Article

Pairwise Likelihood Estimation of the 2PL Model with Locally Dependent Item Responses

Alexander Robitzsch [1,2]

1 IPN–Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de
2 Centre for International Student Assessment (ZIB), 24118 Kiel, Germany

Abstract: The local independence assumption is crucial for the consistent estimation of item parameters in item response theory models. This article explores a pairwise likelihood estimation approach for the two-parameter logistic (2PL) model that treats the local dependence structure as a nuisance in the optimization function. Hence, item parameters can be consistently estimated without explicit modeling assumptions of the dependence structure. Two simulation studies demonstrate that the proposed pairwise likelihood estimation approach allows nearly unbiased and consistent item parameter estimation. Our proposed method performs similarly to the marginal maximum likelihood and pairwise likelihood estimation approaches, which also estimate the parameters for the local dependence structure.

Keywords: item response model; pairwise likelihood estimation; local dependence

1. Introduction

Item response theory (IRT) models [1,2] are central to analyzing multivariate dichotomous random variables. IRT models summarize a high-dimensional contingency table by a low-dimensional latent factor variable of interest. In many cases, a reduction to a unidimensional variable is required. The application of IRT models in educational large-scale assessments [3] like the Programme for International Student Assessment (PISA; [4]) and the testing industry is of particular importance.

Let $X = (X_1, \ldots, X_I)$ be the vector of I dichotomous items $X_i \in \{0, 1\}$. A unidimensional IRT model [5,6] is a statistical model for the probability distribution $\mathrm{P}(X = x)$ for $x \in \{0, 1\}^I$, where

$$\mathrm{P}(X = x; \gamma) = \int_{-\infty}^{\infty} \prod_{i=1}^{I} P_i(x_i, \theta; \gamma_i) f(\theta) \, \mathrm{d}\theta. \tag{1}$$

In Equation (1), a latent variable θ with density function f is involved . The variable (also referred to as a trait or ability) can be interpreted as a unidimensional summary of the test items X. The distribution of θ is frequently fixed to a normal distribution (but see [7,8]). The item response functions (IRF) $P_i(x, \theta; \gamma_i) = \mathrm{P}(X_i = x|\theta)$ model the relationship of the dichotomous item with θ. Let $\gamma = (\gamma_1, \ldots, \gamma_I)$. Note that $P_i(0, \theta; \gamma_i) + P_i(1, \theta; \gamma_i) = 1$.

Importantly, item responses X_i are assumed to be conditionally independent on θ in (1). Therefore, pairs of items X_i and X_j are conditionally uncorrelated after controlling the latent ability θ. More formally, it is assumed that

$$\mathrm{P}(X|\theta) = \prod_{i=1}^{I} \mathrm{P}(X_i|\theta). \tag{2}$$

The property (2) is also known as the local independence assumption that can be statistically tested [6,9].

The vector of item parameters γ of the IRFs in Equation (1) can be estimated by marginal maximum likelihood (MML) using an expectation–maximization (EM) algorithm [10,11]. A variety of statistical software packages exist to estimate IRT models [12].

This article focuses on the estimation of IRT models in the violation of the local independence condition in (2). This situation refers to locally dependent item responses. For example, in an educational test, students respond to items referring to several reading texts. However, items that refer to the same reading text are likely more dependent on each other than items from different reading texts. Hence, the local dependence of item responses of items within the same item stimulus (i.e., the reading text) can be expected. In this article, we confine ourselves to estimating the two-parameter logistic (2PL) IRT model [13]. The IRF for item X_i in the 2PL model is defined as

$$P_i(x, \theta; a_i, b_i) = \Psi((2x - 1)(a_i\theta - b_i)) = \frac{\exp(x(a_i\theta - b_i))}{1 + \exp(a_i\theta - b_i)} \text{ for } x = 0, 1, \tag{3}$$

where the vector of item parameters $\gamma_i = (a_i, b_i)$ contains the item discrimination a_i and the item intercept b_i. The logistic distribution function is denoted by Ψ and it holds $1 - \Psi(x) = \Psi(-x)$ for real-valued x.

This article reviews different estimation techniques for unidimensional IRT models under local dependence. It is demonstrated that MML estimation will lead to biased item parameter estimates if local dependence is ignored in the estimation. A variant of pairwise likelihood estimation is presented that treats the local dependence structure as a nuisance structure in the estimation and effectively removes these pieces of information in model estimation. Hence, our proposed modified estimation approach is unaffected by the presence of local dependence. Our proposed method is compared with other estimation methods in which the parameters of local dependence are simultaneously estimated alongside the item parameters.

The rest of the article is structured as follows. In Section 2, we review the pairwise likelihood estimation approach. We describe a particular choice of weights in pairwise likelihood estimation and the contribution of local dependence in model estimation. Section 3 presents the results from two simulation studies in which we compare estimation methods that ignore local dependence with our modified pairwise likelihood method and other methods that handle local dependence. Three empirical datasets with local dependence are analyzed in Section 4. Finally, the paper closes with a discussion in Section 5.

2. Pairwise Maximum Likelihood (PML) and Maximum Marginal Likelihood (MML) Estimation with Local Dependence

2.1. Pairwise Likelihood Estimation (PML)

In this section, we review pairwise (maximum) likelihood (PML) estimation for IRT models [14–17]. For item responses X_i with IRFs $P_i(x, \theta; \gamma_i)$ ($i = 1, \ldots, I$), marginal univariate probabilities are given by

$$L_{1,X_i}(x; \gamma_i) = \mathsf{P}(X_i = x) = \int P_i(x, \theta; \gamma_i)\phi(\theta)\, d\theta, \tag{4}$$

where ϕ denotes the density of the standard normal distribution. PML estimation also relies on the evaluation of bivariate probabilities for an item pair (X_i, X_j)

$$L_{2,X_iX_j}(x, y; \gamma_i, \gamma_j) = \mathsf{P}(X_i = x, X_j = y) = \int P_i(x, \theta; \gamma_i)P_j(y, \theta; \gamma_j)\phi(\theta)\, d\theta. \tag{5}$$

In PML estimation, item parameters $\gamma = (\gamma_1, \ldots, \gamma_I)$ are computed by maximizing the weighted sum of the likelihood contributions of univariate and bivariate frequencies. In more detail, the PML optimization function is given by

$$l(\gamma) = \sum_{i=1}^{I} w_{1,X_i}\left(\sum_{x=0}^{1} n_{X_i,x} \log L_{1,X_i}(x; \gamma_i)\right) + \sum_{i=1}^{I-1}\sum_{j=i}^{I} w_{2,X_iX_j}\left(\sum_{x=0}^{1}\sum_{y=0}^{1} n_{X_iX_j,xy} \log L_{2,X_iX_j}(x, y; \gamma_i, \gamma_j)\right), \tag{6}$$

where $n_{X_i,x}$ is the univariate frequency with which item X_i takes the value x. Moreover, $n_{X_iX_j,xy}$ denotes the bivariate frequency in the sample with which item X_i has the value

$x \in \{0, 1\}$ and item X_j has the value $y \in \{0, 1\}$. The choice of the weights w_{1,X_i} and w_{2,X_iX_j} can be determined by the researcher's objectives (see also [18]). In the simulation study and the empirical example, we focus on PML estimation that involves all item pairs with the weights $w_{1,X_i} = 1/I$ and $w_{2,X_iX_j} = 2I^{-1}(I-1)^{-1}$. By doing so, the univariate and bivariate frequencies equally contribute to the analysis.

2.2. Local Dependence in IRT Models

Assume that there is local dependence in an IRT model, and the items are arranged in testlets that reflect the dependency structure of items in groups of items [19]. In other words, there exist testlets t such that items X_{it} are locally independent between testlets but dependent within testlets. Define $\boldsymbol{X}_t = (X_{1t}, \dots, X_{I_t t})$, where I_t denotes the number of items in testlet t.

