A boundary mixture approach to violations of conditional independence

Braeken, J.

Published in:
Psychometrika

DOI:
10.1007/s11336-010-9190-4

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

PSYCHOMETRIKA—VOL. 76, NO. 1, 57–76
JANUARY 2011
DOI: 10.1007/S11336-010-9190-4

A BOUNDARY MIXTURE APPROACH TO VIOLATIONS OF CONDITIONAL INDEPENDENCE

JOHAN BRAEKEN

TILBURG UNIVERSITY

Conditional independence is a fundamental principle in latent variable modeling and item response theory. Violations of this principle, commonly known as local item dependencies, are put in a test information perspective, and sharp bounds on these violations are defined. A modeling approach is proposed that makes use of a mixture representation of these boundaries to account for the local dependence problem by finding a balance between independence on the one side and absolute dependence on the other side. In contrast to alternative approaches, the nature of the proposed boundary mixture model does not necessitate a change in formulation of the typical item characteristic curves used in item response theory. This has attractive interpretational advantages and may be useful for general test construction purposes.

Key words: Fréchet–Hoeffding bounds, copula function, local item dependencies, conditional independence.

1. Introduction

Item response theory makes use of statistical probability models to explain the pattern of observed item responses on a test. The key property of these item response models is that both persons as well as items have a position on a latent dimension underlying the test. A fundamental principle behind most item response models is that this latent dimension is considered sufficient to explain the heterogeneity among sample units, as well as the homogeneity among item responses. In other words, the latent proficiency explains why there are individual differences between persons in test performance and why the item responses of a given person interrelate.

The local stochastic or conditional independence assumption (LSI) formalizes this principle into the statistical model.

Definition 1 (LSI).

$$\Pr(\boldsymbol{Y}_p = \boldsymbol{y}_p|\theta_p) = \prod_{i=1}^{I} \Pr(Y_{pi} = y_{pi}|\theta_p). \tag{1}$$

Let Y_{pi} be the outcome on item i ($i = 1, \dots, I$) of person p, then, given θ_p the position of this person on the latent dimension, the item responses can be assumed to be independent. Hence, the joint conditional probability of the item response pattern $\boldsymbol{Y}_p = [Y_{p1}, \dots, Y_{pi}, \dots, Y_{pI}]$ can be conveniently factorized in a mathematical sense and written as a simple product of the marginal conditional probabilities of the individual items $\Pr(Y_{pi} = y_{pi}|\theta_p)$.

However, this assumption formulates such a strict requirement—conditional upon θ_p there is no remaining dependence between the item responses—that it is unlikely to be met completely in most applications. In practice some subsets of items appeal to the same specific background theme, use the same stimulus material, are subquestions of the same problematic case, or in other

Requests for reprints should be sent to Johan Braeken, Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands. E-mail: j.braeken@uvt.nl

words, share some common ground that is not directly relevant to the more general construct underlying the whole test. It can be expected that responses on such subset items will partially show dependence due to their shared idiosyncratic features and not only because they relate to the skill or ability intended to be measured by the test as a whole (Ferrara, Huynh, & Michaels, 1999). Such residual local item dependencies (LID) indicate that the model fails to correctly account for the item dependence structure, resulting in an unwanted negative impact on model results and related inferences.

1.1. Organization of the Paper

In the remainder of the paper, the nature of the LID problem is explicitly situated within an information perspective, theoretical boundary cases for LID are defined, and a model approach to account for the LID problem is presented based upon a mixture representation of these boundaries.

Note that this paper can be seen as a more accessible prequel to Braeken, Tuerlinckx, and De Boeck (2007), yet the concepts and statistics used are perhaps more fundamental, but also more primitive than in the latter paper. It makes the transition to the bigger class of copula models by making use of boundary distributions and traditional mixture techniques everyone can relate to, instead of using specific copula functions that arise less naturally. By making this connection to copula functions as an aside, the previous paper will hopefully be partially demystified. After all, copula functions are merely just another tool to build multivariate distributions, mixing distributions is yet another, and the proposed approach in this paper can be seen as the result of either or both techniques. The distinguishing feature of the model approach is that it changes the formulation of the joint conditional model, yet leaves the traditional item characteristic curves of IRT, i.e., the marginal conditional model part, intact.

The primary goal of the paper is to illustrate the attractive interpretational properties of such a marginal modeling approach for existing residual dependencies (i.e., LID) above the dependence induced by the latent trait of focus. With this purpose the key differences with a well-known alternative, the testlet model (Wainer, Bradlow, & Wang, 2007), are briefly illustrated. The testlet model belongs to the bigger class of factor analytic models such as the bifactor model (Gibbons & Hedeker, 1992) and multidimensional IRT models, which all rely on the introduction of additional latent traits in the traditional item characteristic curves of IRT, thus changing the marginal conditional model part. It will be shown that this also has consequences for comparability and interpretation that people are often not aware of or that are simply overlooked.

2. LID and Information Redundancy

The key problem with locally dependent items can be made more explicit when one recognizes that this is in essence an information issue. This aspect is also more clearly pronounced in an equivalent definition of the LSI assumption given by Lazarsfeld (1950).

Definition 2 (LSI).

$$\Pr(Y_{pi} = y_{pi} | \theta_p) = \Pr\big(Y_{pi} = y_{pi} | \theta_p, Y_{pj};\ \forall j \neq i, j \in [1, \dots, I]\big). \tag{2}$$

Conceptually, if the latent proficiency θ_p is known, one cannot learn anything more from Y_{pj} (the response on item j) about Y_{pi} (the response on item i). All information is sufficiently summarized in θ_p.

Statistically, the information provided by a response pattern y_p on the latent trait θ_p is related to the Fisher information, which is reciprocal to the variance of an estimate, in our case of $\hat{\theta}_p$. The latter variance is denoted by $se^2(\hat{\theta}_p)$. Samejima (1969, 1972) offers the most general applicable definition of information in the context of item response theory. A response pattern information function can be formulated as

$$I_{\boldsymbol{y}}(\theta_p) = -\frac{\partial^2}{\partial\theta_p^2} \log \Pr(\boldsymbol{Y}_p = \boldsymbol{y}|\theta_p) = 1/se^2(\hat{\theta}_p).$$

The expected value of this response pattern information function is then known as the test information function:

$$I(\theta_p) = \sum_{\boldsymbol{y}\in\omega} I_{\boldsymbol{y}}(\theta_p) \Pr(\boldsymbol{Y}_p = \boldsymbol{y}|\theta_p),$$

where the sum is over all possible item response patterns for the test.

Given the simple factorization of the joint conditional probability provided by LSI, it can be shown that the test information is a simple sum over the item information functions

$$I(\theta_p) = \sum_{i=1}^{I} I_i(\theta_p),$$

and hence, each item is assumed to provide unique information on the latent trait. LSI has the practical implication that each response pattern information function is equivalent, and hence corresponds to its expected value, the test information function

$$I_{\boldsymbol{y}}(\theta_p) \equiv I(\theta_p); \quad \forall \boldsymbol{y} \in \omega.$$

In sum, given LSI the contribution of each item to the test information does not depend on other items.

However, for items i and j of a locally dependent subset, one can still learn something extra about the response on one item from the response on the other item. Their interdependency is not fully captured by θ_p alone. Hence, the information provided by locally dependent items i and j is not simply additive. From a conceptual perspective, the item subset $J_s = \{i, j\}$ provides not only unique item information parts, denoted by $I_i^*(\theta_p)$ and $I_j^*(\theta_p)$, but also an overlapping redundant part $I_{ij}^*(\theta_p)$ due to the irrelevant subset-specific common ground. When this redundancy is ignored, the subset information $I_{ij}(\theta_p)$ falsely adds up to

$$\begin{aligned} I_i(\theta_p) &= I_i^*(\theta_p) + I_{ij}^*(\theta_p), \\ I_j(\theta_p) &= I_j^*(\theta_p) + I_{ij}^*(\theta_p), \\ I_{ij}(\theta_p) &= I_i^*(\theta_p) + I_j^*(\theta_p) + 2I_{ij}^*(\theta_p). \end{aligned} \tag{3}$$

This double counting of information in the locally dependent items inflates the reliability of the test (see e.g., Junker, 1991). Hence, measurement of the latent proficiency θ_p is artificially more precise than the test instrument allows for.[1]

To further illustrate the misspecification issue, consider the extreme example of three duplicate item responses Y_{pi}, Y_{pj}, and Y_{pk}, such that $y_{pi} = y_{pj} = y_{pk} = y_{p*}$ always. By definition these items provide the exact same information and adding up their information would lead to an artificially high share and, thus, artificial increase in the test information. The joint conditional

[1] For negative local dependence the effects reverse: information gain instead of redundancy.

probability of these duplicate responses is, of course, not the product of the marginal conditional probabilities as a conditional independence model would state, but merely equals the conditional probability of a single duplicated item:

$$\begin{aligned}\Pr(Y_{pi} = y_{p*}, Y_{pj} = y_{p*}, Y_{pk} = y_{p*}|\theta_p) &= \Pr(Y_{pi} = y_{p*}|\theta_p)\\ &= \Pr(Y_{pj} = y_{p*}|\theta_p)\\ &= \Pr(Y_{pk} = y_{p*}|\theta_p)\\ &\neq \Pr(Y_{pi} = y_{p*}|\theta_p) \times \Pr(Y_{pj} = y_{p*}|\theta_p)\\ &\quad \times \Pr(Y_{pk} = y_{p*}|\theta_p).\end{aligned} \tag{4}$$

In reality, the local dependence is of course less extreme, but the misspecification of the joint conditional model might still potentially bias the model parameters and inferences. This is especially an issue for the discrimination parameters, the standard errors of the latent trait estimate and related test reliability (see e.g., Sireci, Thissen, & Wainer, 1991; Chen & Thissen, 1997; Masters, 1988; Yen, 1984).

