Identifying Local Dependence with a Score Test Statistic Based on the Bifactor 2-Parameter Logistic Model

Yang Liu

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology

Chapel Hill
2011

Approved by:

David M. Thissen

Robert C. MacCallum

Abigail T. Panter

Abstract

Yang Liu: Identifying Local Dependence with a Score Test Statistic Based on the Bifactor 2-Parameter Logistic Model
(Under the direction of Dr. David M. Thissen)

Local dependence (LD) refers to the violation of the local independence assumption of most item response models. Statistics that indicate LD between a pair of items on a test or questionnaire that is being fitted with an item response model can play a useful diagnostic role in applications of item response theory. In this paper a new score test statistic, S_b , for underlying LD (ULD) is proposed based on the bifactor 2-parameter logistic model. To compare the performance of S_b with the score test statistic (S_t) based on a threshold shift model for surface LD (SLD), and the LD X^2 statistic, we simulated data under null, ULD, and SLD conditions, and evaluated the null distribution and power of each of these test statistics. The results summarize the null distributions of all three diagnostic statistics, and their power for approximately matched degrees of ULD and SLD. Future research directions are discussed, including the straightforward generalization of S_b for polytomous item response models, and the challenges involved in the corresponding generalizations of S_t and LD X^2.

Acknowledgement

I would like to show my gratitude to my advisor Dr. David M. Thissen, former and current members of his lab, and committee members–Dr. Robert C. MacCallum and Dr. Abigail T. Panter–for their advices in developing the idea and programming. Also this thesis would not have been possible without my parents and my fiancée "Flora" Meng Xu's encouragement and support.

Table of Contents

Page

List of Tables

List of Figures

Introduction

1. Local dependence

Item response theory (IRT) provides a collection of latent variable models and statistical procedures (including parameter estimation techniques, model evaluation methods and diagnostics) used for item analysis and test scoring (Thissen & Steinberg, 2009). One basic assumption of IRT models is local independence (LI), which requires that responses to different items are independent conditional on the latent variable of interest[1]. Formally, LI implies the probability of observing responses $\mathbf{x} = \{x_1, \ldots, x_j\}$ to all J items given $\boldsymbol{\theta}$ to be the product of each item's trace line function $T_j(\cdot)$:

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{j=1}^{J} T_j(x_j|\boldsymbol{\theta}) \tag{1}$$

Equation 1 formalizes the strong version of LI (SLI; McDonald, 1982). SLI serves as the basis for both estimating item parameters and scoring, where the likelihood function of the overall response pattern for a given $\boldsymbol{\theta}$ is usually treated as though the contribution of each item is independent. Although SLI is important, there is no feasible statistical procedure to test it, for there must be no interaction among item responses in all marginal tables (from two-way to J-way). Consequently, only violations of pairwise independence in certain parametric forms, that will be introduced in the next section, are tested in practice.

Thissen et al. (1992) distinguished two substantive types of local dependence: underlying local dependence (ULD) and surface local dependence (SLD). ULD refers to the scenario where an additional latent variable can be employed to explain the error covariance within each locally dependent set of items. A reading comprehension test is a typical example of ULD: Responses to items following the same text will be more correlated due to context similarity. In contrast, SLD arises when the response to an item is fostered (i.e. positive SLD) or hampered (i.e. negative SLD) by responses to previous items. One frequently cited example of positive SLD involves redundant items: that is, (nearly) the same question asked more than once in a test. Negative SLD, on the other hand, is in general less common and may be difficult to explain.

[1] The latent variable is not necessarily unidimensional

2. Existing methods to identify LD

2.1 Models

The difference between underlying and surface LD can be described using a path diagram[2]. The path diagram in Figure 1 illustrates the bifactor model (left), which implies ULD, and the threshold shift model (right), which implies SLD.

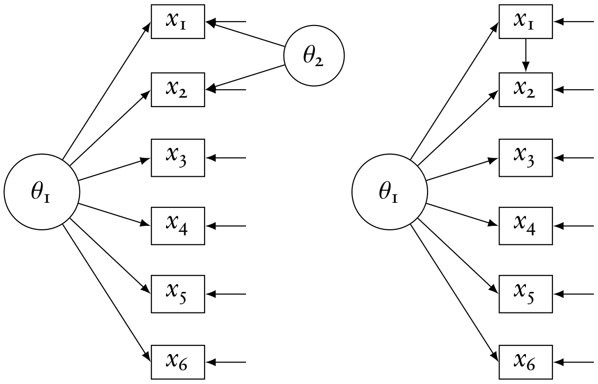

Figure 1: Path diagrams: (a) Bifactor model; (b) Threshold-shift model

The bifactor model is a special multidimensional factor analysis model (Gibbons & Hedeker, 1992, for the more general two-tier model, see Cai (2010)). The specification of the item bifactor model parallels the notion of ULD in the sense that an extra latent variable θ_2 produces residual covariances (for a certain subset of items) when the data are fitted with a unidimensional model. A two-parameter logistic (2PL) version of the bifactor model for item pair p and q is:

$$\begin{cases} T_p(1|\theta_1,\theta_2) = \dfrac{1}{1+\exp(-a_p\theta_1 - a_{pq}\theta_2 - c_p)} \\ T_q(1|\theta_1,\theta_2) = \dfrac{1}{1+\exp(-a_q\theta_1 \pm a_{pq}\theta_2 - c_q)} \end{cases} \tag{2}$$

In order to identify the model when the LD subset comprises only two items, some constraint must be imposed on the two secondary dimension slope parameters. A convenient restriction may be equality of the absolute value of bifactor slopes. The negativity/positivity of the $a_{pq}\theta_2$ term in the second formula is determined by positive/negative LD to be modeled.

The 2PL model is a special case of bifactor model obtained by setting $a_{pq} = 0$ in Equation 2. In addition, it can be proved (see Appendix A) that the bifactor model is equivalent to the error covariance model (see below) in the continuous and probit case. Another parameterization of a restricted bifactor model was given

[2] Only unidimensional models are shown here; however, one can make a generalization to the multidimensional case.

by Bradlow, Wainer & Wang (1999), in which bifactor slopes are fixed to be equal to their corresponding primary slope, and the variance of the secondary factor is estimated.

The development of **a threshold shift model** may be traced back to the work of Kelderman (1984) and Jannarone (1986) on the Rasch model. It also appeared later in work by Hoskens & De Boeck (1997), where it was called an "ordered-constant" interaction model, among four types of item response models employing additional pairwise interaction parameters. Glas & Suárez Falcón (2003) derived a score test for this model's interaction term (see also van der Linden & Glas, 2010).

In the 2PL case, for example, the trace line function of the second item in a locally dependent item pair p and q ($p < q$) may be written as:

$$T_q(1|\theta_1; x_p) = \frac{1}{1 + \exp(-a_q\theta_1 - c_q - \delta_{pq}x_p)} \tag{3}$$

where δ_{pq} can be considered a "threshold shift" for the second item when the first item response is positive, which is in accordance with the scenario of SLD. δ_{pq} is also an ANOVA-like interaction term when the log odds of the pairwise response pattern is considered[3]. Notice that if $\delta_{pq} = 0$, the threshold shift model reduces to 2PL model as well.

Apart from these models, there are other limited or full information parametric models for LD. For example, both LISREL and M*plus* can handle error covariances in a factor analysis model with ordered categorical indicators based on the polychoric correlation matrix (Christoffersson, 1975; Muthén, 1978; see Wirth & Edwards (2007) for a review); Braeken, Tuerlinckx, and De Boeck (2007) proposed to estimate the dependency parameter of the bivariate logistic distribution constructed with copula functions.

