 32-55

Abstract

Items that are clustered according to shared content may violate the principle of conditional independence commonly used in Item Response Theory. This paper investigates the capabilities of a logistic item response model in relation to locally dependent item responses. The model includes main effect and interaction parameters that are computed as linear functions of the latent trait. The paper explains the interpretation of the parameters, the maximum likelihood estimation algorithm, the information matrix and some results concerning parameter identifiability. The problem of over-fitting the data is addressed in a simulation study, and two real data examples are described to illustrate the approach, one from the context of a sample survey and the other from ability testing using testlets.

Key words: Locally dependent items, local independence, logistic models, item response theory.

1. Introduction

Local stochastic independence is a common assumption in Item Response Theory. However, there are applied contexts in which the items are organized in clusters and the responses to different items may show dependencies even after conditioning on the latent trait. Items are clustered, for example, when they are related to a common stem (Wainer, Bradlow & Wang, 2007) or learning during a test occurs (Verhelst & Glas, 1993; Hoskens & de Boeck, 1997).

Logistic response models for locally dependent items have regularly appeared in the psychometric literature. Jannarone (1986, 1997), Kelderman (1984) and Kelderman and Rijkes (1994) proposed interaction models in the Rasch family of models. Hoskens and de Boeck (1997) and Ip, Smits and de Boeck (2009) proposed a number of interaction models based on a two-parameter logistic model (2PL), which are thus unidimensional and applicable to dichotomous data. Multidimensional extensions appeared in Hoskens and de Boeck (2001).

This paper explores the capabilities of logistic models to parameterize interactions. The statistical framework is called the generalized logit linear item response model (GLLIRM), and introduces a number of novelties: 1) it is a general formulation that can accommodate different patterns of interactions and/or covariates through linearly constrained parameters; 2) main effects and interactions are represented by scale and intercept parameters; 3) parameters are interpreted by investigating the invariants of item response functions; 4) the model is applicable to dichotomous and polytomous data; 5) marginal maximum likelihood estimation equations are described, and 6) the paper includes new results for identifiability and its relation to the log-linear model (Agresti, 2002), which is also a particular case of the proposed formulation.

The GLLIRM belongs to the so-called canonical kernels (Ip, Wang, de Boeck & Meulders, 2004), and some properties of the GLLIRM that are not obvious from the canonical kernel formulation will be described below. In particular, one problem with logistic interaction models is that item response functions are not reproducible (Ip, 2002; Ip et al., 2004), which means that the item response functions are altered by the inclusion of interaction parameters. For these reasons, this paper investigates which properties of the item response functions are invariant with respect to the number of items in the cluster and the relationship of these properties with the parameters.

2. Model for item interactions

2.1. The statistical model

Consider a contingency table that cross-classifies a multinomial sample on several categorical variables. The cells of the table are defined by the different response patterns that can be given to a set of T items, and K_t is the number of categories for item t. The number of cells is $C = \prod_{t=1}^{T} K_t$. Cell c is defined by a response pattern $\boldsymbol{k}_c = (k_{c1}, \ldots, k_{ct}, \ldots, k_{cT})$, where $k_{ct} = 1, \ldots, K_t$. A log-linear model (LLM) under multinomial sampling assumes that the probability for the realization $\mathbf{K} = \boldsymbol{k}_c$ is given by (e.g., Laird, 1991; Agresti, 2002):

$$P(\boldsymbol{k}_c) = \frac{\exp(z_c)}{\sum_{c'=1}^{C} \exp(z_{c'})},$$

where z_c is the propensity, or utility in econometric terms, towards response pattern $\boldsymbol{k}_c$ and is given by:

$$z_c = \sum_{t=1}^{T} \delta_{k_{ct}}^{(t)} + \sum_{t_1=1}^{T} \sum_{t_2=t_1+1}^{T} \delta_{k_{ct_1}k_{ct_2}}^{(t_1t_2)} + \sum_{t_1=1}^{T} \sum_{t_2=t_1+1}^{T} \sum_{t_3=t_2+1}^{T} \delta_{k_{ct_1}k_{ct_2}k_{ct_3}}^{(t_1t_2t_3)} + \cdots + \delta_{k_{c1}\cdots k_{cT}}^{(1\cdots T)} .$$

The parameter $\delta_k^{(t)}$ is the main effect of category k of item t, $\delta_{k_1k_2}^{(t_1t_2)}$ is the second-order interaction for categories k_1 and k_2 of the item pair $t_1 < t_2$ and so on up to the *T*-th-order interaction $\delta_{K_1..K_T}^{(1..T)}$. The constant $\Lambda = -\log \sum_{c=1}^{C} \exp(z_c)$ will be used to simplify notation, so that the model can be written as:

$$\log P(\boldsymbol{k}_c) = z_c + \Lambda.$$

The parameters are subject to identifiability constraints, two of them seem especially useful for interpretation: 1) the sum of the parameters over categories is zero, and 2) every item has a baseline category whose parameters are set to zero. The first type of constraint imposes the equalities:

$$\begin{gathered}
\sum_{k=1}^{K_t} \delta_k^{(t)} = 0 \ \text{ for any } t, \\
\sum_{k_1=1}^{K_{t_1}} \delta_{k_1k_2}^{(t_1t_2)} = \sum_{k_2=1}^{K_{t_2}} \delta_{k_1k_2}^{(t_1t_2)} = 0 \ \text{ for any } t_1 < t_2, \\
\vdots \\
\sum_{k_1=1}^{K_{t_1}} \delta_{k_1k_2\cdots k_T}^{(12\cdots T)} = \sum_{k_2=1}^{K_{t_2}} \delta_{k_1k_2\cdots k_T}^{(12\cdots T)} = \cdots = \sum_{k_T=1}^{K_{t_T}} \delta_{k_1k_2\cdots k_T}^{(12\cdots T)} = 0.
\end{gathered}$$

Without loss of generality, we assume that category 1 is the baseline category for every item. Then, constraints in relation to the baseline impose the equalities:

$$\begin{gathered}
\delta_1^{(t)} = 0 \ \text{ for any } t, \\
\delta_{1k_2}^{(t_1t_2)} = \delta_{k_11}^{(t_1t_2)} = 0 \ \text{ for any } t_1 < t_2, \\
\vdots \\
\delta_{1k_2\cdots k_T}^{(12\cdots T)} = \delta_{k_11\cdots k_T}^{(12\cdots T)} = \cdots = \delta_{k_1k_2\cdots 1}^{(12\cdots T)} = 0.
\end{gathered}$$

LLMs are based on the assumption that individuals are homogeneous because the parameters are constant across individuals. In contrast to this, GLLIRM assumes that the parameters are functions of a latent trait, θ, and thus z_c becomes a latent propensity. Let V be a nonempty set of items, and let $\boldsymbol{k}(V)$ be the vector of responses to V. The parameters are given by the linear function:

$$\delta^{(V)}_{\boldsymbol{k}(V)} = \alpha_j \theta + \beta_j .$$

LLM is the particular case $\alpha_j = 0$, whereas in general, α_j can take on arbitrary real values for GLLIRM. This has the effect of doubling the number of parameters. A saturated LLM contains all the parameters up to T-th-order interaction $\delta^{(1\cdots T)}_{k_{ct_1}\cdots k_{ct_T}}$ and has exactly $C-1$ parameters and zero degrees of freedom. A GLLIRM model of the same order would have $2(C-1)$ parameters and would run out of degrees of freedom. Thus, not all of the interaction parameters for GLLIRM can be estimated.

Example. Consider a cluster composed of a dichotomous and a trichotomous item. The probability of the response pattern $\boldsymbol{k}_c = (k_{c1}, k_{c2})$, with $k_{c1} = 1, 2$ and $k_{c2} = 1, 2, 3$, is given by:

$$\log P(k_{c1}, k_{c2}) = \delta^{(1)}_{k_{c1}} + \delta^{(2)}_{k_{c2}} + \delta^{(12)}_{k_{c1}k_{c2}} + \Lambda,$$

where $\Lambda = -\log \sum_{c=1}^{C} \exp(\delta^{(1)}_{k_{c1}} + \delta^{(2)}_{k_{c2}} + \delta^{(12)}_{k_{c1}k_{c2}})$. The GLLIRM parameters are $\delta^{(t)}_{k} = \alpha^{(t)}_{k}\theta + \beta^{(t)}_{k}$ and $\delta^{(12)}_{k_1k_2} = \alpha^{(12)}_{k_1k_2}\theta + \beta^{(12)}_{k_1k_2}$. The parameter constraints can be visualized by writing the model in a matrix form. If the sum over categories is zero, the probabilities of the response patterns can be expressed as:

$$\begin{pmatrix} \log P(1,1) \\ \log P(1,2) \\ \log P(1,3) \\ \log P(2,1) \\ \log P(2,2) \\ \log P(2,3) \end{pmatrix} = \begin{pmatrix} -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_2^{(1)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \\ \delta_{22}^{(12)} \\ \delta_{23}^{(12)} \end{pmatrix} + \begin{pmatrix} \Lambda \\ \Lambda \\ \Lambda \\ \Lambda \\ \Lambda \\ \Lambda \end{pmatrix},$$

where the following constraints are implicit: $\delta_1^{(1)} = -\delta_2^{(1)}$, $\delta_1^{(2)} = -\delta_2^{(2)} - \delta_3^{(2)}$, $\delta_{11}^{(12)} = \delta_{22}^{(12)} + \delta_{23}^{(12)}$, $\delta_{12}^{(12)} = -\delta_{22}^{(12)}$, $\delta_{13}^{(12)} = -\delta_{23}^{(12)}$, and $\delta_{21}^{(12)} = -\delta_{22}^{(12)} - \delta_{22}^{(12)}$. If the first category is a baseline category, the model reads in matrix form:

$$\begin{pmatrix} \log P(1,1) \\ \log P(1,2) \\ \log P(1,3) \\ \log P(2,1) \\ \log P(2,2) \\ \log P(2,3) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_2^{(1)} \\ \delta_2^{(2)} \\ \delta_3^{(2)} \\ \delta_{22}^{(12)} \\ \delta_{23}^{(12)} \end{pmatrix} + \begin{pmatrix} \Lambda \\ \Lambda \\ \Lambda \\ \Lambda \\ \Lambda \\ \Lambda \end{pmatrix},$$

