# Modeling local dependence in longitudinal IRT models

**Maja Olsbjerg · Karl Bang Christensen**

**Abstract** Measuring change in a latent variable over time is often done using the same instrument at several time points. This can lead to dependence between responses across time points for the same person yielding within person correlations that are stronger than what can be attributed to the latent variable. Ignoring this can lead to biased estimates of changes in the latent variable. In this paper we propose a method for modeling local dependence in the longitudinal 2PL model. It is based on the concept of item splitting, and makes it possible to correctly estimate change in the latent variable.

## Introduction

In item response theory (IRT,) a set of items, the instrument, measures a latent variable describing a person. The latent variable could be for instance math skills, the level of depression, or quality of life. In IRT models it is assumed that items are conditionally independent given the latent variable. This technical requirement that the items should only be correlated through the latent trait that the test is measuring is referred to as local independence and is well described in the literature (Lord & Novick, 1968; Lazarsfeld & Henry, 1968).

M. Olsbjerg · K. B. Christensen (✉)
Department of Biostatistics, University of Copenhagen,
Copenhagen, Denmark
e-mail: kach@sund.ku.dk

M. Olsbjerg
e-mail: mola@sund.ku.dk

The assumption of local independence can be violated in different ways. Firstly, instruments are often composed of item bundles each measuring their own aspect of the latent variable and the higher-order latent variable alone might not account for correlations between items in the same bundle. This type of local dependence can be interpreted as a violation of unidimensionality. Secondly, the assumption of local independence can be violated if the response given to one item directly influences the response given to another item. This may happen due to similarities in item content or in response format or, in an educational test, if the correct answer on the first item contains a clue as to the correct answer for the second item.

Both of these situations yield inter-item correlations beyond what can be attributed to the latent variable, but for very different reasons. In order to distinguish between these two types of dependence, the first one is sometimes termed trait dependence and the second one response dependence. In general, trait and response dependence are not clearly distinguished in the literature. Nevertheless, using algebraic formulations of the two phenomena, Marais and Andrich (2008a, b) have demonstrated that the implications of the two types of dependence point in opposite directions. One of the important observations is that reliability indices (the person separation index and Cronbach's coefficient alpha) decrease for data with trait dependence, but increase for data with response dependence. Thus, the reliability of an instrument should be interpreted with caution if the assumption of local independence has not been carefully checked.

Methods for detecting both types of local independence in Rasch models have been proposed by Kelderman (1984). He expressed the dichotomous Rasch model as a log-linear model and then showed that local dependence corresponds to interactions between items in the resulting log-linear Rasch model. Kelderman's (1984) original model was for

dichotomous item only, but log-linear Rasch models for polytomous item formats also exist (Kelderman, 1997).

Log-linear Rasch models have also been considered by Kreiner and Christensen (2004, 2007). Motivated by Tjur (1982), they evaluated partial correlations between item pairs conditionally on rest scores. This approach is similar to the Mantel–Haenszel analysis of differential item functioning (DIF) (Holland & Thayer, 1988; Holland & Wainer 1993) and is readily implemented in standard software. Kreiner & Christensen argue that models that incorporate local dependence still provide essentially valid and objective measurement and describe the measurement properties of such models.

Methods for detecting local dependence in IRT models that are more complicated than Rasch models include the use of conditional covariances (Douglas et al., 1998), Mantel–Haenszel type tests (Ip, 2001), or specification of models that incorporate local dependence (Hoskens & De Boeck, 1997; Ip, 2002). Furthermore, papers have addressed consequences of local dependence (Scott & Ip, 2002) and ways of adjusting for it (Ip, 2000).

Jiao and colleagues proposed a three-level hierarchical generalized linear model (HGLM) to model clustered data like this and compared it to the Rasch-equivalent two-level HGLM that ignores the nested structure of items (Jiao et al. 2005). The result of fitting the too simple model was imprecise estimation of item difficulties and underestimation of the variance in the distribution of the latent variable. Modeling the structure estimated the variance in the distribution of the latent variable correctly, while the two-level HGLM increasingly underestimated the variance as the magnitude of dependence increased.

Another formalization of local dependence in IRT models is the notion of testlets, i.e., groups of items that are placed on a test as a unit, typically reading passages followed by a group of questions (Bradlow et al., 1999; Wang & Wilson, 2005). When evidence of local dependence turns up, constructing testlets post hoc can be a solution that permits the use of conventional unidimensional IRT models even in the presence of local dependence (Yen, 1993).

Confirmatory factor analysis (CFA) models are applied to describe correlation structures, and have been formulated for latent response variables measured using ordinal categorical observed variables (Muthén, 1979, 1984). Confirmatory factor analysis can be used to test local dependence across time points by considering models with an added covariance parameters.

Common to the HGLM, the testlet model and the CFA approach is that they all model dependence using random effects. This means that they are methods for modeling trait dependence. As for response dependence, a way of quantifying this has been proposed by Andrich and Kreiner (2010) for two dichotomous items. It is based on splitting one of the items into two new items according to the responses to the other item. The magnitude of dependence is then estimated as half the distance between the estimated item locations of the new items. A generalization of this approach to polytomous items was later proposed (Andrich et al., 2012).

Beyond local item dependence, local person dependence can also occur in case of cluster sampling (Jiao et al., 2012), however this is not discussed in this paper.

Longitudinal extensions of IRT models have been proposed (Andersen, 1985; Embretson, 1991; Liu & Hedeker, 2006; Bousseboua & Mesbah 2010). Many of these impose an additional requirement of local independence across time points. In these models, correlations between responses for the same person are modeled by including a latent correlation matrix. However, when the same measurement instrument is used at two time points, correlations between responses for the same person might be stronger than what the latent variable accounts for. If this is the case, the requirement of local independence across time points is violated by response dependence. Marais (2009) has shown that ignoring response dependence can either mask or exaggerate changes in both items and persons potentially leading to wrong conclusions. It is therefore important to be able to deal with this type of dependence.

Another key assumption when measuring trends over time is that the item parameters do not change over time (Wells et al., 2002; Miller and Fitzpatrick 2009). This can be considered differential item functioning with respect to time and is often referred to as item parameter drift. In many applications, persons are followed over time in order to measure change in some latent variable. For that purpose, the measurement instrument should be somewhat stable in the sense that there is no item drift. Tests of this assumption have been proposed and implemented in SAS by Olsbjerg and Christensen (2013a, b) for the special case of the Rasch (1PL) model. DIF across time, or item parameter drift, describes the situation where the parameters of an item changes for everybody in the population, while local dependence across time, as operationalized here, describes the situation where the parameters of an item at the second time point depends on the response given to the item at the first time point. Thus, these are different phenomena, but it should be noted that in the case of local dependence across time points, spurious evidence of item parameter drift can turn up: Consider for example a situation where a test item becomes much easier for the 80 % of subjects who answer correctly at time point 1, but retains its difficulty for the remaining 20 % of the population. In this situation, the item will appear to have item parameter drift.