2.2 Statistical procedures

Asymptotic tests: Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_0, \boldsymbol{\eta}_1)$ be the model parameters defined in some parameter space Θ which can be factored into the direct product of subspaces $\Theta_0 \times \Theta_1$ such that $\boldsymbol{\eta}_0 \in \Theta_0$ and $\boldsymbol{\eta}_1 \in \Theta_1$. Consider the hypothesis testing problem based on a partition of the subspace $\Theta_1 = \{\boldsymbol{\vartheta}_1\} \cup \{\boldsymbol{\vartheta}_1\}^c$, in which $\{\boldsymbol{\vartheta}_1\}^c$ is the relative complement of $\{\boldsymbol{\vartheta}_1\}$ in Θ_1.

$$H_0 : \boldsymbol{\eta}_1 = \boldsymbol{\vartheta}_1 \qquad \text{(Null model)}$$

$$H_1 : \boldsymbol{\eta}_1 \neq \boldsymbol{\vartheta}_1 \qquad \text{(Alternative model)}$$

Three asymptotic test statistics are commonly used for this testing problem. The likelihood ratio statistic

[3] This is Hoskens & De Boeck's interpretation; however, it is less intuitive. So I will call it the "threshold shift model" for the rest of this paper.

(Neyman & Pearson, 1928) is defined as

$$\Lambda(\hat{\boldsymbol{\eta}}) = \frac{\sup_{\boldsymbol{\eta}_0 \in \Theta_0} L(\boldsymbol{\eta}_0, \boldsymbol{\vartheta}_1; \mathbf{x})}{\sup_{(\boldsymbol{\eta}_0, \boldsymbol{\eta}_1) \in \Theta} L(\boldsymbol{\eta}_0, \boldsymbol{\eta}_1; \mathbf{x})} = \frac{L_{\Theta_0}(\tilde{\boldsymbol{\eta}}_0, \boldsymbol{\vartheta}_1; \mathbf{x})}{L_{\Theta_0 \times \Theta_1}(\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\eta}}_1; \mathbf{x})} \tag{4}$$

where $(\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\eta}}_1) = \hat{\boldsymbol{\eta}}$ is the maximum likelihood estimate (MLE) in Θ, while $(\tilde{\boldsymbol{\eta}}_0, \boldsymbol{\vartheta}_1) = \tilde{\boldsymbol{\eta}}$ is the conditional MLE obtained under the restriction $\boldsymbol{\eta}_1 = \boldsymbol{\vartheta}_1$. The Wald statistic is (Wald, 1943)

$$W(\hat{\boldsymbol{\eta}}_1) = (\hat{\boldsymbol{\eta}}_1 - \boldsymbol{\vartheta}_1)' \mathbf{H}_{\Theta_1}(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_1)(\hat{\boldsymbol{\eta}}_1 - \boldsymbol{\vartheta}_1) \tag{5}$$

where $\mathbf{H}_{\Theta_1}(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_1)$ is the lower-right block of the Hessian matrix of the log-likelihood. Finally, the score test statistic (Rao, 1948) is

$$S(\boldsymbol{\eta}) = \nabla'_{\Theta}(\tilde{\boldsymbol{\eta}}; \mathbf{x}) \mathbf{H}_{\Theta}^{-1}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\eta}}) \nabla_{\Theta}(\tilde{\boldsymbol{\eta}}; \mathbf{x}) \tag{6}$$

where $\nabla_{\Theta}(\tilde{\boldsymbol{\eta}}; \mathbf{x})$ is the Fisher's score function

$$\nabla_{\Theta}(\tilde{\boldsymbol{\eta}}; \mathbf{x}) = \left.\frac{\partial \ell_{\Theta}(\boldsymbol{\eta}; \mathbf{x})}{\partial \boldsymbol{\eta}}\right|_{\eta = \tilde{\eta}} = \left.\frac{\partial \log L_{\Theta}(\boldsymbol{\eta}; \mathbf{x})}{\partial \boldsymbol{\eta}}\right|_{\eta = \tilde{\eta}} \tag{7}$$

and $\mathbf{H}_{\Theta}^{-1}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\eta}})$ is the inverse of the Hessian matrix (evaluated at $\tilde{\boldsymbol{\eta}}$). It can be proved (see Buse, 1982) that likelihood ratio, Wald, and score statistics are all asymptotically χ^2 distributed with degrees of freedom equal to the number of constraints imposed by H_0. However, the score test statistic does not require the computation of the MLE which is an advantage over the other two tests when applied to LD models. Because in practice we need to test the hypotheses for each pair of items, the process of obtaining MLEs based on each LD model can be very time consuming for long tests.

Residual measures: In this class of procedures, the IRT model with the desired number of dimensions is fitted first, and then diagnostic statistics for residual association are calculated. One popular residual measure uses the LD χ^2 statistics for two-way marginal tables (Chen & Thissen, 1997). Using the dichotomous case as an example, for each pair of items the 2 × 2 tables as shown in Table 1 can be constructed for observed and expected frequencies:

Table 1: Two-way marginal tables for items p and q

		Item q	
		0	1
Item p	0	O_{00}	O_{01}
	1	O_{10}	O_{11}

		Item q	
		0	1
Item p	0	E_{00}	E_{01}
	1	E_{10}	E_{11}

Here, O denotes the observed frequencies and E the expected ones. Expected cell counts are calculated according to the locally independent item response model:

$$E_{x_p x_q} = N \int T_p(x_p|\boldsymbol{\theta}) T_q(x_q|\boldsymbol{\theta}) \phi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{8}$$

Basically, LD χ^2 statistics reflect the discrepancy between the observed and the model-implied expected counts. Chen and Thissen proposed the Pearson X^2 and likelihood ratio G^2 for this purpose:

$$X^2 = \sum_{x_p=0}^{1} \sum_{x_q=0}^{1} \frac{(O_{x_p x_q} - E_{x_p x_q})^2}{E_{x_p x_q}} \tag{9}$$

$$G^2 = -2 \sum_{x_p=0}^{1} \sum_{x_q=0}^{1} O_{x_p x_q} \log \left(\frac{E_{x_p x_q}}{O_{x_p x_q}} \right) \tag{10}$$

The theoretical null distributions for both X^2 and G^2 remain unclear so far. Chen and Thissen suggested using the χ^2-distribution with one degree of freedom as an approximation (for dichotomous items) based on their simulation results. They claimed that the degree of freedom is one for test of independence, whereas estimating slope parameters from the relationships among items could be regarded as imposing fractional loss of the one degree of freedom.

There are other residual diagnostics. For example, Yen's Q_3 (1984) is defined as the sample Pearson correlation between paired residuals. However, it has the same problem as LD X^2 and G^2, that the null distribution of the residual correlation is not clear for categorical data[4]. Another residual diagnostic, the DIMTEST T statistic, was proposed by Stout (1987) for testing unidimensionality. As compared to Yen's Q_3, DIMTEST is a nonparametric procedure testing whether the average residual covariance in a given item subset is significantly larger than zero.

The research described here investigates the feasibility and usefulness of a score test for ULD, and compares its performance to the score test for SLD proposed by Glas & Suárez Falcón (2003) and LD X^2 statistics (Chen & Thissen, 1997).

3. Theory of score test

3.1 Derivatives in general form

Let the response for each item be dichotomous ($x_{ij} \in \{0, 1\}$). Bock (unpublished) gave the general form of first order derivatives taken on the marginal loglikelihood function with respect to a general item parameter

[4] Yen (1984) suggested to use Fisher r-to-z transformation, and then use standard normal distribution as an approximation. However, subsequent studies suggested that it is actually not a good approximation (see Chen & Thissen, 1997; Ip, 2001).