3. Boundaries to LID

Assessing the degree of violation of the LSI assumption requires knowledge about the boundaries between which the local item dependence can vary. An obvious choice would be to adopt a dependence measure, such as the correlation coefficient or odds ratio, of which the boundaries are known, and try to incorporate this into the item response model to assess the severity of the LID. Unfortunately, in the case of discrete item responses, boundaries of traditional dependence measures are a function of the marginal distributions of the individual items. For instance, the default definition of a correlation states that it can take values throughout the $[-1, 1]$ interval, however for categorical variables possible values for a correlation are often constrained to a much narrower interval (see e.g., Cureton, 1959; Joe, 1997, p. 210). Considering the role of these margins, boundary cases of local item dependence can be established by making use of a fundamental result in the study of multivariate distributions with given margins.

Definition 3 (Fréchet–Hoeffding bounds). Let $\mathbb{F}(F_{X_i}\ \forall i \in [1, \ldots, I])$ be a Fréchet class (see e.g., Joe, 1997, p. 57), a class of distributions containing every possible multivariate distribution $F_{\boldsymbol{X}}(\boldsymbol{x})$ of a set of variables $\boldsymbol{X}$, of which each individual variable X_i is necessarily fixed to be distributed according to $F_{X_i}(x_i)$. The limiting boundary distributions, $W_{\boldsymbol{X}}(\boldsymbol{x})$ and $M_{\boldsymbol{X}}(\boldsymbol{x})$, for the Fréchet class $\mathbb{F}(F_{X_i}\ \forall i \in [1, \ldots, I])$ are given by the inequalities

$$W_{\boldsymbol{X}}(\boldsymbol{x}) < \Pi_{\boldsymbol{X}}(\boldsymbol{x}) < M_{\boldsymbol{X}}(\boldsymbol{x}),$$
$$W_{\boldsymbol{X}}(\boldsymbol{x}) \leq F_{\boldsymbol{X}}(\boldsymbol{x}) \leq M_{\boldsymbol{X}}(\boldsymbol{x}).$$

The function Π is the product function and merely defines the independence case

$$\Pi_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{I} F_{X_i}(x_i). \tag{5}$$

The functions M and W are known as the Fréchet–Hoeffding bounds (Fréchet, 1951; Hoeffding, 1940), and demarcate the dependence space that is possible given the set of margins $F_{X_i}(x_i)$; M

is the upper bound to monotone increasing (i.e., positive) dependence

$$M_{\boldsymbol{X}}(\boldsymbol{x}) = \min\bigl(F_{X_i}(x_i);\ \forall i \in [1,\ldots,I]\bigr), \tag{6}$$

and W is the lower bound to monotone decreasing (i.e., negative) dependence

$$W_{\boldsymbol{X}}(\boldsymbol{x}) = \max\left(\sum_{i=1}^{I} F_{X_i}(x_i) - I + 1, 0\right). \tag{7}$$

The upper bound M is always a proper multivariate distribution, yet the lower bound W is only guaranteed to be a proper distribution in the bivariate case, but not necessarily in the multivariate case. In any case, these bounds are sharp, thus every multivariate distribution $F_{\boldsymbol{X}}(\boldsymbol{x})$ of a set of variables $\boldsymbol{X}$, of which each individual variable X_i is fixed to be distributed according to $F_{X_i}(x_i)$, will be necessarily located in between these boundaries.

In the case of item response models, the margins of the multivariate distribution are the cumulative distributions based upon the conditional item probabilities, with $F_{Y_{pi}|\theta_p}(y_{pi}) = \Pr(Y_{pi} \leq y_{pi}|\theta_p)$. The distribution defined by the product function Π is then the regular conditional independence model. The joint distribution defined for a locally dependent subset $J_s = \{i, j\}$ consisting of two duplicate items i and j can then be formulated in terms of the Fréchet–Hoeffding upper bound $M = \min(F_{Y_{pi}|\theta_p}(y_{pi});\ \forall i \in J_s)$. Hence, making use of simple quadrant rules[2] (see e.g., Mood, Graybill, & Boes 1974), this gives rise to the following subset response pattern probabilities when $F_{Y_{pi}|\theta_p}(0) < F_{Y_{pj}|\theta_p}(0)$:

$$\begin{aligned}
\Pr(0,0|\theta_p, M) &= \min\bigl(F_{Y_{pi}|\theta_p}(0), F_{Y_{pj}|\theta_p}(0)\bigr),\\
&= F_{Y_{pi}|\theta_p}(0),\\
\Pr(0,1|\theta_p, M) &= F_{Y_{pi}|\theta_p}(0) - \min\bigl(F_{Y_{pi}|\theta_p}(0), F_{Y_{pj}|\theta_p}(0)\bigr),\\
&= 0,\\
\Pr(1,0|\theta_p, M) &= F_{Y_{pj}|\theta_p}(0) - \min\bigl(F_{Y_{pi}|\theta_p}(0), F_{Y_{pj}|\theta_p}(0)\bigr),\\
&= F_{Y_{pj}|\theta_p}(0) - F_{Y_{pi}|\theta_p}(0),\\
\Pr(1,1|\theta_p, M) &= 1 - F_{Y_{pi}|\theta_p}(0) - F_{Y_{pj}|\theta_p}(0) + \min\bigl(F_{Y_{pi}|\theta_p}(0), F_{Y_{pj}|\theta_p}(0)\bigr),\\
&= 1 - F_{Y_{pj}|\theta_p}(0),
\end{aligned}$$

$$\sum_{y_{pi}=0}^{1}\sum_{y_{pj}=0}^{1} \Pr(y_{pi}, y_{pj}|\theta_p, M) = 1.$$

The corresponding subset response pattern information functions are then

$$\begin{aligned}
&I_{\mathbf{y}^{(s)}=(0,0)}(\theta_p) = I_i(\theta_p),\\
&I_{\mathbf{y}^{(s)}=(0,1)}(\theta_p) = 0,\\
&I_{\mathbf{y}^{(s)}=(1,0)}(\theta_p) = I_i(\theta_p) + I_j(\theta_p),\\
&I_{\mathbf{y}^{(s)}=(1,1)}(\theta_p) = I_j(\theta_p)
\end{aligned}$$

[2] E.g., $F_{Y_{pi}|\theta_p}(0) = \Pr(Y_{pi} = 0|\theta_p) = \Pr(Y_{pi} = 0, Y_{pj} = 0|\theta_p) + \Pr(Y_{pi} = 0, Y_{pj} = 1|\theta_p)$.

such that the subset information under maximal positive local dependence simplifies to

$$\begin{aligned}I_{(s)}(\theta_p) &= I_i(\theta_p)\big[\Pr\big(\boldsymbol{Y}_p^{(s)} = (0,0)|\theta_p\big) + \Pr\big(\boldsymbol{Y}_p^{(s)} = (1,0)\big)\big] \\ &\quad + I_j(\theta_p)\big[\Pr\big(\boldsymbol{Y}_p^{(s)} = (1,1)|\theta_p\big) + \Pr\big(\boldsymbol{Y}_p^{(s)} = (1,0)\big)\big] \\ &= I_i(\theta_p)\Pr(Y_{pj} = 0|\theta_p) + I_j(\theta_p)\Pr(Y_{pi} = 1|\theta_p).\end{aligned}$$

Thus, the information provided by the easier item j (i.e., the item less likely to be correctly answered) is downweighted by the probability of correctly answering the more difficult item i (i.e., the item more likely to be correctly answered), and the information provided by the more difficult item i is downweighted by the probability of incorrectly answering the easier item j. Note that this exact relation only holds when item discriminations are equal within the subset, otherwise the weighting process is a bit less insightful. The corresponding subset response pattern information functions remain the same except for $I_{\boldsymbol{y}^{(s)}=(1,0)}(\theta_p)$, which unfortunately is not a straightforward function of $I_i(\theta_p)$ and $I_j(\theta_p)$.