where it is implicit that $\delta_1^{(1)} = 0$, $\delta_1^{(2)} = 0$, $\delta_{11}^{(12)} = 0$, $\delta_{12}^{(12)} = 0$, $\delta_{13}^{(12)} = 0$, and $\delta_{21}^{(12)} = 0$. Using baseline constraints, the marginal probability of passing the dichotomous item can be written as:

$$P(2) = \frac{\sum_{k'=1}^{3} \exp(\delta_2^{(1)} + \delta_{k'}^{(2)} + \delta_{2k'}^{(12)})}{\sum_{k=1}^{2} \sum_{k'=1}^{3} \exp(\delta_k^{(1)} + \delta_{k'}^{(2)} + \delta_{kk'}^{(12)})}$$

$$= \frac{\exp(\delta_2^{(1)}) \sum_{k'=1}^{3} \exp(\delta_{k'}^{(2)} + \delta_{2k'}^{(12)})}{[\exp(\delta_1^{(1)}) \sum_{k'=1}^{3} \exp(\delta_{k'}^{(2)} + \delta_{1k'}^{(12)})] + [\exp(\delta_2^{(1)}) \sum_{k'=1}^{3} \exp(\delta_{k'}^{(2)} + \delta_{2k'}^{(12)})]}$$

$$= \frac{\exp(\delta_2^{(1)}) \sum_{k'=1}^{3} \exp(\delta_{k'}^{(2)} + \delta_{2k'}^{(12)})}{[\sum_{k'=1}^{3} \exp(\delta_{k'}^{(2)})] + [\exp(\delta_2^{(1)}) \sum_{k'=1}^{3} \exp(\delta_{k'}^{(2)} + \delta_{2k'}^{(12)})]}. \tag{1}$$

If the two interaction parameters were zero ($\delta_{22}^{(12)} = \delta_{23}^{(12)} = 0$), the marginal probability would be $P(2) = \exp(\delta_2^{(1)}) / (1 + \exp(\delta_2^{(1)}))$, which is the 2PL model. However the item characteristic curves are non-reproducible because the probability function in equation (1) is not the 2PL in general.

Model parameters can be interpreted using the highest-order terms as described in Agresti (2002). However, in that interpretation, the statistical meaning of the lower order terms depends on the presence of interactions of higher order. For example, the main effects of the items would depend on the interactions with other items and in consequence of the cluster size. For these reasons, sections 2.2 and 2.3 describe the properties of the response probabilities that are invariant with respect to cluster size and have a simple relationship with the parameters. For example, consider the following probability ratios:

$$\log \frac{P(2,1)}{P(1,1)} = \delta_2^{(1)}, \quad \log \frac{P(2,2)P(1,1)}{P(2,1)P(1,2)} = \delta_{22}^{(12)} \text{ and } \log \frac{P(2,3)P(1,1)}{P(2,1)P(1,3)} = \delta_{23}^{(12)}.$$

These quotients depend on a single log-linear parameter irrespective of the value of the other parameters.

Because $\boldsymbol{\delta} = \boldsymbol{\alpha}\theta + \boldsymbol{\beta}$, GLLIRM can be expressed in matrix form as $\log \mathbf{P} = \mathbf{W} + \mathbf{1}\Lambda = \mathbf{W}(\boldsymbol{\alpha}\theta + \boldsymbol{\beta}) + \mathbf{1}\Lambda$, where $\mathbf{W}$ is a matrix of fixed coefficients that imposes the constraints. However, allowing for different weights for the scale and intercept parameters is a common psychometric practice. Thus, the most general formulation of GLLIRM allows for separate weights and is given by:

$$\log \mathbf{P} = \mathbf{A}\boldsymbol{\alpha}\theta + \mathbf{B}\boldsymbol{\beta} + \mathbf{1}\Lambda,$$

where $\mathbf{A}$ and $\mathbf{B}$ are the matrices of coefficients. This formulation can be used to impose the two types of constraints described above and any other linear constraint.

2.2. Interpretation of parameters that sum to zero

The interpretation of the parameters is based on their relationship with the conditional probabilities of the response patterns. First of all, the constant Λ should be eliminated to clarify this relationship. When the sum of the parameters is set to 0, from the model equation $\log P(\boldsymbol{k}_c) = z_c + \Lambda$, it follows that $\sum_{c=1}^{C} \log P(k_c) = C\Lambda$, and in consequence, Λ is equal to the log geometric mean of the probabilities of the response patterns conditional on θ:

$$\Lambda = \frac{1}{C}\sum_{c=1}^{C} \log P(\boldsymbol{k}_c) = \log\left(\prod_{c=1}^{C} P(\boldsymbol{k}_c)\right)^{1/C}.$$

The interpretation of parameters depends on a number of geometric means:

$$\tilde{P} = \left(\prod_{c=1}^{C} P(\boldsymbol{k}_c)\right)^{1/C} = \left(\prod_{k_1=1}^{K_1}\cdots\prod_{k_T=1}^{K_T} P(k_1,\ldots,k_T)\right)^{1/(K_1\times\cdots\times K_T)},$$

$$\tilde{P}_{k_1} = \left(\prod_{k_2=1}^{K_2}\cdots\prod_{k_T=1}^{K_T} P(k_1,\ldots,k_T)\right)^{1/(K_2\times\cdots\times K_T)},$$

$$\tilde{P}_{k_1k_2} = \left(\prod_{k_3=1}^{K_3}\cdots\prod_{k_T=1}^{K_T} P(k_1,\ldots,k_T)\right)^{1/(K_3\times\cdots\times K_T)}, \quad \text{etc.}$$

Then $\tilde{P}$ is the geometric mean of the probabilities of all response patterns, $\tilde{P}_{k_1}$ is the geometric mean for the category k_1 of Item 1 and $\tilde{P}_{k_1k_2}$ is the geometric mean for the pair of responses k_1 and k_2 for Items 1 and 2. From the equation $\log P(\boldsymbol{k}_c) = z_c + \Lambda$, it follows that $\sum_{k_2=1}^{K_2}\cdots\sum_{k_T=1}^{K_T} \log P(\boldsymbol{k}_c) = (K_2\times\cdots\times K_T)(\delta_{k_1}^{(1)} + \Lambda)$; that is:

$$\frac{1}{K_2 \times \cdots \times K_T} \sum_{k_2=1}^{K_2} \cdots \sum_{k_T=1}^{K_T} \log P(k_1, \ldots, k_T) = \delta_{k_1}^{(1)} + \Lambda.$$

In consequence, the main effect parameter for category k_1 of Item 1 depends on the log ratio:

$$\delta_{k_1}^{(1)} = \log \frac{\tilde{P}_{k_1}}{\tilde{P}}.$$

By a similar argument, second-order interaction parameters are given by:

$$\delta_{k_1 k_2}^{(12)} = \log \frac{\tilde{P}_{k_1 k_2} \tilde{P}}{\tilde{P}_{k_1} \tilde{P}_{k_2}}.$$

The relationship between parameters and geometric means follows a recursive relation. Let $S(V)$ be the set of all subsets of V excluding the *empty* and the *universal* sets. Then, parameters of order two and above are defined by:

$$\delta_{\boldsymbol{k}(V)}^{(V)} = \log \frac{\tilde{P}_{\boldsymbol{k}(V)}}{\tilde{P}} - \sum_{W \in S(V)} \delta_{\boldsymbol{k}(W)}^{(W)}.$$

For example:

$$\begin{aligned} \delta_{k_1 k_2 k_3}^{(123)} &= \log \frac{\tilde{P}_{k_1 k_2 k_3}}{\tilde{P}} - \delta_{k_1 k_2}^{(12)} - \delta_{k_1 k_3}^{(13)} - \delta_{k_2 k_3}^{(23)} - \delta_{k_1}^{(1)} - \delta_{k_2}^{(2)} - \delta_{k_3}^{(3)} \\ &= \log \frac{\tilde{P}_{k_1 k_2 k_3} \tilde{P}_{k_1} \tilde{P}_{k_2} \tilde{P}_{k_3}}{\tilde{P}_{k_1 k_2} \tilde{P}_{k_1 k_3} \tilde{P}_{k_2 k_3} \tilde{P}}. \end{aligned}$$

The interpretation of the parameters of a given order is not altered by the inclusion in the model of parameters for higher-order interactions (by the size of the cluster of items, using the terminology in test construction).

2.3. Interpretation of parameters in relation to a baseline category

From the equation $\log P(\boldsymbol{k}_c) = z_c + \Lambda$, it follows that Λ is the logarithm of the probability of the baseline response pattern under this type of constraint:

$$\log P(\mathbf{1}) = \Lambda.$$

The equations for interpreting parameters have a similar structure to those in section 2.2, but the geometric means are substituted by probabilities of the baseline categories. Because $\log P(k_1, 1, \ldots, 1) = \delta_{k_1}^{(1)} + \Lambda$, the main effect parameters depend on the ratio of probabilities for the category alone and the baseline response pattern:

$$\delta_{k_1}^{(1)} = \log \frac{P(k_1, 1, \ldots, 1)}{P(\mathbf{1})}.$$

Second- and higher-order parameters are defined by the formula:

$$\delta_{\boldsymbol{k}(V)}^{(V)} = \log \frac{P(\boldsymbol{k}(V), \mathbf{1})}{P(\mathbf{1})} - \sum_{W \in S(V)} \delta_{\boldsymbol{k}(W)}^{(W)},$$

where $P(\boldsymbol{k}(V), \mathbf{1})$ is the probability for a response vector in which the responses to the items not included in V are set to 1. For example, the interaction between Items 1 and 2 is given by:

$$\delta_{k_1 k_2}^{(12)} = \log \frac{P(k_1, k_2, 1, \ldots, 1)}{P(\mathbf{1})} - \delta_{k_1}^{(1)} - \delta_{k_2}^{(2)}$$
$$= \log \frac{P(k_1, k_2, 1, \ldots, 1) P(\mathbf{1})}{P(k_1, 1, \ldots, 1) P(1, k_2, 1, \ldots, 1)}.$$

The interpretation of the highest-order parameters depends only on the responses to the interacting items. For example, suppose that a model has second-order interactions only. Then, the interaction between items 1 and 2 is the log odds

$$\delta_{k_1 k_2}^{(12)} = \log \frac{P(k_1, k_2, \boldsymbol{k}) P(1, 1, \boldsymbol{k})}{P(k_1, 1, \boldsymbol{k}) P(1, k_2, \boldsymbol{k})},$$

so that the responses to the other items, $\boldsymbol{k} = (k_3, \ldots, k_T)'$, are arbitrary and need not be fixed to the baseline category.

