UvA ECONOMETRICS

Discussion Paper: 2008/01

Exact tests for some latent traits

Jan G. de Gooijer and Ao Yuan

www.fee.uva.nl/ke/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Roetersstraat 11
1018 WB AMSTERDAM
The Netherlands

UvA UNIVERSITEIT VAN AMSTERDAM

Exact Tests For Some Latent Traits

Jan G. De Gooijer[1*] and Ao Yuan[2]

[1] Department of Quantitative Economics and Tinbergen Institute, University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: j.g.degooijer@uva.nl

[2] Statistical Genetics and Bioinformatics Unit
National Human Genome Center, Howard University
Washington DC, USA
e-mail: ayuan@howard.edu

Abstract: Item response theory is one of the modern test theory with many applications in the educational and psychological testing field. Recent developments made it possible to characterize some desired latent properties in terms of a collection of manifest ones, so that hypothesis tests on these latent traits can, in principle, be performed. But the existing test methodology is based on asymptotic approximation, which is impractical in most applications since the required sample sizes are often unrealistically huge. In this paper, we study a class of tests for making exact statistical inference about four manifest properties (CSN, MM, CA, and VCD). One major advantage is that these exact tests do not require large sample sizes and hence can be routinely adopted in empirical studies. Some numerical examples with applications of the exact tests are also provided.

AMS 2000 subject classification: Primary 62P15; secondary 60P03.

Key words and phrases: Conditional distribution; Exact test; Latent; Monte Carlo; Markov chain Monte Carlo.

*Corresponding author

1 Introduction

Item response theory (IRT, also called latent trait theory), as opposed to the classical statistical test methodology, is a modern theory of standardized tests which are commonly used in educational and psychological measurement settings. IRT provides a basis for the analysis of a collection of test items assigned to many subjects or examinees. Using various (non)parametric IRT models, the goal is to estimate a latent trait (parameter) such as an examinee's ability, attitude, or skill, that is measured by a particular test item. The traits are latent in the sense that they are not directly observable. Some basic references to the historical literature include Birnbaum (1968), Lord and Novick (1968), Fisher (1974), Cressie and Holland (1983), Joag-Dev and Proschan (1983), Holland and Rosenbaum (1986), Rosenbaum (1987), Stout (1987, 1990), and van der Linden and Hambleton (1997).

The majority of IRT models assume that the response data are unidimensional in the reference population, i.e. the item response probabilities are a function of a single underlying latent trait. Another condition of the models is monotonicity, i.e. the item response functions are nondecreasing functions in the latent variables. In this context the works by Junker (1991, 1993) and Junker and Ellis (1997) are worth mentioning. Their main results include an asymptotic characterization of monotone unidimensional latent trait models for dichotomously-scored items in terms of a collection of physically meaningful manifest properties. This is useful because manifest properties are amenable to conventional hypothesis testing. Recently, Yuan and Clarke (2001) developed asymptotic test methods for four manifest properties CSN, MM, CA, and VCD (see Section 2 for a brief introduction). However, since the desired latent properties are characterized by a (usually large) collection of statistics, the joint asymptotic validity requires unrealistically huge sample sizes. So, the proposed tests are mainly of theoretical interest, and have limited practical value. The objective of the current paper is to construct exact tests of certain latent traits, so that they can be carried out in practice.

The concept of exact test, originally proposed by Fisher (1935) for the inference of contingency tables, has received much attention and been extended to various settings since then. Under the null hypothesis the table usually has some kind of row or column or in both way independence, so that one conditional on the sufficient statistics of the parameters of interests, all the unknown parameters are left out, and the P-value of some test statistic can be computed under the parameter free exact distribution. Usually, direct computation of the P-value under the conditional distribution is difficult in practice. Instead, various Monte Carlo sampling methods are used for accurate approximations. Although an enumeration method is possible in some special cases, it is generally computationally infeasible. Based on permutations, for large tables, a simple Monte Carlo method may become a problem in sampling. In this case, Markov chain Monte Carlo sampling can be employed, which only updates a sub-table at each iteration. Hence the computation won't be limited by the table size.