In this paper the focus is on the upper bound because it is always guaranteed to be a proper multivariate distribution and because negative dependence in the multivariate case does not have a straightforward interpretation. In practice, severe negative local item dependencies are also more indicative of more general scale problems (e.g., items not at all measuring something in common).

4. Boundary Mixture Model for LID

The theoretical results on the LID boundary distributions can be used to formulate a model that can accommodate local item dependencies within an information-oriented interpretational framework.

4.1. Model

A new item response model can be constructed by redefining $F_{\mathbf{Y}_p^{(s)}|\theta_p}(\boldsymbol{y}_p^{(s)})$, the joint distribution of the response vector of an LID subset J_s, as a mixture of the joint distribution under independence (i.e., Π) and the joint distribution under absolute monotone increasing dependence (i.e., M), such that

$$\begin{aligned}\Pi_{\mathbf{Y}_p^{(s)}|\theta_p}\big(\boldsymbol{y}_p^{(s)}\big) &= \prod_{i\in J_s} F_{Y_{pi}|\theta_p}(y_{pi}), \\ M_{\mathbf{Y}_p^{(s)}|\theta_p}\big(\boldsymbol{y}_p^{(s)}\big) &= \min\big(F_{Y_{pi}|\theta_p}(y_{pi});\ \forall i \in J_s\big), \\ F_{\mathbf{Y}_p^{(s)}|\theta_p}\big(\boldsymbol{y}_p^{(s)}\big) &= \delta_0^{(s)}\Pi_{\mathbf{Y}_p^{(s)}|\theta_p}\big(\boldsymbol{y}_p^{(s)}\big) + \delta_1^{(s)}M_{\mathbf{Y}_p^{(s)}|\theta_p}\big(\boldsymbol{y}_p^{(s)}\big),\end{aligned} \tag{8}$$

where the usual mixture constraints hold, $\sum_{k=0}^{1}\delta_k^{(s)} = 1$ and $\delta_k^{(s)} \in [0,1]$. The parameter set $\boldsymbol{\delta}^{(s)} = [\delta_0^{(s)}, \delta_1^{(s)}]$ can be seen as weights balancing the two boundary distributions, conditional independence and absolute positive conditional dependence.

Besides being a mixture distribution, the resulting joint conditional distribution is also an instance of the model class proposed by Braeken et al. (2007), in which copula functions are proposed as a tool to deal with LID without having to change the formulation of the marginal conditional distributions, that is, the item response functions characterizing the traditional IRT models (e.g., 1-, 2-, or 3-parameter logistic models) (see also Braeken & Tuerlinckx, 2009). Both boundary distributions Π and M are in fact copula functions, and any convex sum of copula functions can be shown to be a copula itself (see e.g., Nelsen, 1998). Hence, this mixture always results in a valid multivariate distribution. Both mixing distributions and copula functions are common techniques to build multivariate distributions, and the proposed modeling approach can be seen as resulting from either of these two techniques. As such it provides a more gentle and intuitive introduction to the latter technique by presenting a natural arising copula function.

It can be seen that copulas are essentially a class of multivariate cumulative distributions with uniform univariate margins that, after transformation by means of an inverse cumulative distribution function, result in multivariate distributions with given margins and a whole range of dependence properties varying according to which copula was used to construct this new joint distribution. In our specific case, the copula function looks like

$$\bigl(1-\delta_1^{(s)}\bigr)\prod_{i\in J_s} u_i + \delta_1^{(s)} \min(u_i;\ \forall i \in J_s),$$

where the uniform univariate margins are given by $u_i = F_{Y_{pi}|\theta_p}(y_{pi})$ (i.e., uniform in the sense that the variable is defined on the interval [0, 1]), and the multivariate cdf is then $F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(\boldsymbol{y}_p^{(s)})$. The copula multivariate construction method and the related theorem that states that any existing multivariate distribution $F_{\boldsymbol{X}}(\boldsymbol{x})$ can be reformulated as a copula of its univariate margins $F_{X_1}(x_1), \dots, F_{X_I}(x_I)$ (Sklar, 1959),

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = C\bigl(F_{X_1}(x_1), \dots, F_{X_I}(x_I)\bigr),$$
$$C\bigl(F_{X_1}(x_1), \dots, F_{X_I}(x_I)\bigr) = F_{\boldsymbol{X}}(\boldsymbol{x}),$$

are fundamental to the use of copula functions in multivariate modeling and the study of dependence. An extensive review, background and theory on copula functions can be found in the reference works by Joe (1997) and Nelsen (1998).

The modeling approach requires partitioning of the item set $\{1, \dots, I\}$ into $S+1$ disjoint subsets J_s, of which J_0 gathers the locally independent items, and the other subsets gather mutual locally dependent items such that the joint probability under the conditional independence model (see Equation (1)) is redefined as

$$\Pr(\boldsymbol{Y}_p = \boldsymbol{y}_p|\theta_p) = \prod_{i\in J_0} \Pr(Y_{pi} = y_{pi}|\theta_p) \prod_{s=1}^{S} \Pr\bigl(\boldsymbol{Y}_p^{(s)} = \boldsymbol{y}_p^{(s)}|\theta_p, \boldsymbol{\delta}^{(s)}\bigr), \tag{9}$$

where $\Pr(\boldsymbol{Y}_p^{(s)} = \boldsymbol{y}_p^{(s)}|\theta_p, \boldsymbol{\delta}^{(s)})$ is the joint probability derived from the joint cumulative distribution in the boundary mixture formulation of Equation (8) with parameters $\boldsymbol{\delta}^{(s)}$. Conditional independence holds between the $S+1$ subsets and within subset J_0, while local dependence is allowed for within each of the other subsets J_s. Notice that subset sizes can be larger than 2, yet within a subset s exchangeability holds, hence the conditional in/dependence is considered to be homogeneous among all items within this subset. These are similar restrictions as in the testlet model (Wainer et al., 2007).

4.2. *Estimation*

In contrast to for instance latent classes, the mixture concept leads in this case to fairly straightforward model optimization, because each component in the convex sum of distributions is built up based upon the same conditional marginal distributions and parameters; this significantly reduces the estimation problem. Hence, model fitting can be done using a common full information marginal maximum likelihood estimation approach. The parameters to be estimated can be divided into three groups: $\boldsymbol{\eta}$, the parameters of the latent distribution for θ_p; $\boldsymbol{\beta}$, the I sets of item parameters defining the marginal conditional probability function of items within the item response model; and $\boldsymbol{\delta}$, the S sets of weights of the convex sum defining the boundary mixture distributions of the S locally dependent disjoint item subsets. The full model likelihood is given by

$$\begin{aligned} &\mathit{likelihood}(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma_\theta; \boldsymbol{Y}) \\ &\quad = \prod_{p=1}^{P} \int_{\theta_p} \prod_{i \in J_0} \Pr(Y_{pi} = y_{pi} | \theta_p) \prod_{s=1}^{S} \Pr\big(\boldsymbol{Y}_p^{(s)} = \boldsymbol{y}_p^{(s)} | \theta_p, \boldsymbol{\delta}^{(s)}\big) h\big(\theta_p; \sigma_\theta^2\big)\, d\theta_p. \end{aligned}$$

In the application a 2PL model (Birnbaum, 1968) is chosen as model for the individual items, the distribution of the latent proficiency $h(\theta_p; \sigma_\theta^2)$ is chosen to be a standard normal distribution (i.e., mean zero and fixed variance $\sigma_\theta^2 = 1$), and the intractable integral with respect to this distribution is approximated using non-adaptive Gauss–Hermite quadrature (20 points). The joint probability under the boundary-mixture formulation can be evaluated using a recursive quadrant-rule function (see Appendix Equation (10)); alternatively a direct formulation of the joint probability can be written out for small subset sizes (e.g., $I_s \leq 4$, larger subsets might lead to impractically long equations) to avoid the recursion. Optimization of the model likelihood is done using a quasi-Newton method within the open-source software environment R.[3] Note that when the initial value of a redundancy parameter $\delta_1^{(s)}$ is chosen too close to the limiting values zero or one, divergence of the likelihood estimate of this parameter can occur, such that the optimization algorithm remains stuck in that limit even when it is not the optimal value. Hence, it is recommended to generate initial values from a uniform distribution between 0.2 and 0.8 to avoid such a local maximum problem.