In this paper we propose a method for modeling response dependence in a longitudinal version of the 2PL model. This way, unbiased estimation of change in the latent

variable becomes feasible. It is based on the idea by of splitting dependent items in unidimensional Rasch models of Andrich and Kreiner (2010), see also Andrich et al. (2012). Henceforth, we use the term local dependence when referring to response dependence.

## The 2PL model

The dichotomous Rasch (or 1PL) model (Rasch, 1960; Fischer and Molenaar, 1995) and the Birnbaum (or 2PL) model (Birnbaum, 1968) are the simplest IRT models. They describe the responses to manifest dichotomous items $X_1, \ldots, X_I$ measuring a latent variable $\theta \in \mathbb{R}$. The response probability for item $i$ for a given value of $\theta$ is modeled as

$$P(X_i = x_i | \theta) = \frac{\exp[x_i \alpha_i (\theta - \beta_i)]}{1 + \exp[\alpha_i (\theta - \beta_i)]} (i = 1, \ldots, I) \quad (1)$$

where the discrimination $\alpha_i$ and the threshold $\beta_i$ are parameters describing the items and $\theta$ a parameter describing the person responding. The special case of Rasch models appears when the discrimination parameter is constant across items $\alpha_1 = \ldots = \alpha_I$. Usually the $\alpha$'s are fixed at 1 and the variance in the distribution of the latent variable is estimated, but alternatively the variance can be fixed and the common value of the discrimination can be estimated. A technical assumption in both of these models is that items are locally independent

$$P(X_1 = x_1, \ldots, X_I = x_I | \theta)$$
$$= \prod_{i=1}^{I} P(X_i = x_i | \theta) \text{ for all } \theta \in \mathbb{R}. \quad (2)$$

and furthermore that persons respond independently of each other. For persons $v = 1, \ldots, N$ with response vectors $\overline{X}_1, \ldots, \overline{X}_N$, these two independence assumptions yields the joint likelihood

$$L(\overline{\beta}, \theta_1, \ldots, \theta_N | \overline{x}_1, \ldots, \overline{x}_n) = \prod_{v=1}^{N} Pr(\overline{X}_v = \overline{x}_v | \theta_v)$$
$$= \frac{\exp[\sum_v \theta_v \sum_{i=1}^{I} \alpha_i x_{vi} - \sum_i \alpha_i \beta_i x_{.i}]}{\prod_{v=1}^{N} \prod_{i=1}^{I} [1 + \exp(\alpha_i (\theta_v - \beta_i))]}. \quad (3)$$

The model is only identified if restrictions are placed on either the item parameters or the latent variable. One option is to assume that either $\sum_{i=1}^{I} \beta_i = 0$ or $\sum_{v=1}^{N} \theta_v = 0$ and $\prod_{i=1}^{I} \alpha_i = 1$ or $\prod_{v=1}^{N} \theta_v = 1$. Estimation based on the likelihood (3) leads to inconsistent estimates (Neyman & Scott, 1948). For this reason, either marginal maximum likelihood (MML) estimation (Bock & Aitkin, 1981; Thissen, 1982; Zwinderman & van den Wollenberg, 1990), or in the

special case of Rasch models conditional maximum likelihood (CML) estimation (Andersen, 1973) can be used.

## The longitudinal 2PL model

For time points $t = 1, \ldots, T$ let $X_{1t}, \ldots, X_{It}$ be a set of dichotomous items measuring a value $\theta_t \in \mathbb{R}$ of the latent variable, where measurements for the same person at two time points $t_1$ and $t_2$ are correlated, $\mathrm{Corr}(\theta_{t_1}, \theta_{t_2}) > 0$. Assume that at all time point $t$ all items $i$ fit the 2PL-model

$$P(X_{it} = x_{it} | \theta_t) = \frac{\exp(x_{it} \alpha_{it} (\theta_t - \beta_{it}))}{1 + \exp(\alpha_i (\theta_t - \beta_{it}))} \quad (4)$$

and that the assumption of local independence (2) holds within time point. It is tempting to further assume that responses to any two items $i$ and $j$ at any two time points $t_1$ and $t_2$ are locally independent

$$P(X_{it_1} = x_{it_1}, X_{jt_2} = x_{jt_2} | \theta_{t_1}, \theta_{t_2}) = P(X_{it_1} = x_{it_1} | \theta_{t_1})$$
$$\times P(X_{jt_2} = x_{jt_2} | \theta_{t_2}) \quad (5)$$

Doing so would lead to the generalization of Eq. 2

$$P(\overline{\overline{X}} = \overline{\overline{x}} | \overline{\theta}) = \prod_{t=1}^{T} \prod_{i=1}^{I} P(X_{it} = x_{it} | \theta_t) \quad (6)$$

where $\overline{\overline{X}} = (X_{it})_{i \in \{1, \ldots, I\}, t \in \{1, \ldots, T\}}$ and $\overline{\theta} = (\theta_t)_{t=1}^{T}$. When (6) in fact holds, then estimation can be done using simple multivariate extensions of the 2PL model. Such extensions have been considered for Rasch models (Andersen, 1985; Embretson, 1991; Adams et al., 1997).

Unfortunately, the assumption (6) might not be justified. It seems plausible that responses to the same item at two different time points are dependent beyond what is explained by the underlying latent variable. Ignoring violations of Eq. 6 can lead to biased estimates of the latent variable (Marais & Andrich, 2008a; Marais, 2009), hence the assumption should be checked.