The joint distribution of all item responses $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_T)$ fulfills

$$\mathsf{P}(\boldsymbol{X}|\theta) = \mathsf{P}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_T|\theta) = \prod_{t=1}^{T} \mathsf{P}(\boldsymbol{X}_t|\theta). \tag{7}$$

However, local dependence means that the items within a testlet are not conditionally independent (i.e., local dependence exists):

$$\mathsf{P}(\boldsymbol{X}_t|\theta) = \mathsf{P}(X_{1t}, \dots, X_{I_t t}|\theta) \neq \prod_{i=1}^{I_t} \mathsf{P}(X_{it}|\theta). \tag{8}$$

Consequently, we obtain, for an item pair (X_{it}, X_{jt}) of items within the same testlet t,

$$\mathsf{P}(X_{it}, X_{jt}|\theta) \neq \mathsf{P}(X_{it}|\theta)\mathsf{P}(X_{jt}|\theta). \tag{9}$$

This property reflects the residual dependence of the items within a testlet and potentially biases the estimation of IRT models that incorrectly attribute the residual dependence to the ability variable θ. In general, item discriminations are distorted and positively biased in many instances [20]. The neglected local dependence entails inflated reliability estimates [21,22].

However, it should be noted that there is local independence among items X_{it} and X_{ju} within different testlets t and u

$$\mathsf{P}(X_{it}, X_{ju}|\theta) = \mathsf{P}(X_{it}|\theta)\mathsf{P}(X_{ju}|\theta), \tag{10}$$

Local dependence can be detected at the level of item pairs by means of standardized correlations of item residuals. This Q_3 statistic [9] can be interpreted as a measure of the deviation from local independence, and positive residual correlations (e.g., larger than 0.1 or 0.3) can be considered non-negligible deviations from local independence [23].

2.3. Different IRT Modeling Strategies to Handle Local Dependence

The IRT literature mainly deals with the issue of local dependence through three alternative strategies.

First, in testlet IRT models, additional latent variables are included to model the testlet structure [19,24]. Most frequently, the local dependence of all items within a testlet is modeled by an additional latent variable. This unidimensionality assumption might be too restrictive to fit item response data.

Second, the dependency structure is covered by additional parameters. In this approach, a polytomous superitem is defined that decodes the combination of values of items within a testlet in their categories [25]. For example, if a testlet consists of three items, the superitem possesses $2^3 = 8$ categories. The dependence is modeled by additional parameters [26–28]. The second approach has the disadvantage that the item parameters do not have a marginal interpretation. This issue also pertains to the first approach when

using testlet models, although the item parameters can be transformed in order to obtain a marginal interpretation [29].

The third approach starts with a marginal unidimensional IRT model and models the local dependence structure by means of copula models [21,30,31]. The latter approach is most attractive for the purpose of the marginal interpretation of item parameters. However, copula models require the specification of the type of copula distribution of the testlets and estimate testlet residual dependence parameters δ along with item parameters γ.

In this article, we only consider modeling strategies that allow a marginal interpretation of the item response parameters.

2.4. IRT Models with Correlated Item Residuals for Modeling of Local Dependence

We now present an IRT model with correlated item responses that accommodate local dependence. Assume that, marginally, a 2PL model holds for items X_i ($i = 1, \dots, I$). It is assumed that there is an underlying latent item response X_i^* for item X_i such that

$$X_i = \mathbf{1}(X_i^* > 0), \tag{11}$$

where $\mathbf{1}$ denotes the indicator function. Specifically, X_i takes a value of 1 if X_i^* exceeds zero. Let the item residual ε_i follow the logistic distribution. Then, we have

$$\mathsf{P}(\varepsilon_i < y) = \Psi(y) \text{ for all } y \in \mathbb{R}. \tag{12}$$

We define the latent item response X_i^* as

$$X_i^* = a_i \theta - b_i - \varepsilon_i. \tag{13}$$

Then, we obtain, using (11), (13), and (12),

$$\mathsf{P}(X_i = 1) = \mathsf{P}(X_i^* > 0) = \mathsf{P}(\varepsilon_i < a_i\theta - b_i) = \Psi(a_i\theta - b_i). \tag{14}$$

Hence, X_i marginally follows the 2PL model. Local dependence can be modeled by letting the logistically distributed item residuals $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_I)$ be dependent. In principle, any dependence structure with a fixed marginal distribution to the logistic one can be assumed. In fact, this leads to copula models [32] of correlated item residuals [21,33].

2.5. Normal Copula Model for Correlated Item Residuals

We now review the normal copula model for correlated item residuals for the 2PL model [31]. There is an underlying latent item response X_i^* for each item X_i (see (11)). The item residuals in (11) follow the logistic distribution. Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_I)$ be the vector of item residuals. In the normal copula model, it is assumed that the vector $\boldsymbol{e} = (e_1, \dots, e_I)$ of transformed item residuals follows a multivariate normal distribution with a zero mean vector $\mathbf{0}$ and a correlation matrix $\boldsymbol{\Sigma}^*$. Note that all variables e_i have standard deviations of 1. In more detail, each component i ($i = 1, \dots, I$) of $\boldsymbol{e}$ is defined as

$$e_i = \Phi^{-1}(\Psi(\varepsilon_i)), \tag{15}$$

where Φ denotes the distribution function of the standard normal distribution.

This definition allows the computation of item response probabilities for subsets of items within a testlet that accommodates testlet dependence. Assume that $\mathrm{Cor}(e_i, e_j) = \rho_{ij}$ for items X_i and X_j within a testlet. Then, we have

$$
\begin{aligned}
&\mathsf{P}(X_i = 1, X_j = 1|\theta) \\
={}& \mathsf{P}(a_i\theta - b_i - \varepsilon_i > 0, \, a_j\theta - b_j - \varepsilon_j > 0 \,|\, \theta) \\
={}& \mathsf{P}(e_i < \Phi^{-1}(\Psi(a_i\theta - b_i)), e_j < \Phi^{-1}(\Psi(a_j\theta - b_j)) \,|\, \theta) \\
={}& \Phi_2\big(\Phi^{-1}(\Psi(a_i\theta - b_i)), \Phi^{-1}(\Psi(a_j\theta - b_j)); \rho_{ij}\big)
\end{aligned} \tag{16}
$$

where Φ_2 is the distribution function of the bivariate normal distribution depending on the correlation ρ_{ij}. For $x, y \in \{0, 1\}$, the item response probabilities can be evaluated as

$$
\begin{aligned}
&P(X_i = x, X_j = y | \theta) \\
&= \Phi_2\left(\Phi^{-1}(\Psi((2x - 1)(a_i\theta - b_i))), \Phi^{-1}(\Psi((2y - 1)(a_j\theta - b_j))); \rho_{ij}^{x+y}\right).
\end{aligned}
\tag{17}
$$

The extension to more than two items in a testlet follows the same principle [31]. However, the multidimensional normal distribution must be evaluated. There is increasing computational complexity in this calculation with an increasing number of items within a testlet.

2.6. Marginal Maximum Likelihood (MML) Estimation for the Normal Copula Model for Item Residuals

The 2PL model with a normal copula for correlated item residuals (see Section 2.5) can be estimated with MML by modeling the vector X_t of item responses X_{it} within a testlet t ($t = 1, \ldots, T$; see (8)) as

$$
P(X = x; \gamma) = \int \prod_{t=1}^{T} P(X_t | \theta; \gamma_t) f(\theta) \, d\theta.
\tag{18}
$$

The vector γ_t contains all item parameters a_i and b_i of the items within the testlet t and the entries of the correlation matrix of item residuals ε within the testlet.

Now, assume that there are $n = 1, \ldots, N$ independent and identically distributed realizations of X from (18), leading to item responses $x_1, \ldots, x_N \in \{0, 1\}^I$, where $x_n = (x_{n1}, \ldots, x_{nT})$ and $x_{nt} \in \{0, 1\}^{I_t}$. The model parameter $\gamma = (\gamma_1, \ldots, \gamma_T)$ can be obtained by maximizing the negative log-likelihood function

$$
l(\gamma) = \sum_{n=1}^{N} \log[P(X = x_n; \gamma)] = \sum_{n=1}^{N} \log\left[\int \prod_{t=1}^{T} P(X_t = x_{nt} | \theta; \gamma_t) f(\theta) \, d\theta\right],
\tag{19}
$$

Note that the probabilities $P(X_t = x_{nt} | \theta; \gamma_t)$ require the evaluation of the multivariate normal distribution of dimension I_t. The variance matrix of the parameter estimate $\widehat{\gamma}$ can be obtained by taking the inverse of the negative observed information matrix that contains the second-order partial derivatives (i.e., the Hessian matrix)

$$
\left. \frac{\partial^2 l}{\partial \gamma_h \gamma_k} \right|_{\gamma = \widehat{\gamma}},
\tag{20}
$$

where γ_h and γ_k are entries of the parameter vector γ.