4.3. *Interpretation*

The main change in the item response model is only in the formulation of the joint conditional distribution, while the marginal conditional part of the model (i.e., the formula for the item response function) is left intact. This in contrast to other approaches such as the testlet models (Wainer et al., 2007) and conditional interaction models (Hoskens & De Boeck, 1997; Verhelst & Glas, 1993) that accommodate for local item dependence by changing the marginal conditional formulation of the model by either adding additional latent traits or higher order terms into the formula for the item response function. This difference in the way of tackling the LID problem—changing the joint or the marginal conditional model—will also show itself in differences in the interpretation of the common item parameters in these different types of item response models.

The parameter $\delta_1^{(s)}$ can be seen as a threshold on a uniform scale going from conditional independence to absolute positive dependence for given item margins. As such, this parameter can be seen as a margin-free effect size measure of LID. From the margin-dependent perspective, the boundary mixture allows for a type of conditional dependence for which the conditional odds ratio of two locally dependent items increases with larger absolute values of the latent trait θ_p. In other words, it gets more likely for high proficiency persons to score all subset items correct, and for low proficiency persons to score all subset items wrong. From a substantive perspective, this appears quite intuitive and attractive. Notice the subtlety, the conditional dependence measured by a statistic that does not account for the margins is a function of the marginal conditional

[3] Pending a fully documented and user-friendly R-package, R-code can be requested from the author by e-mail.

probabilities—and hence of θ_p—yet the degree of conditional dependence given these margins remains constant (cf. $\delta_1^{(s)}$ parameter).

To further illustrate the interpretation of the boundary-mixture model, it will be compared to a testlet model (Wainer et al., 2007). While it can be anticipated that the two models will not differ too much in performance of dealing with the LID problem, they will differ largely in interpretation and consequences for further applications, equating and test construction. This will be clarified in more practical terms with an example in the application section, yet let us briefly outline the theoretical interpretational differences that are brought along by the introduction of an additional random effect/latent trait to account for the LID as in the testlet model.

The likelihood of a traditional 2PL testlet model can be defined as

$$\begin{aligned}
&\mathit{likelihood}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}; \boldsymbol{Y}) \\
&\quad = \prod_{p=1}^{P} \int_{\theta_p} \prod_{i \in J_0} \Pr(Y_{pi} = y_{pi} | \theta_p) \prod_{s=1}^{S} \int_{\zeta_{ps}} \Pr\big(\boldsymbol{Y}_p^{(s)} = \boldsymbol{y}_p^{(s)} | \theta_p, \zeta_{ps}\big) h\big(\zeta_{ps}; \sigma_s^2\big)\, d\theta_p h\big(\theta_p; \sigma_\theta^2\big)\, d\theta_p \\
&\quad = \prod_{p=1}^{P} \int_{\theta_p} \prod_{i \in J_0} \Pr(Y_{pi} = y_{pi} | \theta_p) \prod_{s=1}^{S} \int_{\zeta_{ps}} \prod_{i \in J_s} \Pr(Y_{pi} = y_{pi} | \theta_p, \zeta_{ps}) \\
&\qquad \times h\big(\zeta_{ps}; \sigma_s^2\big)\, d\theta_p h\big(\theta_p; \sigma_\theta^2\big)\, d\theta_p
\end{aligned}$$

with

$$\Pr(Y_{pi} = y_{pi} | \theta_p) = \frac{\exp(y_{pi}\alpha_i[\theta_p - \beta_i])}{1 + \exp(\alpha_i[\theta_p - \beta_i])} \quad \text{if } i \in J_0,$$

$$\Pr(Y_{pi} = y_{pi} | \theta_p, \zeta_{ps}) = \frac{\exp(y_{pi}\alpha_i[\theta_p - \beta_i + \zeta_{ps}])}{1 + \exp(\alpha_i[\theta_p - \beta_i + \zeta_{ps}])} \quad \text{if } i \in J_s.$$

The item response function for non-subset items remains the same as in a regular 2PL model, but for subset items an additional random effect ζ_{ps} has been added. Hence, the original fixed item effect β_i is now decomposed into a fixed part β_i^* and a random part ζ_{ps} common to the subset. In practice it is easily overlooked that the item parameter β_i in the testlet model is in fact this item parameter β_i^*, and hence adjusted for individual unobserved heterogeneity on the subset level (i.e., the person-specific subset effect ζ_{ps}). To sum it up, β_i and β_i^* are using a different reference frame due to conditioning on either θ_p or on both θ_p and ζ_{ps}. In fact a duality arises, the mixing distributions $h(\zeta_{ps}; \sigma_s^2)$ and $h(\theta_p; \sigma_\theta^2)$ give rise to a different compound logit-normal link that is used for subset items rather than for the non-subset items; the normal part follows a different scale—either $\sigma_\theta^2 + \sigma_s^2$ or σ_θ^2. In a similar fashion, the traditional interpretation of the discrimination parameter α_i for subset items is affected as well by the incorporation of the additional random effect ζ_{ps} and the corresponding difference in reference frame/link function. These interpretational differences are often mistakenly ignored in practice, where the testlet item parameters are confounded with the traditional item parameters. This makes comparison between testlet and non-testlet items less straightforward; the same can be said for anticipating the use of a testlet item within a testlet context or separated from its testlet context.

In contrast, the alternative construction method, using boundary mixtures or copula functions, benefits from its marginal reproducibility property when it concerns interpretation. Item parameters can be interpreted in a traditional fashion, without the disturbance of confounding reference frames. Preservation of the univariate conditional margins based upon the joint conditional distribution (i.e., reproducibility) is easily shown as follows for the example with two locally dependent items i and j:

$$\begin{aligned}
F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(1,0) &= \Pr(0,0|\theta_p) + \Pr(1,0|\theta_p) \\
&= \delta_1^{(s)} F_{Y_{pi}|\theta_p}(0) + \delta_0^{(s)} \Pr(Y_{pi}=0|\theta_p)\Pr(Y_{pj}=0|\theta_p) \\
&\quad + \delta_1^{(s)}\big(F_{Y_{pj}|\theta_p}(0) - F_{Y_{pi}|\theta_p}(0)\big) + \delta_0^{(s)} \Pr(Y_{pi}=1|\theta_p)\Pr(Y_{pj}=0|\theta_p) \\
&= \delta_1^{(s)} F_{Y_{pj}|\theta_p}(0) + \delta_0^{(s)} \Pr(Y_{pj}=0|\theta_p) \\
&= F_{Y_{pj}^{(s)}|\theta_p}(0), \\
F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(0,1) &= \Pr(0,0|\theta_p) + \Pr(0,1|\theta_p) \\
&= \delta_1^{(s)} F_{Y_{pi}|\theta_p}(0) + \delta_0^{(s)} \Pr(Y_{pi}=0|\theta_p)\Pr(Y_{pj}=0|\theta_p) \\
&\quad + \delta_1^{(s)} 0 + \delta_0^{(s)} \Pr(Y_{pi}=0|\theta_p)\Pr(Y_{pj}=1|\theta_p) \\
&= \delta_1^{(s)} F_{Y_{pi}|\theta_p}(0) + \delta_0^{(s)} \Pr(Y_{pi}=0|\theta_p) \\
&= F_{Y_{pi}^{(s)}|\theta_p}(0)
\end{aligned}$$

with $F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(y_{pi}, y_{pj})$ being the boundary mixture distribution as defined in Equation (8).