Formalization of local dependence across time

Henceforth, we consider the situation where responses to item $i$ at different time points $t_1 < t_2$ lead to a violation of Eq. 5. In this case, all we know is that

$$P(X_{i1} = x_{i1}, X_{i2} = x_{i2} | \theta_1, \theta_2) = P(X_{i1} = x_{i1} | \theta_1)$$
$$\times P(X_{i2} = x_{i2} | X_{i1} = x_{i1}; \theta_2) \quad (7)$$

where indices 1 and 2 are short for $t_1$ and $t_2$, respectively. Hence, taking account of the local dependence means

finding a way of modeling the conditional probabilities in Eq. 7. One option is to again turn to the 2PL model and assume that

$$
\begin{aligned}
&P(X_{i2} = x_{i2}|X_{i1} = x_{i1}; \theta_2) \\
&= \frac{\exp[x_{i2}\alpha_{i2}^*(x_{i1})(\theta_2 - \beta_{i2}^*(x_{i1}))]}{1 + \exp[\alpha_{i2}^*(x_{i1})(\theta_2 - \beta_{i2}^*(x_{i1}))]}
\end{aligned}
\tag{8}
$$

where $\alpha_{i2}^*(x_{i1})$ and $\beta_{i2}^*(x_{i1})$ are new item parameters depending on the response given to the item at time $t_1$. If no local dependence across time points is present the parameters will coincide $\alpha_{i2}^*(0) = \alpha_{i2}^*(1)$ and $\beta_{i2}^*(0) = \beta_{i2}^*(1)$.

### Detection of local dependence across time

Methods for detecting local dependence across time points in longitudinal IRT models have not received much attention in the literature. One exception is a paper by Olsbjerg and Christensen (2013a) where two tests in Rasch models are suggested. One of the tests discussed is to consider more general models that include interaction terms to account for local dependence between items at different time points. The longitudinal Rasch model can then be tested against these using likelihood ratio tests. This test can easily be extended to the 2PL model. The second test exploits sufficiency of the sum score and use Mantel–Haenszel tests for association between items conditioning on this. This works only for Rasch models.

## Modeling local dependence

The formalization (8) results in joint probabilities that are products of 2PL model probabilities

$$
\begin{aligned}
&P(X_{i1} = x_{i1}, X_{i2} = x_{i2}|\theta_1, \theta_2) \\
&= P(X_{i1} = x_{i1}|\theta_1)P(X_{i2} = x_{i2}|X_{i1} = x_{i1}; \theta_2) \\
&= \frac{\exp[x_{i1}\alpha_{i1}(\theta_1 - \beta_{i1})]}{1 + \exp[\alpha_{i1}(\theta_1 - \beta_{i1})]} \frac{\exp[x_{i2}\alpha_{i2}^*(x_{i1})(\theta_2 - \beta_{i2}^*(x_{i1}))]}{1 + \exp[\alpha_{i2}^*(x_{i1})(\theta_2 - \beta_{i2}^*(x_{i1}))]}
\end{aligned}
\tag{9}
$$

In the special case of Rasch, Eq. 9 looks very similar to the longitudinal Rasch models of Andersen (1985) and Embretson (1991) where items are locally independent across time points. Except that in Eq. 9 the item parameter at time $t_2$ depends on the observed response at time $t_1$. Fitting this model is a matter of including the response at time point $t_1$ as a covariate that requires a modification of existing IRT software. An alternative way to go about it, well known in the framework of IRT models, is to recode (split) the items as illustrated in section "Item splitting".

### Item splitting

Item splitting is a standard method for handling differential item functioning (DIF), where the response probabilities of an item differ across subpopulations such as males and females. The idea of splitting items was proposed for Rasch models by Andrich and Kreiner (2010) in the context of quantifying the magnitude of local dependence.

Item splitting, used as a tool for modeling local dependence, can be adapted to longitudinal IRT models as follows. A dependent time $t_2$ item $X_{i2}$ is split into two new items $X_{i2}^*(0)$ and $X_{i2}^*(1)$ representing the group of persons with a response at time $t_1$ of 0 and 1, respectively. Each person only contributes with an observed value to one of the new items, cf. Fig. 1.

### Estimation of item parameters

Estimation of item parameters can be done using marginal maximum likelihood (MML) estimation assuming that $\overline{\theta} = (\theta_1, \theta_2)^T$ follows a two-dimensional normal distribution

$$
\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim \mathcal{N}_2\left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix} \right)
\tag{10}
$$

where $\rho = \text{Corr}(\theta_1, \theta_2)$ represents the latent correlation. Often the main interest is to estimate changes in the mean and variance leading to the reparameterization

$$
\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim \mathcal{N}_2\left( \begin{bmatrix} 0 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_2\rho \\ \sigma_2\rho & \sigma_2^2 \end{bmatrix} \right).
$$

Let $\varphi_{\overline{\mu}, \Sigma}$ denote the density of Eq. 10. Then, for persons $v = 1, ..., N$ with response vectors



**Fig. 1** Illustration of item splitting. To the left is a response matrix before the item splitting. To the right is the response matrix after splitting the dependent item $X_{i2}$ into $X_{i2}^*(0)$ and $X_{i2}^*(1)$

$\overline{X}_v = (X_{v11}, ..., X_{vI1}, X_{v12}, ..., X_{vI2})$, the marginal likelihood has the form

$$
L_M(\overline{\alpha}, \overline{\beta}, \overline{\mu}, \Sigma \,|\, \overline{x}_1, ..., \overline{x}_N)
$$

$$
= \prod_{v=1}^{N} \int_{\mathbb{R}^2} P(\overline{X}_v = \overline{x}_v | \overline{\theta}) \varphi_{\overline{\mu}, \Sigma}(\overline{\theta}) d\overline{\theta} \tag{11}
$$

where $\overline{\alpha}$ and $\overline{\beta}$ denote the vectors of item discriminations and thresholds respectively, for both split and unsplit items. The probabilities in Eq. 11 are given by

$$
P(\overline{X}_v = \overline{x}_v | \overline{\theta}) = \prod_{i=1}^{I} P(X_{vi1} = x_{vi1} | \theta_1) P(X_{vi2}
$$

$$
= x_{vi2} | X_{vi1} = x_{vi1}; \theta_2) \prod_{i=1}^{I} \left( \frac{\exp[x_{vi1}\alpha_{i1}(\theta_1 - \beta_{i1})]}{1 + \exp[\alpha_{i1}(\theta_1 - \beta_{i1})]} \right)
$$

$$
\left( \frac{\exp[x_{vi2}\alpha_{i2}^*(x_{vi1})(\theta_2 - \beta_{i2}^*(x_{vi1}))]}{1 + \exp[\alpha_{i2}^*(x_{vi1})(\theta_2 - \beta_{i2}^*(x_{vi1}))]} \right). \tag{12}
$$

We can not measure change in the latent variable with a measurement instrument that changes completely. For that reason, it is important that some of the items remains the same, across time points, in the sense that

$$
\alpha_{i1} = \alpha_{i2}^*(0) = \alpha_{i2}^*(1) \qquad \text{and} \qquad \beta_{i1} = \beta_{i2}^*(0) = \beta_{i2}^*(1)
$$

for $i$ in some (reasonably sized) subset $\mathcal{I}_0 \subseteq \{1, ..., I\}$. Again, restrictions on the parameters are needed in order for the model (11) to be identified. At each time point we have to put restrictions on either the item or population parameters. One option is to require that either $\sum_{i=1}^{I} \beta_{i1} = 0$ and $\prod_{i=1}^{I} \alpha_{i1} = 1$, or that $\mu_1 = 0$ and $\sigma_1^2 = 1$, and similarly at time $t_2$.