3. Statistical inference

3.1. Parameter estimation

Let $\boldsymbol{\varepsilon}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ be the vector of parameters. The sample contains the response patterns from *n* individuals, *n_c* is the observed frequency of response pattern $\boldsymbol{k}_c$ and $\boldsymbol{k}_{c_i}$ is the pattern given by individual i. We assume that θ is a random effect and that the distribution $F(\theta_1,\ldots,\theta_n)$ is fully specified in order to avoid the introduction of more parameters into the model. Moreover, the variables θ_i are independent and identically distributed ($F(\theta_1,\ldots,\theta_n) = \prod_{i=1}^{n} F(\theta_i)$). Then, the maximum-likelihood estimate, $\hat{\boldsymbol{\varepsilon}}$, is the value that maximizes the marginal distribution:

$$\begin{aligned} L(\boldsymbol{\varepsilon}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} P(\boldsymbol{k}_{c_i}) \right) dF(\theta_1,\ldots,\theta_I) \\ &= \prod_{c=1}^{C} \left(\int_{-\infty}^{\infty} P(\boldsymbol{k}_c)\, dF(\theta) \right)^{n_c}. \end{aligned}$$

The integral involved in the marginal probability $\pi_c(\boldsymbol{\varepsilon}) = \int_{-\infty}^{\infty} P(\boldsymbol{k}_c)\, dF(\theta)$ is solved in practice using numerical quadrature (e.g., Stoer & Bulirsch, 1980). The log-likelihood function can be given as $l(\boldsymbol{\varepsilon}) = \log L(\boldsymbol{\varepsilon}) = \boldsymbol{n}' \log \boldsymbol{\pi}$, where $\boldsymbol{n} = (n_1,\ldots,n_C)'$ and $\log = (\log \pi_1(\boldsymbol{\varepsilon}),\ldots,\log \pi_C(\boldsymbol{\varepsilon}))'$.

The maximum likelihood estimate is the solution to the estimation equation $\partial l(\boldsymbol{\varepsilon}) / \partial \boldsymbol{\varepsilon} = \mathbf{0}$, which is solved iteratively using the EM algorithm (Bock & Aitkin, 1981). The E step of the EM consists of computing the numerical value of the matrix $\mathbf{F} = \mathrm{diag}(f(\theta \mid \boldsymbol{k}_1),\ldots, f(\theta \mid \boldsymbol{k}_C))$, where $f(\theta \mid \boldsymbol{k}_c)$ is the posterior distribution of θ conditional on $\boldsymbol{k}_c$, for the grid of values of θ used in numerical quadrature. The M step consists of iteratively solving the likelihood equations while keeping $\mathbf{F}$ constant to the values obtained in the E step (e.g., Nocedal & Wright, 2006). The first-order derivatives involved in the M step can be written as:

$$\frac{\partial}{\partial\boldsymbol{\alpha}} l(\boldsymbol{\varepsilon}) = \left[\int_{-\infty}^{\infty} \theta(\mathbf{A}' - \mathbf{A}'\boldsymbol{p}1')\mathbf{F}\, d\theta\right]\boldsymbol{n}, \text{ and}$$
$$\frac{\partial}{\partial\boldsymbol{\beta}} l(\boldsymbol{\varepsilon}) = \left[\int_{-\infty}^{\infty} (\mathbf{B}' - \mathbf{B}'\boldsymbol{p}1')\mathbf{F}\, d\theta\right]\boldsymbol{n},$$

where $\boldsymbol{p} = (P(\boldsymbol{k}_1),\ldots,P(\boldsymbol{k}_C))'$ is the vector of conditional probabilities and the matrix integral is the matrix formed by the integral of each element. The EM iterates between the E step (computing $\mathbf{F}$) and the M step (solving the equation $\partial l(\boldsymbol{\varepsilon})/\partial\boldsymbol{\varepsilon} = \mathbf{0}$ with respect to $\boldsymbol{\varepsilon}$ and keeping $\mathbf{F}$ fixed) until convergence.

3.2. Information matrix and identifiability

The asymptotic distribution of the maximum likelihood estimate is $\sqrt{n}(\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}) \xrightarrow{L} N(\mathbf{0}, I^{-1})$ under regularity conditions (Birch, 1964; Bishop, Fienberg &Holland, 1975; Cox, 1984). The symbol I represents the Fisher information matrix and is given by (Andersen, 1980, p. 96):

$$I = \boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta}. \tag{3}$$

where $\mathbf{D} = \text{diag}(\boldsymbol{\pi})$, $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\varepsilon})/\partial\boldsymbol{\varepsilon}'$ is the Jacobian matrix of $\boldsymbol{\pi}$ and:

$$\boldsymbol{\Delta} = \mathbf{EC}, \quad \text{with} \quad \mathbf{E} = (\mathbf{M};\mathbf{N}) \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}.$$

Moreover, let $\boldsymbol{\Sigma}$ be the variance–covariance matrix of $\mathbf{K}$ conditional on θ; that is, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{p}) - \boldsymbol{pp}'$. Then, $\mathbf{M} = \int_{-\infty}^{\infty} \theta\boldsymbol{\Sigma}\, dF(\theta)$ and $\mathbf{N} = \int_{-\infty}^{\infty} \boldsymbol{\Sigma}\, dF(\theta)$. Inserting the expression for $\boldsymbol{\Delta}$ into equation (3), the information matrix takes the form

$$I = \mathbf{C}'\mathbf{E}'\mathbf{D}^{-1}\mathbf{EC}.$$

Identifiability of the parameters depends on the rank of I. A point in parameter space is locally identifiable if I is nonsingular (Rothenberg, 1971; Wansbeek & Meijer, 2000; Bekker & Wansbeek, 2003). A necessary identifiability condition is that the vector **1** is not in the column space of $\mathbf{A}$ and $\mathbf{B}$. To prove this, we first note that $\mathbf{M}$ and $\mathbf{N}$ are rank deficient. This is demonstrated by seeking the solution, $\boldsymbol{v}$, to the equation $\mathbf{0} = \mathbf{N}\boldsymbol{v}$, which reads:

$$\mathbf{0} = \mathbf{N}\boldsymbol{v} = \int_{-\infty}^{\infty} (\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}')\boldsymbol{v}\, dF(\theta).$$

All the vectors of the form $\boldsymbol{v} = \mathbf{1}v$, with v arbitrary, satisfy $\mathbf{0} = (\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}')\boldsymbol{v}$, and thus $\mathbf{N}$ is rank deficient because $\mathbf{0} = \mathbf{N}\boldsymbol{v}$. A similar argument applies to $\mathbf{M}$, and therefore $\mathbf{0} = \mathbf{E1}$.

Secondly, suppose that the vector **1** is in the column space of $\mathbf{A}$ and $\mathbf{B}$. In that case, there would be a vector $\boldsymbol{y}$ that satisfies the equation $\mathbf{1} = \mathbf{C}\boldsymbol{y}$, and because $\mathbf{0} = \mathbf{E1}$, one gets that $\boldsymbol{y}'I\boldsymbol{y} = \boldsymbol{y}'\mathbf{C}'\mathbf{E}'\mathbf{D}^{-1}\mathbf{E}\mathbf{C}\boldsymbol{y} = \mathbf{1}'\mathbf{E}'\mathbf{D}^{-1}\mathbf{E1} = 0$. Thus I has a zero eigenvalue and so is singular.

In summary, the necessary identifiability condition is full column rank of the matrix $(\mathbf{1};\mathbf{C})$, which reduces to full column rank of $(\mathbf{1};\mathbf{A})$ and $(\mathbf{1};\mathbf{B})$. Moreover, because the rank of $\mathbf{E}$ is smaller than C and the rank of a matrix product cannot exceed the rank of any of its members, the number of parameters in $\boldsymbol{\varepsilon}$ cannot be higher than $C-1$ for I to be nonsingular. The identifiability condition for the GLLIRM is a generalization of the identifiability condition for an LLM under multinomial sampling (Agresti, 2002, chapter 14). In particular, the matrix $\mathbf{A}$ is dropped in an LLM, and the identifiability condition reduces to full column rank of $(\mathbf{1};\mathbf{B})$.

Example. Consider two LLMs based on the matrices:

$$\mathbf{B}^{(1)} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \text{ and } \mathbf{B}^{(2)} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 2 \\ 4 & 3 \end{pmatrix}$$

The first model satisfies the identifiability condition, whereas the second does not because the equation $\mathbf{1} = \mathbf{B}^{(2)}\boldsymbol{y}$ has the solution $\boldsymbol{y} = (1, -1)'$. The identifiability problem for these models may be visualized using the information divergence or Kullback–Leibler divergence (Bowden; 1973). The divergence $H(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ is computed conditional on a fixed point in the parameter space, $\boldsymbol{\beta}_0$, and is a function of $\boldsymbol{\beta}$. $H(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ measures the difference between the distributions $\boldsymbol{\pi}(\boldsymbol{\beta})$ and $\boldsymbol{\pi}(\boldsymbol{\beta}_0)$, so that $H(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = 0$ when $\pi_c(\boldsymbol{\beta}) = \pi_c(\boldsymbol{\beta}_0)$ for any c, and $H(\boldsymbol{\beta}, \boldsymbol{\beta}_0) < 0$ otherwise. Thus $H(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = 0$ indicates that the points $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ are observationally equivalent and not identifiable. The definition of the information divergence is:

$$H(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = \sum_{c=1}^{C} \left(\log \frac{\pi_c(\boldsymbol{\beta})}{\pi_c(\boldsymbol{\beta}_0)} \right) \pi_c(\boldsymbol{\beta}_0).$$

The upper panel of Figure 1 shows the value of $H(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ in a neighborhood of $\boldsymbol{\beta}_0 = \mathbf{0}$ for the first model. $H(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ takes the value 0 only when $\boldsymbol{\beta} = \mathbf{0}$; this means that there are no points $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ that are observationally equivalent to $\boldsymbol{\beta}_0$.