In IRT inference, with data typically in the form of a table with binary entries, the null hypotheses are often composite. But for some latent hypotheses, tests can be performed on hypotheses on the parameter boundary, on which some kind of conditional independence can be achieved, so that parameter free exact tests can be performed. These simpler hypothesis tests are also tests for the original ones with the same significance level. We elaborate on these exact tests in subsequent sections. First, in the Section 2, we provide the key definitions, and notations. Next, in Section 3, we give four exact tests for four different manifest conditions. In Section 4 we present Monte Carlo test results for two manifest conditions using several unidimensional IRT models. As an illustration we also apply two tests to a familiar item response dataset. Section 5 concludes.

2 Notation and preliminaries

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. with $\mathbf{X} = (X_1, \ldots, X_J)$, a random vector of length J. Typically, in the educational testing context, it represents an examinee's testing scores on J items. $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})$ with X_{ij}'s be the binary (zero for wrong and one for correct) score of the i-th participant. The corresponding observations will be denoted by lower case letters. Let $X^+ = \sum_{j=1}^J X_j$, and $X^+(-j) = X^+ - X_j$. For an observed data table $\mathbf{t} = (x_{ij})$, with the i-th row $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})$ be the scores of the i-th participant over all the J items. Denote $\mathbf{T}$ the corresponding random table of $\mathbf{t}$. Let $x_i^+ = \sum_{j=1}^J x_{ij}$ be the i-th row total, $x_j^+ = \sum_{i=1}^n x_{ij}$ be the j-th column total, $\mathbf{x}^+ = (x_1^+, \ldots, x_J^+)$ be the vector of all the column totals, and $x^{++} = \sum_{i=1}^n \sum_{j=1}^J x_{ij}$ be the grand total.

In the exact test, conditional on the sufficient statistics $\mathbf{S}$ of the parameters of interests, one computes the P-value of some reasonably chosen test statistic $h(\mathbf{T})$ under the parameter free exact distribution, i.e.

$$P(h(\mathbf{T}) \geq h(\mathbf{t})|\mathbf{S}). \tag{1}$$

This test can also be derived from the Lehmann-Pearson framework as conditioning on the nuisance parameter, and under some regularity conditions it is Uniformly Most Powerful Unbiased, though not necessarily Uniformly Most Powerful (UMP) (Lehmann, 1986). Usually direct computation of (1) is difficult, instead, various sampling methods can be used. That is, sample $\mathbf{t}^{(n)}$ $(n = 1, \ldots, N)$ from the conditional distribution $P(\mathbf{T}|\mathbf{S})$, and (1) is approximated by

$$\hat{P}_N = \frac{1}{N} \sum_{n=1}^{N} \chi(h(\mathbf{t}^{(n)}) \geq h(\mathbf{t})),$$

where $\chi(\cdot)$ is the indicator function.

A general form for the joint probability of $\mathbf{X}$ is given by (Cox, 1972; Fitzmaurice and Laird, 1993; Zhao and Prentice, 1990)

$$P(\mathbf{X}) = \exp\{\Psi'\mathbf{X} + \Omega'\mathbf{W} - A(\Psi, \Omega)\}, \tag{2}$$

where Ψ and Ω are parameters and $\exp\{-A(\Psi, \Omega)\}$ is the normalizing constant, $\mathbf{W}$ is all the cross-product terms of $\mathbf{X}$, including all the second and higher order terms. Computation of the P-value of the test statistic under the observed data is infeasible, since there are too many unknown parameters in the above distribution. However, under the latent traits (i.e. hypotheses) of interest, model (2) often has a much simpler form. Then, conditioning on a suitable statistic $\mathbf{S}$, we can get the parameter-free exact distribution, based on which the tests will be performed.

Now we state the latent traits we want to test, and in Section 3 we discuss the corresponding testing statistics $h(\cdot)$, the conditioning statistic $\mathbf{S}$, the conditional distributions, and the sampling scans.

Junker (1993) introduced the notion of covariances given the sum are nonpositive (CSN) to characterize the general dependence nature between pairs of testing items. For self-content, we restate its definition below.

Definition (CSN): The covariances given the sum are nonpositive, if and only if for any $i < j \leq J$ the covariance between items i and j, given the mean, is negative. That is,

$$\mathrm{Cov}(X_i, X_j | X^+) \leq 0.$$

Also from Junker (1993), we have the following.