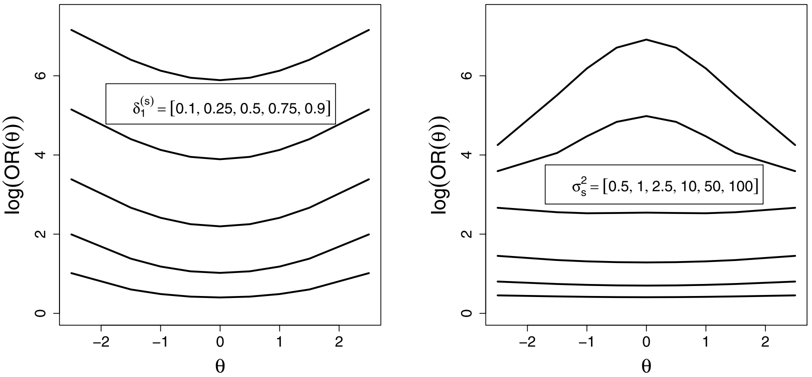
For the testlet model, the original conditional margins are not preserved but are instead put within a different reference frame/scale as implied by the integral over the testlet latent trait ζ_{ps}

$$\begin{aligned}
F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(1,0) &= \int_{\zeta_{ps}} \Pr\big(\boldsymbol{Y}_p^{(s)} = 0,0|\theta_p,\zeta_{ps}\big) h\big(\zeta_{ps};\sigma_s^2\big)\, d\zeta_{ps} \\
&\quad + \int_{\zeta_{ps}} \Pr\big(\boldsymbol{Y}_p^{(s)} = 1,0|\theta_p,\zeta_{ps}\big) h\big(\zeta_{ps};\sigma_s^2\big)\, d\zeta_{ps} \\
&= \int_{\zeta_{ps}} F_{Y_{pj}^{(s)}|\theta_p,\zeta_{ps}}(0) h\big(\zeta_{ps};\sigma_s^2\big)\, d\zeta_{ps}, \\
F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(0,1) &= \int_{\zeta_{ps}} \Pr\big(\boldsymbol{Y}_p^{(s)} = 0,0|\theta_p,\zeta_{ps}\big) h\big(\zeta_{ps};\sigma_s^2\big)\, d\zeta_{ps} \\
&\quad + \int_{\zeta_{ps}} \Pr\big(\boldsymbol{Y}_p^{(s)} = 0,1|\theta_p,\zeta_{ps}\big) h\big(\zeta_{ps};\sigma_s^2\big)\, d\zeta_{ps} \\
&= \int_{\zeta_{ps}} F_{Y_{pi}^{(s)}|\theta_p,\zeta_{ps}}(0) h\big(\zeta_{ps};\sigma_s^2\big)\, d\zeta_{ps}.
\end{aligned}$$

To further illustrate the boundary-mixture implied local dependence structure from a margin-dependent viewpoint, the conditional log odds ratios $\log(OR(\theta_p))$ for two locally dependent items were computed under both a boundary mixture model and a testlet model,

$$\log\big(OR(\theta_p)\big) = \frac{\Pr(Y_{pi}=0, Y_{pj}=0|\theta_p)\Pr(Y_{pi}=1, Y_{pj}=1|\theta_p)}{\Pr(Y_{pi}=0, Y_{pj}=1|\theta_p)\Pr(Y_{pi}=1, Y_{pj}=0|\theta_p)},$$

ranging over θ_p values between -2.5 and 2.5, with the item response functions taken to be one-parameter logistic functions with difficulty parameter equal to 0, such that the local dependence was only a function of the latent trait ($F_{Y_{pi}|\theta_p}(0) = F_{Y_{pj}|\theta_p}(0) = [1 + \exp(\theta_p)]^{-1}$).



(a) Boundary-mixture C (b) Testlet RE

FIGURE 1.
Conditional odds ratios for 2 neutral locally dependent items ($\alpha_i = \alpha_j = 1; \beta_i = \beta_j = 0$) under the boundary-mixture model (C) and the testlet random effect model (RE). Increasing parameter values of both δ_1^s and σ_s^2 correspond to higher log odds ratio lines.

For the boundary mixture model, the dependence threshold parameter $\delta_1^{(s)}$ varied between 0.10 and 0.90. Each line in the left panel of Figure 1 represents the conditional log odds ratio function under a boundary mixture model with fixed value of the dependence parameter $\delta_1^{(s)}$. Notice the typical flat-U shape of the function and that the log odds ratio increases with increasing $\delta_1^{(s)}$ illustrating its function as a dependence parameter. The right panel of Figure 1 contains a similar graphic for the testlet model. To allow this model to span a similar range of dependence given the neutral item characteristics, the variance σ_s^2 of the testlet-specific latent trait needed to vary from 0.5 to 100. The log odds ratio remained relatively constant over the latent trait of focus θ_p, except for large values of σ^2. Note that to compute $\log(OR(\theta_p))$ the testlet specific latent trait ζ_p needed to be integrated out using Gauss–Hermite quadrature. This means that the interpretation of the odds ratio in principle is only applicable for an individual p with an average value on this testlet specific latent trait ζ_p, whereas in the boundary-mixture case the model-implied conditional odds ratio applies to the whole population. This is similar to the distinction between fixed and random effects in multilevel modeling. The fact that the testlet model appears to result in a near-constant conditional odds ratio might appear as a nice characteristic of the model, yet matters are a bit more complex. Notice that for similar conditional margins, the testlet model has to rely on extreme testlet variances σ_s^2 compared to the variance of θ_p (which is fixed at 1) to account for conditional odds ratios that are as high as those for a similar boundary mixture model. In practice the testlet model will account for a local dependence problem by also making changes to the marginal conditional probabilities through the item parameters, whereas the boundary mixture accounts for the local dependence by merely changing the joint conditional probability by means of the $\delta_1^{(s)}$ parameter (see further illustrations in the application section). Thus, the difference in model building strategy will also surface here!

4.4. Information

Let $se(\hat{\theta}_p; \Pi)$ be the standard error of the estimated value $\hat{\theta}_p$ of the latent trait for person p under the conditional independence model Π (i.e., ignoring LID). Let $se(\hat{\theta}_p; C)$ be the standard error of the estimated value $\hat{\theta}_p$ of the latent trait for person p under the boundary mixture model (denoted by C). To assess the misspecification error of ignoring the LID in terms of estimation precision, these two quantities (or the width of related confidence intervals based upon these standard errors) can be compared. Large differences indicate more severe impact of ignoring the LID issues. Let $se(\hat{\theta}_p; \Pi|C)$ be the standard error of the estimated value $\hat{\theta}_p$ of the latent trait for person p under a conditional independence model with item parameters fixed at the values obtained under the boundary mixture model (denoted $\Pi|C$). A comparison of the two quantities $se(\hat{\theta}_p; C)$ and $se(\hat{\theta}_p; \Pi|C)$ can be seen as comparing the test with the LID subsets with an equivalent test not suffering from LID issues, hence it is a comparison to the ideal case scenario. Large differences indicate more severe impact of the LID issues on the efficiency of the test. These comparisons offer a proper framework to assess the consequences of LID in terms of test precision/information and might prove useful to support an informed decision on the degree of LID violation and for general test construction purposes. Note that the diagnostic value of these quantities is conditional upon the adequacy of the defined boundary mixture model C for dealing with the most severe LID issues in the test. Further studies need to evaluate the usefulness of such diagnostics in practice.

5. Application

As an illustration of the boundary mixture approach to LID modeling, a dataset from a small reading test, previously analyzed in Tuerlinckx and De Boeck (2001) and Braeken et al. (2007), is re-examined. The data are binary coded responses from a group of high school students interested in studying law in college ($P = 441$) on items ($I = 6$) referring to a text on the president and the separation of powers in the United States of America.

5.1. *LID Screening*

As a starting point, a one-subset boundary mixture 2PL model is fitted repeatedly with each possible item pair functioning in turn as the potential LID affected subset. Because the boundary mixture model is permutation symmetric, results for subset $J_s = \{i, j\}$ are equivalent to results for subset $J_s = \{j, i\}$, the results for each of these $I * (I - 1)/2$ models with respect to the $\delta_1^{(s)}$ redundancy parameter can be summarized in the upper-triangle of an I-by-I matrix

$$\log(\boldsymbol{likelihood}) \setminus \boldsymbol{\delta}_1 \begin{pmatrix} . & 0.000 & 0.000 & 0.000 & 0.000 & 0.453^{***} \\ 1555 & . & 0.000 & 0.267^{*} & 0.494^{***} & 0.000 \\ 1555 & 1555 & . & 0.230^{*} & 0.213^{*} & 0.000 \\ 1555 & 1553 & 1553 & . & 0.267^{***} & 0.000 \\ 1555 & 1549 & 1553 & 1549 & . & 0.000 \\ 1536 & 1555 & 1555 & 1555 & 1555 & . \end{pmatrix}$$

***: $p < .0001$ and *: $p < .05$,

supported by a lower-triangle containing the loglikelihood of the corresponding one-subset boundary mixture models. This matrix provides helpful information to support a specification search. Patterns of highly redundant item pairs can direct the researcher in choosing a more complete model. Alternatively a variety of diagnostics are also available to support this type of specification search (for an overview see, e.g., Tate, 2003). In either case, when looking at LID diagnostics, one has to be aware of the problem of safeguarding the type-I error rate in this type of multiple testing situation. With respect to the improvement in model likelihood the item subsets $\{1, 6\}$, $\{2, 5\}$, and $\{4, 5\}$ stand out, with the first two sets showing the largest redundancy parameter $\delta_1^{(s)}$-values.

TABLE 1.
Model fit results from a range of boundary mixture models.

