Behavior Research Methods
2009, 41 (4), 1127-1137
doi:10.3758/BRM.41.4.1127

Investigating latent constructs with item response models: A MATLAB IRTm toolbox

JOHAN BRAEKEN
National Institute of Educational Measurement (Cito), Arnhem, The Netherlands and Tilburg University, Tilburg, The Netherlands

AND

FRANCIS TUERLINCKX
Katholieke Universiteit Leuven, Leuven, Belgium

Item response theory (IRT) models are the central tools in modern measurement and advanced psychometrics. We offer a MATLAB IRT modeling (IRTm) toolbox that is freely available and that follows an explicit design matrix approach, giving the end user control and flexibility in building a model that goes beyond standard models, such as the Rasch model (Rasch, 1960) and the two-parameter logistic model. As such, IRTm allows for a large variety of unidimensional IRT models for binary responses, the incorporation of additional person and item information, and deviations from common model assumptions. An exclusive key feature of the toolbox is the inclusion of copula IRT models to handle local item dependencies. Two appendixes for this report, containing example code and information on the general copula IRT in IRTm, may be downloaded from brm.psychonomic-journals.org/content/supplemental.

In the behavioral sciences, many constructs that are of theoretical or practical interest cannot be observed directly. The basic procedure for measuring such constructs involves gathering observable variables, which ought to provide indirect evidence for the construct of interest. Usually, this means that a test is developed by creating items on which the responses function as indicators for the construct of interest.

The current standard to ground and support this type of inference is item response theory (IRT). Whereas classical test theory merely has a descriptive nature (see, e.g., Nunnally & Bernstein, 1994), IRT posits a mathematical model to explain the pattern of observed item responses on a test (see, e.g., Birnbaum, 1968, or, for a recent introduction, Embretson & Reise, 2000). The key idea is that tests are designed to measure an unobservable variable of interest, a latent trait. IRT assumes that both persons and test items have a position on this latent trait. A well-known item response model is the *two-parameter logistic model* (2PL; Birnbaum, 1968), in which the probability of a binary response Y_{pi} of person p ($p = 1, 2, \ldots, P$) on item i ($i = 1, 2, \ldots, I$) is characterized as a function of three parameters θ_p, β_i, and α_i:

$$Pr\left(Y_{pi} = y_{pi} \mid \theta_p\right) = \frac{\exp\left[y_{pi}\alpha_i\left(\theta_p - \beta_i\right)\right]}{1 + \exp\left[\alpha_i\left(\theta_p - \beta_i\right)\right]}. \qquad (1)$$

The parameters θ_p and β_i are exactly the positions of person p and of item i, respectively. Note that the value of β_i, often called the *item difficulty*, corresponds to the location on the latent trait where a person would have a chance of .5 to answer the item correctly [i.e., $Pr(Y_{pi} = y_{pi} | \theta_p = \beta_i)$, the point of inflection of the logistic S-shaped curve]. For an item with a larger β_i, a larger proficiency θ_p is needed to answer it correctly, and vice versa: The presence of a larger θ_p means that one has more chance of answering an item correctly. The parameter α_i controls the steepness of the logistic curve and is often called the *item discrimination*. The higher the α_i value, the better the item is able to differentiate between high- and low-proficiency persons. Thus, the logistic shape of the probability function $Pr(Y_{pi} = y_{pi} | \theta_p)$ is determined by two fixed item parameters, β_i and α_i, hence the name 2PL.

In relating persons and items to the latent trait, restrictions are imposed on the probability model for the whole test. For instance, a common assumption of the 2PL model (and most other item response models) is conditional independence or local stochastic independence, which would imply that the latent trait accounts for all the dependencies among a person's responses on the test. Thus, the latent proficiency explains why there are performance differences among persons and why a given person's item responses interrelate. Hence, given the proficiency of a person p, the joint (conditional) probability

J. Braeken, j.braeken@uvt.nl

 © 2009 The Psychonomic Society, Inc.

of the item response pattern of person p on the full test [$\mathbf{y}_p = (y_{p1}, \ldots, y_{pI})$] can be written as

$$Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_p\right) = \prod_{i=1}^{I} Pr\left(Y_{pi} = y_{pi} \mid \theta_p\right). \quad (2)$$

Item responses are assumed to be independent, given the proficiency of the person p on the latent trait.

Furthermore, it is often assumed that the latent proficiency θ_p is a draw from a normal population distribution with a given mean μ and variance σ^2:

$$\theta_p \sim N\left(\mu, \sigma^2\right). \quad (3)$$

This normality assumption is also used in the multilevel literature (e.g., Snijders & Bosker, 1999), and, in fact, item response models can be shown to belong to the class of generalized or nonlinear multilevel (or mixed) models (see, e.g., Agresti, Booth, Hobert, & Caffo, 2000; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003).

Established commercial IRT software packages, such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) and MULTILOG (Thissen, Chen, & Bock, 2003), can perform large-scale assessment using standard item response models. Various researchers have contributed research-specific custom-made software (see, e.g., de Leeuw & Mair, 2007), and most general-purpose statistical software can now fit item response models as a result of the connection with multilevel models.

We developed an IRT modeling (IRTm) toolbox for MATLAB, which can be downloaded without cost at http://ppw.kuleuven.be/okp/software/IRTm. MATLAB is a general-purpose matrix programming language with a large research and business user base. The toolbox can both fit and simulate a wide variety of unidimensional[1] item response models for binary test data. Full-information marginal maximum likelihood (MML; Bock & Lieberman, 1970) provides the means for model estimation.

Furthermore, IRTm follows an explicit design matrix approach (see, e.g., De Boeck & Wilson, 2004), giving the end user control and flexibility in building a model that goes beyond standard models, such as the 2PL model. To handle deviations from the conditional independence assumption, the toolbox includes recent copula IRT models (Braeken, Tuerlinckx, & De Boeck, 2007), which are not yet implemented elsewhere. With IRTm, we offer practitioners a small, integrated IRT toolbox for explanatory research and for exploring the potential of the copula approach within IRT.

In the present article, we show (1) how standard item response theory models can be modified within a design matrix framework to accommodate research questions involving additional information on test items (e.g., an experimental design) and on test takers (e.g., person group differences) and (2) how one can account for deviations of the two common IRT model assumptions (see Equations 2 and 3). We provide example code for how to conduct such applications in the IRTm toolbox. Note that the applications are brief and concise, serving as appetizers for what IRT can offer for researching latent constructs. Finally, in a short discussion section, we point out the direction of the potential development of the IRTm toolbox.

