# Using Conditional Association to Identify Locally Independent Item Sets

Straat, J.H.; van der Ark, L.A.; Sijtsma, K.

**DOI**
10.1027/1614-2241/a000115

**Publication date**
2016

**Document Version**
Final published version

**Published in**
Methodology

Original Article

# Using Conditional Association to Identify Locally Independent Item Sets

J. Hendrik Straat,[1] L. Andries van der Ark,[2] and Klaas Sijtsma[3]

[1]Cito, Arnhem, The Netherlands
[2]University of Amsterdam, The Netherlands
[3]Tilburg University, The Netherlands

**Abstract:** The ordinal, unidimensional monotone latent variable model assumes unidimensionality, local independence, and monotonicity, and implies the observable property of conditional association. We investigated three special cases of conditional association and implemented them in a new procedure that aims at identifying locally dependent items, removing these items from the initial item set, and producing an item subset that is locally independent. A simulation study showed that the new procedure correctly identified 89.5% of the model-consistent items and up to 90% of the model-inconsistent items. We recommend using this procedure for selecting locally independent item sets. The procedure may be used in combination with Mokken scale analysis.

**Keywords:** conditional association, local independence, model-fit assessment, monotonicity, nonparametric item response theory, unidimensionality

## Using Conditional Association to Identify Local Dependence

The Unidimensional Monotone Latent Variable Model (UMLVM; Holland & Rosenbaum, 1986) is a nonparametric Item Response Theory (IRT) model. Let $j$ index items, $X_j$ be a polytomous item-score variable having scores $x = 0, \ldots m$, $J$ be the test length, and $\theta$ be the latent variable. Expectation $E(X_j|\theta)$ is the item response function. The three assumptions of the UMLVM are:

1. Unidimensionality: one latent variable $\theta$ explains the data;
2. Local independence: item scores are independent given $\theta$:

$$P(X_1 = x_1, \ldots, X_J = x_J|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta); \quad (1)$$

3. Monotonicity:

$$E(X_j|\theta) \text{ is nondecreasing in } \theta. \quad (2)$$

Item response functions are monotone nondecreasing but not parametrically defined, such as by means of the logistic function. This means that one cannot estimate latent variable $\theta$ numerically. However, for dichotomous items, scored $x = 0, 1$, Grayson (1988) proved that total score $X_+ = \sum_{j=1}^{J} X_j$ ($x_+ = 0, \ldots, J$) can be used to stochastically order persons on $\theta$, thus providing an ordinal scale for $\theta$. Hence, the $X_+$ scale can be used to compare people with respect to their $\theta$ levels. For polytomous items, a weaker form of the stochastical ordering of persons holds (Van der Ark & Bergsma, 2010).

The UMLVM implies an ordinal person scale. Sijtsma and Molenaar (2002) discussed methods for goodness-of-fit research for the UMLVM. The methods either aim at identifying unidimensional item sets or estimating the item response functions for assessing the monotonicity assumption. Methods assessing local independence are rare (Douglas, Kim, Habing, & Gao, 1998; Zhang & Stout, 1999). In this study, we explore the UMLVM property of conditional association (CA; Holland & Rosenbaum, 1986; Rosenbaum, 1984, 1988) for assessing local independence in a set of items.

In what follows, first we discuss three special CA cases for investigating UMLVM data fit. A computational study showed how the three cases may be used to assess UMLVM fit. Second, we discuss a procedure for identifying locally independent sets of items. Third, a simulation study shed light on the specificity and the sensitivity of the new procedure. Fourth, the new procedure was applied to real data.

## Conditional Association

Vector **X** contains the $J$ item scores $X_j$ and is divided in two mutually exclusive but not necessarily exhaustive sets **Y** and **Z**. Let $f_1$ and $f_2$ be nondecreasing functions, $h$ be any function, and let σ and $s$ denote the population and sample covariances, respectively. Holland and Rosenbaum (1986, Theorem 6) proved that the UMLVM implies CA, which is defined as

$$\sigma[f_1(\mathbf{Y}), f_2(\mathbf{Y})|h(\mathbf{Z}) = \mathbf{z}] \geq 0. \qquad (3)$$

Let subscripts $a$, $b$, $c$, $d$ and $i$, $j$, $k$ identify items. Items in **Y** have subscript $a$ or $b$ (i.e., $X_a$ or $X_b$); items in **Z** have subscript $c$; items neither in **Y** nor **Z** have subscript $d$; and for generic use subscripts $i$, $j$, and $k$ are used. We discuss three special cases of CA that illustrate its potential for assessing UMLVM fit (Sijtsma, 2003). We choose $\mathbf{Y} = (X_a, X_b)$ and $f_1(\mathbf{Y}) = X_a$ and $f_2(\mathbf{Y}) = X_b$, and for **Z** a third item $X_c$ (Case 2), the sum score on the other $J - 2$ items excluding $a$ and $b$ (Case 3), while Case 1 starts ignoring **Z**, thus refraining from conditioning. Hence, correlations between items $a$ and $b$ must be nonnegative in the whole group (Case 1), in each subgroup having a particular score on third item $c$ (Case 2), and in each subgroup having a particular sum score on the other $J - 2$ item (Case 3). Formally, the three cases are:

1. Case 1: Ignore $h(\mathbf{Z})$ (Rosenbaum, 1984), then Equation 3 reduces to

$$\sigma(X_a, X_b) \equiv \sigma_{ab} \geq 0. \qquad (4)$$

2. Case 2: Let $h(\mathbf{Z}) = X_c$, then

$$\sigma(X_a, X_b|X_c = x) \equiv \sigma_{ab|c(x)} \geq 0. \qquad (5)$$

3. Case 3: Let $R_{(ab)}$ be the sum score on $J - 2$ items, also known as the rest score; that is, $h(\mathbf{Z}) = R_{(ab)} = \sum_{c \neq a,b} X_c$, then

$$\sigma(X_a, X_b|R_{(ab)} = r) \equiv \sigma_{ab|R(r)} \geq 0. \qquad (6)$$

Positive inter-item covariances in Equations 4–6 support UMLVM fit, but negative values are inconsistent with the UMLVM. A practical problem is the huge number of covariances one has to inspect. For example, for $J = 20$ Likert items with $m + 1 = 5$ ordered answer categories, there are $\binom{J}{2} = 190$ covariances, $\sigma_{ab}$ (Equation 4); $(m+1)\binom{J}{3} = 5,700$ covariances conditional on an item score, $\sigma_{ab|c(x)}$ (Equation 5); and $m(J-2)\binom{J}{2} = 13,680$ covariances conditional on the rest score, $\sigma_{ab|R(r)}$ (Equation 6). The large number of covariances together build a strong case for or against the UMLVM but also raise the question how to combine all the information into one conclusion about UMLVM fit. The problem increases if one also considers other choices for **Y**, **Z**, $f_1(\mathbf{Y})$, $f_2(\mathbf{Y})$ and $h(\mathbf{Z})$, such as $\mathbf{Y} = (X_a, X_b, X_{a^*}, X_{b^*})$, $f_1(\mathbf{Y}) = X_a + X_b$, $f_2(\mathbf{Y}) = X_{a^*} + X_{b^*}$, and defines $h$ on the set of remaining items or a subset thereof. We ignore such complicating cases.

How do Equations 4–6 relate to model violations of unidimensionality and local independence (Equation 1) and monotonicity (Equation 2)? We distinguish two violations of local independence: positive local dependence (PLD; $\sigma_{jk|\theta} > 0$) and negative local dependence (NLD; $\sigma_{jk|\theta} < 0$) (Chen & Thissen, 1997; Rosenbaum, 1988). One needs to know the underlying cause of, for example, $\sigma_{ab|c(x)} < 0$. The cause may be items $a$ and $b$ being PLD, items $a$ and $c$ being PLD, two items being NLD, item $a$ being non-monotone (NM), and so on. We investigated questions like this to be able to decide which conditional covariances can be used for assessing particular model violations.

## Computational Study: Detecting Violations of UMLVM Assumptions

We distinguished 11 true scenarios involving model violations of the ULVM (Table 1). Four types of true scenarios involved PLD item pairs: Type 1: Both PLD items are in **Y**; Type 2: one PLD item is in **Y** and the other is in **Z**; Type 3: one PLD item is in **Y** and the other is neither in **Y** nor in **Z**; and Type 4: one PLD item is in **Z** and the other is neither in **Y** nor in **Z**. Four similar types of true scenarios involved NLD item pairs (types 5–8). Finally, two true scenario types involved a non-monotone item: Type 9: The NM-item is in **Y**; Type 10: the NM-item is in **Z**; and Type 11: the NM-item is neither in **Y** nor in **Z**.

The use of (conditional) covariances in Equations 4–6 to detect model violations requires caution, because Rosenbaum (1988, Theorem 4; superscript 1 in Table 1), Rosenbaum (1988, Theorem 1; superscript 2), and Holland and Rosenbaum (1986, Equation 5; superscript 3) proved analytically that several of these covariances are also positive when the UMLVM fails, thus providing misleading information about model fit. For example, if $\sigma_{ab|\theta} > 0$ (Table 1, row 1), which is inconsistent with the UMLVM, then $\sigma_{ab} > 0$, incorrectly suggesting UMLVM support. Such covariances are useless for misfit detection. Results like this also raise the question whether negative covariances of the types in Equation 4–6 have enough power to detect model

 