η_s:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \eta_s} &= \sum_{i=1}^{N} \frac{1}{\Pr(\mathbf{x}_i)} \frac{\partial \Pr(\mathbf{x}_i)}{\partial \eta_s} \\
&= \sum_{i=1}^{N} \frac{1}{\Pr(\mathbf{x}_i)} \int_{\boldsymbol{\theta}} \frac{\partial L(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \eta_s} \phi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \sum_{i=1}^{N} \frac{1}{\Pr(\mathbf{x}_i)} \int_{\boldsymbol{\theta}} \Big[\sum_{j=1}^{J} \frac{x_{ij}}{T_j(x_{ij}|\boldsymbol{\theta})} \frac{\partial T_j(x_{ij}|\boldsymbol{\theta})}{\partial \eta_s} \Big] L(\mathbf{x}_i|\boldsymbol{\theta}) \phi(\boldsymbol{\theta}) d\boldsymbol{\theta}
\end{aligned}
\tag{11}
$$

where $L(\mathbf{x}_i|\boldsymbol{\theta}) = \prod_{j=1}^{J} T_j(x_{ij}|\boldsymbol{\theta})$, and $\Pr(\mathbf{x}_i) = \int_{\boldsymbol{\theta}} L(\mathbf{x}_i|\boldsymbol{\theta})\phi(\boldsymbol{\theta})d\boldsymbol{\theta}$. The analytical expression of the Hessian is quite convoluted; in practice, however, the calculation of second order derivatives can be avoided by invoking the cross-product approximation (Bock & Lieberman, 1970; Kendall & Stuart, 1961):

$$
\frac{\partial^2 \ell}{\partial \eta_s \partial \eta_t} \approx -N \sum_{\{\mathbf{x}_i\}} \frac{1}{\Pr(\mathbf{x}_i)} \frac{\partial \Pr(\mathbf{x}_i)}{\partial \eta_s} \frac{\partial \Pr(\mathbf{x}_i)}{\partial \eta_t} \tag{12}
$$

in which the summation is over all 2^J possible response patterns. This is often further approximated by limiting the summation to observed response patterns.

3.2 Multidimensional 2PL model

For 2PL model, the trace line functions for binary responses can be written as (subscript i's for subjects are dropped):

$$
T_j(x_j|\boldsymbol{\theta}) = \frac{1}{1 + \exp[(-1)^{x_j}(\mathbf{a}_j'\boldsymbol{\theta} + c_j)]} \tag{13}
$$

The first derivatives of trace line with respect to slope and intercept parameters are listed respectively as:

$$
\frac{\partial T_j(x_j|\boldsymbol{\theta})}{\partial \mathbf{a}_j} = \boldsymbol{\theta} T_j(x_j|\boldsymbol{\theta})[x_j - T_j(1|\boldsymbol{\theta})] \tag{14}
$$

$$
\frac{\partial T_j(x_j|\boldsymbol{\theta})}{\partial c_j} = T_j(x_j|\boldsymbol{\theta})[x_j - T_j(1|\boldsymbol{\theta})] \tag{15}
$$

3.3 Glas' score statistic for threshold shift

In Glas' threshold shift model, the trace line for the second item in an item pair is:

$$
T_q(x_q|\boldsymbol{\theta}, x_p) = \frac{1}{1 + \exp[(-1)^{x_j}(\mathbf{a}_q'\boldsymbol{\theta} + c_q + \delta_{pq}x_p)]} \tag{16}
$$

and the corresponding first order derivative with respect to δ_{pq} is:

$$\frac{\partial T_q(x_q|\boldsymbol{\theta}, x_p)}{\partial \delta_{pq}} = x_p T_q(x_q|\boldsymbol{\theta})[x_q - T_q(1|\boldsymbol{\theta})] \tag{17}$$

3.4 The score statistic for bifactor slope

The bifactor model is a special case of multidimensional IRT model, so the derivatives of the trace line function are described by Equation 14 and 15.

First, we note that we cannot compute the score test from exactly $a_{pq} = 0$. Let $(\boldsymbol{\theta}_1, \theta_2)$ be a partition of $\boldsymbol{\theta}$ where $\boldsymbol{\theta}_1$ represents the primary factor(s)[5] and θ_2 the specific factor to item p and q. To evaluate $\frac{\partial \ell}{\partial a_{pq}}$ at $a_{pq} = 0$, notice that when $a_{pq} = 0$, $T_j(\cdot)$ and thus $L(\mathbf{x}_i|\boldsymbol{\theta})$ are constant with respect to θ_2; additionally, $\phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}_1)\phi(\theta_2)$ due to the fact that $\boldsymbol{\theta}_1$ and θ_2 are orthogonal. Therefore, we can write the joint integral as successive integrals according to Fubini's theorem[6]:

$$\begin{aligned}\frac{\partial \ell}{\partial a_{pq}}\Big|_0 = \sum_{i=1}^{N} \frac{1}{\Pr(\mathbf{x}_i)}\Big\{ &\int_{\theta_2} \theta_2 \phi(\theta_2) d\theta_2 \int_{\boldsymbol{\theta}_1} [x_{ip} - T_p(a_{pq}|\boldsymbol{\theta}, x_{ip} = 1) \\ &\pm x_{iq} - T_q(a_{pq}|\boldsymbol{\theta}, x_{iq} = 1)] L(\mathbf{x}_i|\boldsymbol{\theta}_1)\phi(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 \Big\} = 0\end{aligned} \tag{18}$$

It zeros out due to the fact that the first integral is the expectation of standard normal distribution, which further indicates that $a_{pq} = 0$ is either a global maximum or a saddle point. To illustrate this, the bivariate profile log-likelihood surfaces[7] of both the positive and the negative LD models as functions of (a_{pq}, a_p) are plotted as Figure 2.

[5] For the proofs and derivations in this section, we consider a more generic situation where there can be more than one primary factors.

[6] To check the prerequisite of Fubini's theorem, define $D(a_{pq}) = \theta_2[x_{ip} - T_p(a_{pq}|\boldsymbol{\theta}, x_{ip})] \pm \theta_2[x_{iq} - T_q(a_{pq}|\boldsymbol{\theta}, x_{iq})]$. Then $|D(a_{pq})| \le 2|\theta_2|$ implies the integrand is dominated.

[7] The profile log-likelihood at a certain point (s, t) shown in the graphs is obtained as the restricted maximum log-likelihood after fixing $a_{pq} = s$ and $a_p = t$.

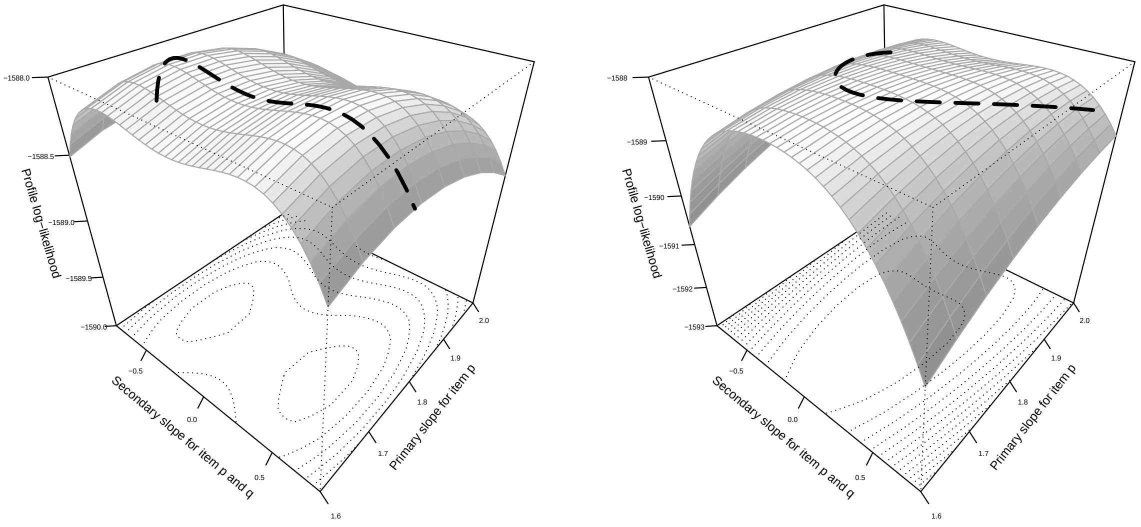

Figure 2: Bivariate profile log-likelihood surfaces for positive LD model (left; correct model for the data used here) and negative LD model (right; wrong model for the data used here): contours (i.e. dotted lines) and univariate profile log-likelihood curves (as a function of a_{pq} only; i.e. thick broken lines) are superimposed

As a result, we test the null hypothesis $a_{pq} = \varepsilon$ instead of $a_{pq} = 0$, with ε some small value (0.0001 in this research). Due to the symmetry of the likelihood function with respect to $a_{pq} = 0$, we only consider the part that $a_{pq} > 0$ (which can be visualized in Figure 2 as the part of log-likelihood surface on the right side of the thick solid line) when we compute the score statistics for both positive and negative LD models. After the two statistics associated with these two models are obtained, we check the sign of $\left.\frac{\partial \ell}{\partial a_{pq}}\right|_{0.0001}$ to determine which model to keep: the model with a positive value of the first derivative will be kept; if both models produce negative value, then we consider the local independence model as the correct model.