Model Building With Design Matrices

The purpose of design matrices is to support the use of both very general models and models that further constrain parameter sets. This provides additional flexibility in modeling and allows researchers to build models that go beyond standard IRT models. In the present article, we first illustrate the main idea by putting the 2PL model within the proposed framework of design matrices. Then we show how to modify this framework to further restrict or generalize the model, depending on the exact research question and on information provided by the study in which the test is used.

The 2PL Model

Consider a small test ($I = 6$) in which an item can be answered correctly ($Y_{pi} = 1$) or incorrectly ($Y_{pi} = 0$) and is assumed to reflect a given skill θ_p. A 2PL model is applied to capture a person's behavior on the test. Assuming that the skill is normally distributed in the population with a mean μ and variance σ^2, the model likelihood given the gathered data is

$$\begin{aligned}
&\ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu, \sigma^2\right) \\
&= \prod_{p=1}^{P} \int_{\theta_p} Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_p\right) \phi\left(\theta_p; \mu, \sigma^2\right) d\theta_p \\
&= \prod_{p=1}^{P} \int_{\theta_p} \prod_{i=1}^{I} \frac{\exp\left[\alpha y_{pi}\left(\theta_p - \beta_i\right)\right]}{1 + \exp\left[\alpha\left(\theta_p - \beta_i\right)\right]} \phi\left(\theta_p; \mu, \sigma^2\right) d\theta_p.
\end{aligned}$$

Because no information is present about the scale of this latent skill, it has to be fixed a priori in order to identify the model. A common convention is to assume a standard normal distribution Φ ($\mu = 0$ and $\sigma^2 = 1$).

The model specification of each theoretical parameter is summarized in a corresponding set of design matrices. IRTm makes a distinction between a general design matrix **D**, a restriction matrix **R**, and an offset matrix **O**. The actual value **V** for the theoretical model parameter **M** is then constructed as $\mathbf{M} = \mathbf{D} \times [\text{diag}(\mathbf{R}) \times \text{parameter estimate} + \mathbf{O}] = \mathbf{V}$.

For instance, the vector of item difficulties $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)$ is modeled as

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \left(\text{diag}\left(\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right) \times \begin{bmatrix} \beta_1^{\text{est}} \\ \beta_2^{\text{est}} \\ \beta_3^{\text{est}} \\ \beta_4^{\text{est}} \\ \beta_5^{\text{est}} \\ \beta_6^{\text{est}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} \beta_1^{\text{est}} \\ \beta_2^{\text{est}} \\ \beta_3^{\text{est}} \\ \beta_4^{\text{est}} \\ \beta_5^{\text{est}} \\ \beta_6^{\text{est}} \end{bmatrix}.$$

The general design matrix **D** takes the form of a logical matrix, number of items $I \times$ number of item parameters K, with 1 indicating that the parameter k needs to be included for item i. Restrictions on the parameters are imposed by the logical vector **R**, which contains the diagonal elements of a matrix and indicates which parameters are free (1) and which are fixed (0). The offset **O** is a vector with K rows, consisting of values that need to be added to the model parameters (here, all offsets are 0). This results in each of the item difficulties β_i in the model being parameterized by a unique difficulty parameter β_i^{est}. However, note that in the next sections, we show how it is possible to structure parameterizations more parsimoniously within this design matrix framework.

An equivalent set of design matrices exists for the vector of item discriminations $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\times \left(\text{diag}(\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}) \times \begin{bmatrix} \alpha_1^{\text{est}} \\ \alpha_2^{\text{est}} \\ \alpha_3^{\text{est}} \\ \alpha_4^{\text{est}} \\ \alpha_5^{\text{est}} \\ \alpha_6^{\text{est}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} \alpha_1^{\text{est}} \\ \alpha_2^{\text{est}} \\ \alpha_3^{\text{est}} \\ \alpha_4^{\text{est}} \\ \alpha_5^{\text{est}} \\ \alpha_6^{\text{est}} \end{bmatrix}.$$

The model for the mean and variance of the latent distribution are constructed as $\mu = [1] \times ([0] \times \mu_{\text{est}} + [0]) = [0]$ and $\sigma^2 = [1] \times ([0] \times \sigma^2_{\text{est}} + [1]) = [1]$.

The design matrix contains one column, indicating that the mean (/variance) of the latent distribution is modeled by one parameter. The restriction matrix **R** is set at 0, indicating that the parameter does not have to be estimated. The fixed value of the model parameter is given in the offset matrix O.

It can be shown that the sufficient statistic for θ_p in the 2PL model is a weighted sum score, in which α_i values function as weights. In other words, this model estimates the optimal scoring rule to be used for the test, because the α_i value indicates the relative value of item i for determining a person's latent proficiency θ_p. This allows us to test the common convention of taking the raw (unweighted) sum score over items as a summary measure for a person's latent proficiency and to formulate explicitly the hidden assumption behind this practice. In the raw sum, each element has the same relative weight, and hence, in a mathematical model formulation, this would imply a rather restrictive assumption corresponding to a model with equal α_i values over items, such that the model likelihood reduces to

$$\ell\left(\alpha, \boldsymbol{\beta}, \mu, \sigma^2\right) = \prod_{p=1}^{P} \int_{\theta_p} \prod_{i=1}^{I} \frac{\exp\left[\alpha y_{pi}\left(\theta_p - \beta_i\right)\right]}{1 + \exp\left[\alpha\left(\theta_p - \beta_i\right)\right]} \phi\left(\theta_p; \mu, \sigma^2\right) d\theta_p,$$

and the design matrices for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$ change to

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \times \left(\text{diag}(\left[1\right]) \times \left[\alpha^{\text{est}}\right] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} \alpha^{\text{est}} \\ \alpha^{\text{est}} \\ \alpha^{\text{est}} \\ \alpha^{\text{est}} \\ \alpha^{\text{est}} \\ \alpha^{\text{est}} \end{bmatrix}.$$

Note that this latter model is, in fact, an equivalent formulation of the well-known Rasch model (Rasch, 1960; the regular formulation differs in measurement scale and sets $\alpha = 1$, freeing σ^2 to be estimated).

Because IRT models can be fitted within a likelihood framework (see, e.g., Severini, 2000), common statistical methodology, such as likelihood-ratio tests and model selection criteria, can be used to compare these two models. In this way, it can be tested whether the implicit assumption that every item is an equally good indicator of the latent trait holds for a given test. This also illustrates that, unlike the descriptive summary provided by classical test theory, IRT follows a constructive theory-driven scientific strategy, in that the test can be revised to fit the model restrictions (see Rasch model properties) or, alternatively, both theory and model can be adapted to better fit the features of the data.