Consistent parameter estimates of γ can be expected if the normal distribution assumption of θ and the multivariate normal distribution of item residuals within a testlet (i.e., the normal copula model) are correct.

2.7. Pairwise Maximum Likelihood (PML) Estimation for the Normal Copula Model for Item Residuals

The MML estimation of the 2PL model with a normal copula model for item residuals has the disadvantage that multidimensional normal probabilities must be evaluated. This is particularly computationally demanding for many items within a testlet. Hence, PML can be applied, which only involves bivariate normal probabilities in the estimation of the 2PL normal copula model.

The PML estimation described in Section 2.1 can now also involve correlation parameters ρ_{ij} that capture the dependence of items i and j. For items X_i and X_j allocated to different testlets, local independence is assumed, and no correlation is estimated.

For items X_i and X_j in different testlets, bivariate probabilities in PML estimation are evaluated according to (see (5))

$$L_{2,X_iX_j}(x,y;\gamma_i,\gamma_j) = \int \Psi(2(x-1)(a_i\theta-b_i))\Psi(2(y-1)(a_j\theta-b_j))\phi(\theta)\,\mathrm{d}\theta \tag{21}$$

with $\gamma_i = (a_i, b_i)$. For items X_i and X_j in the same testlet, the bivariate probabilities are given as (see (17))

$$L_{2,X_iX_j}(x,y;\gamma_i,\gamma_j,\rho_{ij}) = \int \Phi_2\Big(\Phi^{-1}(\Psi((2x-1)(a_i\theta-b_i))), \Phi^{-1}(\Psi((2y-1)(a_j\theta-b_j))); \rho_{ij}^{x+y}\Big)\phi(\theta)\,\mathrm{d}\theta, \tag{22}$$

where the term (22) now additionally includes the correlation ρ_{ij}.

The model parameter γ contains all item parameters a_i and b_i, as well as all correlation parameters ρ_{ij} for the item pairs within a testlet. A parameter estimate is obtained by maximizing the PML optimization function (6) with respect to γ. The standard errors of the model parameters can be obtained with Huber–White standard errors by formalizing PML as an M-estimation problem [14,34]. More formally, PML maximizes

$$l(\gamma) = \sum_{n=1}^{N} l_n(\gamma;\boldsymbol{x}_n), \tag{23}$$

where l_n are the PML contributions of case n. The parameter estimate $\widehat{\gamma}$ is obtained by setting the partial derivative of l with respect to all components γ_h of γ equal to 0:

$$\frac{\partial l}{\partial \gamma_h} = \sum_{n=1}^{N} \frac{\partial l_n}{\partial \gamma_h}(\gamma;\boldsymbol{x}_n) = \boldsymbol{0}. \tag{24}$$

The estimated variance matrix $\widehat{V}$ of $\widehat{\gamma}$ is obtained with the sandwich formula [34]

$$\widehat{V} = \widehat{I}^{-1}\widehat{J}\widehat{I}^{-1}, \text{ where} \tag{25}$$

$$\widehat{J} = \left(\sum_{n=1}^{N} \frac{\partial l_n}{\partial \gamma_h}(\widehat{\gamma};\boldsymbol{x}_n)\frac{\partial l_n}{\partial \gamma_k}(\widehat{\gamma};\boldsymbol{x}_n)\right)_{hk} \text{ and} \tag{26}$$

$$\widehat{I} = \left(\sum_{n=1}^{N} \frac{\partial^2 l_n}{\partial \gamma_h \gamma_k}(\widehat{\gamma};\boldsymbol{x}_n)\right)_{hk}. \tag{27}$$

The partial derivatives in (24), (26), and (27) can be obtained by analytical derivations or numerical approximations.

Like MML, PML estimation relies on the normal distribution assumption for θ. From the first view, it seems that the normal copula model assumption for item residuals is also crucial. However, note that a single parameter ρ_{ij} is estimated for all item pairs (X_i, X_j) of items within the same testlet.

2.8. Pairwise Maximum Likelihood (PML) Estimation with Excluded Item Pairs to Handle Local Dependence

It has been pointed out in Section 2.4 that the specification of the dependency structure for item residuals ε is indeterminate. Hence, the estimated item discriminations and item difficulties will depend on the distributional assumption of the item residuals. Hence, consistent parameter estimates of the 2PL normal copula model with MML estimation (Section 2.6) or PML estimation (Section 2.7) cannot be guaranteed if the multivariate vector item residuals follow a different distribution. In this section, we describe a variant of PML estimation that does not require the specification of the dependence structure of item residuals.

As highlighted in (10), the item responses of items from different testlets are conditionally independent. Hence, it can be expected that marginal item response functions

can be consistently estimated if weights w_{2,X_iX_j} in pairwise likelihood estimation in (6) are set to 0 for an item pair with items within the same testlet but set to 1 for items within different testlets.

Let us illustrate the particular PML specification utilizing a toy example. The test contains seven items arranged in three testlets. Testlet 1 contains Items 1, 2, and 3; Testlet 2 contains Items 4 and 5; and Testlet 3 contains Items 6 and 7. The weight matrix $W_2 = (w_{2,X_iX_j})$ for PML estimation required in (6) can be defined by

$$W_2 = (w_{2,X_iX_j}) = \begin{pmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ 1 & 1 & 1 & & & \\ 1 & 1 & 1 & 0 & & \\ 1 & 1 & 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \tag{28}$$

By applying PML estimation based on only evaluating the bivariate frequencies of item pairs for items within different testlets, the local independence assumption for the bivariate probabilities holds. Hence, consistent item parameter estimates from this modified PML estimation can be expected because the ability variable θ receives a marginal interpretation, which is used in the univariate frequency PML terms L_{1,X_i} and bivariate frequency PML terms L_{2,X_iX_j} that do not receive a weight of zero.

The standard errors can be obtained in the same way using the sandwich formula as for PML estimation with correlation parameters in the normal copula model (see (25) in Section 2.7).

The PML estimation with excluded item pairs is attractive because no specification of the dependence structure and its distribution is required. In fact, the dependence parameters ρ are treated as nuisance parameters that are not of interest and are therefore eliminated from the estimation. Note that a misspecified distribution of the local dependence structure in the IRT model can also bias the item parameters. In this regard, our proposed modified PML estimation approach offers advantages over the copula modeling approach. However, there is still a normal distribution assumption for θ, which might be required to obtain consistent parameter estimates.

2.9. Software

The proposed MML and PML estimations in this paper have been implemented by the author in the R (Version 4.3.1, [35]) software. The code has been made available at https://osf.io/xjp54 (accessed on 18 March 2024). Various 2PL copula models have been implemented in the R package sirt [36] in the `sirt::rasch.copula2()` function and in the Matlab IRTm toolbox [37]. The R package lavaan [38] also allows PML estimation. Users have to properly adapt the weight matrix W_2 (see (28)) in this software. Moreover, the R package pln [39] also contains PML estimation for IRT models. However, it seems that users are not allowed to specify the weight matrix W_2.

3. Simulation Study

The performance of MML with PML estimation under local dependence is compared in various simulation conditions in two simulation studies in which the number of items per testlet is varied.

3.1. Simulation Study 1: Two Items per Testlet
3.1.1. Method

In this simulation study, item responses with local dependence are simulated according to the 2PL model with a normal copula model for item residuals. We study the situation containing $I = 12$ items. All items are arranged in testlets consisting of two items. Hence, there are six testlets for the case of 12 items.