Method

1. Simulation design

To investigate the performance of bifactor LD score statistics, dichotomous item response data were generated using R (R Development Core Team, 2010) in three scenarios: the null case (i.e. local independence), a ULD case, and an SLD case.

For the locally independent case, we generated test data with 10, 25, and 50 items for the sample sizes of 200, 500, and 1000. Item parameters were sampled from the same distributions as Chen and Thissen (1997) used, which were believed to resemble the empirical distributions seen in practice. Specifically, the slopes are drawn from log-normal distribution $\log a \sim \mathcal{N}(0, 0.5)$, and the thresholds ($b = -c/a$) are drawn from $\mathcal{N}(0, 1.5)$. The data generation procedure was replicated 1000 times for each condition; LD statistics (i.e. bifactor score statistic S_b, threshold shift score statistic S_t, LD X^2) were extracted only for the first item pair of each replication.

Table 2: Conditions under the null case

No. of items	Sample sizes
10	200, 500, 1000
25	200, 500, 1000
50	200, 500, 1000

Design: No. of items × Sample sizes

ULD data were simulated using a bifactor version of the 2PL model, with only one LD pair for each replication. Apart from sample size and number of items, we also introduced variations of the strength and direction of LD (as shown in Table 3). The strength of LD was manipulated by changing the magnitude of secondary slopes sampled after being transformed to loadings (see Wirth & Edwards, 2007):

$$\lambda_{pq} = \frac{a_{pq}/1.702}{\sqrt{1 + (a_{pq}/1.702)^2}}; \tag{19}$$

$\lambda_{pq} \sim \mathcal{N}(\mu_\lambda, 0.01)$ where $\mu_\lambda = 0.3, 0.5, 0.7$, indicating low, moderate, and strong LD, respectively. In order to obtain valid values for loadings and control the dispersion of item parameters, all these sampling distributions were truncated to the interval $(\mu_\lambda - 0.2, \mu_\lambda + 0.2)$. As for direction, although negative LD does not have the same substantive interpretation as positive LD, they are mathematically the same. Thus, we generate 500 negative LD pairs out of the total 1000 replications under each simulated condition.

Table 3: Conditions under the ULD and SLD cases

No. of items	Sample sizes	Properties of LD pairs	
		Strength	Direction
10	200, 500, 1000	$\mu_\lambda = 0.3, 0.5, 0.7$	(500)+, (500)−
25	200, 500, 1000	$\mu_\lambda = 0.3, 0.5, 0.7$	(500)+, (500)−

50	200, 500, 1000	$\mu_\lambda = 0.3, 0.5, 0.7$	(500)+, (500)−

Design: No. of items × Sample sizes × Strength

Simulated conditions for SLD data are the same as ULD; however, the transformation between δ_{pq} and λ_{pq} are not exact because no apparently equivalent factor analysis model is defined for the threshold shift model. One approximation is to use the result proved (in Appendix A) with continuous indicators:

$$\lambda_{pq} = \sqrt{\delta_{pq}(1 - \lambda_p^2)} \tag{20}$$

where λ_p is the primary slope for the first item in each LD pair.

A modified version of the computational engine of the software IRTPRO was used for estimating item parameters and computing LD statistics.

2. Evaluation of LD statistics

The distribution of S_b, S_t, and X^2 under the null hypothesis can be obtained by pooling over all replications within each cell. For the locally independent case, we compare the empirical distributions to the χ_1^2 distribution[8] by means of quantiles. This shows the Type I error rates and empirical *p*-values of the LD statistics.

The receiver operating characteristic (ROC) curves of all three statistics for all conditions are presented to provide information about the power of these tests. The horizontal axis of each of ROC curve represents the false positive rate, or the alpha level in the setting of hypothesis testing; the vertical axis represents the true positive rate which is the power of the statistics.

Results

1. Locally independent data

Several important quantiles (i.e. 0.25, 0.5, 0.75, 0.9, 0.95, and 0.99[9]) of the empirical distributions are tabulated for each condition of the simulation in Table 4; the corresponding quantiles of the χ_1^2 distribution are included in the footnote of the table.

[8] χ_1^2 is the asymptotic distribution of S_b and S_t according to the large-sample theory, as well as the approximate distribution of LD X^2 (Chen & Thissen, 1997).

[9] The 0.99 quantiles estimated out of 1000 replications are not very reliable.

Table 4: Empirical quantiles of the null distribution of the three statistics

	10 items							25 items							50 items					
	25%	50%	75%	90%	95%	99%		25%	50%	75%	90%	95%	99%		25%	50%	75%	90%	95%	99%
N=200							N=200							N=200						
S_b	0.13	0.49	1.48	3.28	4.95	10.63	S_b	0.14	0.65	2.16	4.32	6.12	19.84	S_b	0.19	0.84	2.88	6.53	10.47	325.45
S_t	0.13	0.50	1.45	3.38	5.16	29.86	S_t	0.15	0.65	2.10	4.28	6.52	45.26	S_t	0.19	0.83	2.83	7.12	10.84	240.26
X^2	0.06	0.25	0.71	1.46	2.21	4.10	X^2	0.08	0.32	1.01	2.34	2.89	4.08	X^2	0.11	0.39	1.07	2.29	3.35	42.08
N=500							N=500							N=500						
S_b	0.11	0.46	1.29	2.56	3.73	8.96	S_b	0.10	0.49	1.61	3.46	5.28	11.27	S_b	0.11	0.54	1.70	3.46	5.26	10.56
S_t	0.10	0.45	1.31	2.67	3.88	8.28	S_t	0.11	0.50	1.68	3.43	5.16	11.44	S_t	0.12	0.58	1.70	3.48	5.38	11.85
X^2	0.05	0.26	0.71	1.52	2.26	4.69	X^2	0.07	0.31	0.96	2.08	2.83	4.78	X^2	0.11	0.39	1.06	1.99	2.89	5.58
N=1000							N=1000							N=1000						
S_b	0.11	0.46	1.34	2.85	4.03	7.46	S_b	0.09	0.48	1.44	2.82	4.37	6.79	S_b	0.11	0.53	1.58	3.34	4.80	8.79
S_t	0.11	0.48	1.33	2.94	4.16	7.35	S_t	0.09	0.48	1.37	2.87	4.46	7.11	S_t	0.13	0.52	1.58	3.34	4.87	10.57
X^2	0.06	0.27	0.75	1.77	2.39	4.03	X^2	0.07	0.32	0.93	1.83	2.67	4.35	X^2	0.14	0.40	1.05	2.03	3.12	5.61

Note: The corresponding quantiles for χ^2_1 is 0.10, 0.45, 1.32, 2.71, 3.84, and 6.63

A general trend is that both S_b and S_t tend to be liberal while LD X^2 is conservative if the χ_1^2 cutoffs are used. Both increasing test length and decreasing sample size result in larger statistics, which makes the distribution of LD X^2 closer to χ_1^2, but which further exacerbates the liberality of both score test statistics. The empirical quantiles of S_b and S_t are nearly identical; that can be seen more clearly in empirical cumulative density function (cdf) plots in Figure 3.