Example. Appendix I of the supplemental materials includes code to fit and compare the two models by means of the IRTm toolbox. The model results for an example data set are shown in Table 1. The likelihood-ratio test favors the Rasch model over the more general 2PL model. Hence, for this example, the use of a raw sum as scoring rule is supported. Notice that the common discrimination parameter α is about the mean of the unique item discriminations α_i in the 2PL model. Figure 1 shows the standard error of θ_p and illustrates that the test is more precise for those positions of the measurement scale where more items are located and, hence, where more information is present about θ_p.

Incorporating Information on the Item Side

Consider that the six items in the example below are realizations from an experiment in which the accuracy of hitting a target is measured in three conditions and in which each participant must run through two trials of these conditions. In the first condition (Items 1 and 4), a loud noise is added to the environment; in the second (Items 2 and 5), the light in the room switches off and on; and in the third (Items 3 and 6), both disturbing effects are added. Hence, for each person, six recordings (item responses) are made, regardless of whether the target is

Table 1
Model Results for the First Example

Two-Parameter Logistic Model (2PL)			Rasch			Linear Logistic Test Model (LLTM)		
Parameter	Estimate	*SE*	Parameter	Estimate	*SE*	Parameter	Estimate	*SE*
α_1	1.386	0.232						
α_2	1.199	0.220						
α_3	1.320	0.220						
α_4	1.548	0.258						
α_5	1.101	0.187						
α_6	1.163	0.196	α	1.284	0.084	α	1.276	0.083
β_1	−0.181	0.088	β_1	−0.190	0.086	β_1	−0.819	0.506
β_2	−1.432	0.199	β_2	−1.372	0.117	β_2	0.522	0.503
β_3	0.049	0.087	β_3	0.049	0.085	β_3	−0.598	0.503
β_4	0.201	0.084	β_4	0.223	0.087	β_4	1.077	0.506
β_5	−0.900	0.140	β_5	−0.817	0.103	β_5	0.628	0.072
β_6	1.022	0.139	β_6	0.961	0.107			
Log likelihood	1,771.39			1,772.68			1,777.5	
No. parameters	12			7			6	

Note—Likelihood-ratio tests: 2PL vs. Rasch, $\chi^2(5) = 2.58$, $p = .765$; Rasch vs. LLTM, $\chi^2(1) = 4.82$, $p = .002$.

missed. This experimental design can be reflected in the model design of the difficulty parameters using a linear logistic test model (LLTM; Fischer, 1973) formulation: $\mathbf{M} = \mathbf{D} \times [\text{diag}(\mathbf{R}) \times \text{parameter estimate} + \mathbf{O}] = \mathbf{V}$

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\times \left(\text{diag}\left(\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right) \times \begin{bmatrix} \beta_1^{\text{est}} \\ \beta_2^{\text{est}} \\ \beta_3^{\text{est}} \\ \beta_4^{\text{est}} \\ \beta_5^{\text{est}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right)$$

$$= \begin{bmatrix} \beta_1^{\text{est}} + \beta_2^{\text{est}} \\ \beta_1^{\text{est}} + \beta_3^{\text{est}} \\ \beta_1^{\text{est}} + \beta_2^{\text{est}} + \beta_3^{\text{est}} + \beta_4^{\text{est}} \\ \beta_1^{\text{est}} + \beta_2^{\text{est}} + \beta_5^{\text{est}} \\ \beta_1^{\text{est}} + \beta_3^{\text{est}} + \beta_5^{\text{est}} \\ \beta_1^{\text{est}} + \beta_2^{\text{est}} + \beta_3^{\text{est}} + \beta_4^{\text{est}} + \beta_5^{\text{est}} \end{bmatrix}.$$

The first parameter β_1^{est} is the intercept constant or average item difficulty, β_2^{est} is the item difficulty due to the loud noise, β_3^{est} is the difficulty corresponding to the light flashes, β_4^{est} is the added difficulty when these effects are jointly present, and β_5^{est} is the training effect of having a rerun in Trial 2. So, in fact, the difficulty of the individual items is decomposed into common difficulties due to the experimental design, leading to a model that is both more parsimonious and more restrictive.

Example. Code to fit the LLTM is provided in Appendix I of the supplemental materials. Table 1 contains the model results of this LLTM for the example data. Statistically significant evidence was found only for the joint experimental effect and the rerun effect. For instance, the value of β_5 is 0.628, which means that there is a negative training effect: The difficulty of a condition increases on the second trial compared with the first. An effect size measure interpretation is that the odds of being accurate at a rerun are twice (i.e., $e^{0.628} = 1.874$) as low compared with the first trial (perhaps the individuals tire or lose concentration). Compared with the less restrictive Rasch model, the Rasch model again receives support from the likelihood-ratio test. However, it is fairly obvious that the experimenter's interests are better served and research questions are better answered by the latter LTTM model alternative.

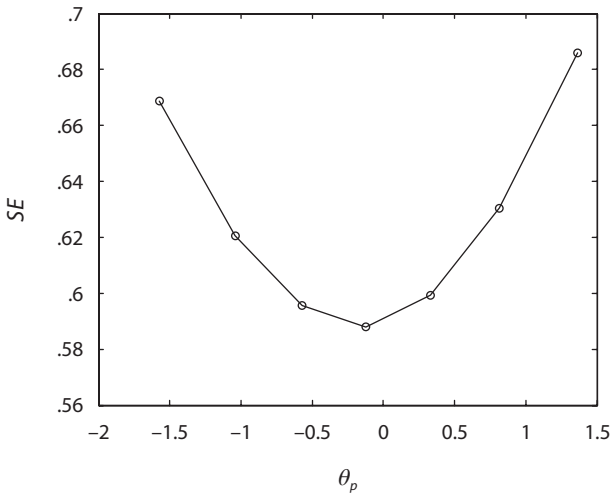

Figure 1. Precision of the test over the measurement scale.

Incorporating Information on the Person Side

Consider a 15-problem mathematics test that is designed for an international study on geometric knowledge. A researcher might be interested in individual differences in geometric knowledge and how they relate to the background information available about the pupils taking the test. This requires the availability of J variables Z_{pj} ($j = 1, \ldots, J$) containing this covariate information. For instance, for each pupil ($N = 1{,}000$) it is known whether they live in Country E ($Z_{pj} = 1, n = 600$), a highly developed and industrial country, or in Country F ($Z_{pj} = 0$, $n = 400$), a third-world country.