The item discriminations a_i for the 12 items are 0.8, 1.5, 1.3, 1.5, 1.1, 1.6, 0.7, 0.9, 1.1, 1.7, 0.8, and 0.9. The item intercepts b_i are chosen as −0.2, 0.0, 0.3, 0.7, −1.2, 2.2, 0.2, 1.1, −0.3, 1.8, 2.1, and −1.1. The item parameters can also be found in the R code for this Simulation Study 1 located at https://osf.io/xjp54 (accessed on 18 March 2024).

Locally dependent item responses are simulated by assuming a normal copula model [31] for the residuals. The residual correlation matrix $\boldsymbol{\Sigma}^*$ is defined by

$$
\boldsymbol{\Sigma}^* =
\begin{pmatrix}
1 & & & & & & & & & & & \\
f & 1 & & & & & & & & & & \\
0 & 0 & 1 & & & & & & & & & \\
0 & 0 & f & 1 & & & & & & & & \\
0 & 0 & 0 & 0 & 1 & & & & & & & \\
0 & 0 & 0 & 0 & f & 1 & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & f & 1 & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & f & 1 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & f & 1
\end{pmatrix}.
\tag{29}
$$

Note that there is a dependency structure in the residuals for items within testlets. Specifically, the first two items are allocated to the first testlet. The correlation of item residuals is assumed to be f. Items 3 and 4 belong to the second testlet. More generally, testlet t ($t = 1, \ldots, 6$) is composed of items $X_{2(t-1)+1}$ and $X_{2(t-1)+2}$. In the simulation, the residual correlation f is varied as 0, 0.35, and 0.7. The condition $f = 0$ corresponds to local independence.

In the simulation, we chose sample sizes $N = 500, 1000$, and 2000. We did not opt for smaller sample sizes because the 2PL model requires a sufficiently large sample size for stable item parameter estimation.

We now describe the simulation of item responses with local dependence in more detail. First, we simulate a normally distributed ability variable θ with a zero mean and a standard deviation of one. The item response functions are defined according to the 2PL model such that they marginally follow

$$
\mathsf{P}(X_i = 1|\theta) = P_i(1, \theta; \gamma_i) = \Psi(a_i\theta - b_i)
\tag{30}
$$

with $\gamma_i = (a_i, b_i)$. Next, we simulate a multivariate random vector $\boldsymbol{R}^* = (R_1^*, \ldots, R_I^*)$ that follows a multivariate normal distribution with zero mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}^*$, defined in (34). Note that all variables R_i^* are standardized and normally distributed. In the next step, we compute the vector $\boldsymbol{R} = (R_1, \ldots, R_I)$ with uniformly distributed correlated variables by

$$
\boldsymbol{R} = (R_1, \ldots, R_I) = (\Phi(R_1), \ldots, \Phi(R_I)),
\tag{31}
$$

where Φ denotes the standard normal distribution. In the last step, dichotomous item responses X_i ($i = 1, \ldots, I$) are determined by

$$
X_i = 1 \text{ if and only if } R_i < P_i(1, \theta; \gamma_i) .
\tag{32}
$$

In fact, one can show that item responses X_i marginally follow the assumed item response function P_i. Due to the simulation procedure, we obtain

$$
\mathsf{P}(X_i = 1|\theta) = \mathsf{P}(R_i < P_i(1, \theta; \gamma_i)) = P_i(1, \theta; \gamma_i)
\tag{33}
$$

because R_i is uniformly distributed.

The simulated datasets are analyzed with the 2PL model assuming a normal θ distribution with a zero mean and a standard deviation of 1. In this Simulation Study 1, five

different analysis models were specified. First, MML and PML estimation applied to all items were specified under the (incorrect) local independence assumption. The resulting estimation methods are denoted as MML1 and PML1, respectively. Second, MML and PML were specified with a normal copula model for item residuals (denoted as MML2 and PML2). Third, PML was specified by excluding item pairs stemming from the same testlet (denoted as PML3). All estimation methods have been implemented in dedicated R functions that can be found at https://osf.io/xjp54 (accessed on 18 March 2024). The R script for the simulation can also be found at this link.

In total, 3000 replications were conducted. We computed the empirical bias, the empirical standard deviation, and the coverage rates at the confidence level of 95% for all estimated item discriminations a_i and item intercepts b_i. To summarize the simulation results, we computed the average absolute bias, the average standard deviation, and the average coverage rate across all item discriminations and item intercepts, respectively.

3.1.2. Results

Table 1 presents the average absolute bias, average standard deviation, and average coverage rate of item discriminations a_i and item intercepts b_i. The condition $f = 0$ corresponds to the situation of local independence, while $f > 0$ indicates situations of local dependence.

First, it can be concluded that all five estimation methods, MML1, PML1, MML2, PML2, and PML3, provided nearly unbiased estimates for item discriminations and item intercepts. Notably, there is a small bias for a sample size of $N = 500$, but this bias vanishes with an increasing sample size. As expected, the standard deviation of the item parameter estimates decreases with the increasing sample size. In the case of $f = 0$, the efficiency of the different estimators can be defined by the standard deviation. We define a percentage efficiency loss by relating the standard deviations of PML1, MML2, PML2, and PML3 to the standard deviation of the MML1 method. It was found that PML based on all items (i.e., PML1) had a very small average efficiency loss of 0.5% for item discriminations a_i. In contrast, the PML2 and PML3 methods had a larger efficiency loss of 2.7%. MML under the normal copula model (i.e., MML2) had a slightly smaller efficiency loss of 2.4% for the a_i parameters. The efficiency loss for item intercepts b_i was smaller than for item discriminations a_i. It was 0.3% for PML1, 0.5% for MML2, 0.7% for PML2, and 0.7% for PML3.

In the conditions with locally dependent item responses (i.e., $f = 0.35$, and 0.7), MML1 and PML1 produced strongly biased estimates, while MML2, PML2, and PML3 were nearly unbiased in all conditions. The bias was larger for larger residual correlations f. Notably, PML2 resulted in estimates almost identical to those of PML3.

The coverage rates were satisfactory for item parameters that were unbiased. In line with the bias findings, the coverage rates were too low for MML1 and PML1 (i.e., assuming local independence) if the item response data were simulated under local dependence.

Table 1. Simulation Study 1: Average absolute bias and average standard deviations of item parameters as a function of sample size N, and the size of the maximum residual correlation f.

		Average Absolute Bias					Average Standard Deviation					Average Coverage Rate				
f	N	MML1	PML1	MML2	PML2	PML3	MML1	PML1	MML2	PML2	PML3	MML1	PML1	MML2	PML2	PML3
		Item discriminations a_i														
	500	0.016	0.016	0.016	0.016	0.016	0.189	0.190	0.194	0.194	0.194	95.4	95.6	95.4	95.5	95.5
0	1000	0.008	0.008	0.008	0.008	0.008	0.132	0.132	0.135	0.135	0.135	95.0	95.1	95.1	95.1	95.1
	2000	0.005	0.005	0.005	0.005	0.005	0.093	0.093	0.095	0.095	0.095	95.1	95.1	95.0	95.0	95.0
	500	**0.075**	**0.076**	0.020	0.021	0.021	0.195	0.195	0.200	0.201	0.201	94.2	95.1	95.2	95.5	95.5
0.35	1000	**0.063**	**0.064**	0.007	0.008	0.008	0.136	0.136	0.139	0.140	0.140	**92.2**	93.2	95.2	95.3	95.3
	2000	**0.060**	**0.061**	0.004	0.004	0.004	0.095	0.095	0.097	0.098	0.098	**87.8**	**89.5**	95.1	95.2	95.2
	500	**0.163**	**0.148**	0.021	0.021	0.021	0.234	0.217	0.209	0.212	0.212	**85.0**	**90.9**	95.1	95.4	95.4
0.7	1000	**0.154**	**0.139**	0.008	0.008	0.008	0.160	0.149	0.143	0.145	0.145	**76.4**	**82.5**	95.1	95.2	95.2
	2000	**0.152**	**0.138**	0.005	0.005	0.005	0.111	0.104	0.099	0.101	0.101	**66.1**	**69.4**	95.1	95.2	95.2
		Item intercepts b_i														
	500	0.009	0.009	0.009	0.009	0.009	0.143	0.144	0.144	0.144	0.144	95.3	95.3	95.3	95.3	95.3
0	1000	0.005	0.005	0.005	0.005	0.005	0.100	0.101	0.101	0.101	0.101	95.1	95.2	95.1	95.2	95.2
	2000	0.003	0.003	0.003	0.003	0.003	0.071	0.071	0.071	0.071	0.071	94.9	95.0	94.9	94.9	94.9
	500	0.020	0.022	0.012	0.012	0.012	0.145	0.146	0.146	0.146	0.146	95.2	95.3	95.2	95.3	95.3
0.35	1000	0.013	0.015	0.005	0.005	0.005	0.101	0.102	0.101	0.102	0.102	95.0	95.0	95.1	95.1	95.1
	2000	0.012	0.013	0.003	0.003	0.003	0.071	0.072	0.071	0.071	0.071	94.7	94.6	95.0	95.0	95.0
	500	**0.031**	0.029	0.013	0.013	0.013	0.150	0.150	0.149	0.150	0.150	94.3	94.7	95.1	95.2	95.2
0.7	1000	0.028	0.026	0.005	0.005	0.005	0.103	0.103	0.102	0.102	0.102	93.9	94.4	95.2	95.2	95.2
	2000	0.028	0.025	0.003	0.003	0.003	0.073	0.073	0.071	0.072	0.072	**92.2**	92.9	95.1	95.0	95.0