**Table 1.** Proportion of negative values for three types of covariances (columns), for 11 types of true scenarios in which a model violation occurred (rows)

| | True scenario | | | Type of covariance | | |
|---|---|---|---|---|---|---|
| Violation | Type | | Description | $\sigma_{ab}$ | $\sigma_{ab\|c(x)}$ | $\sigma_{ab\|R(r)}$ |
| PLD | 1 | PLD (*a*, *b*) | Both PLD items in **Y** | 0[1] | 0[1] | 0[1] |
| | 2 | PLD (*a*, *c*) or PLD (*b*, *c*) | One PLD item in **Y**, the other in **Z** | 0[1,2] | 314[8] | .318[9] |
| | 3 | PLD (*a*, *d*) or PLD (*b*, *d*) | One PLD item in **Y**, the other neither in **Y** nor **Z** | 0[1,2] | 0[1,2] | [4] |
| | 4 | PLD (*c*, *d*) | One PLD item in **Z**, the other neither in **Y** nor **Z** | 0[1,2] | 0[1,2] | [4] |
| NLD | 5 | NLD (*a*, *b*) | Both PLD items in **Y** | .497[5] | .652[6] | .774[7] |
| | 6 | NLD (*a*, *c*) or NLD (*b*, *c*) | One NLD item in **Y**, the other in **Z** | .0[2] | .000[8] | .000[9] |
| | 7 | NLD (*a*, *d*) or NLD (*b*, *d*) | One NLD item in **Y**, the other neither in **Y** nor **Z** | 0[2] | 0[2] | [4] |
| | 8 | NLD (*c*, *d*) | One NLD item in **Z**, the other neither in **Y** nor **Z** | 0[2] | 0[2] | [4] |
| NM | 9 | NM (*a*) or NM (*b*) | The NM-item in **Y** | .000[10] | .000[11] | .000[12] |
| | 10 | NM (*c*) | The NM-item in **Z** | 0[3] | 0[3] | 0[3] |
| | 11 | NM (*d*) | The NM-item neither in **Y** nor **Z** | 0[3] | 0[3] | [4] |

*Notes.* 1 = Power equals 0 by definition (Rosenbaum, 1988, Theorem 4);
2 = Power equals 0 by definition (Rosenbaum, 1988, Theorem 1);
3 = Power equals 0 by definition (Holland & Rosenbaum, 1986, Equation 5);
4 = Case is impossible under the true scenario type;
5 = Power estimated by averaging $P(\sigma_{12} < 0)$ across all levels of discrimination;
6 = Power estimated by averaging $P(\sigma_{(12|c(x)} < 0)$ across $c$ = 3, 4, 5; $x$ = 0,. . .,4; and all levels of discrimination;
7 = Power estimated by averaging $P(\sigma_{12|R(r)} < 0)$ across $r$ = 0,. . .,12; and all levels of discrimination;
8 = Power estimated by averaging $P(\sigma_{1b|2(x)} < 0)$ and $P(\sigma_{2b|1(x)} < 0)$ across $b$ = 3,4,5; $x$ = 0,. . .,4; and all levels of discrimination;
9 = Power estimated by averaging $P(\sigma_{1b|R(r)} < 0)$ and $P(\sigma_{2b|1(R)} < 0)$ across $b$ = 3, 4, 5; $r$ = 0,. . .,12; and all levels of discrimination;
10 = Power estimated by averaging $P(\sigma_{3b} < 0)$ across $b$ = 1, 2, 4, 5; and all levels of discrimination;
11 = Power estimated by averaging $P(\sigma_{ab|3(x)} < 0)$ across $a$, $b$ = 1, 2, 4, 5, $a \neq b$; $x$ = 0,. . .,4; and all levels of discrimination;
12 = Power estimated by averaging $P(\sigma_{ab|R(r)} < 0)$ across $a$, $b$ = 1, 2, 4, 5, $a \neq b$; $r$ = 0,. . .,12; and all levels of discrimination.

violations. We did a computational study to answer this question. In particular, for each case of CA (Equations 4–6), and for each true scenario, we estimated the power of a negative (conditional) covariance to detect a model violation.

## Method

We assumed $J = 5$, and items scored $x = 0, \ldots, 4$. Population covariances were derived from a two-dimensional graded response model (De Ayala, 1994). Vector $\boldsymbol{\theta} = (\theta, \theta^*)$ contained nuisance latent variable $\theta^*$ used to model local dependence. Let $\delta_{jx}$ ($j = 1, \ldots, 5$; $x = 1, \ldots, 4$) be location parameters, and $\boldsymbol{\alpha}_j = (\alpha_j, \alpha_j^*)$ contain discrimination parameters. The two-dimensional graded response model is defined as

$$P(X_j \geq x|\theta) = \frac{\exp[\alpha_j(\theta - \delta_{jx}) + \alpha_j^*(\theta^* - \delta_{jx})]}{1 + \exp[\alpha_j(\theta - \delta_{jx}) + \alpha_j^*(\theta^* - \delta_{jx})]}. \quad (7)$$