In our first example, we assumed that the latent trait followed a normal distribution in the population with a given mean and variance. In this case, it is not unreasonable to think that pupils from Country E would, on average, have a higher geometric knowledge (i.e., the latent trait the test is measuring), such that $\theta_p \sim \Phi(\mu + Z_{pj}\lambda_j, \sigma^2)$, where λ_j is a person-covariate parameter indicating this expected change in geometry knowledge θ_p depending on a person's (p's) country Z_{pj}. This type of model is often referred to as a *latent regression model* (LRM; see Andersen & Madsen, 1977), and the model likelihood is

$$\begin{aligned}
&\ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \mu, \sigma^2\right) \\
&= \prod_{p=1}^{P} \int_{\theta_p} Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_p\right) \phi\left(\theta_p; \mu + Z_{pj}\lambda_j, \sigma^2\right) d\theta_p \\
&= \prod_{p=1}^{P} \int_{\theta_p} \prod_{i=1}^{I} \frac{\exp\left[y_{pi}\alpha_i\left(\theta_p - \beta_i\right)\right]}{1 + \exp\left[\alpha_i\left(\theta_p - \beta_i\right)\right]} \\
&\quad \cdot \phi\left(\theta_p; \mu + Z_{pj}\lambda_j, \sigma^2\right) d\theta_p.
\end{aligned}$$

In practice, the issue of such country effects is a sensitive one and is a topic of much discussion (cf., culture-free tests). Perhaps pupils with the same ability, but from a different country, do not have an equally fair chance of answering correctly on a particular item due to some contextual or culture-specific information in that item. This type of phenomenon is commonly referred to as *differential item functioning* (DIF; Holland & Wainer, 1993), because how the item works depends on the group to which it is presented.

To accommodate for DIF, a new parameter set ζ can be introduced, leading to the following new model likelihood:

$$\begin{aligned}
&\ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \mu, \sigma^2\right) \\
&= \prod_{p=1}^{P} \int_{\theta_p} Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_p\right) \phi\left(\theta_p; \mu, \sigma^2\right) d\theta_p \\
&= \prod_{p=1}^{P} \int_{\theta_p} \prod_{i=1}^{I} \frac{\exp\left[y_{pi}\alpha_i\left(\theta_p - \beta_i - z_{pj}\zeta_{ji}\right)\right]}{1 + \exp\left[\alpha_i\left(\theta_p - \beta_i - z_{pj}\zeta_{ji}\right)\right]} \\
&\quad \cdot \phi\left(\theta_p; \mu, \sigma^2\right) d\theta_p.
\end{aligned}$$

The parameter ζ_{ji} represents the change in item difficulty on the basis of the value of the person covariate Z_{pj}—in this case, the country the pupil lives in. This corresponds to a set of design matrices similar to those for unique item difficulties: $\mathbf{M} = \mathbf{D} \times [\text{diag}(\mathbf{R}) \times \text{parameter estimate} + \mathbf{O}] = \mathbf{V}$

$$\begin{bmatrix} \zeta_{j1} \\ \vdots \\ \zeta_{ji} \\ \vdots \\ \zeta_{jI} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & \mathbf{0} & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\times \left(\text{diag}\left(\begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right) \times \begin{bmatrix} \zeta_{j1}^{\text{est}} \\ \vdots \\ \zeta_{ji}^{\text{est}} \\ \vdots \\ \zeta_{jI}^{\text{est}} \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) = \begin{bmatrix} \zeta_{j1}^{\text{est}} \\ \vdots \\ \zeta_{ji}^{\text{est}} \\ \vdots \\ \zeta_{jI}^{\text{est}} \end{bmatrix}.$$

It is clear that, if one considers the country effect to be constant over items, this essentially reduces the model to an LRM

$$\begin{bmatrix} \zeta_{j1} \\ \vdots \\ \zeta_{ji} \\ \vdots \\ \zeta_{jI} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} \times \left(\text{diag}\left(\begin{bmatrix} 1 \end{bmatrix} \right) \times \begin{bmatrix} \zeta_{j}^{\text{est}} \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \right) = \begin{bmatrix} \zeta_{j}^{\text{est}} \\ \vdots \\ \zeta_{j}^{\text{est}} \\ \vdots \\ \zeta_{j}^{\text{est}} \end{bmatrix}$$

with the following equivalent model likelihood:

$$\begin{aligned}
&\ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \mu, \sigma^2\right) \\
&= \prod_{p=1}^{P} \int_{\theta_p} Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_p\right) \phi\left(\theta_p; \mu, \sigma^2\right) d\theta_p \\
&= \prod_{p=1}^{P} \int_{\theta_p} \prod_{i=1}^{I} \frac{\exp\left[y_{pi}\alpha_i\left(\theta_p - \beta_i - z_{pj}\zeta_{j}\right)\right]}{1 + \exp\left[\alpha_i\left(\theta_p - \beta_i - z_{pj}\zeta_{j}\right)\right]} \\
&\quad \cdot \phi\left(\theta_p; \mu, \sigma^2\right) d\theta_p.
\end{aligned}$$

This result can be seen as the decomposition of the geometric knowledge in a fixed part $Z_{pj}\zeta_j$ due to the country effect and a random residual part θ_p. This model is an equivalent formulation of the LRM and shows that it is, in fact, the person-side analogue of the LLTM (i.e., instead of the item parameter, the person parameter is decomposed). Differentiating between such a general country effect, called *impact*, and an item-specific country effect (i.e., DIF) can be done by sequentially checking whether an item functions differently by adding an item-covariate interaction parameter ς_{ji} to the LRM. The parameter ς_{ji} can then be interpreted as indicating the item-specific deviation from the general country effect λ_j, and the DIF hypothesis can be tested by means of a likelihood-ratio test between this model and a regular LRM.