Note: MML1 = marginal maximum likelihood estimation under local independence; PML1 = pairwise likelihood estimation based on all item pairs under local independence; MML2 = marginal maximum likelihood estimation under the normal copula model; PML2 = pairwise likelihood estimation under the normal copula model; PML3 = pairwise likelihood estimation based on item pairs with items from different testlets. The residual correlation matrix $\mathbf{\Sigma}^*$ as a function of f is given in (29). Average absolute bias values larger than 0.03 and coverage rates smaller than 92.5 are printed in bold font.

3.2. Simulation Study 2: Four Items per Testlet

3.2.1. Method

In Simulation Study 2, we also considered $I = 12$ items. The items were arranged in testlets consisting of four items. Hence, there were three testlets in the case of $I = 12$ items.

The same item parameters as in Simulation Study 1 were used (see Section 3.1.1). The item parameters can be found in the simulation code located at https://osf.io/xjp54 (accessed on 18 March 2024).

Locally dependent item responses were again simulated using the normal copula model. The residual correlation matrix $\mathbf{\Sigma}^*$ was defined by

$$\mathbf{\Sigma}^* = \begin{pmatrix} 1 & & & & & & & & & & & \\ f & 1 & & & & & & & & & & \\ f & f & 1 & & & & & & & & & \\ f & f & f & 1 & & & & & & & & \\ 0 & 0 & 0 & 0 & 1 & & & & & & & \\ 0 & 0 & 0 & 0 & f & 1 & & & & & & \\ 0 & 0 & 0 & 0 & 0.3f & 0.3f & 1 & & & & & \\ 0 & 0 & 0 & 0 & 0.3f & 0.3f & f & 1 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & f & 1 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & f & f & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{34}$$

This specification induced a dependency structure in the residuals for items within testlets. Specifically, the first four items were allocated to the first testlet, and there was a constant residual correlation f. Items 5 to 8 referred to the second testlet, and the residual correlations varied depending on the item pair. The third testlet was composed of Items 9 to 12. Note that Item 12 was locally independent of Items 9, 10, and 11. The maximum residual correlation f was varied as 0, 0.35, and 0.7. The condition $f = 0$ corresponded to local independence.

As in Simulation Study 1, sample sizes $N = 500$, 1000, and 2000 were chosen.

The same estimation methods except for MML2 (i.e., MML with normal copula model for item residuals) as in Simulation Study 1 were specified. MML2 was computationally demanding as four-dimensional normal probabilities had to be extensively evaluated because the dependence structure consisted of testlets of size 4. The R code for the simulation can be found at https://osf.io/xjp54 (accessed on 18 March 2024).

In total, 2000 replications were conducted. As in Simulation Study 1, the average absolute bias, average standard deviation, and average coverage rates of item discriminations a_i and item intercepts b_i were investigated.

3.2.2. Results

Table 2 presents the average absolute bias, average standard deviation, and average coverage rates of item discriminations a_i and item intercepts b_i. Again, the condition $f = 0$ corresponds to local independence, while $f > 0$ corresponds to local dependence.

Table 2. Simulation Study 2: Average absolute bias and average standard deviations of item parameters as a function of sample size N, and the size of the maximum residual correlation f.

f	N	Average Absolute Bias				Average Standard Deviation				Average Coverage Rate			
		MML1	PML1	PML2	PML3	MML1	PML1	PML2	PML3	MML1	PML1	PML2	PML3
		Item discriminations a_i											
	500	0.014	0.014	0.017	0.017	0.189	0.189	0.208	0.208	95.3	95.5	95.6	95.6
0	1000	0.008	0.008	0.010	0.010	0.131	0.132	0.144	0.144	95.2	95.2	95.4	95.4
	2000	0.004	0.004	0.005	0.005	0.093	0.093	0.102	0.102	94.9	95.0	94.9	94.9
	500	**0.220**	**0.200**	0.025	0.025	0.207	0.204	0.230	0.230	**74.3**	**79.5**	95.9	95.9
0.35	1000	**0.216**	**0.196**	0.010	0.010	0.144	0.142	0.159	0.159	**62.6**	**66.4**	95.3	95.3
	2000	**0.214**	**0.194**	0.005	0.005	0.100	0.099	0.111	0.111	**57.1**	**58.5**	95.1	95.1
	500	**0.827**	**0.710**	0.028	0.028	0.296	0.290	0.260	0.260	**47.1**	**53.2**	95.5	95.5
0.7	1000	**0.818**	**0.703**	0.012	0.012	0.195	0.187	0.173	0.173	**36.1**	**44.7**	95.5	95.5
	2000	**0.816**	**0.701**	0.007	0.007	0.139	0.133	0.121	0.121	**25.7**	**34.0**	95.3	95.3
		Item intercepts b_i											
	500	0.010	0.010	0.011	0.011	0.142	0.142	0.145	0.145	95.4	95.5	95.6	95.6
0	1000	0.005	0.005	0.005	0.005	0.100	0.100	0.102	0.102	95.3	95.3	95.4	95.4
	2000	0.002	0.002	0.002	0.002	0.070	0.071	0.072	0.072	95.0	95.0	95.0	95.0
	500	**0.035**	**0.032**	0.012	0.012	0.148	0.148	0.152	0.152	94.1	94.4	95.4	95.4
0.35	1000	**0.037**	**0.033**	0.006	0.006	0.103	0.103	0.105	0.105	92.7	93.2	95.4	95.4
	2000	**0.037**	**0.033**	0.003	0.003	0.073	0.073	0.074	0.074	**90.1**	**91.0**	95.0	95.0
	500	**0.132**	**0.114**	0.016	0.016	0.169	0.166	0.158	0.159	**84.8**	**87.5**	95.6	95.6
0.7	1000	**0.135**	**0.116**	0.007	0.007	0.119	0.116	0.109	0.109	**75.7**	**79.8**	95.2	95.2
	2000	**0.135**	**0.116**	0.004	0.004	0.082	0.081	0.075	0.075	**66.0**	**71.2**	95.1	95.1

Note: MML1 = marginal maximum likelihood estimation under local independence; PML1 = pairwise likelihood estimation based on all item pairs under local independence; PML2 = pairwise likelihood estimation under the normal copula model; PML3 = pairwise likelihood estimation based on item pairs with items from different testlets. The residual correlation matrix Σ^* as a function of f is given in (34). Average absolute bias values larger than 0.03 and coverage rates smaller than 92.5 are printed in bold font.