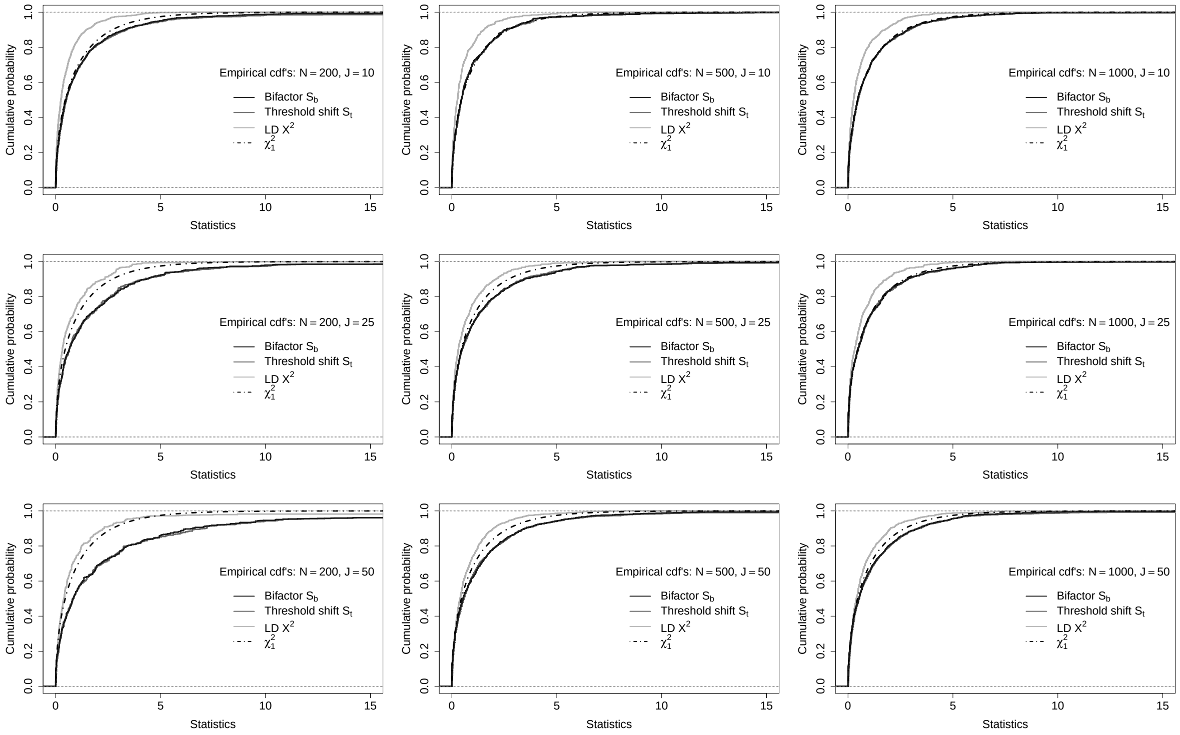

Figure 3: Empirical cdf plots

From Figure 3, it can be concluded that χ_1^2 serves as a good approximation of the null distribution for the score statistics S_b and S_t in a short test (i.e. 10 items) with medium (i.e. $N = 500$) or large (i.e. $N = 1000$) samples, as well as a medium length test (i.e. 25 items) with large samples. Otherwise, both score statistics are liberal. In all the conditions of this study, the empirical distributions of LD X^2 are always stochastically smaller than χ_1^2, which indicates conservativeness.

2. Surface LD data

Figures 4 to 6 show the ROC curves for the three statistics with data generated by the surface LD model.

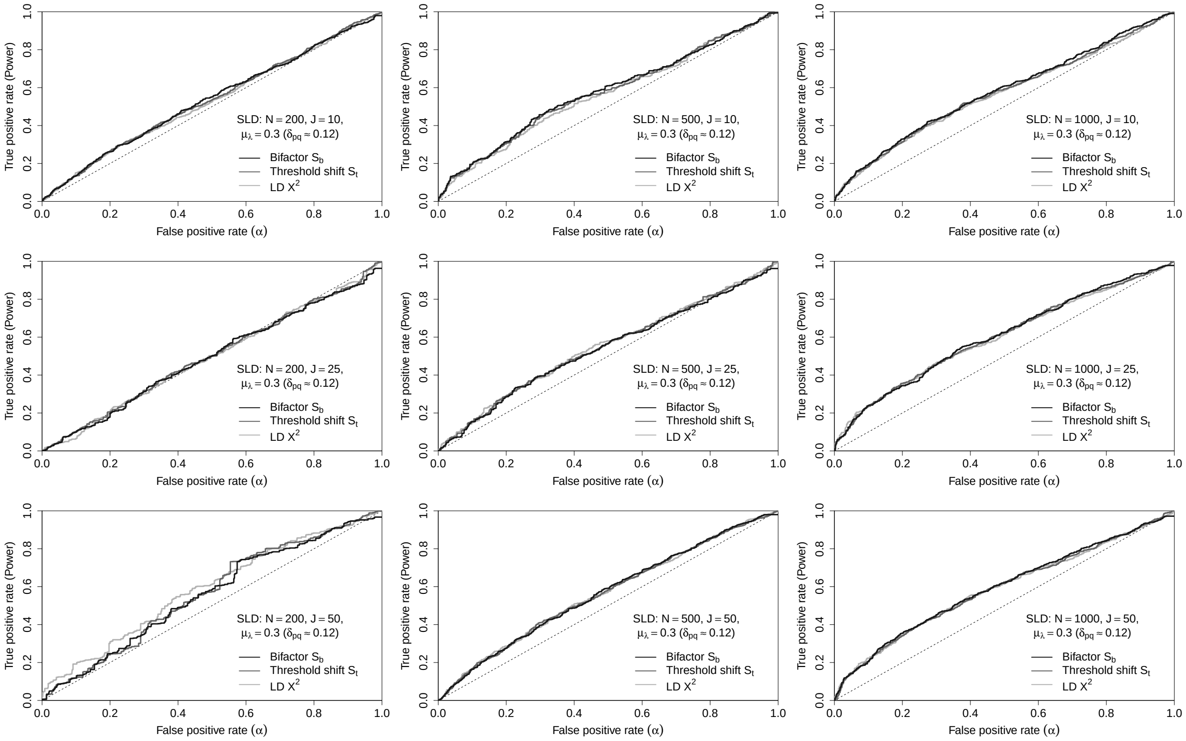

Figure 4: ROC curves for weak SLD ($\delta_{pq} = 0.16$)

In Figure 4, the results for weak SLD, with a threshold shift of only 0.16 standard units, show that there is very little difference between the empirical curves and the diagonal line of the ROC plots (i.e. random guess at rejecting null hypothesis) for any of the statistics, which indicates low power. Even though we gain some power with increasing sample size, it is no more than 0.2 with 1000 observations if 0.05 is used as the nominal level.

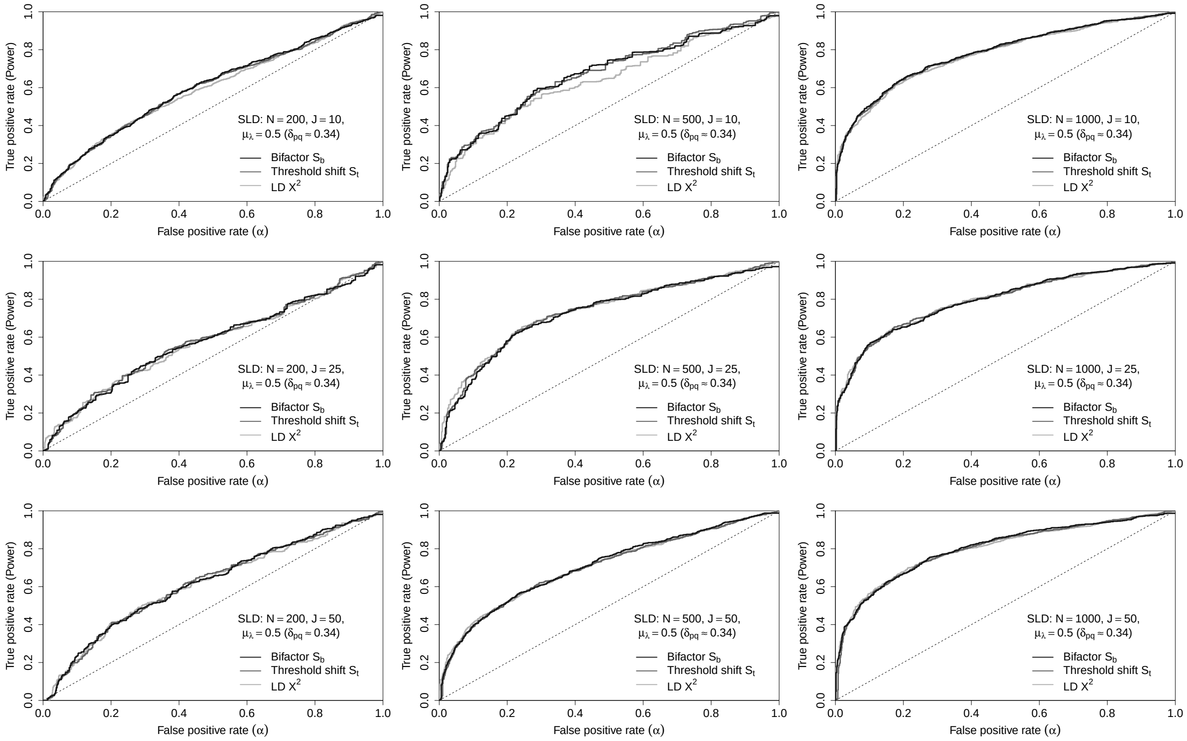

Figure 5: ROC curves for medium SLD ($\delta_{pq} = 0.34$)