(Figure 1)

The lower panel of Figure 1 shows the value of the divergence for the second model. The function $H(\boldsymbol{\beta},\boldsymbol{\beta}_0)$ takes the value 0 along the line $\boldsymbol{\beta}=(1,-1)'t$ for t arbitrary, and all the points in this line are observationally equivalent.

These findings illustrate a general result: all the points in the parameter space that are proportional to the solution of $\mathbf{1}=\mathbf{C}\boldsymbol{y}$ are observationally equivalent. This can be demonstrated by observing that the directional derivative of $H(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}_0)$ at the direction of $\boldsymbol{y}$ is zero. More specifically, consider a line segment in the parameter space between two points $\boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon}_1$; that is, $\boldsymbol{\varepsilon}_2=\boldsymbol{\varepsilon}_0+(\boldsymbol{\varepsilon}_1-\boldsymbol{\varepsilon}_0)t$ where $t\in[0,1]$. The derivative of $H(\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_0)$ with respect to t is given by:

$$H'(\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_0)=\mathbf{1'D}[\boldsymbol{\varepsilon}_0]\mathbf{D}^{-1}[\boldsymbol{\varepsilon}_2]\mathbf{E}[\boldsymbol{\varepsilon}_2]\mathbf{C}(\boldsymbol{\varepsilon}_1-\boldsymbol{\varepsilon}_0),$$

where $\mathbf{D}[\boldsymbol{\varepsilon}_0]$, $\mathbf{D}^{-1}[\boldsymbol{\varepsilon}_2]$ and $\mathbf{E}[\boldsymbol{\varepsilon}_2]$ represent matrix functions and the term in brackets is the argument, whereas $\mathbf{C}(\boldsymbol{\varepsilon}_1-\boldsymbol{\varepsilon}_0)$ represents a matrix product. Therefore, if $\boldsymbol{y}$ is the solution to $\mathbf{1}=\mathbf{C}\boldsymbol{y}$ and because $\mathbf{E1}=\mathbf{0}$, then $H'(\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_0)=0$ in the direction of the vector $(\boldsymbol{\varepsilon}_1-\boldsymbol{\varepsilon}_0)=\boldsymbol{y}$. The consequence is that $H(\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_0)$ and $H'(\boldsymbol{\varepsilon}_2,\boldsymbol{\varepsilon}_0)$ are constant with value 0 along the line segment, and all the points that lie on it are observationally equivalent.

4. Simulation study

GLLIRMs can be heavily parameterized. Even models that satisfy the identifiability conditions may present problems of estimation if the data does not contain enough information to estimate with precision all the parameters. This section presents a simulation study conducted to investigate the problems raised by an incorrect specification of the estimated model -in particular the dangers involved in estimating a model with fewer or more interaction parameters than in the true model- and the increase of model complexity.

An artificial test was created with eight items. Items 1 to 4 form one cluster and items 5 to 8 constitute the second cluster. Four models were defined, all with scale and intercept parameters for the main effects. The interaction parameters for each model are: 1) conditional independence (CI), which has no interactions; 2) second order (2B) includes intercept parameters for the second-order interactions; 3) third order (3B) with intercepts for the third-order interactions, and 4) fourth order (4AB) with scale and intercept parameters for the fourth-order interactions. Model CI is nested in 2B, 3B and 4AB, and the models 2B, 3B and 4AB are not nested within each other. True values of the intercepts were sampled from a normal (0, 1) distribution and the scale parameters were sampled from a lognormal distribution with a mean of 1.2 and standard deviation of 0.4. Ability had a standard normal distribution.

The design of the simulation study has three factors: simulated model (four levels), estimated model (four levels) and sample size (three levels of 500, 1000 and 5000 subjects). One thousand samples were simulated for each cell of the design from the simulated model, and the estimated model was fitted to those data.

Table 1 contains the mean value of the likelihood ratio test statistic (G^2) and the empirical proportion of rejection (EPR) for the 48 cells of the design. Several conclusions can be drawn from this table: when the simulated and estimated models coincide, the resulting statistical test is conservative. The same result is obtained when the investigator fits a model with the same parameters as in the true model and a number of unnecessary interaction parameters (the simulated model is nested in the estimated one). If the patterns of interactions in the simulated and estimated models do not match (they are not nested models), G^2 has high statistical power with a large sample.

(Table 1)

Tables 2 and 3 contain the bias of the estimated parameters for the 16 conditions with n = 500, which constitutes the worst-case scenario. The tables also contain the true parameter values. A Wald statistic (Buse, 1982) was computed by squaring the standardized bias, which is the bias divided by the simulated standard deviation of the parameter estimates. The p-value for this statistic was taken from a χ_1^2 distribution. Figures in boldface in Tables 2 and 3 indicate statistical significance at a nominal Type I error rate of 0.05.

(Table 2)

(Table 3)

Tables 2 and 3 show that estimates are unbiased when the simulated model is nested in the estimated one. However, bias arises in the estimates when the models are not nested. Thus, ignoring interaction parameters may distort the estimation of the remaining parameters. Moreover, the results for the remaining sample sizes (not shown in Tables 2 and 3) show that the biases do not vanish when the sample size increases.

Finally, the correlation matrix of the parameter estimates was computed to investigate dependencies; the results are summarized in Table 4. The table shows the correlations between the estimates of the location parameters for item cluster 1 when the simulated model is CI and the estimated models are CI or 2B. The table shows that the inclusion of interactions increases the correlation of all of the parameters, and these correlations are not reduced by large sample size. Thus, the distortions in the estimates in heavily parameterized models are not independent from one parameter to another.

(Table 4)

In summary, the general conclusion of the simulation study is that it is safer to err in the direction of over-fitting the data to avoid biases in the estimates, a significant G^2 statistic may indicate a lack of interaction parameters, and heavily parameterized models may have high standard errors and correlations between estimates. Thus, in the absence of information about which interactions should be estimated, models can be fitted using a two step process that is illustrated below in Section 5.1. First, a relatively general model is estimated, and then this model is simplified to eliminate unnecessary interactions.

5. Empirical examples

This section presents two empirical examples in which the model was estimated assuming that $F(\theta)$ is standard normal.

5.1. Sample survey with dichotomous data

A survey of religious feelings was applied to 1333 individuals. We have analyzed a cluster of five items that share a common stem. The stem reads *I would sacrifice my life*, and the five items are: 1) *for my country*; 2) *to save another's life*; 3) *for democracy*; 4) *for God* and 5) *for my family*. The response categories are Yes or No. The observed data appear in Table 5, which shows the observed frequencies and the response patterns.

The values 0 and 1 indicate No and Yes, respectively. The columns labeled *Pattern* and *Interactions* contain the columns of $\mathbf{W}$ for a model with main effects and second-order interactions in which the first category is the baseline.

(Table 5)

Several models were estimated: the one- and two-parameter logistic models; the 2PL with location parameters for the second-order interactions (2PL-2B) and with scale and location parameters for the interactions (2PL-2AB); LLM with main effects, second-order and third-order interactions (LLM-1, LLM-2 and LLM-3); and a categorical factor analysis model for two latent dimensions. Results for goodness of fit appear in Table 6. Chi square indicates that all the models fit the data, except for the 1PL and LLM-1. The 2PL model with second-order interactions (2PL-2B) minimizes the AIC.

A process of backward elimination was applied to the 2PL-1B to eliminate unnecessary interaction parameters. A Wald statistic was computed for each interaction parameter and the model re-estimated after the non-significant interactions were eliminated. This process was repeated until the model that minimizes the AIC was identified. This model was labeled 2PL-1B-Backward and the results for its goodness of fit appear in Table 6.

(Table 6)

Table 7 contains the parameter estimates for the five models. The different models offer different pictures of the data. The LLM-2 does not take into account individual differences. The factorial model includes two factors: the first is specific for items 1 and

4, and the second relates to the other items. The 2PL includes a single latent dimension that can be regarded as a tendency to sacrifice, and the 2PL-2B includes a latent dimension and interaction parameters between item pairs to explain their associations after conditioning on θ. The 2PL-2B-Backward model only includes interactions between items 1 and 2, and between 1 and 3.

(Table 7)

One interesting finding in Table 7 is that the interactions for LLM-2 are all positive, whereas the interactions for the 2PL-2B model are mostly negative. This is because LLM-2 ignores individual differences in the disposition to sacrifice and attributes all the associations in the data to second-order interactions. In contrast, the interactions for 2PL-2B are due to the associations that appear after the effect of the latent variable is removed.

The item characteristic functions for the 2PL and the 2PL-1B-Backwards appear in Figure 2. The figure illustrates that interaction parameters may produce deviations from monotonicity, which in this example affect mainly Item 3.

The lower row of Figure 2 contains the item characteristic functions of items 2 and 3 and the probability of endorsing these items conditional on the response to Item 1. The probability of endorsing items 2 and 3 is lower for those individuals who endorse Item 1 than for those who do not. Apparently, some individuals attempt to compensate for a negative response to Item 1 by responding positively to items 2 and 3.

(Figure 2)

Overall, this example shows that some associations remain in the data even after the effect of the latent variable is removed. Increasing the latent dimensionality is a way of dealing with these associations, but a unidimensional model is preferable because there is no a priori reason to seek other latent dimensions apart from the general tendency to sacrifice.

5.2. Ability testing with polytomous data

The second example uses data from the USA database for Population 3 (final year of secondary school) of the 1995 edition of the TIMSS (Gonzalez, Smith & Sibberns, 1998). Sample size is 10834. Twelve items from the mathematics literacy scale were used. The items and the responses can be retrieved from the internet address *http://timss.bc.edu/timss1995.html*.