Definition (MM): Manifest monotonicity holds if

$$E(X_i | X^+(-i)) \ \text{ is nondecreasing as a function of } \ X^+(-j)$$

for all $i \leq J$ and all J.

The following concept, conditionally associated (CA) is from Holland and Rosenbaum (1986).

Definition (CA): The components in $\mathbf{X}$ are conditionally associated, if and only if for every pair of disjoint, finite response vectors $\mathbf{Y}$ and $\mathbf{Z}$ in $\mathbf{X}$, and for every pair of coordinatewise non-decreasing functions $f(\mathbf{Y})$ and $g(\mathbf{Y})$, and for every function $h(\mathbf{Z})$, and for every $c \in \text{range}(h)$ we have that

$$\text{Cov}(f(\mathbf{Y}), g(\mathbf{Y})|h(\mathbf{Z}) = c) \geq 0.$$

Let $\mathbf{X}_{J,k} = (X_{J+1}, \ldots, X_{J+k})$ be a k-vector of future items after $\mathbf{X}$. The following definition of vanishing conditional dependence (VCD) is from Junker and Ellis (1997).

Definition (VCD): $\mathbf{X}$ has vanishing conditional dependence, if and only if for any partition $(\mathbf{Y}, \mathbf{Z})$ of the response vector $\mathbf{X}$, and any measurable functions f and g (and any J) we have that

$$\lim_{k\to\infty} \text{Cov}(f(\mathbf{Y}), g(\mathbf{Z})|\mathbf{X}_{J,k}) = 0$$

almost surely.

Junker (1993), Junker and Ellis (1997) characterized the relationships among CSN, CA, MM, VCD and some other latent properties. To perform exact tests of the above latent properties, the key is to derive the conditional distributions for each of the properties and the corresponding sampling methods. In Section 3 we consider these one by one.

3 Construction of the tests

The exact tests below are based on the condition that the level α test is determined by the boundary condition under which all the J items are independent. Without this condition, the conditional distributions and related samplings will be difficult to handle. Since the tests will be non-standard, and often the number of parameters involved is huge, to test the null hypothesis H, we first simplify the conditions to be tested on the boundary of the parameter set, and a simpler hypothesis H_0, such that any level α test for H_0 is also a level α test for H, although these two hypotheses are not equivalent. Then tests for H_0 are constructed in a much easier way instead of those for H.

3.1 Test for CSN

Since the data are binary, X^+ can only take the values $0, 1, \ldots, J$ (values 0 and J are trivial, implying all the scores are 0 or 1), and CSN can be formulated as

$$r(i, j|k) := \mathrm{Cov}(X_i, X_j|X^+ = k) \leq 0, \quad 0 \leq i < j \leq J; \;\; 0 \leq k \leq J.$$

Given $X^+ = k$, the joint probability of $\mathbf{X}$ can be specified by (2) for each k.

Let $\mathbf{t}(k)$ be the $n_k \times J$ sub-table of all $\mathbf{x}_i$'s in $\mathbf{t}$ with $x_i^+ = k$. A natural estimate of $r(i, j|k)$ is (only for those with $n_k > 0$)

$$\hat{r}(i, j|k) = \frac{1}{n_k} \sum_{\mathbf{x}_s \in \mathbf{t}(k)} (x_{si} - \overline{x}_i)(x_{sj} - \overline{x}_j), \qquad (k = 1, \ldots, J-1) \tag{3}$$

where $\overline{x}_i$ and $\overline{x}_j$ are the means of the i-th and j-th item across all subjects in $\mathbf{t}(k)$. A reasonable choice for a test statistic for CSN is

$$h(\mathbf{t}) = r := \sum_{k=1}^{J-1} \frac{n_k}{n} \max_{i,j} \hat{r}(i, j|k). \tag{4}$$