If $\alpha_j^* = 0$ for all $j$, then the UMLVM holds; if $\alpha_j^* \neq 0$ and $\alpha_k^* \neq 0$, then item $j$ and item $k$ are locally dependent (Ip, 2010). We assumed $\theta \sim \mathcal{N}(0, 1)$, $\theta^* \sim \mathcal{N}(0, 1)$, and $\rho_{\theta\theta^*} = 0$. Variation of $\delta_{jx}$s did not affect the sign of the covariances; hence, $\delta_{jx}$s were fixed: $\delta_1 = (-1.5, -0.75, 0.25, 1)$, $\delta_2 = (-1.25, -0.5, 0.5, 1.25)$, $\delta_3 = (-1, -0.25, 0.75, 1.5)$, $\delta_4 = (-0.75, 0, 1, 1.75)$, and $\delta_5 = (-0.5, 0.25, 1.25, 2)$.

### Positive Local Dependence

In the four true scenarios involving a PLD item pair, items 1 and 2 were PLD and items 3–5 were UMLVM consistent. Relative to $\theta$, $\alpha_j > 0$ ($j = 1, \ldots, 5$), and relative to $\theta^*$, $\alpha_1^* > 0$ and $\alpha_2^* > 0$, so that $\sigma_{12|\theta} > 0$; and $\alpha_3^* = \alpha_4^* = \alpha_5^* = 0$, so that $\sigma_{jk|\theta} = 0$ ($j \neq k$; $j, k \in \{3,4,5\}$). The independent variables were the true scenario (four levels), and the values of parameters $\alpha_1$, $\alpha_2$, $\alpha_1^*$, and $\alpha_2^*$, each parameter having 13 values equally spaced between 0.25 and 3.25 (Table 2, 1st column), yielding $13^4$ = 28,651 combinations in total. A pilot showed that variation of other $\alpha$s had negligible effect on the sign of the covariances; hence, $\alpha_3 = \alpha_4 = \alpha_5 = 1.5$.

The dependent variables were the proportion of negative values for $\sigma_{ab}$, $\sigma_{ab|c(x)}$, and $\sigma_{ab|R(r)}$, which can be interpreted as the power to detect a model violation for a true scenario. For some cells in Table 1, no computational study was done. These include the cells for which it was proven that the power equals 0, and the cells that cannot exist. For example, for $\sigma_{ab|R(r)}$ (Table 1, column 3) sets **Y** and **Z** are exhaustive and an item $d$ does not exist, so a PLD item pair with one item neither in **Y**, nor in **Z** (true scenario types 3 and 4) is impossible. For the remaining cells, the notes in Table 1 provide details on the computation of the power, where $P(\sigma_{12} < 0)$ denotes the proportion of negative covariances.

 

### Negative Local Dependence

The true scenarios involving an NLD item pair differed from the PLD scenarios only because $\alpha_2^* < 0$, so that $\sigma_{12|\theta} < 0$.

### Non-Monotonicity

Only item 3 violated monotonicity (Table 1, third panel). Item response function $E(X_3|\theta)$ decreased between $(-0.5; 0.5)$ and $(0.5; 1.5)$. The independent variables were true scenario (three levels: item 3 in $\mathbf{Y}$ [NM($a$)]; item 3 in $\mathbf{Z}$ [NM($c$)]; and item 3 neither in $\mathbf{Y}$, nor in $\mathbf{Z}$ [NM($d$)]) and the values of parameters $\alpha_1$, $\alpha_2$, and $\alpha_3$, each parameter having 13 values (Table 2). Variation of $\alpha_4$ and $\alpha_5$ had negligible effect on the signs of the covariances, hence they were fixed to $\alpha_4 = \alpha_5 = 1.5$.

## Results and Conclusions

Cells in Table 1 that result from the computational study show the proportions of negative values for covariances $\sigma_{ab}$, $\sigma_{ab|c(x)}$, and $\sigma_{ab|R(r)}$ (columns) attained for the 11 true scenario types. Higher proportions represent higher power to identify PLD, NLD, or NM. We found that design choices have little effect on the proportions and thus on the identification of particular powerful covariances. Table 1 shows that conditional covariances can detect model violations only in true scenario types 2 (PLD) and 5 (NLD). NM could not be detected. Three results from Table 1 best identify UMLVM violations:

Result 1 for PLD: $\sigma_{ab|c(x)} < 0$ identifies PLD in true scenario Type 2 (Table 1: PLD item pair, one item in $\mathbf{Y}$, and one item in $\mathbf{Z}$). The estimated power equals .314.

Result 2 for PLD: $\sigma_{ab|R(r)} < 0$ also identifies PLD in true scenario Type 2. The estimated power equals .318 (Table 1).