The test contains multiple-choice and open-ended items. Moreover, it includes two testlets, each composed of two open-ended items. The structure of the test and the psychometric models applied to it appear in Table 8. The open-ended items were scored in three categories (nonresponse, incorrect and correct), according to the manual of the TIMSS, and were parameterized using the partial credit model (PCM; Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992) and the nominal categories model (NCM; Bock, 1972). Two versions of these models were applied: the original version (no interactions) and a model supplemented with location parameters for the second-order interactions within each testlet. Finally, a bifactor model was estimated. The bifactor model includes three factors, the general one and a cluster-specific factor for each testlet. Thus, each item in a testlet loads only on two factors. In this way, the latent structure of the bifactor model is the same as in the factor analytic tradition (Harman, 1976), although the measurement model used in this particular example is a multidimensional nominal categories model (Bock & Gibbons, 2010) instead of the traditional linear factor analysis. The nonresponse category is the baseline.

(Table 8)

Goodness-of-fit statistics appear in Table 9. The AIC and BIC always favor the interaction models with respect to their respective independence model. Secondly, the overall best-fitting model is NCM with interactions (NCM-I).

(Table 9)

Table 10 contains the parameter estimates for the testlets and the three best fitting models. The log odds and log odds ratios used to interpret parameters are also included. The high and positive interaction parameters suggest that the lack of fit of the conditional independence model is because of a tendency to repeat the same response within the two items of each testlet. Moreover, the tendency to respond with adjacent categories seems higher than the tendency to omit the first item and pass the second one, and vice versa.

The parameterization in relation to a baseline category can be modified to obtain information about the relationship between the probabilities of successive categories. This is appropriate when the response categories have an order, such as in rating scales or partial credit items, where categories are compared with the previous one instead of comparing all of them with the baseline. Consider the following reparameterization of the main effect and second-order interactions: $\delta_k^{(i)} = \sum_{k'=1}^{k} \tau_{k'}$ and $\delta_{k_1 k_2}^{(ij)} = \sum_{k'=1}^{k_1} \sum_{k''=1}^{k_2} \tau_{k'k''}$, where $\tau_{k(V)}^{(V)} = \alpha_j \theta + \beta_j$. Moreover, the parameters for the first category are set to zero: $\tau_1 = \tau_{1k} = \tau_{k1} = 0$. Then, the main effect parameters depend on the odds of a category and the previous one.

$$\tau_k^{(1)} = \log \frac{P(k_1, \mathbf{1})}{P(k_1 - 1, \mathbf{1})} \text{ for } k_1 = 2, \ldots, K_1.$$

The second-order parameters contain information about the odds ratio between successive categories.

$$\tau_{k_1 k_2}^{(12)} = \log \frac{P(k_1, k_2, \boldsymbol{k}) P(k_1 - 1, k_2 - 1, \boldsymbol{k})}{P(k_1 - 1, k_2, \boldsymbol{k}) P(k_1, k_2 - 1, \boldsymbol{k})} \text{ for } k_1 = 2, \ldots, K_1, k_2 = 2, \ldots, K_2.$$

The values of α_j and β_j used for computing $\tau_{k(V)}^{(V)}$ also appear in Table 10. The interactions between categories 2-2 and 3-3 are high and positive, whereas most of the interactions between 2-3 and 3-2 are negative. This result confirms a tendency to repeat a response to the two items of each testlet.

(Table 10)

Finally, the item response functions for the two items of testlet 16, and the NCM, NCM-I and bifactor models appear in Figure 3. Comparing the NCM and NCM-I reveals that the introduction of the interaction parameters had a small effect in the response functions, which was a decrease in the probability of passing the items D-16a and D-16b.

The item response functions for the bifactor model were computed conditional on the values –1 and 1 of the cluster-specific factor. The figure shows that the cluster-specific factor determines the distribution of responses between categories 2 and 3, and has little effect on the nonresponse category. Because of this and the factor loadings in Table 10, the cluster-specific factor appears to be an ability factor that is not shared with the other items.

(Figure 3)

6. Final remarks

This paper explores the capabilities of the GLLIRM to address the problem of local dependencies within item clusters. Several issues must be considered in applications. First, the model suffers from a lack of reproducibility, so that the item characteristic functions are altered by the inclusion of interaction parameters. For that reason, the paper explains the interpretations of the parameters that remain unchanged when other parameters are incorporated into the model.

Second, there is a limitation of the number of items that can be analyzed together. Each effect can potentially be described using scale and intercept parameters. In contrast to LLMs that admit interaction up to the highest order, the GLLIRM would run out of degrees of freedom if all parameters were included. The number of parameters increases rapidly and the interpretation becomes more complicated as the number of interacting items and the order of interaction increase. Thus, in practice, only a few of all possible interactions are estimable. Although the test can be long and may have many item clusters, the number of items in each cluster should not be high to keep the model reasonably simple.

A third caveat is related to over-fitting of the data. A simulation study has shown that parameter estimates show biases when the model includes fewer interaction parameters than the true model. However, no biases arose when the model included unnecessary interaction. These results suggest a two-step strategy in model fitting: a relatively general model can be estimated first to capture all relevant interactions, and unnecessary interactions are eliminated in a second step based on a Wald test.

Finally, there are other competing models for analyzing the same kind of data. LLMs are aimed at the analysis of interactions but are psychometrically uninteresting because of the lack of parameters representing individual differences. The paper shows by example that the interaction parameters for an LLM can be inflated because of the absence of a latent variable. One more tenable alternative within the context of psychometrics is a multidimensional model. Ip (2010) showed that multidimensional models may be observationally equivalent to unidimensional models with interaction terms. However, multidimensional models suffer from their own list of problems. Estimation of categorical factorial models relies on numerical integration or simulation techniques and may be unfeasible when there are many latent dimensions. In the present context, if the factorial model includes a general latent trait and a specific factor for each cluster of interacting items, there would be a sharp limitation on the number of clusters that could be analyzed simultaneously. Thus factorial models may be useful in those situations where GLLIRM is inapplicable: with few clusters composed of a large number of items. However, the cluster specific factor may be difficult to interpret if there are not a priori reasons to conceive item clusters as multidimensional.

In closing, different models may obtain an acceptable fit if either the order of the interactions or the number or latent dimensions are increased. Model choice would be conditioned by technical estimation problems and should ultimately be based on the structure of the test, the intended interpretation and the judgment of the investigator.

References

Agresti, A. (2002). *Categorical data analysis*. Second edition. New York. Wiley.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam. North–Holland.

Birch, M. W. (1964). A new proof of the Pearson–Fischer theorem. *Annals of mathematical statistics, 35,* 818–824.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete multivariate analysis. Theory and practice*. Cambridge, MA. The MIT Press.

Bekker, P. & Wansbeek, T. (2003). Identification in parametric models. In B. H. Baltagi (Ed.). *A companion to theoretical econometrics*. Malden, MA. Blackwell Publishing.

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.

Bock, R. D. & Gibbons, R. (2010). Factor analysis of categorical item responses. In M. L. Nering and R. Ostini (Eds.). *Handbook of polytomous item response theory models*. Routledge. New York.

Bowden, R. (1973). The theory of parametric identification. *Econometrica, 41,* 1061–1074.

Buse, A. (1982). The likelihood ratio, Wald and Lagrange multiplier tests: An expository note. The American Statistician, 36, 153-157.

Cox, C. (1984). An elementary introduction to maximum likelihood estimation for multinomial models: Birch’s theorem and the delta method. *The American Statistician, 38,* 283–287.

Gonzalez, E. J., Smith, T. A. & Sibberns, H. (1998). *User guide for the TIMSS international database*. Chestnut Hill, MA. TIMSS International Study Center. Boston College.

Harman, H. (1976). Modern factor analysis; third edition. University of Chicago Press. Chicago.

Hoskens, M. & de Boeck, P. (1997). A parametric model for local dependencies among test items. *Psychological Methods, 2*, 261–277.

Hoskens, M. & de Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement, 25*, 19–37.

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika, 67,* 367–386.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63*, 395-416.

Ip, E. H., Smits, D. J. M. and de Boeck, P. (2009). Locally Dependent Linear Logistic Test Model With Person Covariates. *Applied Psychological Measurement, 33*, 555-569.

Ip, E. H., Wang, Y. J., de Boeck, P. & Meulders, M. (2004). Locally dependent latent trait model for polytomous responses with application to inventory of hostility. *Psychometrika, 69,*, 191–216.

Jannaronne, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357-373.

Jannarone, R. J. (1997). Models for locally dependent responses. Conjunctive item response theory. In W. J. van der Linden and R. K. Hambleton (Eds.). *Handbook of modern item response theory*. New York. Springer.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223–245.

Kelderman, H. & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 147–177.

Laird, N. M. (1991). Topics in likelihood–based methods for longitudinal data analysis. *Statistica Sinica, 1,* 33–50.

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

Muraki, E. (1992). A generalized partial credit model. Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*. New York. Springer.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica, 39,* 577–591.

Stoer, J. & Bulirsch, R. (1980). *Introduction to numerical analysis*. New York: Springer-Verlag.

Verhelst, N. D. & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58,* 395–415.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge. Cambridge University Press.

Wansbeek, T. & Meijer, E. (2000). *Measurement error and latent variables in econometrics*. Amsterdam. North-Holland.

TABLE 1. Mean value of the likelihood ratio test statistic (G^2) and empirical proportion of rejection (EPR) for the different combinations of true vs. estimated model and sample size in the simulation study

n = 500	CI	2B	3B	4AB
CI	197.3(.00)	185.7(.00)	189.2(.00)	194.1(.00)
2B	149.9(.00)	102.8(.00)	124.1(.00)	136.5(.00)
3B	316.4(.92)	185.5(.00)	164.8(.00)	240.6(.09)
4AB	233.9(.03)	197.1(.00)	190.5(.00)	179.3(.00)
n = 1000	CI	2B	3B	4AB
CI	218.3(.00)	206.8(.00)	210.6(.00)	214.3(.00)
2B	203.1(.00)	123.0(.00)	161.3(.00)	182.5(.00)
3B	489.9(1.00)	235.5(.14)	188.0(.00)	335.1(.99)
4AB	308.2(.91)	248.2(.25)	228.4(.03)	203.4(.00)
n = 5000	CI	2B	3B	4AB
CI	231.1(.01)	222.3(.02)	223.9(.00)	227.6(.01)
2B	515.6(1.00)	156.9(.00)	225.3(.98)	425.8(1.00)
3B	1638.8(1.00)	458.5(1.00)	206.2(.01)	927.0(1.00)
4AB	731.4(1.00)	486.0(1.00)	367.8(1.00)	226.1(.01)
df	239	227	231	235

Note: The rows of the table indicate the model used to simulate the data (true model) and the columns indicate the estimated model. The cells contain the mean of G^2 and the EPR between brackets. df is the number of degrees of freedom for the estimated model under the hypothesis that the model fits.