Note that $\hat{r}(i, j|0) = \hat{r}(i, j|J) \equiv 0, \;\; \forall i, j$. Let

$$\Theta = \{r(i, j|k) : 0 \leq i < j \leq J; \;\; 1 \leq k \leq J-1\}$$

be the collection of all the $r(i, j|k)$'s, CSN can be written as $H : \Theta \leq \mathbf{0}$ (here "$\leq$" in the sense of componentwise). The rejection rule of a level α test of CSN has the form $h(\mathbf{t}) \geq h_0$ for some h_0 satisfies

$$\sup_{\Theta} P(h(\mathbf{t}) \geq h_0|\Theta) \leq \alpha.$$

Apparently, the above $\sup_\Theta$ is attained at $\Theta = \mathbf{0}$. Thus to get a level α test for CSN, we only need to construct a level α test for $H_0 : \Theta = \mathbf{0}$ vs. $K : \sup_{\theta \in \Theta} > \mathbf{0}$. When H_0 is true, $h(\cdot)$ tends to be close to zero, otherwise large.

Now we describe the exact test for H_0 vs. K. For this we first need the distribution of the data $\mathbf{t}$ under H_0, and then conditioning on a sufficient statistic of the parameters in the distribution to get a parameter free conditional distribution. Based on the conditional distribution, i.i.d. samples are drawn to evaluate the observed statistic given in (4), and to compute the Monte Carlo P-value under H_0. Conditional on $\mathbf{x}^+$ we have the following.

Proposition 1. Under H_0,

$$P(\mathbf{t}|\mathbf{x}^+) = \frac{\prod_{j=1}^{J} x_j^+!}{x^{++}!}, \tag{5}$$

Proof: Under H_0, for $i \neq j$ we have

$$Cov(X_i, X_j) = \sum_{j=k}^{J} Cov(X_i, X_j | \sum_{l=1}^{J} X_l = k) P(\sum_{l=1}^{J} X_l = k) = 0.$$

Since the X_i's are binary, we have

$$0 = Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = P(X_i = 1, X_j = 1) - P(X_i = 1)P(X_j = 1). \quad (6)$$

By (6) we get

$$\begin{aligned} P(X_i = 1, X_j = 0) &= P(X_i = 1) - P(X_i = 1, X_j = 1) \\ &= P(X_i = 1) - P(X_i = 1)P(X_j = 1) = P(X_i = 1)P(X_j = 0). \end{aligned}$$

Similarly

$$P(X_i = 0, X_j = 1) = P(X_i = 0)P(X_j = 1), \qquad P(X_i = 0, X_j = 0) = P(X_i = 0)P(X_j = 0).$$

Thus under H_0, X_i and X_j are independent for all $i \neq j$.

Let $p_j = P(X_j = 1)$ $(j = 1, \ldots, J)$ and $\mathbf{p} = (p_1, \ldots, p_J)$. Under H_0 the mass function of $\mathbf{t}$ is

$$P(\mathbf{T} = \mathbf{t}) = \prod_{i=1}^{n} \prod_{j=1}^{J} p_j^{x_{ij}} = \prod_{j=1}^{J} p_j^{x_j^+}.$$

Now we show $\mathbf{x}^+$ is a sufficient statistic for $\mathbf{p}$. For this we only need to show the conditional distribution of $\mathbf{t}$ given $\mathbf{x}^+$ is free of parameters and is given in (5). In fact, let $\mathbf{X}^+$ be the corresponding random variable for the observation $\mathbf{x}^+$, then under H_0, $\mathbf{X}^+$ is distributed as the multinomial $M(x^{++}, \mathbf{p})$, so

$$\begin{aligned} P(\mathbf{T} = \mathbf{t} | \mathbf{X}^+ = \mathbf{x}^+) &= \frac{P(\mathbf{T} = \mathbf{t}, \mathbf{X}^+ = \mathbf{x}^+)}{P(\mathbf{X}^+ = \mathbf{x}^+)} = \frac{P(\mathbf{T} = \mathbf{t})}{P(\mathbf{X}^+ = \mathbf{x}^+)} \\ &= \frac{\prod_{j=1}^{J} p_j^{x_j^+}}{\frac{x^{++}!}{\prod_{j=1}^{J} x_j^+!} \prod_{j=1}^{J} p_j^{x_j^+}} = \frac{\prod_{j=1}^{J} x_j^+!}{x^{++}!}. \qquad \square \end{aligned}$$

Proposition 1 tells us how to sample from (5). However, our purpose is to compute the test statistic from (3) or/and (4) for each new sample. Specifically, the Monte Carlo samples are drawn as follows.