The model (11) can be estimated in SAS using the NLMIXED procedure for fitting nonlinear mixed models. MML estimation is carried out by maximizing an approximation to the likelihood (11) integrating out the random effects. The SAS macro %LRASCH_MML (Olsbjerg & Christensen, 2013b) is an implementation of the special case of Rasch models. In this implementation, adaptive Gaussian quadrature is used for integral approximation and the Newton-Raphson algorithm for optimization. The SAS macro %LRASCH_MML (Olsbjerg & Christensen, 2013b) works for incomplete data, and for this reason, local dependence across time points can be modeled according to Eq. 8. Whether the effect of local dependence is significant can be evaluated by comparing the likelihood in Eq. 11 to the likelihood of the simple model based on the assumption of local independence (6) in a likelihood ratio test.

Estimation of person location parameters

Usually, it is of interest to estimate change at the individual level. In the previous section, we described how the item parameters can be estimated by assuming a certain distribution for the latent vector and then maximizing an approximation to the marginal likelihood. Estimation of the person parameters can be carried out in a similar manner by substituting the item parameters in Eq. 12 by their MML estimates $\widehat{\overline{\alpha}}$ and $\widehat{\overline{\beta}}$ resulting in the likelihood function

$$
L_M(\overline{\theta} \,|\, \overline{x}_1, ..., \overline{x}_N) = L(\widehat{\overline{\alpha}}, \widehat{\overline{\beta}}, \overline{\theta} \,|\, \overline{x}_1, ..., \overline{x}_N) \tag{13}
$$

where

$$
L(\widehat{\overline{\alpha}}, \widehat{\overline{\beta}}, \overline{\theta} \,|\, \overline{x}_1, ..., \overline{x}_N) = \prod_{v=1}^{N} P(\overline{X}_v = \overline{x}_v | \overline{\theta}_v)
$$

is the joint likelihood with the estimated item parameters inserted. This corresponds to assuming a distribution of the item parameters which is degenerate in their estimated values. Estimates of $\overline{\theta}_v$ can be derived from Eq. 13 by numerical optimization such as Newton–Raphson.

Thus, we can compute estimates $\widehat{\overline{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$ and change scores $\hat{\theta}_2 - \hat{\theta}_1$ for each person. This can be done in the model assuming local independence across time points and in a model that takes local dependence across time points into account. It is then possible to evaluate whether local dependence across time points affects the conclusions about individuals.

## Simulation study

A simulation study was conducted to illustrate the implications of local dependence across time points and to illustrate the advantage of splitting items. Responses were simulated by (i) simulating person parameters from a two-dimensional normal distribution

$$
\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim \mathcal{N}_2 \left( \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right)
$$

(ii) simulating responses at time 1 from a 2PL model given the person and item parameters, and (iii) simulating responses at time 2 from the same dichotomous Rasch model with the exception that item thresholds for the locally dependent items were shifted by 1. More specifically, item thresholds for these items were given as

$$
\beta_{i2}^*(0) = \beta_{i1} + 1 \qquad \text{and} \qquad \beta_{i2}^*(1) = \beta_{i1} - 1.
$$

This means that at time 2 the item becomes more difficult for those with a wrong response at time 1 and easier for those with a correct response at time 1 in the case where 0

and 1 represent an incorrect and correct response, respectively. In each case we simulated data sets with no change in the population mean across time points ($\mu = 0$) and data sets where a change in the population mean ($\mu = 0.5$) was present. We simulated 100 data sets with responses from $N = 500$ persons at two time points, to 24 dichotomous items with equidistant thresholds ranging from $-2$ to 2. We simulated data sets where the assumption of local independence across time points was violated for two, four, six, eight, and ten items, respectively. Furthermore, we simulated data sets with dependence for difficult items and dependence for easy items, respectively. An overview of the simulation setups is provided in Table 1.

For each simulated data set, we considered two ways of splitting items, no splitting and splitting of the item(s) with dependence. The effects of splitting an item without local dependence across time points were studied in some but not all simulations. Overview of the simulation study:

**Table 1** Overview of the simulation study

| LD item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Simulation setups | | | |
| 1 | | | | x | | x | x | | |
| 2 | | | | | x | | | | |
| 3 | | | x | x | | | x | | |
| 4 | | x | | | | | | | |
| 5 | x | | | x | x | | x | | |
| 6 | | | | | | x | | | |
| 7 | | | | | | | x | | |
| 8 | | x | x | x | x | | | | |
| 9 | | | | | | | x | | |
| 10 | | | | | | | | | |
| 11 | | | x | x | x | x | x | | |
| 12 | | | | | | | | | |
| 13 | | | | | | | | | |
| 14 | | | x | x | x | | | x | x |
| 15 | | | | | | | | | |
| 16 | | | | | | | | | x |
| 17 | | x | x | x | x | | | | |
| 18 | | | | | | | | | x |
| 19 | | | | | | | | x | |
| 20 | x | | | x | x | | | | x |
| 21 | | x | | | | | | | |
| 22 | | | x | | x | | | | x |
| 23 | | | | | x | | | | |
| 24 | | | | | x | | | x | x |

Setups 1 through 5 illustrate the effect of the number of affected items. Simulation setups 6 and 7 and simulation setups 6 and 7 illustrate the impact when the local dependence is for easy and difficult items, respectively

setups 1 through 5 illustrate that the effect of local dependence across time increases with the number of affected items. Simulation setups 6 and 7 and simulation setups 6 and 7 illustrate the impact when the local dependence is for easy and difficult items, respectively. Each simulated data set was, after recoding locally dependent items (cf. Fig. 1), analyzed in SAS using PROC NLMIXED. The results are shown in Table 2.

The results clearly show that the latent correlation $\rho$ was only recovered in the rows labeled (ii), where the dependent item(s) were split. In rows labeled (i), where the dependent items were not split, $\rho$ was consistently overestimated. When the number of dependent items increases the extent to which the latent correlation $\rho$ is overestimated increases. In the part of Table 2 where the population mean changes over time ($\mu = 0.5$) the change in mean was only recovered in where dependent item(s) were split. In setup (i), where no items were split, the mean change was underestimated. The extent to which the mean change was underestimated increased with the number of dependent items. There appeared to be little impact on the item parameter estimates (results not shown). The results of this simulation study indicate that local dependence across time points has an impact on estimates of the mean change in the latent variable, and on estimates of the latent correlation. The simulation study did not provide evidence of different impact of local dependence for difficult and easy items, respectively.