Example code illustrating the equivalence of the models discussed above is included in Appendix I of the supplemental materials. The summary of the models in Table 2 shows that the LRM and the LRM formulated as common DIF have the same log likelihood and that there is a one-to-one relationship between their parameters. The results show that pupils of the industrial Country E have more chance of answering an item correctly than do pupils of the third-world Country F. The difference is 2.755 on the latent trait scale, which is an increase in the odds of 16 times (i.e., $e^{2.755}$). Hence, the unexplained variance σ^2 in geometry proficiency differences decreased severely, with the country effect accounting for 63% of the total variance (i.e., the sum of the variance of the fixed country effect and the residual latent trait variance: The former can be computed as the variance of Z_j multiplied by its squared effect λ_j^2; the latter is simply σ^2). To interpret the item-difficulty parameters of the LRM, the mean of the person effects has to be used as reference, which here is the sum of the average μ of the residual part and the average of the fixed country indicator Z_j. Adding this constant to the item difficulties brings them back on a range close to the original Rasch scale. The results show that there is no real indication of DIF. The increase in goodness-of-fit is only marginal, and the unique DIF parameters ζ_{ji} fluctuate closely around the common country effect ζ_j. However, there is one exception: The parameter ζ_{j1} has an extreme high value. This is, in fact, a result of an empirical underidentification, because, in the current data sample, every pupil of Country E responds correctly on this item; hence, little information is present to locate exactly the relative position of the unique DIF parameter. A more specific DIF analysis, in which a sequence of likelihood-ratio tests (see above) is performed, confirms these observations, because none of the tests were significant (see Table 3). Note that, as a result of multiple testing, a Bonferroni correction was applied to control the Type I error.

Accounting for Deviations of Common Assumptions

From a scientific point of view, formalizing a test in terms of a mathematical model helps to clarify exactly what is being measured. The test can be revised to fit the model restrictions (thereby obtaining specific measurement properties, e.g., the Rasch model), and the theory and corresponding model can be adapted to better fit the features of the data. The latter option has the logical implication that even the common main assumptions of normality of the latent trait population distribution (see Equation 3) and conditional independence (see Equation 2) are not always plausible for every application. First, we show how to account for deviations from the normality assumption by means of a finite mixture and how this can be used to investigate the potential existence of person subgroups in the sample. Then we show how to account for deviations of the conditional independence assumption by using copula functions.

A Finite Mixture Population Distribution and Latent Groups

For many human skills and properties, it is quite reasonable to assume that the latent trait θ_p is distributed normally in the population. However, in some cases, this might still not hold true, and this assumption would put too stringent a constraint on the shape of the distribution of the latent trait. A serious misspecification of the distribution can lead to biased parameter estimates in the model (see e.g., Agresti, Caffo, & Ohman-Strickland, 2004; Neuhaus, Hauck, & Kalbfleisch, 1992). For instance, with respect to depression, only a few people would be expected to be strongly affected, and most people would feel relatively unaffected. This would resemble a nonnormal, highly skewed population distribution for θ_p.

Instead of assuming that the latent trait is distributed normally with a given mean and variance, it is possible to assume that the latent trait is, for instance, a combination of two component distributions, $\theta_p \sim \pi_1 N(\mu_1,\sigma_1^2) + \pi_2 N(\mu_2,\sigma_2^2)$, in which the component weights π_g necessarily sum to 1 and are each restricted to the interval [0,1]. This type of distribution is called a *finite mixture* (McLachlan & Peel, 2000), and the new model likelihood is formulated as

$$\ell\left(\alpha,\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}^2\right) = \prod_{p=1}^{P}\sum_{g=1}^{G}\pi_g \int_{\theta_{pg}} Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_{pg}\right) \cdot \phi\left(\boldsymbol{\theta}_{pg};\mu_g,\sigma_g^2\right)d\theta_{pg}.$$

A finite mixture distribution has two advantages. First, from a technical point of view, the latent trait need not be restricted to a simple normal distribution, and nonnormal distributions can be considered instead, because these mixtures can, for instance, give rise to skewed or multimodal distributions. Second, from a substantive point of view, the mixture suggests two subpopulations of persons (i.e., latent groups). For instance, an alternative hypothesis for the positively skewed distribution of depression is that there are two subpopulations—a small clinically depressed group and a larger nondepressed group—instead of one homogeneous population. When such a subpopulation effect is hypothesized and membership of these subpopulations is known, an LRM can be applied. However, when membership information is not a priori available or the groups are unknown, a finite mixture analysis might provide suggestive evidence for the existence of subpopulations. The existence of the two subpopulations should then be reflected in the posterior allocation of persons among the mixture components and in the component means for these latent groups.

Estimation of this type of model is a bit more involved and is done using a generalized expectation–maximization algorithm (EM; McLachlan & Krishnan, 1997). A multistart procedure is advisable for this algorithm, because it only guarantees to find at least a local maximum. Thus, unfortunately, the advantages of the mixture approach are

Table 2
Model Results for the Geometry Example

Rasch			Latent Regression Model (LRM)			Differential Item Functioning (DIF)			LRM–DIF			Mixture		
Parameter	Estimate	*SE*	Parameter	Estimate	*SE*	Parameter	Estimate	*SE*	Parameter	Estimate	*SE*	Parameter	Estimate	*SE*
β_1	−4.835	0.205	β_1	−3.072	0.215	β_1	−2.943	0.208	β_1	−3.072	0.199	β_1	−3.368	1.016
β_2	−4.582	0.190	β_2	−2.828	0.198	β_2	−2.759	0.195	β_2	−2.828	0.189	β_2	−3.117	1.015
β_3	−3.668	0.149	β_3	−1.946	0.154	β_3	−1.964	0.161	β_3	−1.946	0.154	β_3	−2.216	1.011
β_4	−3.202	0.133	β_4	−1.494	0.147	β_4	−1.461	0.138	β_4	−1.494	0.142	β_4	−1.756	1.009
β_5	−3.146	0.133	β_5	−1.439	0.143	β_5	−1.444	0.142	β_5	−1.439	0.148	β_5	−1.701	1.009
β_6	−2.849	0.124	β_6	−1.150	0.134	β_6	−1.177	0.135	β_6	−1.150	0.128	β_6	−1.408	1.008
β_7	−2.514	0.118	β_7	−0.821	0.123	β_7	−0.809	0.123	β_7	−0.821	0.127	β_7	−1.075	1.007
β_8	−1.959	0.110	β_8	−0.269	0.114	β_8	−0.209	0.114	β_8	−0.269	0.115	β_8	−0.520	1.007
β_9	−1.463	0.103	β_9	0.230	0.109	β_9	0.196	0.117	β_9	0.230	0.113	β_9	−0.022	1.007
β_{10}	−0.670	0.098	β_{10}	1.036	0.109	β_{10}	0.876	0.122	β_{10}	1.036	0.112	β_{10}	0.785	0.997
β_{11}	−1.374	0.103	β_{11}	0.320	0.111	β_{11}	0.319	0.116	β_{11}	0.320	0.112	β_{11}	0.068	1.007
β_{12}	−1.165	0.101	β_{12}	0.532	0.112	β_{12}	0.546	0.118	β_{12}	0.532	0.114	β_{12}	0.278	1.007
β_{13}	−0.340	0.097	β_{13}	1.370	0.112	β_{13}	1.444	0.149	β_{13}	1.370	0.116	β_{13}	1.090	0.732
β_{14}	−0.923	0.098	β_{14}	0.778	0.111	β_{14}	0.689	0.127	β_{14}	0.778	0.111	β_{14}	0.522	1.008
β_{15}	−0.340	0.097	β_{15}	1.370	0.111	β_{15}	1.461	0.139	β_{15}	1.370	0.116	β_{15}	1.104	0.976
σ^2	3.333	0.220	λ	2.755	0.100	ζ_1	−9.794	1.016	ζ	−2.755	0.105	μ_2	2.534	0.859
λ			σ^2	1.094	0.100	ζ_2	−3.496	0.740	σ^2	1.094	0.099	σ^2	1.575	1.491
						ζ_3	−2.637	0.369				π_1	0.427	0.032
						ζ_4	−2.947	0.361				π_2	0.573	
						ζ_5	−2.725	0.321						
						ζ_6	−2.628	0.287						
						ζ_7	−2.803	0.244						
						ζ_8	−2.959	0.224						
						ζ_9	−2.668	0.178						
						ζ_{10}	−2.460	0.172						
						ζ_{11}	−2.754	0.183						
						ζ_{12}	−2.788	0.177						
						ζ_{13}	−2.880	0.192						
						ζ_{14}	−2.573	0.177						
						ζ_{15}	−2.908	0.184						
Log Likelihood	5,836.5			5,471.39			5,463.48			5,471.39			5,832.69	
No. Parameters	16			17			31			17			18	
AIC	11,705.00			10,976.77			10,988.95			10,976.77			11,701.38	
BIC	11,783.52			11,060.21			11,141.09			11,060.21			11,789.72	