Get the sub-tables $\mathbf{t}(k)$, $(k = 1, \ldots, J-1)$ from the observation $\mathbf{t}$, and compute the $\hat{r}(i, j)|k)$'s by (3), then compute $r_0 = h(\mathbf{t})$ by (4). To draw the Monte Carlo samples, we first compute the column totals $\mathbf{x}^+ = (x_1^+, \ldots, x_J^+)$. Now the Monte Carlo sampling is performed below: Specify an integer M, and let a sequence $z_1, \ldots, z_M$ to be assigned in the sampling process. For $m = 1, \ldots, M$ do the following steps:

(i) Draw a sample $\mathbf{t}^{(m)}$ from (5), which is realized by a random permutation of the j-th column $\mathbf{t}_j$ of $\mathbf{t}$, for each $j = 1, \ldots, J$ independent of each other.

(ii) For $k = 1, \ldots, J-1$, compute $\mathbf{t}^{(m)}(k)$, which is composed of all the row vectors in $\mathbf{t}^{(m)}$ with row total k. The size $n_k^{(m)}$ is the number of rows in $\mathbf{t}^{(m)}(k)$.

(iii) Compute the $r^{(m)}(i,j|k)$'s by (3) based on $\mathbf{t}^{(m)}(k)$, for each k, then compute $r^{(m)} = h(\mathbf{t}^{(m)})$ using the $r^{(m)}(i,j|k)$'s and $n_k^{(m)}$'s by (4). If $r^{(m)} \le r_0$, let $z_m = 1$ otherwise $z_m = 0$.

The Monte Carlo P-value is $\hat{\alpha} = \frac{1}{M}\sum_{m=1}^{M} z_m$, its estimated standard error sd is given by

$$sd^2 = \frac{1}{M(M-1)}\sum_{m=1}^{M}(z_m - \overline{z})^2 = \frac{1}{M-1}\overline{z}(1-\overline{z}),$$

and $\overline{z} = \frac{1}{M}\sum_{m=1}^{M} z_m$; its $(1-\alpha)\%$ confidence interval is estimated by $[\overline{z} \pm \Phi^{-1}(1-\alpha/2)sd/\sqrt{M}]$ (Mehta, et al. 1988), where $\Phi^{-1}(1-\alpha/2)$ is the upper $(1-\alpha/2)\%$ quantile of the standard normal distribution. Since $sd^2 \approx \frac{1}{M-1}\hat{\alpha}(1-\hat{\alpha}) \le 1/4$, to estimate $\hat{\alpha}$ within accuracy β, one should choose $M \ge \Phi^{-2}(1-\alpha/2)/(4\beta^2)$. For $\alpha = 0.05$, $\beta = 0.01$, we have $M \ge (\frac{2.576}{2\times 0.01})^2 \approx 17,000$. If $\hat{\alpha}$ is smaller than some prespecified level α, H_0 and hence CSN is rejected.

Remark: The sampling above is based on permutation of data with size n. It is known that the amount of computation for permutation increases rapidly with n, and may result in computation overflow. In this case, instead of a full updating of the original data table in the sampling process, we only update a sub-table of it at each sampling step. Let n_ℓ be the number of examinees with ℓ ($\ell = 0, 1, \ldots, J$) scores. Then, replace step (i) above by

(i') For each $j = 1, \ldots, J$ draw an index vector $\mathbf{i}_j = (i_{j1}, \ldots, i_{jn_\ell})$ of length n_ℓ from $\{1, \ldots, n\}$, uniformly without replacement (so that all the i_{jn_ℓ}'s are different). This can be done as follows: divide $[0,1]$ into non-overlapping sub-intervals $I_1, \ldots, I_n$ with equal lengths. Draw $u_1 \sim U[0,1]$, if $u_1 \in I_{s_1}$, assign $i_{j1} = s_1$. Then draw $u_2 \sim U[0,1]$, if $u_2 \in I_{s_2}$ and $s_2 \ne s_1$, assign $i_{j2} = s_2$; if $s_2 = s_1$ (the possibility is zero), redraw $u_2 \sim U[0,1]$, if $u_2 \in I_{s_2}$ and $s_2 \ne s_1$, assign $i_{j2} = s_2$. Continue until all the i_{jn_ℓ}'s are assigned. Given this $\mathbf{i}_j$, let $\mathbf{t}_j(\mathbf{i}_j)$ be the sub-vector of length n_ℓ of $\mathbf{t}_j$ with indices in $\mathbf{i}_j$, do a permutation within $\mathbf{t}_j(\mathbf{i}_j)$ for $j = 1, \ldots, J$. Merge the results in a new table $\mathbf{t}^{(m)}$.