Figure 5 shows that when $\delta_{pq} = 0.34$, there is, again, more power to detect LD in larger samples; the power is affected very little by test length. To illustrate, choosing 0.05 as the nominal level and only considering 50-item tests, the power of bifactor statistics is 0.11 for sample size 200, 0.28 for sample size 500, and 0.43 for sample size 1000. The empirical ROC curves for all three statistics are very similar.

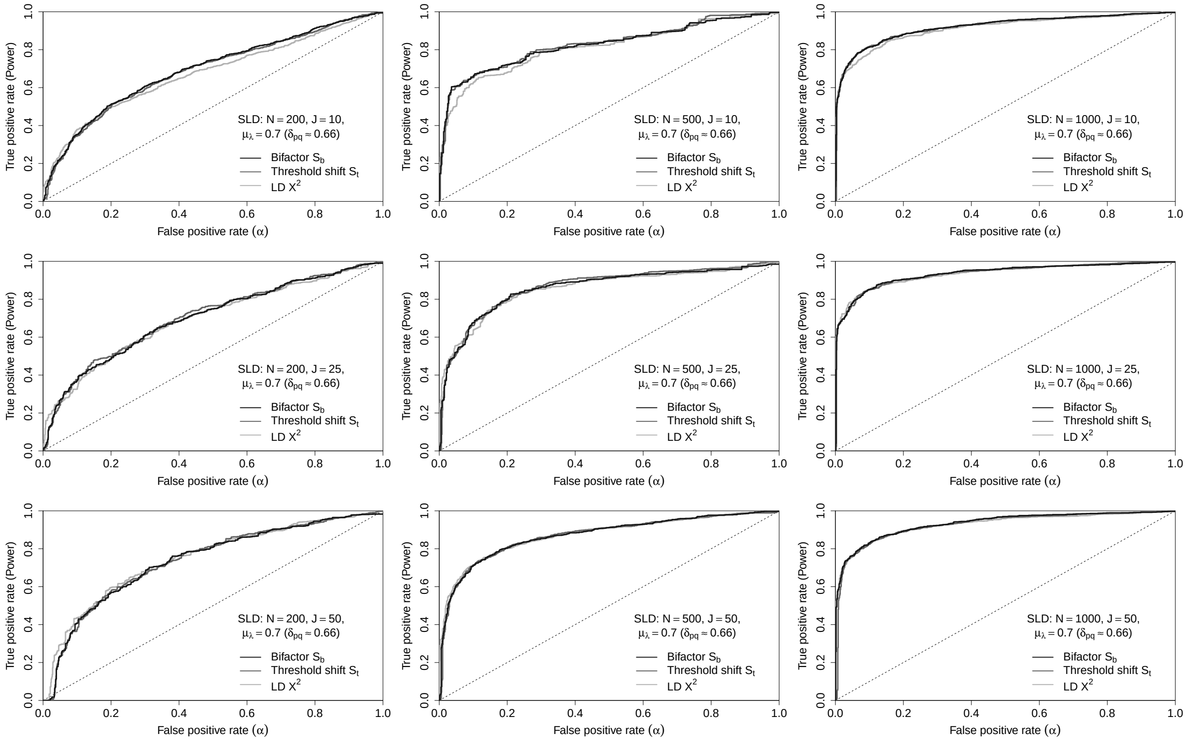

Figure 6: ROC curves for strong SLD (δ_{pq} = 0.66)

Figure 6 shows that as the threshold shift parameter increases to 0.66, the power of all statistics becomes higher (i.e. greater than 0.6) if sample size is 500 or 1000, but remains low/moderate (i.e. about 0.25) for small samples.

To summarize, the ROC curves of all three statistics are roughly identical, which reflects their similar performance in all the SLD conditions. The power increases when sample size increases, but remains about the same when test length changes.

3. Underlying LD data

Figures 7 to 9 show the ROC curves for the three statistics with data generated by underlying LD model.

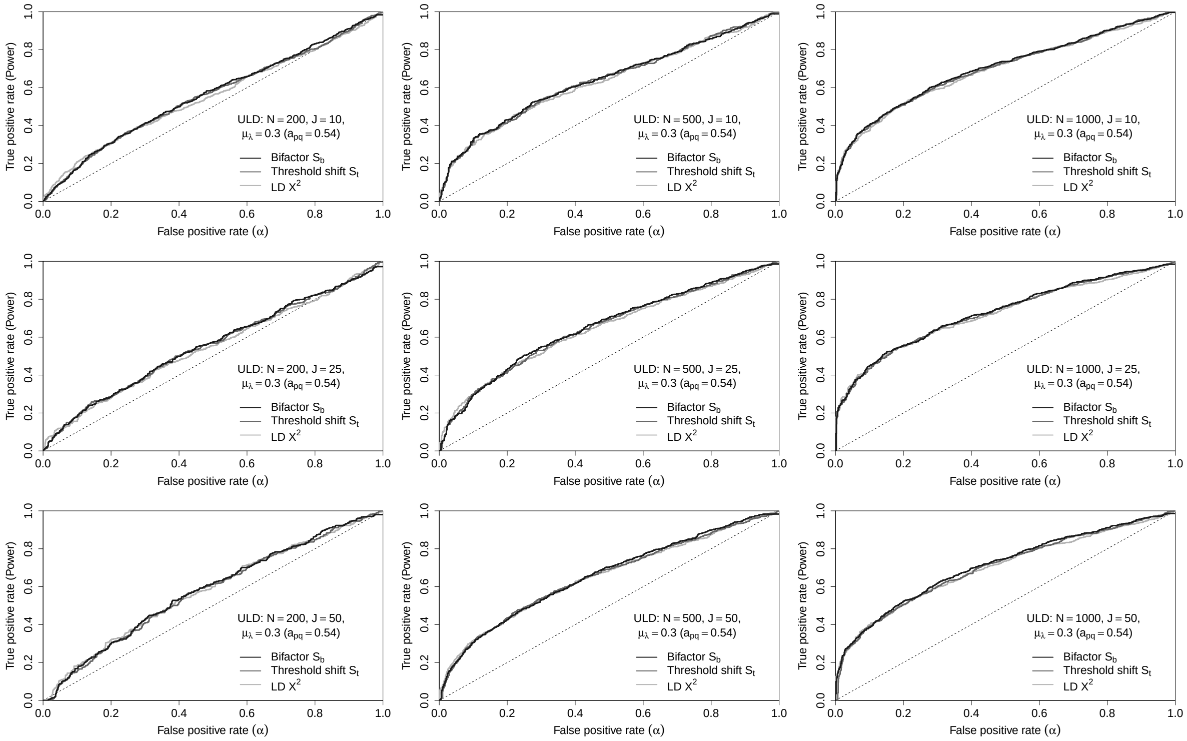

Figure 7: ROC curves for weak ULD ($a_{pq} = 0.54$)

Figure 7 shows that for the weak ULD condition ($a_{pq} = 0.54$; error covariance is 0.09), the power is relatively low or moderate for all cells; the power at $\alpha = 0.05$ ranges from (approximately) 0.1 to 0.3. However, the pattern of results is similar: Power increases as the number of observations increases, and is not influenced very much by the number of items; all three statistics have nearly the same ROC curves.