# Example

The methods proposed in this paper will be illustrated using data collected in the Bradford Metropolitan Health District. A total of 113 persons with episodes of low back pain responded to the Roland and Morris (1983) disability questionnaire, consisting of 24 dichotomous items, at two time points. Further details about the data collection can be found elsewhere (Waxman et al., 1998).

We first analyzed the responses by fitting the unidimensional 1PL and 2PL models at each time point separately. Likelihood ratio tests were used to compare the two models. The results, summarized in Table 3, provide no justification for fitting the simpler 1PL model at any of the two time points.

To take into account the longitudinal nature of the data, we fitted the generalized longitudinal 2PL model (9), which handles both items with time-varying parameters and local dependence across time points. In order to identify items with such characteristics likelihood ratio tests were conducted. These tests are described for the 1PL model (Olsbjerg & Christensen, 2013a), but are easily implemented in SAS for both the 1PL and the 2PL model. The

**Table 2** Simulation study results, 100 simulated data sets, N = 500 persons, 24 items (equidistant $\beta_1 = -2, \ldots, \beta_{24} = 2$)

| Simulation setup | | True value $\mu=0$ | | | | True value $\mu=0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\mu}$ | | $\hat{\rho}$ | | $\hat{\mu}$ | | $\hat{\rho}$ | |
| | | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 1 | (i) | 0.01 | 0.06 | 0.50 | 0.04 | 0.51 | 0.06 | 0.51 | 0.04 |
| | (ii) | 0.01 | 0.05 | 0.52 | 0.04 | 0.50 | 0.06 | 0.56 | 0.04 |
| 2 | (i) | 0.04 | 0.06 | 0.51 | 0.04 | 0.51 | 0.05 | 0.51 | 0.05 |
| | (ii) | -0.02 | 0.05 | 0.55 | 0.05 | 0.50 | 0.05 | 0.57 | 0.04 |
| 3 | (i) | 0.00 | 0.06 | 0.51 | 0.05 | 0.50 | 0.05 | 0.50 | 0.05 |
| | (ii) | 0.00 | 0.05 | 0.60 | 0.04 | 0.47 | 0.05 | 0.59 | 0.05 |
| 4 | (i) | 0.00 | 0.06 | 0.50 | 0.05 | 0.52 | 0.06 | 0.50 | 0.05 |
| | (ii) | 0.00 | 0.05 | 0.62 | 0.04 | 0.49 | 0.06 | 0.62 | 0.04 |
| 5 | (i) | 0.00 | 0.06 | 0.50 | 0.05 | 0.50 | 0.06 | 0.50 | 0.05 |
| | (ii) | 0.00 | 0.05 | 0.63 | 0.04 | 0.46 | 0.04 | 0.64 | 0.04 |
| 6 | (i) | 0.00 | 0.05 | 0.50 | 0.05 | 0.51 | 0.05 | 0.50 | 0.04 |
| | (ii) | -0.02 | 0.05 | 0.54 | 0.05 | 0.48 | 0.06 | 0.54 | 0.05 |
| 7 | (i) | 0.00 | 0.04 | 0.50 | 0.05 | 0.53 | 0.06 | 0.50 | 0.05 |
| | (ii) | 0.00 | 0.06 | 0.56 | 0.06 | 0.50 | 0.06 | 0.56 | 0.06 |
| 8 | (i) | 0.00 | 0.05 | 0.51 | 0.06 | 0.53 | 0.06 | 0.51 | 0.06 |
| | (ii) | 0.00 | 0.05 | 0.60 | 0.06 | 0.51 | 0.06 | 0.60 | 0.06 |
| 9 | (i) | 0.01 | 0.06 | 0.51 | 0.05 | 0.54 | 0.06 | 0.51 | 0.05 |
| | (ii) | -0.01 | 0.05 | 0.64 | 0.05 | 0.50 | 0.06 | 0.63 | 0.05 |

Average estimate of change in the latent mean $\mu$ and latent correlation $\rho$ (true value 0.5). Assumption of local independence across time points is violated for items 5, 20 (1), items 4, 8, 17, 21 (2), items 3, 8, 11, 14, 17, 22 (3), items 2, 5, 8, 11, 14, 17, 20, 23 (4), items 1, 3, 5, 8, 11, 13, 17, 20, 22, 24 (5), items 1, 6, 11 (6), items 1, 3, 5, 7, 9, 11 (7), items 14, 19, 24 (8), items 14, 16, 18, 20, 22, 24 (9), (i): correct item splitting, (ii): no item splitting

assumption of time-invariant item parameters was first investigated graphically using the estimates derived from the two unidimensional models. Centralized item thresholds at the two time points are plotted against each other in Fig. 2 and standardized item discriminations in Fig. 3.

Further investigation of the assumption of time-invariant item parameters was done using a likelihood ratio test proposed by Olsbjerg and Christensen (2013a). This test is for DIF with respect to time based on MML estimation where, for each person, only responses at a single (randomly chosen) time point are used in order to get rid of possible local dependence. These tests suggested that item parameters for

**Table 3** Likelihood ratio tests of the 1PL model against the 2PL model, separately at each time point

| Time point | Model | −2 loglikelihood | p-value |
|---|---|---|---|
| 1 | 2PL | 2348.3 | < 0.0001 |
| | 1PL | 2429.3 | |
| 2 | 2PL | 2279.1 | 0.0003 |
| | 1PL | 2334.9 | |
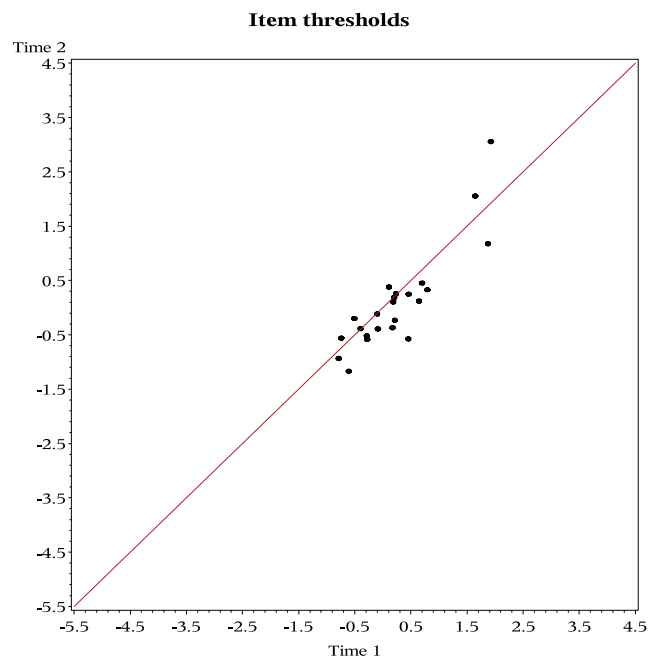