All four estimation methods, MML1, PML1, PML2, and PML3, provided approximately unbiased results for item discriminations a_i and item intercepts b_i. There was a small bias for a sample size of $N = 500$, but this bias vanished with an increasing sample

size. The average standard deviation of the item parameter estimates decreased with the increasing sample size.

In the condition $f = 0$, indicating local independence, the efficiency loss was assessed. It was 0.4% for PML1 compared to MML1 for item discriminations. However, the efficiency loss for PML2 and PML3 was 10.1%, much higher than in Simulation Study 1, because more item pairs were excluded from the PML estimation. The efficiency loss for item intercepts was smaller for PML2 and PML3, with a value of 2.1%.

If the item response data were generated under local dependence, biased item parameter estimates under MML1 and PML1 resulted. For methods PML2 and PML3, there was only a small bias for $N = 500$. However, the bias was negligible for larger sample sizes.

Finally, the coverage rates for PML2 and PML3 were acceptable. In the conditions with biased estimates for MML1 and PML1, the coverage rates were too low.

4. Empirical Examples

In this section, three datasets with item responses with a testlet structure from the R package sirt [36] are analyzed. Estimated item parameters from the different MML and PML estimation methods (i.e., MML1, PML1, PML2, and PML3; see Section 3.1.1) are compared.

4.1. Dataset `data.read`

The first dataset, `data.read`, contains the item responses of 328 subjects on 12 items that stem from a reading comprehension test of Austrian students. The 12 items are arranged into three testlets, each testlet containing of four items. Table 3 contains the estimated item discriminations a_i and estimated item intecepts b_i and their standard error estimates for the four different estimation methods MML1, PML1, PML2, and PML3. The average item discrimination was higher for the methods that ignored the testlet structure (MML: 1.241, PML1: 1.184) than PML2 or PML3 with 0.983, which took the testlet structure into account in the estimation.

Table 3. Dataset `data.read`: Estimated item discriminations a_i and item intercepts b_i (standard errors in parentheses) for different estimation methods.

Item	Testlet	a_i				b_i			
		MML1	PML1	PML2	PML3	MML1	PML1	PML2	PML3
A1	1	0.96 (0.25)	0.97 (0.26)	0.85 (0.30)	0.85 (0.30)	−2.09 (0.23)	−2.04 (0.21)	−1.97 (0.22)	−1.97 (0.22)
A2	1	1.44 (0.28)	1.36 (0.31)	1.54 (0.51)	1.54 (0.51)	−1.45 (0.21)	−1.38 (0.20)	−1.46 (0.28)	−1.46 (0.28)
A3	1	1.14 (0.21)	1.08 (0.27)	1.07 (0.33)	1.07 (0.33)	−0.28 (0.14)	−0.33 (0.14)	−0.33 (0.14)	−0.33 (0.14)
A4	1	0.83 (0.18)	0.85 (0.21)	0.88 (0.28)	0.88 (0.28)	0.19 (0.13)	0.18 (0.13)	0.19 (0.13)	0.19 (0.13)
B1	2	0.58 (0.17)	0.60 (0.17)	0.67 (0.23)	0.67 (0.23)	−1.04 (0.14)	−0.98 (0.14)	−1.00 (0.14)	−1.00 (0.14)
B2	2	0.67 (0.17)	0.67 (0.16)	0.80 (0.24)	0.80 (0.24)	−0.09 (0.12)	−0.03 (0.12)	−0.03 (0.13)	−0.03 (0.13)
B3	2	1.22 (0.32)	1.11 (0.29)	1.18 (0.42)	1.17 (0.42)	−2.85 (0.33)	−2.76 (0.30)	−2.80 (0.37)	−2.80 (0.37)
B4	2	1.11 (0.22)	1.11 (0.23)	1.51 (0.45)	1.51 (0.45)	−0.93 (0.16)	−0.95 (0.16)	−1.08 (0.21)	−1.08 (0.21)
C1	3	2.40 (0.63)	2.31 (0.90)	0.91 (0.36)	0.91 (0.36)	−4.30 (0.75)	−4.37 (1.00)	−2.97 (0.33)	−2.97 (0.33)
C2	3	1.53 (0.29)	1.48 (0.31)	1.05 (0.25)	1.05 (0.25)	−1.32 (0.21)	−1.27 (0.20)	−1.11 (0.16)	−1.11 (0.16)
C3	3	1.95 (0.49)	1.56 (0.48)	0.72 (0.26)	0.72 (0.26)	−3.07 (0.49)	−2.65 (0.40)	−2.10 (0.21)	−2.10 (0.21)
C4	3	1.08 (0.23)	1.11 (0.28)	0.63 (0.20)	0.63 (0.20)	−1.36 (0.18)	−1.26 (0.18)	−1.11 (0.14)	−1.11 (0.14)

Note: MML1 = marginal maximum likelihood estimation under local independence; PML1 = pairwise likelihood estimation based on all item pairs under local independence; PML2 = pairwise likelihood estimation under the normal copula model; PML3 = pairwise likelihood estimation based on item pairs with items from different testlets.

The average absolute difference in the estimated item discriminations was 0.454 for PML3 and MML, and it was 0.345 for PML3 and PML1. The average absolute difference between PML1 and MML was substantially smaller at 0.130. Note that PML2 and PML3 resulted in practically identical item parameter estimates.

4.2. Dataset `data.pisaRead`

The second dataset, `data.pisaRead`, consisted of item responses from 623 Austrian students in a PISA study [40] to one item cluster of reading items. The 12 items in the dataset were arranged into four testlets that each contained three items.

Table 4 presents the estimated item discriminations a_i and estimated item intercepts b_i for the three estimation methods. The four estimation methods resulted in similar average item discriminations (MML1: 1.518, PML1: 1.555, PML3: 1.518). As for the first dataset, PML2 and PML3 resulted in practically identical parameter estimates. The average absolute differences in item discrimination between PML3 and MML1 with 0.091, as well as between PML3 and PML1 with 0.089, were only slightly larger than between PML1 and MML1 with 0.068.

Table 4. Dataset `data.pisaRead`: Estimated item discriminations a_i and item intercepts b_i (standard errors in parentheses) for different estimation methods.

Item	Testlet	a_i				b_i			
		MML1	PML1	PML2	PML3	MML1	PML1	PML2	PML3
R432Q01	1	2.33 (0.34)	2.64 (0.43)	2.83 (0.54)	2.83 (0.54)	−3.39 (0.36)	−3.67 (0.46)	−3.86 (0.56)	−3.86 (0.56)
R432Q05	1	2.04 (0.26)	2.02 (0.26)	2.13 (0.31)	2.13 (0.31)	−1.70 (0.19)	−1.72 (0.20)	−1.78 (0.22)	−1.78 (0.22)
R432Q06	1	1.22 (0.25)	1.09 (0.27)	1.07 (0.28)	1.07 (0.28)	2.99 (0.26)	2.90 (0.26)	2.88 (0.26)	2.88 (0.26)
R456Q01	2	1.19 (0.30)	1.59 (0.49)	1.38 (0.43)	1.38 (0.43)	−4.30 (0.41)	−4.81 (0.68)	−4.56 (0.57)	−4.56 (0.57)
R456Q02	2	0.88 (0.15)	0.94 (0.17)	0.83 (0.16)	0.83 (0.16)	−1.89 (0.14)	−1.94 (0.15)	−1.88 (0.14)	−1.88 (0.14)
R456Q06	2	1.90 (0.26)	2.00 (0.30)	1.87 (0.27)	1.87 (0.27)	−2.82 (0.26)	−2.90 (0.30)	−2.78 (0.27)	−2.78 (0.27)
R460Q01	3	1.53 (0.19)	1.51 (0.19)	1.48 (0.20)	1.48 (0.20)	−0.71 (0.12)	−0.76 (0.13)	−0.75 (0.12)	−0.75 (0.12)
R460Q05	3	1.71 (0.24)	1.79 (0.27)	1.67 (0.27)	1.67 (0.27)	−2.68 (0.24)	−2.74 (0.27)	−2.64 (0.26)	−2.64 (0.26)
R460Q06	3	1.24 (0.16)	1.24 (0.16)	1.16 (0.16)	1.16 (0.16)	−0.50 (0.11)	−0.49 (0.11)	−0.47 (0.11)	−0.47 (0.11)
R466Q02	4	1.32 (0.17)	1.29 (0.15)	1.29 (0.16)	1.29 (0.16)	0.09 (0.11)	0.11 (0.11)	0.11 (0.11)	0.11 (0.11)
R466Q03	4	0.97 (0.17)	1.00 (0.16)	0.97 (0.17)	0.97 (0.17)	1.66 (0.14)	1.66 (0.14)	1.64 (0.14)	1.64 (0.14)
R466Q06	4	1.52 (0.20)	1.53 (0.21)	1.55 (0.22)	1.55 (0.22)	−1.76 (0.16)	−1.79 (0.17)	−1.80 (0.18)	−1.80 (0.18)

Note: MML1 = marginal maximum likelihood estimation under local independence; PML1 = pairwise likelihood estimation based on all item pairs under local independence; PML2 = pairwise likelihood estimation under the normal copula model; PML3 = pairwise likelihood estimation based on item pairs with items from different testlets.