TABLE 2. True parameter values and bias of the estimated parameters for different combinations of simulated vs. estimated model in the simulation study

	True	Estimated				True	Estimated			
	CI	CI	2B	3B	4AB	2B	CI	2B	3B	4AB
β_1	0.46	0.01	0.23	0.01	0.02	0.46	**-2.92**	-0.51	**-2.68**	**-3.04**
β_2	2.02	0.04	0.32	0.09	0.05	2.02	0.45	1.32	-0.12	0.13
β_3	0.09	0.00	0.13	0.00	0.01	0.09	**-3.06**	-0.32	**-2.76**	**-3.21**
β_4	-1.60	-0.04	0.16	-0.04	-0.04	-1.60	**-3.77**	-0.75	**-3.47**	**-3.70**
β_{12}			-0.18			0.89		0.14		
β_{13}			-0.08			-1.24		0.06		
β_{14}			-0.20			-1.24		0.18		
β_{23}			-0.08			-0.27		-0.24		
β_{24}			-0.23			0.31		-0.92		
β_{34}			-0.11			-1.52		0.13		
β_{123}				0.03					0.93	
β_{124}				-0.03					1.37	
β_{134}				0.00					-0.40	
β_{234}				-0.02					0.12	
β_{1234}					0.17					2.36
β_5	-1.21	-0.02	-0.54	-0.05	-0.02	-1.21	**0.53**	-0.39	**0.50**	**0.52**
β_6	1.27	0.01	-0.31	0.01	0.01	1.27	**2.13**	-0.09	**1.40**	**1.93**
β_7	-0.22	0.00	-0.28	-0.02	0.00	-0.22	**0.35**	-0.25	**0.32**	**0.33**
β_8	-0.67	-0.01	-0.22	-0.02	-0.01	-0.67	-0.05	-0.19	-0.09	-0.06
β_{56}			0.33			1.74		0.43		
β_{57}			0.25			0.82		0.30		
β_{58}			0.17			0.03		0.18		
β_{67}			0.16			1.01		0.29		
β_{68}			0.11			1.83		0.30		
β_{78}			0.09			-0.28		0.10		
β_{567}				0.02					1.54	
β_{568}				0.01					2.23	
β_{578}				0.02					0.02	
β_{678}				0.00					1.78	
β_{5678}					0.08					4.29
α_1	1.83	0.04	-0.14	0.04	0.03	1.83	1.96	0.57	1.94	2.08
α_2	1.77	0.06	-0.11	0.07	0.04	1.77	-0.53	0.24	1.15	0.32
α_3	0.77	0.01	-0.09	0.01	0.00	0.77	**2.05**	0.26	**1.85**	**2.18**
α_4	2.43	0.06	-0.09	0.09	0.09	2.43	2.69	0.63	2.59	**2.52**
α_{1234}					0.14					1.01
α_5	1.70	0.04	0.62	0.08	0.04	1.70	**-0.44**	0.51	**-0.41**	**-0.47**
α_6	0.99	0.02	0.28	0.04	0.03	0.99	**-0.93**	0.26	-0.44	**-0.78**
α_7	0.84	0.01	0.21	0.03	0.01	0.84	**-0.31**	0.22	-0.29	**-0.30**
α_8	0.54	0.00	0.15	0.02	0.01	0.54	0.00	0.13	0.02	-0.01
α_{5678}					0.07					-0.02

Note: Sample size is 500. Empty cells are structural zeroes. Boldface indicates statistical significance using a Wald test.

TABLE 3. True parameter values and bias of the estimated parameters for different combinations of simulated vs. estimated model in the simulation study

	True	Estimated				True	Estimated			
	3B	CI	2B	3B	4AB	4AB	CI	2B	3B	4AB
β_{1}	0.46	**-0.78**	1.02	-0.01	**-1.13**	0.46	-0.05	0.61	-0.03	0.01
β_{2}	2.02	0.20	0.37	0.08	-0.24	2.02	-0.05	0.83	0.11	0.08
β_{3}	0.09	**-0.43**	0.37	0.00	**-0.75**	0.09	-0.06	0.31	-0.03	0.01
β_{4}	-1.60	-0.38	0.66	-0.14	-1.17	-1.60	0.12	0.59	-0.19	-0.05
β_{12}			**-0.34**					-0.59		
β_{13}			-1.19					-0.24		
β_{14}			**-1.28**					-0.59		
β_{23}			1.89					-0.46		
β_{24}			-0.63					-0.59		
β_{34}			-0.56					-0.24		
β_{123}	2.09			-0.01					-0.40	
β_{124}	-0.84			0.02					0.19	
β_{134}	-1.54			-0.02					0.07	
β_{234}	0.52			0.05					0.02	
β_{1234}					2.32	0.29				0.30
β_{5}	-1.21	**-0.86**	-2.54	-0.12	-0.06	-1.21	**-0.70**	-4.53	-0.03	-0.09
β_{6}	1.27	**-1.90**	-0.55	-0.02	**-1.01**	1.27	**-1.02**	-1.83	0.37	0.02
β_{7}	-0.22	**-0.79**	-1.66	-0.04	-0.12	-0.22	**-0.81**	-2.73	0.05	-0.01
β_{8}	-0.67	-0.12	-1.07	-0.03	**0.44**	-0.67	**-0.75**	-2.04	-0.01	-0.01
β_{56}			-0.43					1.15		
β_{57}			1.05					1.41		
β_{58}			0.81					1.10		
β_{67}			-0.96					0.29		
β_{68}			-0.39					-0.30		
β_{78}			0.75					0.46		
β_{567}	-1.96			0.00					-0.68	
β_{568}	-0.70			0.01					-0.58	
β_{578}	0.95			0.03					-0.16	
β_{678}	-0.58			0.00					**-1.09**	
β_{5678}					**-1.67**	-1.85				0.10
α_{1}	1.83	0.20	-0.40	0.08	**1.82**	1.83	-0.10	-0.55	0.16	0.01
α_{2}	1.77	-0.80	-0.29	0.09	0.57	1.77	0.15	-0.42	0.32	0.13
α_{3}	0.77	0.18	-0.17	0.01	**0.89**	0.77	0.10	-0.21	0.08	0.00
α_{4}	2.43	0.15	-0.44	0.28	1.76	2.43	-0.29	-0.68	0.36	0.10
α_{1234}					-0.56	0.68				0.25
α_{5}	1.70	0.91	2.71	0.18	0.04	1.70	0.97	5.08	0.18	0.15
α_{6}	0.99	**1.97**	1.87	0.09	0.57	0.99	**1.92**	2.80	0.43	0.03
α_{7}	0.84	**0.92**	1.34	0.05	0.17	0.84	**1.17**	2.64	0.12	0.02
α_{8}	0.54	0.14	0.67	0.03	**-0.34**	0.54	**0.90**	1.61	0.11	0.01
α_{5678}					-0.45	0.96				0.26

Note: Sample size is 500. Empty cells are structural zeroes. Boldface indicates statistical significance using a Wald test.

Table 4. Correlation matrix between some parameter estimates in the simulation study for different sample sizes. The simulated model is CI and the estimated ones are CI or 2B

		β_1	β_2	β_3	β_4	β_{12}	β_{13}	β_{14}	β_{23}	β_{24}	β_{34}
n=500	β_1	1.00	0.09	0.16	0.46						
	β_2	0.67	1.00	-0.01	0.03						
	β_3	0.58	0.52	1.00	0.16						
	β_4	0.85	0.73	0.55	1.00						
	β_{12}	-0.68	-0.66	-0.46	-0.61	1.00					
	β_{13}	-0.59	-0.46	-0.78	-0.40	0.43	1.00				
	β_{14}	-0.91	-0.57	-0.36	-0.81	0.58	0.30	1.00			
	β_{23}	-0.45	-0.58	-0.72	-0.43	0.43	0.57	0.28	1.00		
	β_{24}	-0.61	-0.90	-0.42	-0.75	0.60	0.36	0.55	0.42	1.00	
	β_{34}	-0.48	-0.45	-0.89	-0.56	0.40	0.51	0.36	0.55	0.39	1.00
n=1000	β_1	1.00	0.00	0.16	0.38						
	β_2	0.66	1.00	-0.02	0.04						
	β_3	0.55	0.53	1.00	0.15						
	β_4	0.87	0.72	0.58	1.00						
	β_{12}	-0.62	-0.59	-0.40	-0.55	1.00					
	β_{13}	-0.54	-0.50	-0.79	-0.43	0.39	1.00				
	β_{14}	-0.90	-0.50	-0.32	-0.81	0.49	0.19	1.00			
	β_{23}	-0.43	-0.63	-0.70	-0.45	0.35	0.56	0.24	1.00		
	β_{24}	-0.62	-0.88	-0.44	-0.72	0.57	0.43	0.49	0.49	1.00	
	β_{34}	-0.50	-0.47	-0.89	-0.61	0.36	0.52	0.39	0.56	0.40	1.00
n=5000	β_1	1.00	0.11	0.06	0.48						
	β_2	0.81	1.00	-0.11	0.08						
	β_3	0.83	0.74	1.00	0.09						
	β_4	0.99	0.79	0.85	1.00						
	β_{12}	-0.88	-0.67	-0.77	-0.88	1.00					
	β_{13}	-0.79	-0.75	-0.93	-0.80	0.74	1.00				
	β_{14}	-	-	-	-	0.82	0.64	1.00			

	0.97	0.75	0.72	0.95						
β_{23}	-0.80	-0.66	-0.90	-0.81	0.82	0.85	0.69	1.00		
β_{24}	-0.86	-0.64	-0.78	-0.87	0.96	0.78	0.79	0.85	1.00	
β_{34}	-0.84	-0.75	-0.97	-0.85	0.76	0.84	0.77	0.85	0.73	1.00

Note: Results for the estimated CI model appear above the diagonal and the estimated 2B mode appears below.