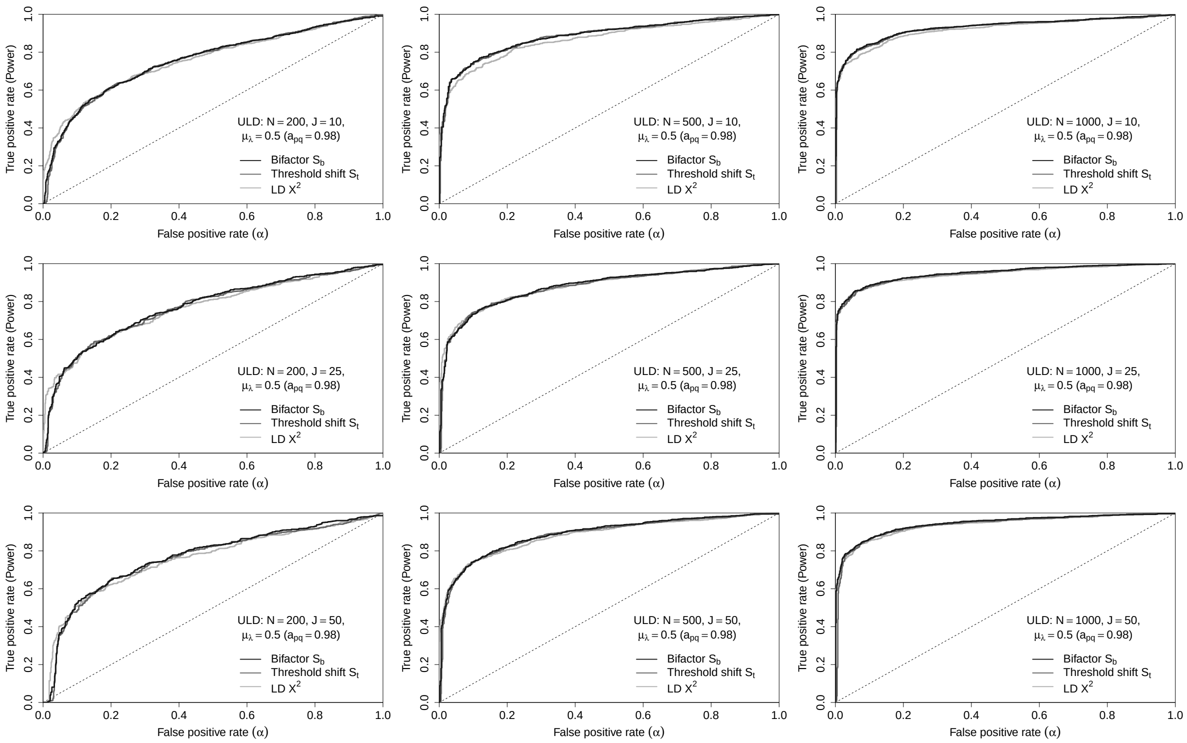

Figure 8: ROC curves for medium ULD ($a_{pq} = 0.96$)

Figure 8 shows that the power at nominal level 0.05 falls between 0.4 to 0.8 when $a_{pq} = 0.96$ (error covariance is 0.25) for all the statistics, which is high compared to the medium SLD conditions. Again, all three statistics yield similar results.

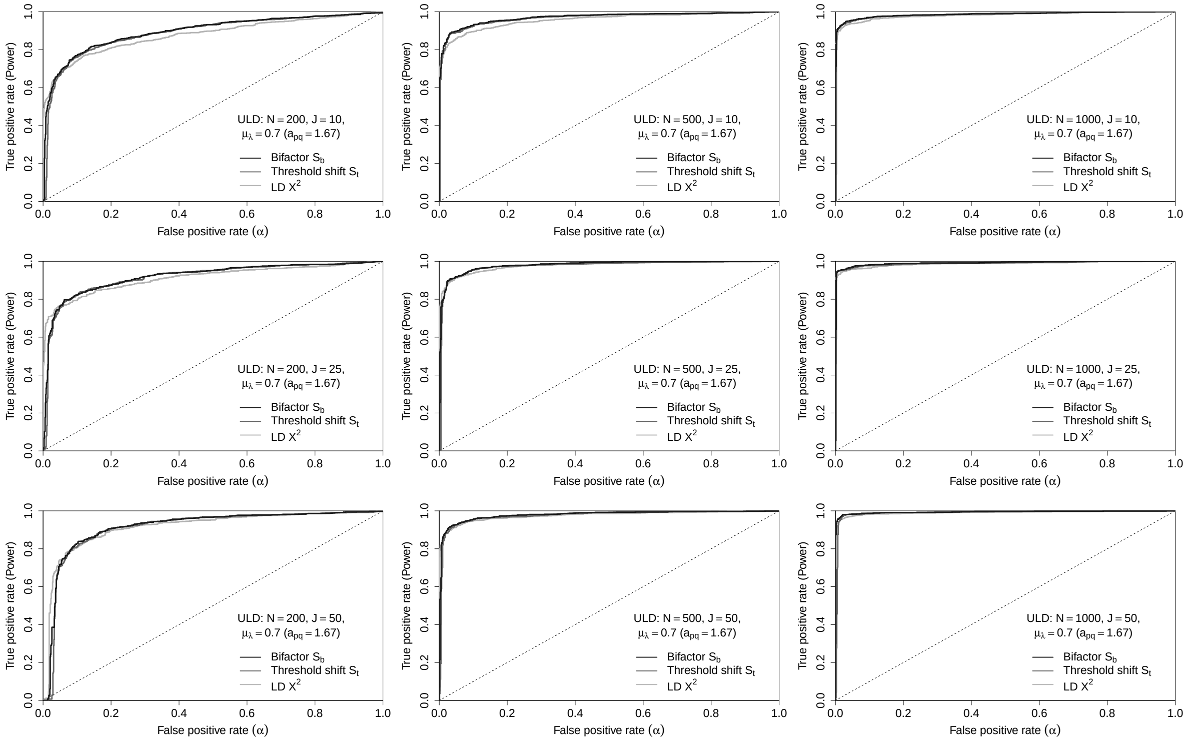

Figure 9: ROC curves for strong ULD (a_{pq} = 1.67)

Figure 9 shows that when a_{pq} increases to 1.67 (error covariance is 0.49), the power for all statistics is high for all cells; especially with large samples, all three procedures have nearly perfect power to detect underlying LD, which is shown by their extremely steep ROC curves.

In all, the results for ULD data are similar to those obtained for SLD data; the only difference is that all statistics seem to be more powerful here than in the corresponding SLD conditions. That may be attributed to the approximate formula (i.e. Equation 20) used to transform loadings to threshold shift parameters, which may lead to the unmatched LD levels between the SLD and ULD data generating models.

Discussion

For locally independent data, the empirical distribution of both score test statistics closely resembles χ^2_1, especially when sample size is large and the number of items is small (which means fewer parameters are estimated). However, we should be cautious when using χ^2_1 as the null distribution in smaller samples and/or with longer tests, because the results suggest that both score test statistics tend to be liberal. In contrast, LD X^2 is conservative if treated as χ^2_1, which is consistent with the results found by Chen & Thissen (1997);

however, it might be a good alternative for small samples and/or long tests.

With respect to power, all three statistics exhibit similar patterns in terms of receiver operating characteristics, whether SLD or ULD is the data generating model. Together with the similarity of their empirical null distributions, we conclude that both score test statistics are very close numerically. This calls the theoretical difference between surface and underlying local dependence into question; however, further investigation is required before reaching any conclusion about that.

In summary, all three statistics considered in this study provide useful diagnostic information about local dependence. LD X^2 is the easiest to compute and works well; its power is comparable to the score statistics, although it is conservative when χ^2_1 is used as its approximate null distribution. Both score statistics behave similarly; however, the computation of the bifactor statistic S_b requires numerically integrating over one more dimension which makes it computationally more expensive than the threshold shift statistic S_t.