	Π	$\{1,6\}$	$\{2,5\}$	$\{4,5\}$	$\{1,6\}$ $\{2,5\}$	$\{1,6\}$ $\{4,5\}$	$\{1,6\}$ $\{2,4,5\}$
α_1	2.133	1.114	2.373	2.882	1.152	1.114	1.240
α_2	0.498	0.666	0.357	0.419	0.487	0.666	0.465
α_3	0.963	1.090	0.928	0.851	1.126	1.090	1.248
α_4	1.640	2.173	1.533	1.121	2.354	2.172	1.723
α_5	1.542	2.130	1.331	1.208	1.863	2.131	1.465
α_6	1.610	0.779	1.758	1.875	0.816	0.779	0.909
β_1	−0.215	−0.308	−0.209	−0.195	−0.303	−0.308	−0.292
β_2	1.581	1.229	2.156	1.851	1.615	1.230	1.690
β_3	−0.011	−0.008	−0.013	−0.014	−0.008	−0.008	−0.008
β_4	−0.937	−0.826	−0.973	−1.160	−0.803	−0.826	−0.918
β_5	−1.055	−0.908	−1.149	−1.208	−0.960	−0.908	−1.061
β_6	−0.017	−0.047	−0.018	−0.016	−0.045	−0.047	−0.039
$J_1 : \delta_1^{(1)}$		0.453	0.494	0.267	0.446	0.453	0.425
$J_2 : \delta_1^{(2)}$					0.413	0.001	0.213
$\log(l)$	1554.9	1535.7	1548.7	1549.2	1532.4	1535.7	1534.0
AIC	3133.8	3097.4	3123.4	3124.4	3092.8	3099.4	3096.0

TABLE 2.
Mean precision of θ_p empirical Bayes estimates for the locally dependent test under the regular conditional independence model (Π) and the boundary mixture alternative (C), and for an equivalent test for which conditional independence does hold ($\Pi|C$).

	Π			$C : J_1 = \{1,6\}, J_2 = \{2,5\}$			$\Pi\|C$		
S_p	θ_p	95% interval	width	θ_p	95% interval	width	θ_p	95% interval	width
0	−1.606	[−2.842, −0.369]	2.473	−1.572	[−2.785, −0.360]	2.425	−1.632	[−2.845, −0.420]	2.425
1	−1.056	[−2.155, 0.044]	2.199	−1.093	[−2.191, 0.005]	2.196	−1.142	[−2.233, −0.051]	2.183
2	−0.649	[−1.695, 0.396]	2.091	−0.726	[−1.814, 0.361]	2.175	−0.751	[−1.814, 0.311]	2.125
3	−0.248	[−1.292, 0.795]	2.087	−0.370	[−1.499, 0.759]	2.258	−0.355	[−1.452, 0.742]	2.195
4	0.128	[−0.966, 1.222]	2.188	0.015	[−1.200, 1.230]	2.430	0.058	[−1.120, 1.237]	2.357
5	0.580	[−0.622, 1.782]	2.404	0.446	[−0.887, 1.779]	2.666	0.520	[−0.779, 1.819]	2.598
6	1.173	[−0.214, 2.559]	2.773	1.067	[−0.441, 2.575]	3.016	1.148	[−0.317, 2.612]	2.929

5.2. *LID Modeling*

Guided by this matrix, a series of models of interest was defined, of which the model fit results are shown in Table 1. The model that takes into account the local dependence between the item pairs $\{1,6\}$ and $\{2,5\}$ gives rise to the largest increase in model fit. Also notice that when the LID in item pair $\{1,6\}$ was already accounted for, a boundary mixture for the item pair $\{4,5\}$ no longer resulted in an improved model fit. Extending the subset J_2 to three items $\{2,4,5\}$ also did not lead to a superior model fit. Thus, taking into account the LID in the two most seriously affected item pairs, seems to lead to a sufficient handling of the LID issue in the data; and, hence, the model of choice is the boundary mixture 2PL model with $J_0 = \{3,4\}$, $J_1 = \{1,6\}$, and $J_2 = \{2,5\}$ (further referred to as model C). This conclusion is supported by a likelihood ratio test between the standard conditional independence model Π and this model C, $LR = 45 \sim \chi^2_{df=2}$, $p < .0001$, and the comparison of AIC values across models. The redundancy parameters $\delta_1^{(s)}$ for the two subsets are equal to about 0.4.

TABLE 3.
Item parameter estimates under conditional independence after item elimination compared to parameter estimates of the testlet model (RE) and the boundary mixture model (C).

	RE	C	Π without $i=6,2$	Π without $i=1,5$	Π without $i=1,2$	Π without $i=5,6$
α_1	2.66	1.15	1.16	.	.	1.13
α_2	0.60	0.49	.	0.50	.	0.57
α_3	1.11	1.13	1.12	1.11	1.16	1.07
α_4	2.32	2.35	2.23	2.37	2.37	2.41
α_5	3.52	1.86	2.07	.	1.89	.
α_6	1.12	0.82	.	0.89	0.85	.
β_1	−0.81	−0.30	−0.31	.	.	−0.32
β_2	0.84	1.61	.	1.57	.	1.41
β_3	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
β_4	−1.87	−0.80	−0.82	−0.81	−0.80	−0.81
β_5	−3.36	−0.96	−0.92	.	−0.96	.
β_6	−0.04	−0.05	.	−0.04	−0.04	.
$J_1:\delta_1^{(1)}$	.	0.45	.	.	.	.
$J_2:\delta_1^{(2)}$	.	0.41	.	.	.	.
$J_1:\sigma_1$	1.30	.	.	.	.	.
$J_2:\sigma_2$	0.78	.	.	.	.	.
$\log(l)$	1533.8	1532.4	.	.	.	.

5.3. *Precision and Interpretation*

The following results illustrate the precision artifact in estimating θ_p in the regular model (Π) that ignores the LID, and the correction by means of the boundary mixture model of choice. Table 2 contains for each of the seven possible sum scores $S_p = \sum_{i=1}^{I} Y_{pi}$, the mean width of the 95% confidence intervals around the resulting empirical Bayes estimates of θ_p for the students in the sample with that score. Taking into account the LID by means of the boundary mixture model, the width of the confidence intervals get upward corrected up to 10%, although for response patterns resulting in a sum score below 2 (i.e., $S_p \leq 1$) confidence intervals are roughly equal. This is as expected, as sum scores above 1 correspond to areas on the latent trait scale where the test has more coverage, and hence the inflated precision artifact will be more pronounced; in contrast differences in precision will fade away when reaching areas of the scale where coverage is limited. Compared to an equivalent test for which conditional independence would hold (i.e., $\Pi|C$ in Table 2), the width of the confidence intervals are about 3% larger (differences again fade out in latent trait scale areas with no coverage). So in sum, in terms of impact, ignoring the LID issue in this small test would result in it being artificially 10% more precise than is warranted, and in terms of efficiency, the LID test is 3% less precise than an equivalent LSI test.

To try to increase the efficiency of the test (see e.g., Chapter 6 in Lord, 1980), one could decide to redesign the test by eliminating items, such that only one item of each of the severe LID item pairs is left. For instance, one could opt to eliminate items 2 and 6, and refit the item response model under conditional independence. The resulting item parameter estimates are displayed in Table 3. Notice how the item parameters of the boundary mixture model closely approximate the item parameters of the conditional independence models in which the major LID issues are removed. These compatibility features of the model offer promise for applications in which different test forms are included and for general test construction and evaluation purposes.

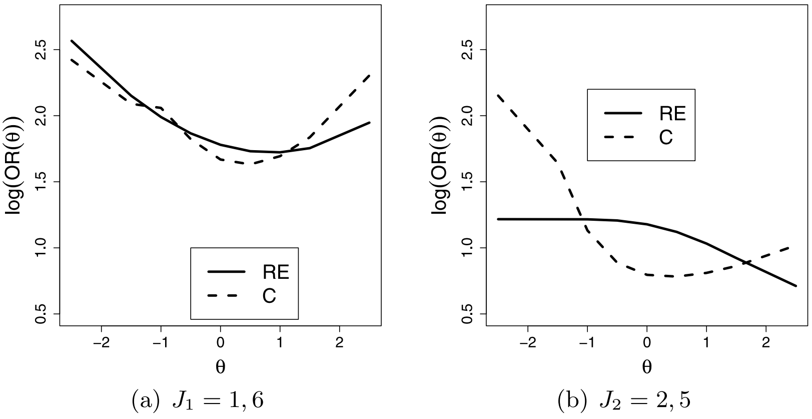


FIGURE 2.
Conditional odds ratios for the 2 locally dependent subsets J_1 and J_2 under the boundary-mixture model (C) and the testlet random effect model (RE).