Result 3 for NLD: $\sigma_{ab|R(r)} < 0$ identifies NLD in true scenario Type 5 (NLD item pairs, both items in $\mathbf{Y}$). The estimated power equals .774. Although $\sigma_{ab}$ and $\sigma_{ab|c(x)}$ also have power to detect NLD in true scenario Type 5, their estimated power is lower.

# Identifying Locally Independent Item Sets

Procedure CA uses results 1, 2, and 3 for flagging items suspected to be locally dependent. The results may be used for removing items from the $J$-item set to obtain a locally independent item subset. For different conditional covariances, indices $W^{(1)}$, $W^{(2)}$, and $W^{(3)}$ quantify the degree to which an item is suspected. An index adds probabilities that particular conditional covariances such as $s_{ij|k(x)}$ are negative. Let $N_{k(x)}$ be the size of the subsample for which $X_k = x$. Then, $N_{k(x)}$ is used to compute $s_{ij|k(x)}$. Using the Fisher-Z transformation (e.g., Hays, 1994, p. 649), we approximated the sampling distribution of $s_{ij|k(x)}$ by a normal distribution with mean

**Table 2.** Discrimination parameters for local dependence and violation of monotonicity conditions

| Discrimination | Type of violation | |
|---|---|---|
| parameter | Local dependence | Violation of monotonicity |
| $\alpha_1$ | 0.25, 0.5, . . ., 3.25 | 0.25, 0.5, . . ., 3.25 |
| $\alpha_2$ | 0.25, 0.5, . . ., 3.25 | 0.25, 0.5, . . ., 3.25 |
| $\alpha_3$ | 1.5 | 0.25, 0.5, . . ., 3.25 |
| $\alpha_4$ | 1.5 | 1.5 |
| $\alpha_5$ | 1.5 | 1.5 |
| $\alpha_1^*$ | 0.25, 0.5, . . ., 3.25 | 0 |
| $\alpha_2^*$ | 0.25, 0.5, . . ., 3.25 | 0 |

$$\mu_{ij|k(x)} = \frac{1}{2}\ln\left(\frac{1+\sigma_{ij|k(x)}/(\sigma_{i|k(x)}\sigma_{j|k(x)})}{1-\sigma_{ij|k(x)}/(\sigma_{i|k(x)}\sigma_{j|k(x)})}\right), \qquad (8)$$

and variance $\sigma^2_{k(x)} = 1/(N_{k(x)} - 3)$. Furthermore, let $Z$ be $\mathcal{N}(0,1)$. The probability of a negative sample covariance equals $P(Z < \frac{0-\mu_{ij|k(x)}}{\sigma_{k(x)}})$ and is estimated by means of $P(Z < \frac{-\hat{\mu}_{ij|k(x)}}{\hat{\sigma}_{k(x)}})$. The three indices are defined as follows.

Index $W^{(1)}$ uses Result 1 and adds probabilities based on inter-item covariances $s_{aj|c(x)}$. For item pair ($a$, $c$),

$$W^{(1)}_{ac} = \sum_{j\neq a,c}\sum_{x} P\left(Z < \frac{-\hat{\mu}_{aj|c(x)}}{\hat{\sigma}_{c(x)}}\right). \qquad (9)$$

It may be noted that indices $W^{(1)}_{ac}$ and $W^{(1)}_{ca}$ usually produce different values because $W^{(1)}_{ac}$ conditions on item $c$ and $W^{(1)}_{ca}$ conditions on item $a$. Hence, $J(J-1)$ values of $W^{(1)}$ are considered. If $W^{(1)}_{ac}$ is large, then item pair ($a$, $c$) likely is PLD.

Index $W^{(2)}$ uses Result 2, and for item $a$ adds probabilities based on covariances $s_{aj|R(r)}$, so that

$$W^{(2)}_{a} = \sum_{j\neq a}\sum_{r} P\left(Z < \frac{-\hat{\mu}_{aj|R(r)}}{\hat{\sigma}_{R(r)}}\right). \qquad (10)$$

If $W^{(2)}_{a}$ is large, then item $a$ likely is in a PLD item pair. Index $W^{(3)}$ uses Result 3, and for item pair ($a$, $b$) adds probabilities based on covariances $s_{ab|R(r)}$ in which item pair ($a$, $b$) is involved, so that

$$W^{(3)}_{ab} = \sum_{r} P\left(Z < \frac{-\hat{\mu}_{ab|R(r)}}{\hat{\sigma}_{R(r)}}\right). \qquad (11)$$

If $W^{(3)}_{ab}$ is large, then item pair ($a$, $b$) likely is NLD.

Let $M$ be the median and $Q_3$ the third quartile of the empirical $W$ distribution, then $W$ is an outlier if $W > Q_3 + 3 \times (Q_3 - M)$; this value is known as a Tukey fence

 

(Tukey, 1977; also Hubert & Vandervieren, 2008) and is used when distributions are skew. Items for which one or more $W$ values are flagged were considered outliers.