Table 5. Response patterns, second-order interactions between items and observed frequencies for the sample survey example

Pattern	Interactions	Frequency	Pattern	Interactions	Frequency
00000	0000000000	55	10000	0000000000	0
00001	0000000000	381	10001	0001000000	7
00010	0000000000	0	10010	0010000000	0
00011	0000000001	20	10011	0011000001	3
00100	0000000000	0	10100	0100000000	0
00101	0000000010	4	10101	0101000010	2
00110	0000000100	0	10110	0110000100	0
00111	0000000111	2	10111	0111000111	1
01000	0000000000	2	11000	1000000000	0
01001	0000001000	423	11001	1001001000	44
01010	0000010000	2	11010	1010010000	0
01011	0000011001	76	11011	1011011001	39
01100	0000100000	1	11100	1100100000	0
01101	0000101010	112	11101	1101101010	50
01110	0000110100	0	11110	1110110100	0
01111	0000111111	38	11111	1111111111	71

Table 6. Goodness of fit statistics for three types of models in the sample survey example: local independence models, unidimensional models with item interactions and log linear-models.

Type	Model	Parameters	G^2	d.f.	p-value	AIC
Conditional	1PL	5	261.2	26	≤ 0.001	5603.1
independence	2PL	10	33.3	21	0.05	5367.5
	Factorial	15	16.2	16	0.44	**5360.5**
Conditional	2PL-2B	20	4.85	11	0.94	5359.1
	2PL-2B-Backward	12	12.5	19	0.93	**5349.2**
dependence	2PL-2AB	30	0.64	1	0.42	5374.9
Log-linear	LLM-1	5	761.4	21	≤ 0.001	6085.7
	LLM-2	15	16.0	16	0.46	**5360.3**
	LLM-3	25	1.3	6	0.97	5365.6

Note: Type refers to the type of structure imposed by the model. CI stands for conditional independence. The boldface indicates the smallest AIC for each type of model.

Table 7. Parameter estimates for two local independence models (2PL and Two-factor) two models with local dependencies (2PL-2B and 2PL-2B-Backward) and a log-linear model applied to the sample survey example

		2PL		Factorial			2PL-2B		2PL-2B-Backward		LLM-2
Item 1	Item 2	α	β	η	λ_{i1}	λ_{i2}	α	β	α	β	β
For my country		2.53	-3.03	4.08	3.75	0.00	7.97	-9.63	8.73	-4.31	-8.87
Another's life		2.88	1.25	-1.39	-0.52	3.78	4.66	2.60	3.66	1.63	-2.57
For democracy		2.26	-2.29	2.73	-0.26	3.12	3.41	-1.25	4.40	-2.88	-4.82
For God		1.57	-2.05	2.12	1.54	0.15	8.71	1.59	1.39	-1.96	-3.55
For my family		2.54	5.29	-5.11	-0.08	2.47	3.28	6.43	2.40	5.09	1.97
For my country	Another's life							-2.76		-3.23	1.42
	For democracy							-0.63		-2.06	1.44
	For God							-5.26			1.53
	For my family							6.90			5.19
Another's life	For democracy							0.30			2.69
	For God							-3.36			1.19
	For my family							-0.37			2.68
For democracy	For God							-1.95			0.59

	For my family	-1.32	0.81
For God	For my family	-3.64	0.66

Note: The column η contains the thresholds for the categorical factor analysis model, λ_{i1} and λ_{i2} are factor loadings; λ_{12} is a structural zero and the correlation between the factors was 0.85.

Table 8. Items from the Mathematics Literacy scale of the TIMSS database and psychometric models

Type	Seq.	Label	Cats.	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Multiple choice	1	A-3	5	NCM	NCM	NCM	NCM	NCM	NCM	NCM
	2	A-4	5							
	3	A-5	5							
	7	D-6	5							
Open ended	4	A-8	4	PCM	PCM	GPCM	GPCM	NCM	NCM	NCM
	5	A-10	4							
	6	A-12	5							
	12	D-17	4							
Testlet	8	D-15a	3	PCM	PCM-I	GPCM	GPCM-I	NCM	NCM-I	Bifactor
	9	D-15b	3							
Testlet	10	D-16a	3	PCM	PCM-I	GPCM	GPCM-I	NCM	NCM-I	Bifactor
	11	D-16b	3							

Note: The column Seq. indicates the sequential order in which the items are applied. Cats. is the number of categories. The models are: PCM (partial credit model), GPCM (generalized partial credit model), NCM (nominal categories model), PCM-I (partial credit item with interactions), GPCM-I (generalized partial credit model with interactions), NCM-I (nominal categories model with interactions) NCM-3D (nominal categories model with a second dimension for the interacting items). The interactions are analyzed within each testlet only. All the items of each type are analyzed with the same model.

Table 9. Goodness of fit statistics for the TIMMS database

Model	Parameters	Log. lik.	AIC	BIC
1	53	-86698.4	173502.8	173889.2
2	61	-81169.1	162460.2	162904.9
3	59	-76991.7	154101.4	154531.6
4	67	-76107.3	152348.5	152837.0
5	74	-73809.7	147767.4	148306.9
6	82	-72906.8	145977.5	146575.4
7	82	-73481.3	147126.6	147724.4

Note: The description of models appears in Table 8.

Table 10. Parameter estimates for the two testlets of the TIMMS database under the NCM, bifactor and NCM-I models

		NCM		Bifactor			NCM-I				
Item	Category	α	β	α_1	α_2	β	$\alpha(\delta)$	$\beta(\delta)$	$\alpha(\tau)$	$\beta(\tau)$	Odds
D-15a	2	1.46	0.18	12.54	-2.79	-13.61	1.03	-1.47	1.03	-	$\frac{P(2,1)}{P(1,1)}$
	3	7.39	-8.88	19.43	-6.14	-18.95	2.66	-7.61	1.63	-	$\frac{P(3,1)}{P(1,1)}$
D-15b	2	0.93	-0.51	13.92	-3.31	-14.02	0.77	-0.59	0.77	-	$\frac{P(2,1)}{P(1,1)}$
	3	8.05	-8.60	15.87	-4.46	-15.86	3.76	-8.25	2.99	-	$\frac{P(3,1)}{P(1,1)}$
D-15a	2-2							7.77		7.77	$\frac{P(2,2,k)P(1,1,k)}{P(2,1,k)P(1,2,k)}$
	2-3							7.11		-	$\frac{P(2,3,k)P(1,1,k)}{P(2,1,k)P(1,3,k)}$
	3-2							9.47		1.70	$\frac{P(3,2,k)P(1,1,k)}{P(3,1,k)P(1,2,k)}$
	3-3							9.95		1.14	$\frac{P(3,3,k)P(1,1,k)}{P(3,1,k)P(1,3,k)}$
D-16a	2	3.93	-6.28	10.01	1.86	-10.74	1.08	-2.33	1.08	-	$\frac{P(2,1)}{P(1,1)}$
	3	6.99	-8.02	13.32	4.70	-16.86	5.41	-6.98	4.33	-	$\frac{P(3,1)}{P(1,1)}$
D-16b	2	3.11	-5.91	6.63	0.30	-8.66	0.86	-2.53	0.86	-	$\frac{P(2,1)}{P(1,1)}$
	3	6.50	-8.52	9.15	3.08	-15.07	2.07	-7.46	1.21	-	$\frac{P(3,1)}{P(1,1)}$
D-16a	2-2							5.85		5.85	$\frac{P(2,2,k)P(1,1,k)}{P(2,1,k)P(1,2,k)}$
	2-3							4.48		-	$\frac{P(2,3,k)P(1,1,k)}{P(2,1,k)P(1,3,k)}$
	3-2							5.64		-	$\frac{P(3,2,k)P(1,1,k)}{P(3,1,k)P(1,2,k)}$
	3-3							7.05		2.78	$\frac{P(3,3,k)P(1,1,k)}{P(3,1,k)P(1,3,k)}$

Note: The column Odds contains the log odds and log odds ratios used for interpretation of the NCM-I parameters. The columns $\alpha(\delta)$ and $\beta(\delta)$ contain the values of α and β used for computing δ. $\alpha(\tau)$ and $\alpha(\tau)$ are used for computing τ. α_1 and α_2 are factor loadings on the general and the cluster-specific factor.