4.3. Dataset `data.pisaMath`

The third dataset, `data.pisaMath`, contains the item responses of 565 Austrian students on 11 mathematics items to an item cluster in a PISA study [40]. The 11 items are arranged into four testlets with two items each, and there are three single items without a testlet structure included in the dataset.

Table 5 contains the estimated item parameters for `data.pisaMath`. The average item discrimination was slightly larger for MML1 and PML1, with 1.269 and 1.263, compared to the PML3 estimation method, with a value of 1.205. Interestingly, the average absolute difference in the item discriminations between PML1 and MML1 was very small at 0.012, but they were substantially larger between PML3 and MML1, with 0.125, and between PML3 and PML1, with 0.132. Again, PML2 and PML3 provided practically identical item parameter estimates.

Table 5. Dataset `data.pisaMath`: Estimated item discriminations a_i and item intercepts b_i (standard errors in parentheses) for different estimation methods.

Item	Testlet	a_i				b_i			
		MML1	PML1	PML2	PML3	MML1	PML1	PML2	PML3
M192Q01	1	1.26 (0.16)	1.27 (0.17)	1.32 (0.18)	1.32 (0.18)	0.23 (0.11)	0.23 (0.11)	0.23 (0.11)	0.23 (0.11)
M406Q01	2	1.77 (0.22)	1.74 (0.23)	1.45 (0.20)	1.45 (0.20)	0.37 (0.13)	0.36 (0.13)	0.33 (0.12)	0.33 (0.12)
M406Q02	2	2.27 (0.32)	2.22 (0.33)	1.83 (0.26)	1.83 (0.26)	1.66 (0.22)	1.63 (0.21)	1.45 (0.18)	1.45 (0.18)
M423Q01	3	0.53 (0.12)	0.52 (0.13)	0.53 (0.13)	0.53 (0.13)	−1.11 (0.10)	−1.11 (0.10)	−1.11 (0.11)	−1.11 (0.11)
M496Q01	4	1.38 (0.18)	1.39 (0.18)	1.24 (0.17)	1.24 (0.17)	−0.29 (0.12)	−0.30 (0.12)	−0.28 (0.11)	−0.28 (0.11)
M496Q02	4	1.13 (0.16)	1.12 (0.17)	0.94 (0.15)	0.94 (0.15)	−1.15 (0.13)	−1.15 (0.13)	−1.09 (0.12)	−1.09 (0.12)
M564Q01	5	0.82 (0.13)	0.82 (0.13)	0.85 (0.13)	0.85 (0.13)	−0.07 (0.10)	−0.07 (0.10)	−0.07 (0.10)	−0.07 (0.10)
M564Q02	5	0.79 (0.12)	0.79 (0.12)	0.82 (0.13)	0.82 (0.13)	−0.12 (0.10)	−0.12 (0.10)	−0.12 (0.10)	−0.12 (0.10)
M571Q01	6	1.42 (0.18)	1.43 (0.19)	1.53 (0.21)	1.53 (0.21)	−0.26 (0.12)	−0.26 (0.12)	−0.27 (0.12)	−0.27 (0.12)
M603Q01	7	1.17 (0.15)	1.18 (0.16)	1.23 (0.17)	1.23 (0.17)	−0.28 (0.11)	−0.29 (0.11)	−0.29 (0.11)	−0.29 (0.11)
M603Q02	7	1.42 (0.18)	1.42 (0.18)	1.51 (0.21)	1.51 (0.21)	0.14 (0.12)	0.14 (0.12)	0.15 (0.12)	0.15 (0.12)

Note: MML1 = marginal maximum likelihood estimation under local independence; PML1 = pairwise likelihood estimation based on all item pairs under local independence; PML2 = pairwise likelihood estimation under the normal copula model; PML3 = pairwise likelihood estimation based on item pairs with items from different testlets.

5. Discussion

In this article, we propose a variant of PML estimation that allows the consistent estimation of item parameters in the 2PL model despite the presence of local dependence. In the modified PML approach, the bivariate frequencies of items within the same testlet do not appear in the optimization function. As such, local dependence is essentially removed from the estimation. Of course, this change brings the disadvantage of efficiency losses compared with MML estimation if local independence holds. However, with large sample sizes, the bias outweighs the higher variance of PML estimation. Our proposed method had equal performance to the MML and PML estimation methods on a 2PL normal copula model that additionally estimated local dependence parameters. This finding was obtained with an unstructured model of local dependence. If local dependence is summarized into a small(er) number of dependence parameters per testlet, the simultaneous estimation of the item parameters and dependence parameters might be more efficient than PML estimation with excluded item pairs of items referring to the same testlet.

The PML estimation approach can be easily adapted to IRT models with polytomous item responses under local dependence. Similarly, multidimensional IRT models with a dependence structure can also be handled with the modified PML estimation. Importantly, if all items only load on one dimension, PML estimation solves the issue of the evaluation of high-dimensional integrals in MML estimation for multidimensional IRT models.

This article did not discuss the estimation of standard errors. However, standard M-estimation theory can be used for statistical inference (see also [14,41]). Future studies could evaluate the accuracy of standard error estimation under PML.

In some applications, the standard errors of person parameter estimates $\hat{\theta}$ of θ are of interest [42,43]. The M-estimation theory for a clustered data structure (i.e., cluster-robust standard errors) can be used to determine the standard errors in the case of local dependence. Alternatively, the resampling of items or testlets can also be used to compute standard errors [44].

This paper focuses on the estimation of the 2PL model with local dependence. However, PML estimation for the Rasch model [45] with local dependence could also be carried out for smaller sample sizes [46]. Alternatively, prior distributions or regularized estimation can stabilize the estimation of the 2PL model in smaller sample sizes [47–49]. In PML estimation, a prior distribution (or penalty terms) could also be included (see [50,51] for a similar approach).

Finally, previous research indicated that the effect of local dependence on the item parameters reduces with an increasing number of items for fixed lengths of testlets. In

this case, the local dependence vanishes for infinitely long tests [52]. It is likely that the negative impacts of local dependence in reporting from tests should not be exaggerated unless confronted with very short tests or an excessive dependence structure.