Note—AIC, Akaike's information criterion; BIC, Bayesian information criterion.

Table 3
Likelihood-Ratio Test Sequence for DIF Detection in the Geometry Example

i	$\chi^2(1)$	*p*	*i*	$\chi^2(1)$	*p*	*i*	$\chi^2(1)$	*p*
1	6.559	.010	6	0.260	.611	11	0.000	.996
2	1.248	.264	7	0.039	.844	12	0.043	.836
3	0.121	.728	8	1.048	.306	13	0.665	.415
4	0.362	.547	9	0.264	.607	14	1.364	.243
5	0.013	.909	10	3.780	.052	15	0.983	.322

Note—Bonferroni correction: Type I error alpha = .05/15 = .003.

slightly counterbalanced by its lesser convergence behavior (speed and local optima).

Example. To illustrate the approach, a finite mixture was fitted on the geometry example, ignoring the country information. Given that there are subpopulations present in the data set, the mixture should be able to retrieve this grouping, even without the membership information otherwise present in Z_{pj}. Example code is given in Appendix I of the supplemental materials, and the model results are displayed in Table 2. Note that the following restrictions are set to identify the finite mixture in a Rasch model context: $\mu_1 = 0$ (i.e., Component 1 is the reference) and $\sigma_1^2 = \sigma_2^2$ (i.e., homoscedasticity).

Allocating pupils on the basis of their maximum posterior component probability performed remarkably well, as can be seen when comparing the known classification Z_{pj} with the retrieved classification in the following confusion matrix:

$$\left(\begin{array}{cc|l} .33 & .07 & 0.4\ \text{F} \\ .06 & .54 & 0.6\ \text{E} \end{array}\right) \text{known country}$$

$$\left(\begin{array}{cc|l} .39 & .61 & \\ 1\text{=F} & 2\text{=E} & \end{array}\right) \text{component} \cong \text{inferred country}.$$

The outer matrix contains the marginal totals, and the inner matrix contains the four possible allocation combinations. For instance, the correct classification rate of the finite mixture is 87% (i.e., 33% + 54%), and 10% of the pupils of Country E are misclassified in Country F (i.e., .06/.6). Relative to the LRM model, the finite mixture results in a slightly smaller average group difference and a larger latent trait variance. Note that standard errors in the finite mixture model are quite large, partly because the convergence criterion was set low to spare computation time, since interest merely went out to the group retrieval and not to the accurate estimation of the item parameters.

Copula Functions for Locally Dependent Items

In the previous sections, the latent trait was considered to account for all the dependency between a person's responses on the test: This is referred to as the *conditional independence assumption*. This assumption allows for the convenient formulation of the joint probability of the item responses on the full test for a person *p* as the product of the marginal item probabilities:

$$Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_p\right) = \prod_{i=1}^{I} Pr\left(Y_{pi} = y_{pi} \mid \theta_p\right).$$

However, in practice, some subsets of items might be more interrelated than can be explained by the general latent trait θ_p underlying the test. Examples include item subsets that appeal to the same specific background knowledge, such as subquestions of the same problematic case or items with the same question format. These item subsets are sometimes called *item bundles* or *testlets* in the literature, and the prototypical examples here are items relating to the same reading passage.

These residual local item dependencies (LIDs) are not relevant for the construct being measured, but do imply that conditional independence does not hold for all items in the test, and, hence, that model estimates and inferences will be distorted if ignored (see, e.g., Chen & Thissen, 1997; Ip, Wang, De Boeck, & Meulders, 2004; Junker, 1991; Sireci, Thissen, & Wainer, 1991; Tuerlinckx & De Boeck, 2001). In other words, another, more adequate formulation of $Pr(\mathbf{Y}_p = \mathbf{y}_p | \theta_p)$, the joint probability of a person's response vector, is needed.

Copula functions. A convenient way of accounting for these residual dependencies within an item subset can be found in the use of copula functions (Braeken et al., 2007). A *copula* is a type of function that is able to connect sets of marginal distributions to form a multivariate distribution that preserves these margins (for reference works on copula theory, see Joe, 1997, and Nelsen, 1999). The main idea here is that, instead of making use of the product function C_Π to couple the marginal distributions $F(Y_{pi} = y_{pi} | \theta_p) = Pr(Y_{pi} \leq y_{pi} | \theta_p)$, forming the multivariate distribution under conditional independence,

$$F_{\mathbf{Y}_p|\theta_p}\left(\mathbf{y}_p\right) = C_\Pi\left[F_{\mathbf{Y}_p|\theta_p}\left(\mathbf{y}_p\right)\right] = \prod_{i=1}^{I} F_{Y_{pi}|\theta_p}\left(y_{pi}\right),$$

a different function *C* is applied to form a multivariate distribution that does allow for a residual dependency structure between the different margins (i.e., items):

$$F_{\mathbf{Y}_p|\theta_p}\left(\mathbf{y}_p\right) = C\left[F_{Y_{p1}|\theta_p}\left(y_{p1}\right), \ldots, F_{Y_{pI}|\theta_p}\left(y_{pI}\right)\right].$$

This function *C* is, in fact, a copula function, making the regular independence case a specific instance of the larger class of copula functions. The biggest advantage of the approach is that it does not require changing the formulation of the model for an individual item (see Equation 1).