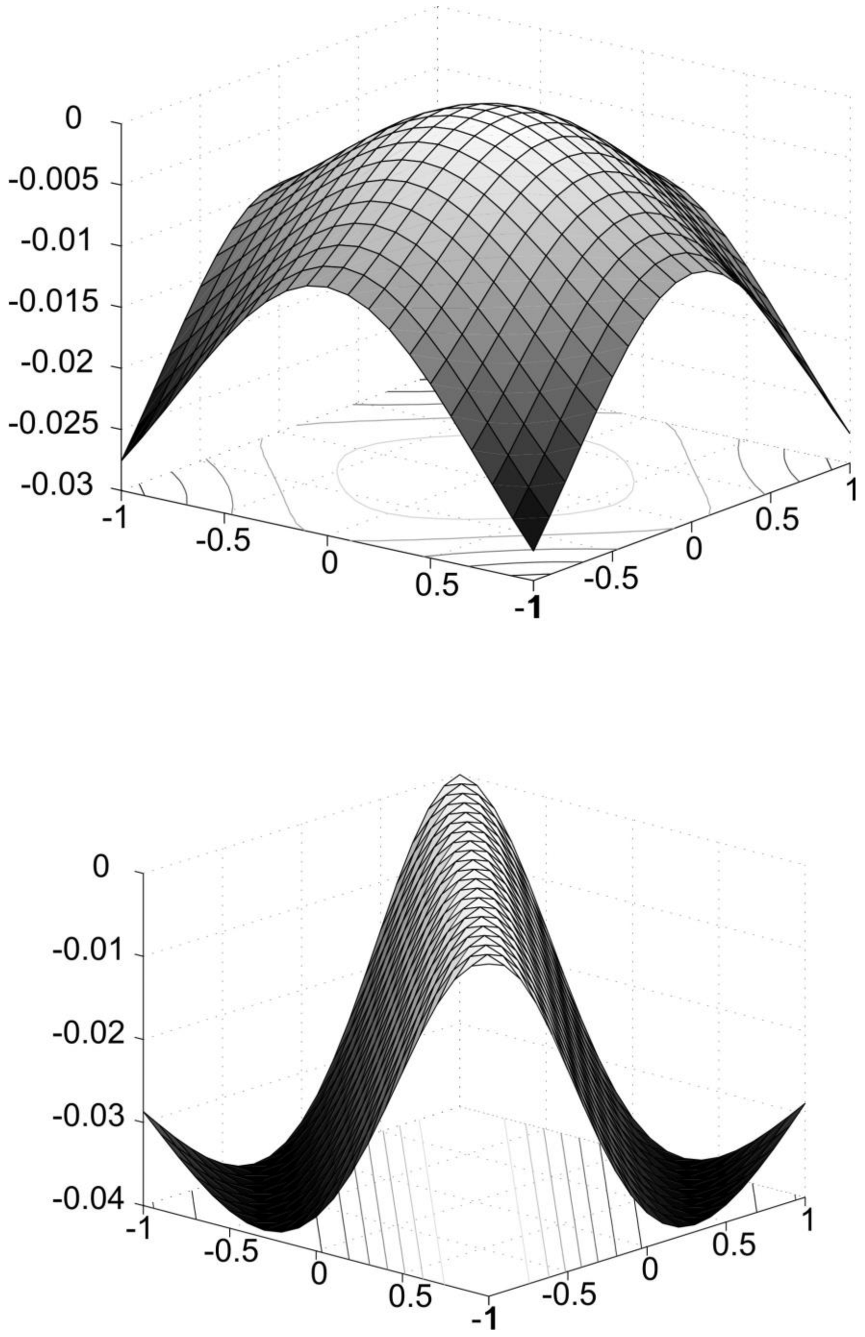

Figure 1. Information divergence between the distributions $\pi(\boldsymbol{\beta})$ and $\pi(\boldsymbol{\beta}_0)$, where $\boldsymbol{\beta}_0 = \mathbf{0}$.

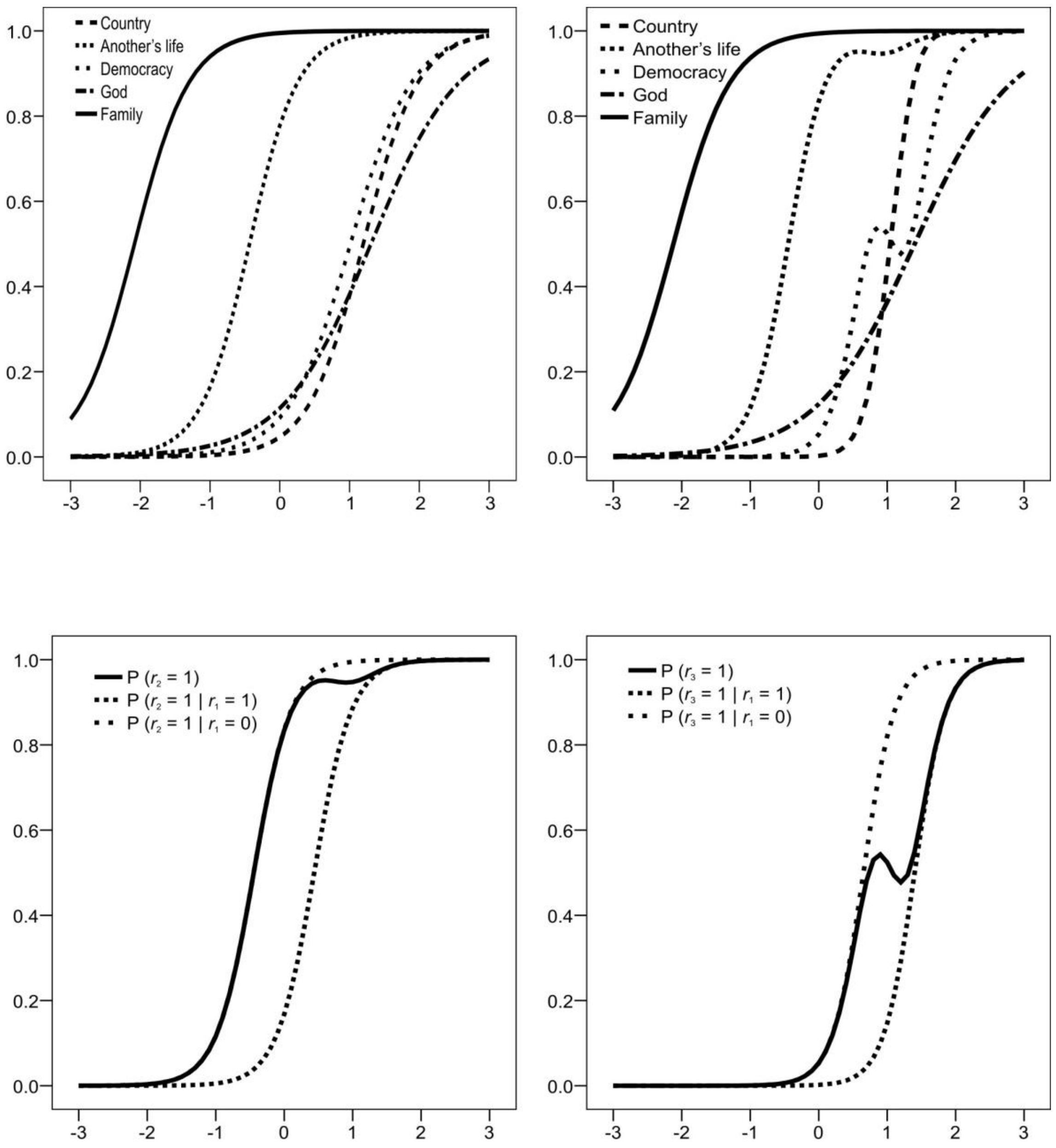

Figure 2. The upper row contains the marginal probabilities of endorsing the items –item characteristic curves- as a function of θ; the left column corresponds to the 2PL and the right column to the 2PL-2B-Backward. The lower row contains the marginal probabilities of endorsing the items 2 and 3 and probabilities of endorsing the items conditional on the response to Item 1 for the 2PL-2B-Backward. The lower left panel stands for Item 2 and lower right is Item 3.

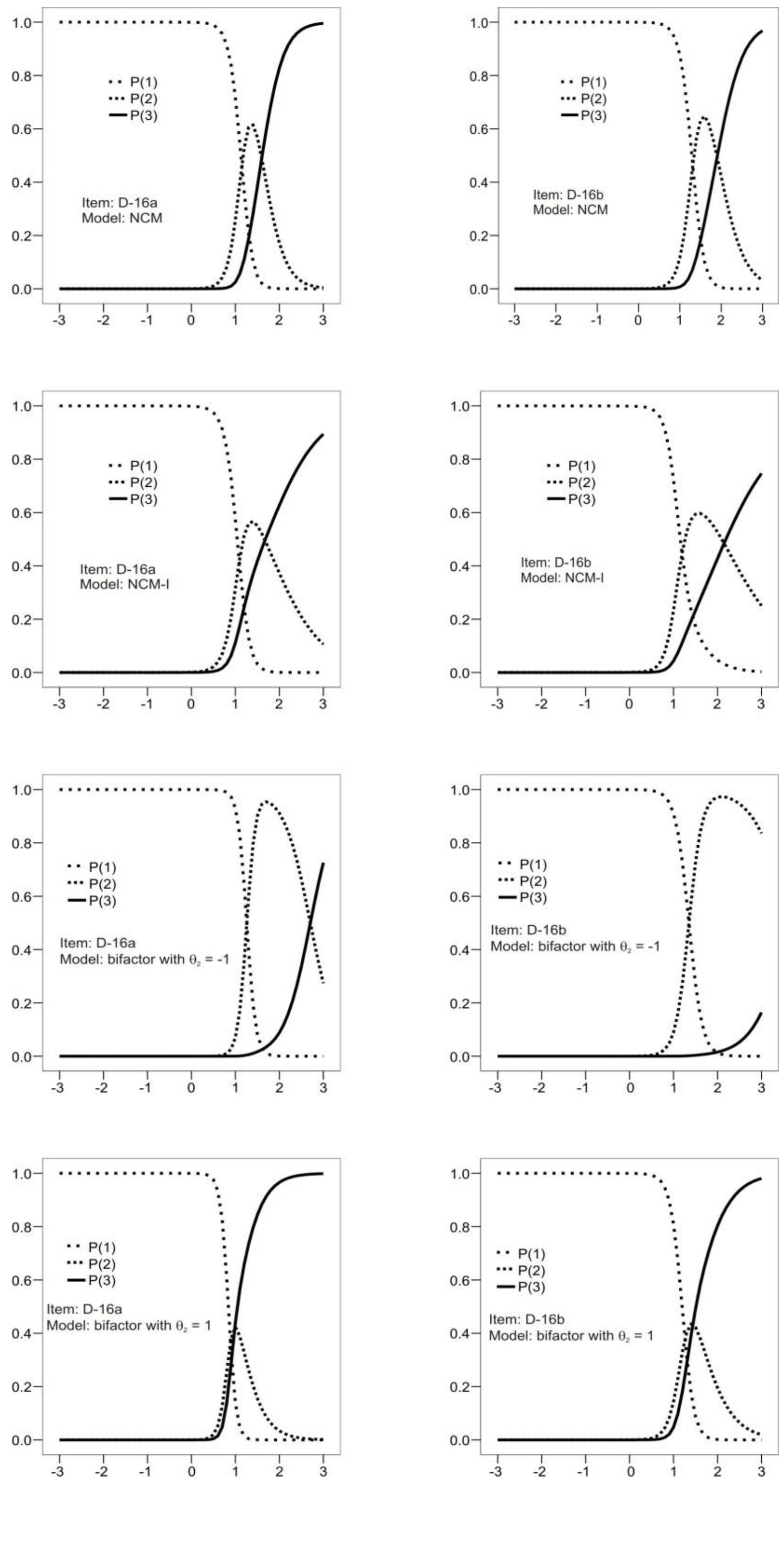

Figure 3. Item response functions for the items D-16a (left column) and D-16b (right column) and the models NCM (first row), NCM-I (second row) and bifactor (third and fourth rows).