For each item, $2(J-1)$ indices $W^{(1)}$, 1 index $W^{(2)}$, and $J-1$ indices $W^{(3)}$ were computed; hence, the number of flags ran from 0 to $3J-2$. Removal of an item may affect the number of flags for the other items. For example, if only index $W^{(1)}$ flags item pair ($a$, $b$), removal of item $b$ clears the flags for item $a$ which then is UMLVM consistent. Hence, flagged items are removed one by one; also, see Ligtvoet, Van der Ark, Te Marvelde, and Sijtsma (2010). In each next selection step, the item with the largest number of $W$ flags was removed until only items remained that had no flags. If items had the same number of flags, the item having the smallest item-scalability coefficient $H_j$ (Mokken, 1971, pp. 151–152) was removed.

## Specificity and Sensitivity of Procedure CA

### Method

Procedure CA identifies and removes items from PLD and NLD item pairs to obtain a locally independent item set. We investigated the specificity and the sensitivity of procedure CA. We used the multidimensional graded response model to simulate data sets also containing PLD and NLD item pairs. Vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_Q)$ contains $Q$ latent variables, $\alpha_{jq}$ is the discrimination parameter of item $j$ with respect to trait $q$, and $\beta_{jx}$ is the location parameter of category $x$ of item $j$. The model is defined as

$$P(X_j \geq x|\theta) = \frac{\exp \sum_q \alpha_{jq}(\theta_q - \beta_{jx})}{1 + \exp \sum_q \alpha_{jq}(\theta_q - \beta_{jx})}. \qquad (12)$$

Local dependence was defined as follows. All items load on $\theta_1$: that is, $\alpha_{j1} > 0$, all $j$. If the UMLVM holds, all other discrimination parameters equal 0. If two items are locally dependent, both load on another latent variable; Ip (2010) discusses equivalence of local dependence and multidimensionality. Based on this model, we sampled 1,000 data sets, each data set containing scores $x = 0, \ldots, 4$ for 1,000 persons on 16 items. For each person, five $\theta$ values were sampled from $\mathcal{N}(0,1)$. Table 3 shows the values of the $\alpha_{jq}$s. All $\beta_{jx}$s were drawn from $\mathcal{N}(0,1)$; and for each item $\beta$s were ordered from smallest to largest and numbered accordingly. Items 1 and 2 were *weak PLD* because they were loaded weakly on $\theta_2$ (Table 3); similarly, items 3 and 4 were *strong PLD*, items 5 and 6 were *weak NLD*, and items 7 and 8 were *strong NLD*. Items 9 to 16 were locally independent.

**Table 3.** Discrimination parameters in the simulation study

| Item | Latent trait | | | | |
|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ |
| 1 | Weak | Weak | 0 | 0 | 0 |
| 2 | Weak | Weak | 0 | 0 | 0 |
| 3 | Weak | 0 | Strong | 0 | 0 |
| 4 | Weak | 0 | Strong | 0 | 0 |
| 5 | Weak | 0 | 0 | Weak | 0 |
| 6 | Weak | 0 | 0 | −Weak | 0 |
| 7 | Weak | 0 | 0 | 0 | Strong |
| 8 | Weak | 0 | 0 | 0 | −Strong |
| 9–16 | Weak | 0 | 0 | 0 | 0 |

*Notes*. Weak = the discrimination parameter equals the natural logarithm of a random draw from $N(.2, .05)$ (mean value $e^{0.2} \approx 1.22$); strong = the discrimination parameter equals the natural logarithm of a random draw from $N(.9, .1)$ (mean value $e^{0.9} \approx 2.46$).

Specificity was the percentage from 1,000 replications in which procedure CA correctly identified a locally independent item (i.e., items 9–16). Sensitivity had two definitions: Type I: percentage from 1,000 replications in which procedure CA correctly removed one item from a locally dependent item pair. Using this definition one strives to retain as many items as possible; Type II: percentage from 1,000 replications in which procedure CA correctly removed both items from a locally dependent item pair. Type-II sensitivity is always higher than Type-I sensitivity. Procedure CA has been implemented in version 2.8.2 of the R package mokken (Van der Ark, 2007, 2012).

### Results

Specificity was 89.5%. Type-I sensitivity was highest for weak NLD item pairs (66.3%) but low in general (Table 4). Type-II sensitivity was high for NLD and strong PLD; at least one item was removed from at least 90% of the locally dependent item pairs. Type-II sensitivity was lower for weak PLD; A violation was detected in 41.9% of the cases.