**Fig. 2** Item thresholds at time 1 and time 2
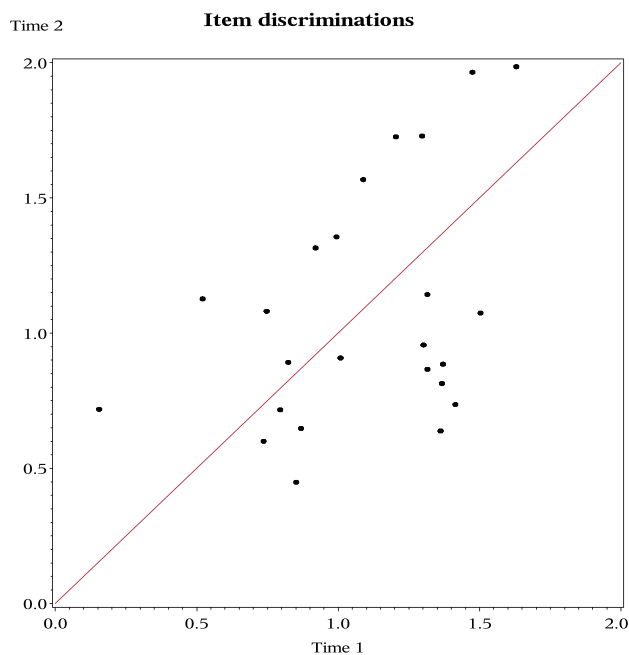
**Fig. 3** Item discriminations at time 1 and time 2

three items (items 2, 15, and 21) change over time. However, in light of the large number of statistical tests and the

inherent risk of type I, we disregarded this evidence and we continued the analyses assuming that all item parameters were the same at both time points. This corresponds well with what was observed in the plots in Figs. 2 and 3, where the estimates were relatively evenly scattered around the diagonal. As expected, the variation in the estimated item discriminations across time points is noticeably larger than for the thresholds.

In order to identify items with local dependence across time points, likelihood ratio tests were conducted. These tests were proposed by (Olsbjerg and Christensen 2013a) for Rasch models but can easily be extended to the 2PL model. A total of eight out of the 24 items showed significant evidence of local dependence across time points. The results are summarized in Table 4.

We chose as our final model the 2PL model that incorporates local dependence for the eight items with $p$-values below 5 %. Estimates from this model can then be compared to the estimates from the simpler 2PL model that ignores the dependence and where all items have equal parameters at the two time points. In Table 5, estimates of the population mean change and the latent correlation derived from the two models are displayed. The estimated population mean change differs only slightly between the two models. The

**Table 4** Item wording, estimated item parameters, and likelihood ratio tests for local independence across time points

| no | Item wording | $\alpha$ (s.e.) | $\beta$ (s.e.) | $\chi^2$ | DF | $p$-value |
|---|---|---|---|---|---|---|
| 1 | stay at home | 2.66 (0.61) | 1.56 (0.21) | 4.1 | 4 | 0.3926 |
| 2 | change position | 0.59 (0.20) | -2.13 (0.73) | 15.8 | 4 | 0.0033 |
| 3 | walk more slowly | 1.96 (0.35) | 0.57 (0.14) | 6.6 | 4 | 0.1586 |
| 4 | not doing jobs | 2.39 (0.46) | 0.96 (0.16) | 8.7 | 4 | 0.0691 |
| 5 | use handrail | 1.58 (0.32) | 1.40 (0.22) | 5.5 | 4 | 0.2397 |
| 6 | lie down | 1.77 (0.33) | 0.89 (0.17) | 14.0 | 4 | 0.0073 |
| 7 | hold on | 1.43 (0.28) | 1.20 (0.21) | 3.6 | 4 | 0.4628 |
| 8 | other people | 2.49 (0.52) | 1.36 (0.19) | 10.5 | 4 | 0.0328 |
| 9 | dressed more slowly | 1.75 (0.33) | 0.84 (0.16) | 5.7 | 4 | 0.2227 |
| 10 | stand up short periods | 1.32 (0.25) | 0.58 (0.17) | 16.8 | 4 | 0.0021 |
| 11 | try not to bend | 1.59 (0.29) | 0.55 (0.15) | 2.0 | 4 | 0.7358 |
| 12 | get out chair | 1.54 (0.30) | 1.20 (0.20) | 4.2 | 4 | 0.3796 |
| 13 | painful all the time | 1.17 (0.23) | 0.14 (0.16) | 7.4 | 4 | 0.1162 |
| 14 | turn over in bed | 1.14 (0.23) | 0.75 (0.19) | 6.5 | 4 | 0.1648 |
| 15 | appetite not good | 0.98 (0.31) | 3.33 (0.84) | 18.9 | 4 | 0.0008 |
| 16 | trouble putting on socks | 1.25 (0.24) | 0.56 (0.17) | 18.6 | 4 | 0.0009 |
| 17 | walk short distances | 2.09 (0.40) | 1.16 (0.18) | 9.6 | 4 | 0.0477 |
| 18 | sleep less well | 0.94 (0.20) | 0.08 (0.18) | 8.4 | 4 | 0.0780 |
| 19 | dress with help | 1.55 (0.46) | 2.88 (0.55) | 3.0 | 4 | 0.5578 |
| 20 | sit down most the day | 2.77 (0.64) | 1.58 (0.21) | 0.8 | 4 | 0.9385 |
| 21 | avoid heavy jobs | 1.54 (0.28) | 0.25 (0.15) | 19.8 | 4 | 0.0006 |
| 22 | more irritable | 1.30 (0.25) | 0.93 (0.20) | 31.3 | 4 | 0.0000 |
| 23 | upstairs more slowly | 1.71 (0.33) | 1.13 (0.19) | 3.5 | 4 | 0.4779 |
| 24 | stay in bed | 1.93 (0.57) | 2.61 (0.44) | 5.2 | 4 | 0.2674 |

**Table 5** Population parameter estimates for Model 1, the simple model where no items are split for dependence and Model 2, the final model where eight items are split for dependence

| Parameter | Model | Estimate | SE |
|---|---|---|---|
| Population mean change $\mu_2 - \mu_1$ | 1 | 0.231 | 0.09 |
| | 2 | 0.229 | 0.11 |
| Latent correlation $\rho$ | 1 | 0.71 | 0.06 |
| | 2 | 0.65 | 0.07 |

estimated latent correlation decreases when items are split for dependence and this corresponds well with what was observed in the simulation study.