Item elimination appears to be a quite straightforward strategy to counter LID, but unfortunately (among other things) it decreases the differentiation power of the test; and, in practice, eliminating items is not always an option because of reasons of face and construct validity, or external obligations. Hence, modeling the LID by means of this boundary mixture approach can offer a good alternative. A more accurate picture of the item characteristics and the test precision is obtained than if LID issues were to be simply ignored, and at the same time the test composition can be left intact.

The results displayed in Table 3 also allow the illustration of the difference in construction method between the boundary-mixture model and the testlet model. Whereas the item parameters under the boundary-mixture model are consistent with the item parameters for the models in which the local item dependence issue was solved by item elimination, this is not the case for the testlet model. This holds both for subset and non-subset items. For instance, the item difficulty β_5 (i.e., part of LID subset J_2) for the testlet model (RE) is three times as large as under the boundary mixture or item-eliminated models, and the item difficulty β_4 (i.e., part of the conditionally independent items in J_0) is two times as large. It is obvious that if one does not take into account the differences in reference frame (i.e., compound link function) for the items, one is comparing apples to oranges. For further illustration of the difference in construction method between the boundary-mixture model and the testlet model, Figure 2 presents similar graphics of conditional log odds ratios as presented earlier in Figure 1, but now based upon the estimated item parameters under both models for the two locally dependent subsets. The boundary mixture model is still characterized by a U-shaped conditional log odds ratio function, in contrast to the testlet model for which the shape depends on the particular subset and corresponding item parameters (see Figure 2). Hence, in practice the testlet model accounts for a local dependence problem by also making changes to the marginal conditional probabilities through the item parameters, whereas the boundary mixture accounts for the local dependence by merely changing the joint conditional probability by means of the $\delta_1^{(s)}$ parameter.

5.4. *Model Evaluation*

To further assess the model approach and the performance of the chosen model, compared to the regular conditional independence alternative, a parametric bootstrap was set up. For both

TABLE 4.
Recovery Monte Carlo simulation for the chosen boundary mixture model.

(a) Data condition: Conditional Independence

	True	MCest	MLSE	MCSE	RMSE
α_1	2.13	2.13	0.38	0.32	0.40
α_2	0.50	0.50	0.07	0.11	0.14
α_3	0.96	0.99	0.17	0.14	0.18
α_4	1.64	1.75	0.19	0.27	0.38
α_5	1.54	1.58	0.17	0.20	0.24
α_6	1.61	1.55	0.27	0.23	0.28
β_1	−0.21	−0.22	0.07	0.06	0.08
β_2	1.58	1.73	0.21	0.44	0.61
β_3	−0.01	−0.01	0.11	0.09	0.11
β_4	−0.94	−0.93	0.09	0.11	0.13
β_5	−1.06	−1.05	0.09	0.10	0.13
β_6	−0.02	−0.01	0.08	0.07	0.10
$J_1 : \delta_1^{(1)}$	0.00	0.04	0.00	0.05	0.08
$J_2 : \delta_1^{(2)}$	0.00	0.07	0.00	0.07	0.12

(b) Data condition: Conditional Dependence

	True	MCest	MLSE	MCSE	RMSE
α_1	1.15	1.18	0.18	0.17	0.21
α_2	0.49	0.51	0.07	0.14	0.17
α_3	1.13	1.12	0.21	0.18	0.22
α_4	2.35	2.62	0.38	0.69	1.69
α_5	1.86	1.94	0.24	0.33	0.44
α_6	0.82	0.81	0.14	0.12	0.16
β_1	−0.30	−0.29	0.09	0.09	0.11
β_2	1.61	1.78	0.21	0.55	0.77
β_3	−0.01	−0.03	0.10	0.08	0.11
β_4	−0.80	−0.80	0.07	0.09	0.11
β_5	−0.96	−0.96	0.08	0.12	0.14
β_6	−0.04	−0.05	0.11	0.11	0.14
$J_1 : \delta_1^{(1)}$	0.45	0.46	0.06	0.05	0.07
$J_2 : \delta_1^{(2)}$	0.41	0.42	0.13	0.13	0.15

True parameter value from Table 1; 100 replications in each data condition.
MCest: Monte Carlo parameter estimate.
MLSE: Original Maximum Likelihood standard error.
MCSE: Monte Carlo standard error.
RMSE: Monte Carlo root mean squared error.

the conditional independence model (Π) and the chosen conditional dependence model C, the estimated maximum likelihood (ML) parameters were used to generate 100 replicated datasets. Table 4 shows the results of a small Monte Carlo recovery study when refitting the conditional dependence model on both types of data. The recovery of the true parameter values is quite accurate and the Monte Carlo standard error (MCSE) is quite close to the ML standard error (cf. Table 1), and this with only a small number of replications. Notice that in the conditional independence condition, the subset redundancy parameter δ_1 indeed reduces to near-zero values, whereas in the conditional dependence condition, the non-zero redundancy is picked up.

To be able to compare models on observable data properties, the unconditional pairwise item odds ratios were considered. The odds ratio between item i and item j is defined as

$$OR_{ij} = \frac{n_{11}n_{00}}{n_{10}n_{01}},$$

where n_{10} is the frequency of occurrence of the response vector $(Y_{pi} = 1, Y_{pj} = 0)$, and n_{00}, n_{11}, and n_{01} are defined in similar fashion. The item-by-item predicted odds ratio matrix under the conditional independence model ($\boldsymbol{OR}_{\mathit{\Pi}}$) shows large deficiencies for the pairs $\{1, 6\}$, $\{4, 5\}$, and $\{2, 5\}$ compared to the observed odds ratios ($\boldsymbol{OR}_{\text{obs}}$), whereas the predictions under the conditional dependence model ($\boldsymbol{OR}_{C}$) are much more accurate (Mean squared error, $\text{MSE}_{\mathit{\Pi}} = 2.44$ vs. $\text{MSE}_{C} = 0.11$).

$$\boldsymbol{OR}_{\text{obs}} = \left(\begin{array}{ccccccc} {}_{j}\backslash^{i} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & . & 1.60 & 2.39 & 3.70 & 3.65 & 8.24 \\ 2 & . & . & 1.37 & 2.56 & 4.00 & 1.22 \\ 3 & . & . & . & 3.69 & 3.54 & 2.18 \\ 4 & . & . & . & . & 7.12 & 3.03 \\ 5 & . & . & . & . & . & 2.43 \\ 6 & . & . & . & . & . & . \end{array}\right)$$

$$\boldsymbol{OR}_{\mathit{\Pi}} = \left(\begin{array}{ccccccc} {}_{j}\backslash^{i} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & . & 1.87 & 2.99 & 5.25 & 4.88 & 4.78 \\ 2 & . & . & 1.52 & 1.87 & 1.86 & 1.74 \\ 3 & . & . & . & 2.79 & 2.67 & 2.54 \\ 4 & . & . & . & . & 4.05 & 4.24 \\ 5 & . & . & . & . & . & 4.01 \\ 6 & . & . & . & . & . & . \end{array}\right),$$

$$\boldsymbol{OR}_{C} = \left(\begin{array}{ccccccc} {}_{j}\backslash^{i} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & . & 1.54 & 2.32 & 3.93 & 3.42 & 8.84 \\ 2 & . & . & 1.57 & 2.05 & 4.29 & 1.43 \\ 3 & . & . & . & 3.76 & 3.37 & 1.93 \\ 4 & . & . & . & . & 6.41 & 2.77 \\ 5 & . & . & . & . & . & 2.56 \\ 6 & . & . & . & . & . & . \end{array}\right)$$

Notice that the conditional dependence model only explicitly accounts for the subsets $\{1, 6\}$ and $\{2, 5\}$, yet improves upon all three of the largest deficiencies under the conditional independence model, including the odds ratio for the item pair $\{4, 5\}$.

6. Discussion

The above illustrated level difference between the unconditional observed data part and the conditional latent aspect of item response models is the key problem with which LID detection methods are confronted, and is similar in nature to the problem of specification searches in structural equation models (SEM; for an interesting discussion see Steiger, 1990; Salhi, 1998; MacCallum, 1986).

Hence, LID modeling strategies necessarily involve issues such as LID screening (see e.g., Chen & Thissen, 1997), subset definition and selection, multiple hypothesis testing (see e.g., Ip, 2001; Shaffer, 1995), and model comparison. As a reviewer rightly pointed out, each of these components constitutes a subject of research on its own. Hence, in practice it is recommended to take a holistic approach and support substantive motivations (see e.g., Yen, 1993) with a combination of various LID detection methods and explicit LID modeling. The proposed modeling approach should be viewed in this perspective as one potential tool or element in the larger context of available methods and approaches.