## Real-Data Example: Type D Scale 14

Procedure CA was used to analyze real data from 3,111 persons who responded to the Type D Scale-14 (DS14) questionnaire (Table 5; Denollet, 2005). DS14 measures distressed personality – Type D, for short – and contains two 7-item scales measuring negative affectivity (NA) and social inhibition (SI). Three subtraits called feelings of dysphoria (items NA1, NA2, and NA3), anxious apprehension (items NA4 and NA5), and irritability (items NA6 and

**Table 4.** Sensitivity of the CA procedure (%)

| | Violation | | | |
|---|---|---|---|---|
| | Weak | Strong | Weak | Strong |
| Type of sensitivity | PLD | PLD | NLD | NLD |
| Type I (one item removed) | 26.2 | 24.0 | 66.3 | 33.8 |
| Type II (at least one item removed) | 41.9 | 89.9 | 97.6 | 99.9 |

**Table 5.** $H_j$ coefficients for the negative affectivity scale and the social inhibition scale

| Item | Content | $H_j$ |
|---|---|---|
| *Negative affectivity scale* | | |
| NA1 | Often feels unhappy | .493 |
| NA2 | Takes gloomy view of things | .570 |
| NA3 | Is often down in the dumps | .606 |
| NA4 | Worries about unimportant things | .442 |
| NA5 | Often worries about something | .548 |
| NA6 | Is easily irritated | .486 |
| NA7 | Is often in a bad mood | .484 |
| *Social inhibition scale* | | |
| SI1 | Inhibited in social interactions | .482 |
| SI2 | Difficulties starting a conversation | .547 |
| SI3 | Does not find things to talk about | .535 |
| SI4 | Closed kind of person | .514 |
| SI5 | Keeps others at a distance | .509 |
| SI6 | Makes contact easily | .545 |
| SI7 | Often talks to strangers | .457 |

NA7) drive NA, and three subtraits called discomfort in social situations (items SI1, SI2, and SI3), reticence (items SI4 and SI5), and lack of social poise (items SI6 and SI7) drive SI (Table 5). We expected sets of items measuring the same subtrait to be PLD. Next, for the DS14 data we discuss the results of procedure-CA analysis.

Table 5 shows the item-scalability $H_j$ values for the complete NA and SI scales. Table 6 (upper panel) shows the $W$ indices for the entire NA scale. The upper fences for the box plots of the $W$ indices were 3.235, 54.905, and 16.727, respectively. Item NA3 had two two flags and was removed first. Repetition of the procedure produced new $W$ values and new upper fences, but none of the remaining items was flagged. Table 6 (lower panel) shows the $W$ indices for the entire SI scale. $W$ indices had upper fences equal to 0.270, 51.780, and 19.221, respectively. All flags pertained to $W^{(1)}$: Item SI6 had six flags, item SI2 three, items SI2 and SI7 two, and items SI3 and SI5 one. Without item SI6, new weights and upper fences were computed: Item SI2 had two $W^{(1)}$ flags (not tabulated), and was removed. For the five remaining items, $W^{(3)}_{1,7}$ exceeded the upper fence. $H_{SI7} = .41$ and $H_{SI1} = .45$, hence item SI7 having the lowest value was removed. Procedure CA proved to be

**Table 6.** $W$ Indices for the negative affectivity (top) and social inhibition (bottom) scales. Flagged values in boldface

| Index | Item | NA1 | NA2 | NA3 | NA4 | NA5 | NA6 | NA7 |
|---|---|---|---|---|---|---|---|---|
| $W_1$ | NA1 | | 1.701 | 2.945 | 0.000 | 0.294 | 0.024 | 1.492 |
| | NA2 | 0.000 | | 2.877 | 0.000 | 0.094 | 0.063 | 4.216 |
| | NA3 | 0.000 | 0.775 | | 0.013 | 0.118 | 0.142 | 1.636 |
| | NA4 | 0.017 | 2.481 | 4.830 | | 1.048 | 0.574 | 2.956 |
| | NA5 | 0.017 | 0.682 | 3.134 | 0.001 | | 0.004 | 2.108 |
| | NA6 | 0.018 | 0.363 | 2.099 | 0.013 | 0.002 | | 1.925 |
| | NA7 | 0.017 | 2.545 | 4.241 | 0.000 | 0.545 | 0.354 | |
| $W_2$ | | 30.142 | 28.028 | 28.682 | 42.616 | 26.442 | 35.511 | 35.573 |
| $W_3$ | NA1 | | | | | | | |
| | NA2 | 3.086 | | | | | | |
| | NA3 | 1.707 | 3.255 | | | | | |
| | NA4 | 10.507 | 8.477 | 7.650 | | | | |
| | NA5 | 4.891 | 2.528 | 0.846 | 1.441 | | | |
| | NA6 | 3.571 | 6.319 | 11.801 | 5.119 | 6.726 | | |
| | NA7 | 6.379 | 4.364 | 3.422 | 9.422 | 10.010 | 1.974 | |
| Index | Item | SI1 | SI2 | SI3 | SI4 | SI5 | SI6 | SI7 |
| $W_1$ | SI1 | | 0.295 | 0.009 | 0.028 | 0.484 | **1.682** | 0.000 |
| | SI2 | 0.005 | | 0.021 | 0.014 | 0.104 | 0.873 | 0.000 |
| | SI3 | 0.000 | 0.209 | | 0.003 | 0.018 | 0.314 | 0.000 |
| | SI4 | 0.000 | 0.001 | 0.000 | | 0.325 | 0.281 | 0.000 |
| | SI5 | 0.004 | 0.004 | 0.009 | 0.043 | | 0.240 | 0.000 |
| | SI6 | 0.000 | 0.007 | 0.000 | 0.008 | 0.000 | | 0.000 |
| | SI7 | 0.000 | 0.425 | 0.021 | 0.010 | 0.406 | **3.377** | |
| $W_2$ | | 35.835 | 33.17 | 25.29 | 36.449 | 35.93 | 29.731 | **44.232** |
| $W_3$ | SI1 | | | | | | | |
| | SI2 | 1.482 | | | | | | |
| | SI3 | 2.691 | 1.955 | | | | | |
| | SI4 | 7.875 | 9.918 | 3.254 | | | | |
| | SI5 | 6.798 | 9.355 | 2.412 | 0.953 | | | |
| | SI6 | 4.606 | 2.940 | 5.665 | 5.928 | 10.255 | | |
| | SI7 | 12.383 | 7.520 | 9.312 | 8.521 | 6.158 | 0.337 | |