To get a sense of how it might affect the individual respondents to ignore local dependence as compared to modeling it, we investigate the estimated person locations in the simple model and in the model taking local dependence across time into account. Keeping the item parameters fixed at their estimated values, we consider person location estimates and estimates of the change

$$\Delta = \theta_2 - \theta_1$$

in the latent variable for Model 1 and Model 2. The results, presented in Table 6, reveal that generally the difference between the two models is not that big.

The two models do not differ with respect to the estimated time 1 person locations, but there appears to be more variability in the incorrect Model 1. Regarding the person locations at the second time point, there is a difference in that Model 1 appears to underestimate the values. Again, there is more variability in the incorrect model. The difference between the person time 2 person location estimates from Model 1 and Model 2 were larger than the differences between time 1 person location estimates.

Regarding change scores there was more variability when using the more correct model. A substantial variation in the difference between $\Delta$ values assigned to individuals was observed. A consequence of this variation is that the two models did not agree about who had a significant change

score. This occurred for five out of the 113 people responding: 3 (2.7 %) who were considered to have a significant $\Delta$ value by Model 2, but not by Model 1 and 2 (1.8 %) who were considered to have a significant $\Delta$ value by Model 1 only.

## Discussion

The assumption of local independence in unidimensional IRT models has been the focus of much research (Hoskens & De Boeck, 1997; Douglas et al., 1998; Ip, 2000, 2001, 2002; Scott & Ip 2002). In unidimensional IRT models, violations of this assumption can be resolved by changing the wording or the response categories of the items. Another solution is to use the sum of the dependent items as a so-called 'subtest' (Andrich, 1985). Alternatively, a testlet model (Bradlow et al., 1999; Wang & Wilson, 2005) or other models taking account of local dependence (Hoskens & De Boeck, 1997; Ip, 2002) can be applied. In Rasch models, a simple way of quantifying local dependence has been proposed (Andrich and Kreiner, 2010).

In longitudinal studies where the same instrument is used at several occasions to measure change in a latent variable, the assumption of local independence across time points may well be violated. In that case, there is usually no desire to change the item content, and collapsing items across time points does not make sense in the context of measuring change.

This paper described how local dependence across time points can be modeled in longitudinal IRT models. Based on the method of item splitting by Andrich and Kreiner (Andrich and Kreiner (2010)), which has so far only been used for quantification of local dependence in unidimensional models, we proposed a general method that can be used to test for and model local dependence across time points. It should be noted that in the Andrich and Kreiner approach, the first item is discarded after recoding the second item, but that in the method proposed here we keep the time 1 item along both versions of the time 2 item. Because the method is based on item splitting, a concept well known

**Table 6** Summary of the estimated person locations ($\theta_1$ and $\theta_2$) and of the individual change scores $\Delta = \theta_2 - \theta_1$ for the 113 persons in the data example. Model 1 is the simple model where no items are split for dependence and Model 2 the true model where the eight items showing signs of local dependence are split

| | Model 1 | | Model 2 | | Difference | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | (95 % ref. interval) |
| $\hat{\theta}_1$ | 0.003 | 1.09 | 0.021 | 1.06 | −0.02 | (−0.32 to 0.15) |
| $\hat{\theta}_2$ | 0.145 | 1.13 | 0.347 | 1.04 | −0.02 | (−0.51 to 0.47) |
| $\hat{\Delta}$ | 0.214 | 1.03 | 0.289 | 1.10 | −0.01 | (−0.554 to 0.488) |
| % significant | 17 % | | 18 % | | | |

for resolving differential item functioning (DIF), it can be used in existing software such as RUMM (Andrich et al., 2010)

The simulations incorporated local dependence in a way that made it more likely for a person to give the same response to an item at the two time points, than it would have been in the case of local independence. By fitting the simple model assuming local independence across time points for all items, we demonstrated some of the effects of ignoring local dependence, one of them being that the dependence was mistakenly accounted for by the latent correlation $\rho$, which as a result was overestimated. These patterns were visible in simulations with only a single dependent item and became even more pronounced when more items were dependent. Estimation of the change in the mean of the latent variable was also affected by local dependence across time points. Results suggest that when local dependence occurs for items located in the middle of the latent continuum (as was the case in the simulation study) the change in the mean of the latent variable is underestimated, whereas dependence for items located in one end of the continuum (as in the simulation study) lead to overestimation of the change in the mean of the latent variable. This finding corresponds well with Marais (2009) who found that in certain circumstances local dependence will exaggerate changes and in other circumstances it will mask them.

In the data example, an effect on the latent correlation, similar to those of the simulations, was found when dependence was ignored, but regarding the estimated mean change of the latent variable, no difference was found when splitting the items with signs of local dependence.

In the data examples and in the simulation study, estimation was done PROC NLMIXED estimating item parameters and the two-dimensional latent distribution by MML estimation. For the special case of the Rasch (1PL) model, estimation can also be carried out in RUMM (Andrich et al., 2010) by splitting dependent items, fitting the model at each time point and then comparing person estimates. We considered dichotomous items administered at two time points. The proposed method is easily generalized to polytomous items because it is based on the simple concept of item splitting. For that reason, it can also be adapted to other IRT models. In principle, it is also straightforward to make extensions to accommodate more than two time points. However, estimating the latent correlation matrix can become computationally challenging. Moreover, the splitting procedure can potentially become quite complex and lead to situations with sparse data if we allow for dependence structures that go beyond items at two consecutive time points.