In future research, it would be interesting to generalize both score test statistics to polytomous IRT models, e.g. the graded response model (Samejima, 1969). The generalization of bifactor statistic is straightforward; that involves adding another dimension to the original model. In contrast, because graded response model has more than one threshold, there are complications in generalizing the threshold shift statistic. Suppose the item pair has K response categories; one possibility is to incorporate different shift parameters for each of $K - 1$ non-zero responses of the first item on each of $K - 1$ thresholds of the second item, which leads the number of additional parameters to increase from one to $(K - 1)^2$. It might be unwieldy to explain each of these shift parameters; nevertheless, it is acceptable to compute a multivariate score test statistic (i.e. with limiting distribution $\chi^2_{(K-1)^2}$) without actually obtaining parameter estimates. Other possible extensions involve imposing constraints on the $(K - 1)^2$ shift parameters; for example, shift parameters on the same threshold of the second item may be constrained to be equal, which results in a $K - 1$ dimensional score test statistic. Further simulation studies are needed to evaluate these possible alternatives.

Based on all the results of the current study, we conclude that (1) LD X^2 is the easiest to compute, performs well, and has been implemented in commercial software; (2) Score test statistics S_t based on the threshold shift model is also easy to compute, and works better than LD X^2 in larger samples and shorter tests; (3) Score test statistics S_b based on the bifactor model is the hardest to compute, and provides no advantage over S_t for dichotomous data.

Appendix 1: Model equivalence

The equivalence of the bifactor model and the error covariance model can be established by comparing their covariance structures. Here we only prove the simplest case in which there is only one primary dimension and one secondary dimension (the proof can be easily generalized to multiple primary dimensions). Consider the bifactor model:

$$\begin{cases} x_1^* = \lambda_{11}\theta_1 + \lambda_{12}\theta_2 + \varepsilon_1 \\ x_2^* = \lambda_{21}\theta_1 \pm \lambda_{12}\theta_2 + \varepsilon_2 \end{cases} \tag{21}$$

which has covariance structure:

$$\begin{cases} \mathrm{Var}(x_1^*) = \lambda_{11}^2 + \lambda_{12}^2 + \mathrm{Var}(\varepsilon_1) \\ \mathrm{Var}(x_2^*) = \lambda_{21}^2 + \lambda_{12}^2 + \mathrm{Var}(\varepsilon_2) \\ \mathrm{Cov}(x_1^*, x_2^*) = \lambda_{11}\lambda_{21} \pm \lambda_{12}^2 \end{cases} \tag{22}$$

Similarly, the error covariance model:

$$\begin{cases} x_1^* = \lambda'_{11}\theta_1 + \varepsilon'_1 \\ x_2^* = \lambda'_{21}\theta_1 + \varepsilon'_2 \end{cases} \tag{23}$$

has covariance structure:

$$\begin{cases} \mathrm{Var}(x_1^*) = \lambda'^2_{11} + \mathrm{Var}(\varepsilon'_1) \\ \mathrm{Var}(x_2^*) = \lambda'^2_{21} + \mathrm{Var}(\varepsilon'_2) \\ \mathrm{Cov}(x_1^*, x_2^*) = \lambda'_{11}\lambda'_{21} + \mathrm{Cov}(\varepsilon'_1, \varepsilon'_2) \end{cases} \tag{24}$$

By equating the covariance structures (that is, Equation sets 24 and 22), we have:

$$\begin{cases} \lambda'_{11} = \lambda_{11} \\ \lambda'_{21} = \lambda_{21} \\ \mathrm{Var}(\varepsilon'_1) = \lambda_{12}^2 + \mathrm{Var}(\varepsilon_1) \\ \mathrm{Var}(\varepsilon'_2) = \lambda_{12}^2 + \mathrm{Var}(\varepsilon_2) \\ \mathrm{Cov}(\varepsilon'_1, \varepsilon'_2) = \pm\lambda_{12}^2 \end{cases} \tag{25}$$

The same procedure can be applied to prove the equivalence between the error covariance model and the threshold shift model. The only difference is that we need to include the covariance structure of a third item other than the item pair of interest. The threshold shift model is:

$$\left\{ \begin{aligned} x_1^* &= \lambda''_{11}\theta_1 + \varepsilon''_1 \\ x_2^* &= \lambda''_{21}\theta_1 + \beta_{21}x_1^* + \varepsilon''_2 \\ x_1^* &= \lambda''_{31}\theta_1 + \varepsilon''_3 \end{aligned} \right. \tag{26}$$

with covariance structure:

$$\left\{ \begin{aligned} &\mathrm{Var}(x_1^*) = \lambda''^2_{11} + \mathrm{Var}(\varepsilon''_1) \\ &\mathrm{Var}(x_2^*) = \lambda''^2_{21} + \beta^2_{21}\lambda''^2_{11} + 2\beta_{21}\lambda''_{11}\lambda''_{21} + \beta^2_{21}\mathrm{Var}(\varepsilon''_1) + \mathrm{Var}(\varepsilon''_2) \\ &\mathrm{Cov}(x_1^*, x_2^*) = \lambda''_{11}\lambda''_{21} + \beta_{21}\lambda''^2_{11} + \beta_{21}\mathrm{Var}(\varepsilon''_1) \\ &\mathrm{Cov}(x_2^*, x_3^*) = \lambda''_{21}\lambda''_{31} + \beta_{21}\lambda''_{11}\lambda_{31} \end{aligned} \right. \tag{27}$$

Augmenting Equation set 24 with

$$\mathrm{Cov}(x_2^*, x_3^*) = \lambda'_{21}\lambda'_{31} \tag{28}$$

and then comparing them with Equation set 27 yields:

$$\left\{ \begin{aligned} &\lambda'_{11} = \lambda''_{11} \\ &\lambda'_{21} = \lambda''_{21} + \beta_{21}\lambda''_{11} \\ &\lambda'_{31} = \lambda''_{31} \\ &\mathrm{Var}(\varepsilon'_1) = \mathrm{Var}(\varepsilon''_1) \\ &\mathrm{Var}(\varepsilon'_2) = \beta_{21}\mathrm{Var}(\varepsilon''_1) + \mathrm{Var}(\varepsilon''_2) \\ &\mathrm{Cov}(\varepsilon'_1, \varepsilon'_2) = \beta_{21}\mathrm{Var}(\varepsilon''_1) \end{aligned} \right. \tag{29}$$

The last equation above together with Equation set 25 justifies the transformation between threshold shift parameter (i.e. β_{21}) and the error covariance, then the secondary loading.

References

Bock, R. D. Notes on parameter estimation for polytomous item responses models. Unpublished manuscript.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for N dichotomously scored items. *Psychometrika*, *35*, 179–197.

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

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

Buse, A. (1982). The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, *36(3)*, 153–157.

Cai, L. (2010). A two-tier full-information item factor analysis model withÂăapplications. *Psychometrika*, *75*, 581–612.

Chen, W.-H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22(3)*, 265–289.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32.

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

Glas, C. A. & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27(2)*, 87–106.

Hoskens, M. & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2(3)*, 261–277.

Ip, E. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66(1)*, 109–132.

Jannarone, R. (1986). Conjunctive item response theory kernels. *Psychometrika*, *51*, 357–373.

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

Kendall, M. & Stuart, A. (1961). *The advanced theory of statistics. Vol. II*. London: Griffin.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, *6(4)*, 379–396.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551–560.

Neyman, J. & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, *20A(1/2)*, pp. 175–240.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, *44*, 50–57.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17*.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Thissen, D., Bender, R., Chen, W., Hayashi, K. & Wiesen, C. A. (1992). Item response theory and local dependence: A preliminary report. Research memorandum, L. L. Thurstone Psychometric Lab, the University of North Carolina at Chapel Hill, Chapel Hill.

Thissen, D. & Steinberg, L. (2009). Item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). London: Sage Publications.

van der Linden, W. & Glas, C. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120–139.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54(3)*, 426–482.

Wirth, R. & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12(1)*, 58 – 79.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8(2)*, 125–145.