Abbreviations

The following abbreviations are used in this manuscript:

2PL	two-parameter logistic
EM	expectation–maximization
IRT	item response theory
MML	marginal maximum likelihood
PISA	Programme for International Student Assessment
PML	pairwise maximum likelihood

References

1. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Item response theory—A statistical framework for educational and psychological measurement. *Stat. Sci.* **2023**. Available online: https://imstat.org/journals-and-publications/statistical-science/statistical-science-future-papers/ (accessed on 18 March 2024).
2. van der Linden, W.J. Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume 1: Models*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30. [CrossRef]
3. Rutkowski, L.; von Davier, M.; Rutkowski, D. (Eds.) *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Chapman Hall/CRC Press: London, UK, 2013. [CrossRef]
4. OECD. *PISA 2009*; Technical Report; OECD: Paris, France, 2012. Available online: https://bit.ly/3xfxdwD (accessed on 18 March 2024).
5. Bock, R.D.; Moustaki, I. Item response theory in a general framework. *Handb. Stat.* **2006**, *26*, 469–513. [CrossRef]
6. Yen, W.M.; Fitzpatrick, A.R. Item response theory. In *Educational Measurement*; Brennan, R.L., Ed.; Praeger Publishers: Westport, WA, USA, 2006; pp. 111–154.
7. Woods, C.M. Empirical histograms in item response theory with ordinal data. *Educ. Psychol. Meas.* **2007**, *67*, 73–87. [CrossRef]
8. Xu, X.; von Davier, M. *Fitting the Structured General Diagnostic Model to NAEP Data*; (Research Report No. RR-08-28); Educational Testing Service: Princeton, NJ, USA, 2008. [CrossRef]
9. Yen, W.M. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Meas.* **1984**, *8*, 125–145. [CrossRef]
10. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
11. Aitkin, M. Expectation maximization algorithm and extensions. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 217–236. [CrossRef]
12. Ünlü, A.; Yanagida, T. R you ready for R? The CRAN Psychometrics task view. *Brit. J. Math. Stat. Psychol.* **2011**, *64*, 182–186. [CrossRef] [PubMed]
13. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*; Lord, F.M., Novick, M.R., Eds.; MIT Press: Reading, MA, USA, 1968; pp. 397–479.
14. Katsikatsou, M.; Moustaki, I.; Yang-Wallentin, F.; Jöreskog, K.G. Pairwise likelihood estimation for factor analysis models with ordinal data. *Comp. Stat. Data An.* **2012**, *56*, 4243–4258. [CrossRef]
15. Katsikatsou, M.; Moustaki, I.; Jamil, H. Pairwise likelihood estimation for confirmatory factor analysis models with categorical variables and data that are missing at random. *Brit. J. Math. Stat. Psychol.* **2022**, *75*, 23–45. [CrossRef]
16. Renard, D.; Molenberghs, G.; Geys, H. A pairwise likelihood approach to estimation in multilevel probit models. *Comp. Stat. Data An.* **2004**, *44*, 649–667. [CrossRef]
17. Varin, C.; Reid, N.; Firth, D. An overview of composite likelihood methods. *Stat. Sin.* **2011**, *21*, 5–42. Available online: https://bit.ly/38lbhom (accessed on 18 March 2024).

18. Vasdekis, V.G.S.; Rizopoulos, D.; Moustaki, I. Weighted pairwise likelihood estimation for a general class of random effects models. *Biostatistics* **2014**, *15*, 677–689. [CrossRef] [PubMed]
19. Bradlow, E.T.; Wainer, H.; Wang, X. A Bayesian random effects model for testlets. *Psychometrika* **1999**, *64*, 153–168. [CrossRef]
20. Tuerlinckx, F.; De Boeck, P. Non-modeled item interactions lead to distorted discrimination parameters: A case study. *Methods Psychol. Res. Online* **2001**, *6*, 2. Available online: http://tinyurl.com/mrydstwz (accessed on 18 March 2024).
21. Braeken, J.; Tuerlinckx, F.; De Boeck, P. Copula functions for residual dependency. *Psychometrika* **2007**, *72*, 393–411. [CrossRef]
22. Ip, E.H. Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika* **2001**, *66*, 109–132. [CrossRef]
23. Debelak, R.; Koller, I. Testing the local independence assumption of the Rasch model with Q_3-based nonparametric model tests. *Appl. Psychol. Meas.* **2020**, *44*, 103–117. [CrossRef] [PubMed]
24. Wang, W.C.; Wilson, M. The Rasch testlet model. *Appl. Psychol. Meas.* **2005**, *29*, 126–149. [CrossRef]
25. Eckes, T. Item banking for C-tests: A polytomous Rasch modeling approach. *Psych. Test Assess. Model.* **2011**, *53*, 414–439. Available online: http://tinyurl.com/bap2z4nh (accessed on 18 March 2024).
26. Hoskens, M.; De Boeck, P. A parametric model for local dependence among test items. *Psychol. Methods* **1997**, *2*, 261–277. [CrossRef]
27. Marais, I.; Andrich, D. Effects of varying magnitude and patterns of response dependence. *J. Appl. Meas.* **2008**, *9*, 105–124. Available online: http://tinyurl.com/yc7mhmkw (accessed on 18 March 2024).
28. Wilson, M.; Adams, R.J. Rasch models for item bundles. *Psychometrika* **1995**, *60*, 181–198. [CrossRef]
29. Ip, E.H. Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *Brit. J. Math. Stat. Psychol.* **2010**, *63*, 395–416. [CrossRef]
30. Braeken, J. A boundary mixture approach to violations of conditional independence. *Psychometrika* **2011**, *76*, 57–76. [CrossRef]
31. Braeken, J.; Kuppens, P.; De Boeck, P.; Tuerlinckx, F. Contextualized personality questionnaires: A case for copulas in structural equation models for categorical data. *Multivar. Behav. Res.* **2013**, *48*, 845–870. [CrossRef] [PubMed]
32. Joe, H. *Dependence Modeling with Copulas*; CRC Press: Boca Raton, FL, USA, 2014.
33. Nikoloulopoulos, A.K.; Joe, H. Factor copula models for item response data. *Psychometrika* **2015**, *80*, 126–150. [CrossRef]
34. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference*; Springer: New York, NY, USA, 2013. [CrossRef]
35. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2023. Available online: https://www.R-project.org/ (accessed on 15 March 2023).
36. Robitzsch, A. *Sirt: Supplementary Item Response Theory Models*; R Package Version 4.1-15. 2024. Available online: https://CRAN.R-project.org/package=sirt (accessed on 6 February 2024).
37. Braeken, J.; Tuerlinckx, F. Investigating latent constructs with item response models: A MATLAB IRTm toolbox. *Behav. Res. Methods* **2009**, *41*, 1127–1137. [CrossRef] [PubMed]
38. Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **2012**, *48*, 1–36. [CrossRef]
39. Falk, C.F.; Joe, H. *pln: Polytomous Logit-Normit (Graded Logistic) Model Estimation*; R Package Version 0.2-2. 2020. Available online: https://CRAN.R-project.org/package=pln (accessed on 30 July 2020).
40. Lietz, P.; Cresswell, J.C.; Rust, K.F.; Adams, R.J. (Eds.). *Implementation of Large-Scale Education Assessments*; Wiley: New York, NY, USA, 2017. [CrossRef]
41. White, H. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
42. Penfield, R.D.; Bergeron, J.M. Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Appl. Psychol. Meas.* **2005**, *29*, 218–233. [CrossRef]
43. Warm, T.A. Weighted likelihood estimation of ability in item response theory. *Psychometrika* **1989**, *54*, 427–450. [CrossRef]
44. Wainer, H.; Morgan, A.; Gustafsson, J.E. A review of estimation procedures for the Rasch model with an eye toward longish tests. *J. Educ. Stat.* **1980**, *5*, 35–64. [CrossRef]
45. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
46. Robitzsch, A. A comprehensive simulation study of estimation methods for the Rasch model. *Stats* **2021**, *4*, 814–836. [CrossRef]
47. Fox, J.P. *Bayesian Item Response Modeling*; Springer: New York, NY, USA, 2010. [CrossRef]
48. König, C.; Alexandrowicz, R.W. Benefits of the curious behavior of Bayesian hierarchical item response theory models—An in-depth investigation and bias correction. *Appl. Psychol. Meas.* **2024**, *48*, 38–56. [CrossRef]
49. Levy, R. The rise of Markov chain Monte Carlo estimation for psychometric modeling. *J. Probab. Stat.* **2009**, *2009*, 537139. [CrossRef]
50. Huang, P.H. Penalized least squares for structural equation modeling with ordinal responses. *Multivar. Behav. Res.* **2022**, *57*, 279–297. [CrossRef] [PubMed]
51. Hui, F.K.C.; Müller, S.; Welsh, A.H. Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *J. Am. Stat. Assoc.* **2018**, *113*, 1759–1769. [CrossRef]
52. Stout, W. A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* **1987**, *52*, 589–617. [CrossRef]