Differential item functioning identified in relation to time is another phenomenon that occurs in longitudinal studies (Specht et al., 2011). This is often called item parameter drift (Wells et al., 2002; Miller & Fitzpatrick, 2009) and the assumption that item parameters are stable over time should also be tested. Different methods for detection of item parameter drift exist (Donoghue & Isham, 1998; DeMars, 2004; Galdin & Laurencelle, 2010). In the data example and in the simulation study, we assumed that item parameters were stable over time, but for the special case of Rasch models, the SAS macro %LRASCH_MML (Olsbjerg & Christensen, 2013b) can be used to test this assumption using likelihood ratio tests. Since local response dependence formulated using item splitting is also implemented, this yields a modeling framework where items that change over time and items with local dependence across time points can be included. This model can be fitted using two-dimensional MML including three types of items: items with local dependence across time points, items with DIF across time points, and items with neither. Items with local dependence across time points are included by splitting the time 2 item and estimating $(\alpha_{i2}^*(0), \beta_{i2}^*(0))$ and $(\alpha_{i2}^*(1), \beta_{i2}^*(1))$, items with DIF across time points are included by the restriction $\alpha_{i2}^*(0) = \alpha_{i2}^*(1)$ and $\beta_{i2}^*(0)) = \beta_{i2}^*(1))$, and items with neither can be included by the further restriction $\alpha_{i2}^*(0) = \alpha_{i2}^*(1) = \alpha_{i1}$ and $\beta_{i2}^*(0) = \beta_{i2}^*(1) = \beta_{i1}$.

For investigating local dependence across time points for a single item, the proposed methodology yields a likelihood ratio test. In a realistic situation with many items in a test, care must be taken to control the type I error rate by adjusting for multiple testing using e.g., the Benjamini–Hochberg (1995) procedure.

The methodology outlined in this paper is a simple way of accounting for local item dependence, while local person dependence was not discussed. However, this can also occur, and quite general methods for handling this using a four-level IRT model to simultaneously account for dual local dependence due to item clustering and person clustering have been proposed by (Jiao et al. 2012). The simple approach of taking dependence into account using item splitting proposed in this enables researchers to include local dependence in simpler multilevel models (Kamata, 2001). However, when an item $X_{i2}$ is split for local dependence, fewer persons will contribute to the estimation the new item parameters $\beta_{i2}^*(0)$ and $\beta_{i2}^*(1)$ than to the original parameter $\beta_{2i}$. Hence, splitting items is at the cost of precision of the item parameter estimates and should only be done when there is evidence of local dependence. In the data analysis, the sample size was 113 and the 2PL model was fitted to the data, but the proposed methods were able to disclose evidence of local dependence across time points, and to model these. However, the proposed methodology should not

uncritically be used in applications with small sample sizes.

# References

Adams, R.J., Wilson, M., Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Andersen, E.B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31–44.

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In Embretson, S. (Ed.) *Test design: Contributions from psychology, education and psychometrics, chapter 9* (pp. 245–273). New York: Academic Press.

Andrich, D., Humphry, S., Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement*, *36*, 309–324.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, *34*, 181–192.

Andrich, D., Sheridan, B., Luo, G. (2010). *RUMM2030 [Computer software and manual]*. Perth, Australia: RUMM Laboratory.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, *57*, 289–300.

Birnbaum, A.L. (1968). Latent trait models and their use in inferring an examinee's ability. In Lord, F.M., & Novick, M.R. (Eds.) *Statistical theories of mental test*. Reading, MA: Addison-Wesley.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an em algorithm. *Psychometrika*, *46*, 443–459.

Bousseboua, M., & Mesbah, M. (2010). Longitudinal latent Markov processes observable through an invariant Rasch model. In Rykov, V.V., Balakrishnan, N., Nikulin, M. (Eds.) *Mathematical and statistical models and methods in reliability* (pp. 87–100). New York: .Springer.

Bradlow, E.T., Wainer, H., Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

DeMars, C.E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*, 265–300.

Donoghue, J.R., & Isham, S.P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*, 33–51.

Douglas, J., Kim, H.R., Habing, B., Gao, E. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*, 129–151.

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–516.

Fischer, G.H., & Molenaar, I.W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Galdin, M., & Laurencelle, L. (2010). Assessing parameter invariance in item response theory's logistic two item parameter model: A Monte Carlo investigation. *Tutorials in Quantitative Methods for Psychology*, *6*, 39–51.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H., & Braun, H.I. (Eds.) *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P.W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

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

Ip, E.H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*, 73–91.

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

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

Jiao, H., Kamata, A., Wang, S., Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82–100.

Jiao, H., Wang, S., Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, *6*(3), 311–321.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

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

Kelderman, H. (1997). Loglinear multidimensional item response models for polytomously scored items. In van der Linden, W., & Hambleton, R. (Eds.) *Handbook of modern item response theory* (pp. 287–304). New York: Springer.

Kreiner, S., & Christensen, K.B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communications in Statistics - Theory and Methods*, *33*, 1239–1276.

Kreiner, S., & Christensen, K.B. (2007). Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models. In von Davier, M., & Carstensen, C.H. (Eds.) *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis*. Houghton Mifflin, Boston, New York, Atlanta, Geneva (IL), Dallas, Palo Alto.

Liu, L.C., & Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, *62*, 261–228.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement*, *10*, 17–29.

Marais, I., & Andrich, D. (2008a). Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*, 105–124.

Marais, I., & Andrich, D. (2008b). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*, 200–215.

Miller, G.E., & Fitzpatrick, S.J. (2009). Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educational and Psychological Measurement*, *69*, 357–368.

Muthén, B. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association*, *74*, 807–811.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Neyman, J., & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16*, 1–32.

Olsbjerg, M., & Christensen, K.B. (2013a). Marginal and conditional approach to longitudinal Rasch models. *Pub. Inst. Stat. Univ. Paris, 57, fasc, 1–2*, 109–126.

Olsbjerg, M., & Christensen, K.B. (2013b). SAS macro for marginal maximum likelihood estimation in longitudinal polytomous Rasch models. Research Report 11, Department of Statistics, University of Copenhagen.

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

Roland, M., & Morris, R. (1983). A study of the natural history of back pain. 1. development of a reliable and sensitive measure of disability in low back pain. *Spine, 8*, 141–144.

Scott, S., & Ip, E.H. (2002). Empirical Bayes and item clustering effects in a latent variable hierarchical model: A case from the national assessment of educational progress. *Journal of the American Statistical Association, 97*, 409–419.

Specht, K., Leonhardt, J.S., Revald, P., Mandøe, H., Andresen, E.B., Brodersen, J., Kreiner, S., Kjaersgaard-Andersen, P. (2011). No evidence of a clinically important effect of adding local infusion analgesia administrated through a catheter in pain treatment after total hip arthroplasty. *Acta Orthopaedica, 82*(3), 315–320.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics, 9*, 23–30.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149.

Waxman, R., Tennant, A., Helliwell, P. (1998). Community survey of factors associated with consultation for low back pain. *British Medical Journal, 317*, 1564–1567.

Wells, C.S., Subkoviak, M.J., Serlin, R.C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77–87.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Zwinderman, A.H., & van den Wollenberg, A.L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement, 14*, 73–81.