The approach has some attractive characteristics. The technical feature that the formulas for the item characteristic curves of single items do not need to be changed with the proposed model makes for straightforward communication. As mentioned earlier, the compatibility features of the model together with the interpretational information framework also offer promise for applications in which different testlet and non-testlet test forms are included and for general test construction and evaluation purposes. Furthermore, the conceptual idea behind the method, balancing two extreme situations (i.e., independence and complete dependence), is also rather intuitive and appealing.

Further research might want to investigate the limitations of disjoint subsets and non-exchangeable within-subset dependence. An approach for such overlapping subsets will need to establish which conditions and inequalities are needed to result in a proper multivariate distribution. These complexities might be costly in terms of model interpretation and clarity.

The fact that the boundary mixture approach readily fits within the copula class illustrates the generality of the latter class of models and suggests a thorough comparison of available LID approaches in search of commonalities and differences on the latent and observed data level, and a way of selecting between the different alternatives. Perhaps an equivalent marginal reformulation of the testlet model is even within reach. The fact that different types of copula functions are available makes it possible to investigate different LID dependence types in a similar way as the random effect in a testlet model does not necessarily have to be normally distributed. This issue of potentially different LID dependence types and the robustness of LID models to misspecification is for now relatively unexplored territory. Based upon Zeger, Liang, and Albert (1988) a conjecture is made that a marginal modeling approach such as the copula functions might be more robust to misspecification than a conditional approach such as the testlet model. Of course, in any case modeling the true dependence structure will increase statistical efficiency of parameter estimates. As such, finding a comprehensive and reliable way to define the dependence structure of a test (i.e., dimensionality and LID assessment) should remain a key research area within psychometrics.

Appendix: Recursive Formula to Compute Probabilities Based upon Cumulative Probabilities

For a subset $J_s = \{i, j, k\}$ with cardinality $I_s = 3$, the joint conditional probabilities can be computed using principles similar to the quadrant rules. The following general formulation is useful for this purpose

$$\begin{aligned}&\Pr\bigl(\boldsymbol{Y}_p^{(s)} = \boldsymbol{y}_p^{(s)} | \theta_p, \boldsymbol{\delta}^{(s)}\bigr)\\&\quad = \sum_{m_1=0}^{1} \dots \sum_{m_{I_s}=0}^{1} (-1)^{m_1+\dots+m_{I_s}} F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(y_{p1} - m_1, \dots, y_{pI_s} - m_{I_s}),\end{aligned} \tag{10}$$

where the arguments $y_{pi} - m_i$ of the conditional cumulative item probabilities stem from the definition of the distribution functions $F_{Y_{pi}|\theta_p}(y_{pi})$

$$F_{Y_{pi}|\theta_p}(y_{pi}) = \begin{cases} 0 & \text{for } y_{pi} < 0, \\ \Pr(Y_{pi} = 0|\theta_p) & \text{for } y_{pi} = 0, \\ 1 & \text{for } y_{pi} = 1. \end{cases}$$

Note that similar algorithms exist in multivariate probit analysis (see e.g., Ashford & Sowden, 1970)

For notational clarity let $F_{\boldsymbol{Y}_p^{(s)}|\theta_p}(y_{pi}, y_{pj}, y_{pk})$, the conditional joint cumulative probability of the subset response vector J_s, be written in the shorthand $F(y_{pi}, y_{pj}, y_{pk})$; then the resulting computations for the joint probabilities of each trivariate subset response pattern are given by the following set of expressions

$$\begin{aligned}
\Pr(0,0,0|\theta_p) &= F(0,0,0) - F(0,0,-1) - F(0,-1,0) + F(0,-1,-1)\\
&\quad - F(-1,0,0) + F(-1,0,-1) + F(-1,-1,0) - F(-1,-1,-1)\\
&= F(0,0,0),\\
\Pr(0,0,1|\theta_p) &= F(0,0,1) - F(0,0,0) - F(0,-1,1) + F(0,-1,0)\\
&\quad - F(-1,0,1) + F(-1,0,0) + F(-1,-1,1) - F(-1,-1,0)\\
&= F(0,0,1) - F(0,0,0),\\
\Pr(0,1,0|\theta_p) &= F(0,1,0) - F(0,1,-1) - F(0,0,0) + F(0,0,-1)\\
&\quad - F(-1,1,0) + F(-1,1,-1) + F(-1,0,0) - F(-1,0,-1)\\
&= F(0,1,0) - F(0,0,0),\\
\Pr(1,0,0|\theta_p) &= F(1,0,0) - F(1,0,-1) - F(1,-1,0) + F(1,-1,-1)\\
&\quad - F(0,0,0) + F(0,0,-1) + F(0,-1,0) - F(0,-1,-1)\\
&= F(1,0,0) - F(0,0,0),\\
\Pr(0,1,1|\theta_p) &= F(0,1,1) - F(0,1,0) - F(0,0,1) + F(0,0,0)\\
&\quad - F(-1,1,1) + F(-1,1,0) + F(-1,0,1) - F(-1,0,0)\\
&= F(0,1,1) - F(0,1,0) - F(0,0,1) + F(0,0,0),\\
\Pr(1,0,1|\theta_p) &= F(1,0,1) - F(1,0,0) - F(1,-1,1) + F(1,-1,0)\\
&\quad - F(0,0,1) + F(0,0,0) + F(0,-1,1) - F(0,-1,0)\\
&= F(1,0,1) - F(1,0,0) - F(0,0,1) + F(0,0,0),\\
\Pr(1,1,0|\theta_p) &= F(1,1,0) - F(1,1,-1) - F(1,0,0) + F(1,0,-1)\\
&\quad - F(0,1,0) + F(0,1,-1) + F(0,0,0) - F(0,0,-1)\\
&= F(1,1,0) - F(1,0,0) - F(0,1,0) + F(0,0,0),\\
\Pr(1,1,1|\theta_p) &= F(1,1,1) - F(1,1,0) - F(1,0,1) + F(1,0,0)\\
&\quad - F(0,1,1) + F(0,1,0) + F(0,0,1) - F(0,0,0).
\end{aligned}$$

References

Ashford, J.R., & Sowden, R.R. (1970). Multivariate probit analysis. *Biometrics*, *26*, 535–546.

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. 397–497). Reading: Addison-Wesley.

Braeken, J., & Tuerlinckx, F. (2009). A mixed model framework for teratology studies. *Biostatistics*, *10*, 744–755.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copulas for residual dependency. *Psychometrika*, *72*, 393–411.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Cureton, E.E. (1959). Note on $\phi/\phi_{\max}$. *Psychometrika*, *24*, 89–91.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement*, *36*, 119–140.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université Lyon: Série 3*, *14*, 53–77.

Gibbons, R.D., & Hedeker, D.R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Hoeffding, W. (1940). Masstabinvariante Korrelations Theorie. *Schriften des Matematischen Instituts und des Instituts für angewandte Mathematik der Universität Berlin*, *5*, 179–223. [Reprinted as Scale-invariant correlation theory in the Collected Works of Wassily Hoeffding, N.I. Fischer, and P.K. Sen (Eds.), New York: Springer.]

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

Ip, E. (2001). Testing for local dependence in dichotomous and polutomous item response models. *Psychometrika*, *66*, 109–132.

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.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Claussen (Eds.), *Measurement and prediction* (pp. 7–56). Princeton University Press: Thousand Oaks.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah: Erlbaum.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107–120.

Masters, G.N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement*, *25*, 15–29.

Mood, A.M., Graybill, F.A., & Boes, D.C. (1974). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Nelsen, R.B. (1998). *An introduction to copulas*. New York: Springer.

Salhi, S. (1998). Heuristic search methods. In G.A. Marcoulides (Ed.), *Modern methods for business research* (pp. 147–175). Mahwah: Lawrence Erlbaum.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 7*.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement, 18*.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584.

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

Sklar, A. (1959). Fonctions de répartition à *n* dimension et leurs marges. *Publications Statistiques Université de Paris*, *8*, 229–231.

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

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

Tuerlinckx, F., & De Boeck, P. (2001). Non-modeled item interactions lead to distorted discrimination parameters: A case study. *Methods of Psychological Research, 6*. [Retrieved May 20, 2005 from http://www.mpr-online.de/issue14/art3/Tuerlinckx.pdf.

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

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

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Zeger, S.L., Liang, K.-Y., & Albert, P.S. (1988). Models for longitudinal data: A generalized estimation equation approach. *Biometrics*, *44*, 1049–1060.

Manuscript Received: 18 AUG 2009
Final Version Received: 18 AUG 2010
Published Online Date: 12 NOV 2010