In practice, this means that we will consider the total set of items $J = \{1,2, \ldots, I\}$ to consist of mutually exclusive item subsets J_s ($s = 1, \ldots, S$), for which conditional independence holds between the different subsets, but where the joint probability of the responses in subset *s* $Pr_s(\mathbf{Y}_p^{(s)} = \mathbf{y}_p^{(s)} | \theta_p)$ is evaluated from a copula function C_s (when subset size $I_s > 1$), allowing for LIDs within the subset:

$$Pr\left(\mathbf{Y}_p = \mathbf{y}_p \mid \theta_p\right) = \prod_{s=1}^{S} Pr_s\left(\mathbf{Y}_p^{(s)} = \mathbf{y}_p^{(s)} \mid \theta_p\right).$$

Note that, when $S = I$ (i.e., $I_s = 1 \forall s$) and, thus, each item is its own subset, this formulation reduces to the regular model formulation under conditional independence. In Table AII.3 (see the supplemental materials), the imple-

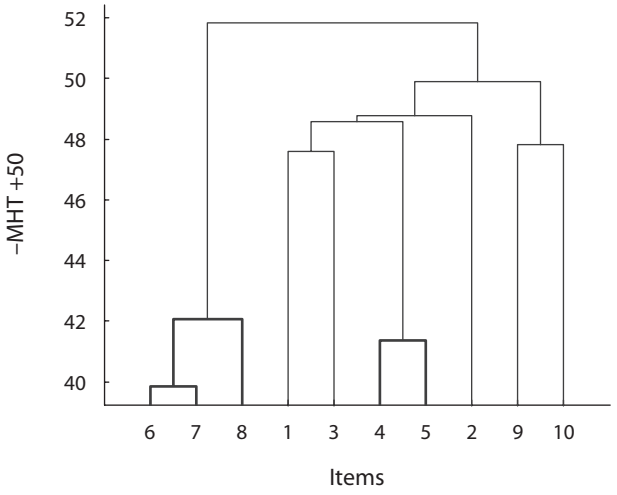

Figure 2. A dendrogram of the item dependency structure.

mented copula functions in IRTm are listed with a description of the type of LID they imply. Note that a new copula C_s can be composed as a convex linear combination of copula functions,

$$
\begin{aligned}
&F\left(\mathbf{Y}_p^{(s)} = \mathbf{y}_p^{(s)} \mid \theta_p\right) \\
&= C_s\left[F\left(Y_{p1} \mid \theta_p\right), \ldots, F\left(Y_{pI_s} \mid \theta_p\right); \boldsymbol{\delta}_s\right] \\
&= \delta_{1,2} C_{s,1} \\
&\quad \cdot \left[F\left(Y_{pi} = y_{p1} \mid \theta_p\right), \ldots, F\left(Y_{pi} = y_{pI_s} \mid \theta_p\right); \delta_{1,1}\right] \\
&\quad + \delta_{2,2} C_{s,2} \\
&\quad \cdot \left[F\left(Y_{pi} = y_{p1} \mid \theta_p\right), \ldots, F\left(Y_{pi} = y_{pI_s} \mid \theta_p\right); \delta_{2,1}\right],
\end{aligned}
$$

with $\delta_{k,1}$, a dependency parameter specific to the kth copula in the linear combination, and $\delta_{k,2}$, a parameter indicating the weight of the kth component, within this combination, for which the usual restrictions for mixture-like models hold:

$$\sum_k^K \delta_{k,2} = 1, \ \delta_{k,2} \in [0,1].$$

Hence, a constrained estimation algorithm is required. For such a convex combination copula, model identification is an issue. A subtle trade-off exists between the degree of association, as is indicated by the parameter $\delta_{k,1}$ and the component weight $\delta_{k,2}$: An increase in either increases the relative importance of the copula $C_{s,k}$ in the convex linear combination and may give rise to similar association patterns. Therefore, in practice, the application of this construction limits itself to combinations of nonparametric copulas or copulas with fixed values for $\delta_{k,1}$.

Exploration. The IRTm toolbox offers a rough tool to assist in exploring the structure of the test, based on the comparison of Mantel–Haenszel (MH; Holland & Rosenbaum, 1986; Mantel & Haenszel, 1959) statistics for each pair of items (for a review of other approaches, see Tate, 2003). This MH statistic essentially measures whether the odds ratio for two items is equal over the different levels of the latent trait, which is implied by the conditional independence assumption. Hence, a large value for this statistic is an indication of LID for the specific item pair. A figure containing the matrix of MH statistics for all item pair combinations and a dendrogram of a rough hierarchical clustering on the basis of these MH statistics can be obtained. Both can assist in determining whether and where one must account for residual dependency issues in the data set. To calculate the MH statistics, the sum score over items for each person can be used as a temporary proxy for the latent trait for the division in latent trait level groups.

Example. Consider a 10-item reading-comprehension test, taken by 1,000 students applying to a university. Some of the questions refer to similar parts in the main text that the students had to read; hence, there may be some concern regarding the potential presence of LIDs. First, to

Table 4
Model Results for the Reading Comprehension Example

Rasch			Copula		
Parameter	Estimate	*SE*	Parameter	Estimate	*SE*
β_1	−2.165	0.106	β_1	−2.072	0.105
β_2	−1.690	0.097	β_2	−1.613	0.089
β_3	−1.125	0.088	β_3	−1.069	0.086
β_4	−0.511	0.084	β_4	−0.484	0.082
β_5	0.096	0.083	β_5	0.098	0.081
β_6	−0.103	0.083	β_6	−0.099	0.079
β_7	0.616	0.085	β_7	0.603	0.080
β_8	1.024	0.088	β_8	0.972	0.081
β_9	1.820	0.099	β_9	1.738	0.090
β_{10}	2.071	0.104	β_{10}	1.978	0.101
			δ_1	1.405	0.072
			δ_2	2.036	0.207
σ^2	1.582	0.119	σ^2	1.219	0.108
Log likelihood	5,371.94			5,179.81	
No. parameters	11			12	
$R_{\theta,\theta}$	.774			.722	

have a base of reference, a Rasch model is formulated under conditional independence and the item dependency structure of the test is explored. The example code to perform these actions in IRTm is provided in Appendix I of the supplemental materials, and a dendrogram of the item dependency structure is shown in Figure 2. On the basis of the lowest level clusters in the dendrogram, one can distinguish two subsets that show the presence of a LID issue, $J_1 = \{4,5\}$ and $J_2 = \{6,7,8\}$. Thus, the items within each subset are joined by means of a copula, thereby correcting for the excess item interdependency, which should result in a more appropriate joint probability of the item responses and in a better model for our test.