effective producing locally independent six-item NA and four-item SI scales. This data set is available from the R package *mokken* as of version 2.8.2.

# Discussion

Procedure CA has good specificity, thus tends to keep UMLVM-consistent items in the scale. When attributes are measured by means of few items, procedure CA has the effect of retaining items that only deviate little from the other items and keeping reliability and trait coverage at an acceptable level. Procedure CA has sensitivity equal to 89.5%, and suggests removing only one item in a locally dependent item pair, again avoiding removing items all too easily. Other methods are greedier, removing pairs of items (e.g., Sijtsma & Molenaar, 2002, chap. 5; Zhang & Stout, 1999).

 

Investigation of local independence concerns only one model assumption and the question is how to embed this aspect in a complete scale analysis in which unidimensionality and item response function monotonicity also have to be assessed. This is a complex methodological issue that should be addressed in future studies. Another topic for future research is joining small adjacent rest-score groups and compute conditional covariances on the joint groups. Joining groups is expected to increase precision of covariance estimates but may also have a positive effect on the reliability of conclusions about CA and local dependence and possibly on the specificity and the sensitivity of the procedure (e.g., Sijtsma & Molenaar, 2002).

## References

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289. doi: 10.3102/10769986022003265

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*, 155–170. doi: 10.1177/014662169401800205

Denollet, J. (2005). DS14: standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine, 67*, 89–97.

Douglas, J., Kim, H., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129–151. doi: 10.3102/10769986023002129

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383–392. doi: 10.1007/BF02294219

Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Holt, Rinehart, & Winston.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543. doi: 10.1214/aos/1176350174

Hubert, M., & Vandervieren, E. (2008). An adjusted box plot for skewed distributions. *Computational Statistics and Data Analysis, 52*, 5186–5201. doi: 10.1016/j.csda.2007.11.008

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63*, 395–416. doi: 10.1348/000711009X466835

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578–595. doi: 10.1177/0013164409355697

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Rosenbaum, P. R. (1984). Testing conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425–435. doi: 10.1007/BF02306030

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349–359. doi: 10.1007/BF02294217

Sijtsma, K. (2003). Developments in practical nonparametric IRT scale analysis. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 183–190). Tokyo, Japan: Springer.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–19. doi: 10.18637/jss.v020.i11

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1–27. doi: 10.18637/jss.v048.i05

Van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*, 272–279. doi: 10.1007/S11336-010-9147-7

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–249. doi: 10.1007/BF02294536

Received March 31, 2015
Revision received May 25, 2016
Accepted July 12, 2016
Published online December 5, 2016

J. Hendrik Straat is senior research scientist at Cito, Institute for Educational Test Development at Arnhem, The Netherlands. His research interests include nonparametric item response theory, the setting of performance standards, multistage testing, and the application of Bayesian networks in an assessment-for-learning context.

L. Andries van der Ark is professor by special appointment in quantitative research methods at the Research Institute of Child Development and Education of the University of Amsterdam. His research interests include reliability analysis, nonparametric item response theory, Mokken scale analysis, and marginal models for the analysis of test and questionnaire data.

Klaas Sijtsma is professor of methods of psychological research at the Tilburg School of Social and Behavioral Sciences, Tilburg University. His research interest is with methods and statistical models for the measurement of individual differences with respect to psychological attributes.

**J. Hendrik Straat**
Cito
Amsterdamseweg 13
6814 CM Arnhem
The Netherlands
hendrik.straat@cito.nl

 