A summary of the model results can be found in Table 4. Both copula parameters appear to be significant ($p < .0001$), indicating that there is indeed a LID issue in the data set. Furthermore, when the regular conditional independence Rasch model is compared with the copula Rasch alternative, the latter model outperforms the former by far. Hence, the suspicions of LID are formally confirmed and are even taken into account by the new model.

Ignoring LID would mean that the set of items are assumed to provide more information on the latent trait than they actually do. Hence, the reliability of the test instrument would be overestimated. A summary measure of reliability is the ratio of the true variance to the total variance (true + error variance), which can be computed as the latent trait population variance σ^2 divided by the sum of σ^2 and the mean squared error of the θ_p estimates in the sample data. The resulting statistics, $R_{\theta,\theta} = \sigma^2/(\text{MSE}(\theta_p) + \sigma^2)$, can be found in Table 4. As is expected, compared with the regular Rasch model, the reliability decreases, in this case with about 7% (i.e., comparable to one item) when accounting for the LID issues by means of the copula model.

DISCUSSION

With the presentation of the IRTm toolbox, we provide a small integrated IRT toolbox for practitioners who wish to make use of IRT for exploratory and explanatory research purposes. Note that IRTm can not only model, but can also simulate, data. The adopted design matrix approach allows for control and flexibility in the model-building stage, and the MATLAB environment allows for convenient postprocessing and graphical presentation of model results. The most general model formulation that the IRTm toolbox can fit is presented in Appendix II of the supplemental materials, together with the basic procedures behind the toolbox. Furthermore, we offer a means of exploring, in practice, the potential of the copula approach to LID, which has rich theoretical and mathematical properties.

Some enhancements can be made to the toolbox: further optimization of speed and memory use, for instance by accelerating the implemented EM algorithm; allowance for data missing completely at random; additional utility functions for standard IRT test statistics; and support for plotting. Some ongoing developments include polytomous item responses and alternative LID models, such as the conditional interaction models (Hoskens & De Boeck, 1997) and testlet models (Wainer, Bradlow, & Wang, 2007).

AUTHOR NOTE

Preparation of this article was supported in part by the Fund for Scientific Research Flanders (F.W.O.) Grant No. G.0148.04, by the K. U. Leuven Research Council Grant No. GOA/2005/04, and by the Cito Psychometric Research and Knowledge Center. Correspondence and requests for reprints should be sent to J. Braeken, Department of Methodology and Statistics, Postbus 90153, 5000 LE Tilburg, The Netherlands (e-mail: j.braeken@uvt.nl).

REFERENCES

AGRESTI, A., BOOTH, J. G., HOBERT, J. P., & CAFFO, B. (2000). Random effects modeling of categorical response data. *Sociological Methodology*, **30**, 27-80.

AGRESTI, A., CAFFO, B., & OHMAN-STRICKLAND, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency and possible remedies. *Computational Statistics & Data Analysis*, **47**, 639-653.

ANDERSEN, E. B., & MADSEN, I. (1977). Estimating the parameters of a latent population distribution. *Psychometrika*, **42**, 357-374.

BIRNBAUM, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

BOCK, R. D., & LIEBERMAN, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, **35**, 179-197.

BRAEKEN, J., TUERLINCKX, F., & DE BOECK, P. (2007). Copula functions for residual dependency. *Psychometrika*, **72**, 393-411.

CHEN, W.-H., & THISSEN, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational & Behavioral Statistics*, **22**, 265-289.

DE BOECK, P., & WILSON, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

DE LEEUW, J., & MAIR, P. (2007). An introduction to the special volume on "Psychometrics in *R*." *Journal of Statistical Software*, **20**, 1-5.

EMBRETSON, S. E., & REISE, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

FISCHER, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, **37**, 359-374.

HOLLAND, P. W., & ROSENBAUM, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, **14**, 1523-1543.

HOLLAND, P. W., & WAINER, H. (EDS.) (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

HOSKENS, M., & DE BOECK, P. (1997). A parametric model for local dependencies among test items. *Psychological Methods*, **2**, 261-277.

IP, E., 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.

JOE, H. (1997). *Multivariate models and dependence concepts.* London: Chapman & Hall.

JUNKER, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, **56**, 255-278.

MANTEL, N., & HAENSZEL, W. (1959). Statistical aspects of the retrospective study of disease. *Journal of the National Cancer Institute*, **22**, 719-748.

MCLACHLAN, G. J., & KRISHNAN, T. (1997). *The EM algorithm and extensions.* New York: Wiley.

MCLACHLAN, G. J., & PEEL, T. (2000). *Finite mixture models*. New York: Wiley.

NELSEN, R. B. (1999). *An introduction to copulas.* New York: Springer.

NEUHAUS, J. M., HAUCK, W. W., & KALBFLEISCH, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effect logistic models. *Biometrika*, **79**, 755-762.

NUNNALLY, J. C., & BERNSTEIN, I. H. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.

RASCH, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, **8**, 185-205.

Severini, T. A. (2000). *Likelihood methods in statistics*. New York: Oxford University Press.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, **28**, 237-247.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, **27**, 159-203.

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). Multilog. [Computer software]. Lincolnwood, IL: Scientific Software International.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, **6**, 181-195.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3: Item analysis and test scoring with binary logistic models [Computer software]. Chicago: Scientific Software.

NOTE

1. A single underlying latent trait is assumed, in contrast to multidimensional extensions, similar to factor analysis.

SUPPLEMENTAL MATERIALS

The two appendixes mentioned in this article, containing example code and information on the general copula IRT in IRTm, may be downloaded from brm.psychonomic-journals.org/content/supplemental.

(Manuscript received August 19, 2008;
revision accepted for publication June 16, 2009.)