In this case the number M for the samples should be much larger to ensure the ergodicity of the Monte Carlo samples, and the convergence of the corresponding P-value. Note that P-values for other latent traits, expressed in terms of covariances $\le (\ge) 0$, can be computed in the way as above.

3.2 Test for MM

Using the notation of Yuan and Clarke (2001), let the total score of the i-th examinee over the J items, but the term for the j-th item be $x_i^+(-j) = \sum_{r=1,r\neq j}^J x_{ir}$. As a generic random variable this is $X^+(-j) = \sum_{i=1,i\neq j}^J X_i$, in which j indexes the item. Now, the quantity we use to test MM is $\Delta_k(-j) := E(X_j|X^+(-j) = k+1) - E(X_j|X^+(-j) = k)$, where $k = 0,\ldots,J-1$ and $j = 1,\ldots,J$. Let $\Theta = \{\Delta_k(-j) : k = 0,\ldots,J-1; j = 1,\ldots,J\}$. MM now is expressed as $H : \Theta \geq \mathbf{0}$ vs. $K : \Theta < \mathbf{0}$. We first get natural estimators of $\Delta_k(-j)$'s and so a test statistic for MM. For this we partition the collection of examinees' binary response vectors based on the values of $x_i^+(-j)$. Let $\mathbf{t}(k,-j) = \{\mathbf{x}_i : \ \ x_i^+(-j) = k\}$ $(k = 0,1,\ldots,J-1;\ j = 1,\ldots,J)$, and $t(k,-j) = |\mathbf{t}(k,-j)|$ is its cardinality. A natural estimate of $\Delta_k(-j)$ is

$$\hat{\Delta}_k(-j) = \frac{1}{t(k+1,-j)} \sum_{\mathbf{x}_i \in \mathbf{t}(k+1,-j)} x_{i,j} - \frac{1}{t(k,-j)} \sum_{\mathbf{x}_i \in \mathbf{t}(k,-j)} x_{i,j}. \tag{7}$$

In the above we use the convention $\sum_{\mathbf{x}_i \in \mathbf{t}(k,-j)} x_{i,j}/t(k,-j) = 0$ if $t(k,-j) = 0$. A reasonable choice for $h(\cdot)$ is

$$h(\mathbf{t}) = \hat{\Delta} := \sum_{0\leq k<J-1;1\leq j\leq J} \frac{t(k,-j)+t(k+1,-j)}{2Jn}\hat{\Delta}_k(-j). \tag{8}$$

When MM is not true $h(\mathbf{t})$ will tend to be small. By the same argument as for CSN, to get a level α test for H, we only need to construct a level α test for $H_0 : \Theta = \mathbf{0}$ vs. $K : \Theta < \mathbf{0}$. For two random variables X and Y, $X \perp Y$ denote X and Y are independent. Let $\mathbf{x}(k,-j) = (x_{+1}(k,-j),\ldots,x_{+J}(k,-j))$ be the vector of the observed column totals in $\mathbf{t}(k,-j)$. We have the following.

Proposition 2. Under H_0, (5) is still true in this case.

Proof: Under H_0, we have

$$\begin{aligned} P(X_j = 1|X^+(-j) = 0) &= E(X_j|X^+(-j) = 0) = E(X_j|X^+(-j) = 1) \\ &= \cdots = E(X_j|X^+(-j) = J-1), \end{aligned}$$

or

$$P(X_j = 1|X^+(-j) = 0) = P(X_j = 1|X^+(-j) = 1) = \cdots = P(X_j = 1|X^+(-j) = J-1),$$

so

$$\begin{aligned} P(X_j = 1) &= \sum_{k=0}^{J-1} P(X_j = 1 | X^+(-j) = k) P(X^+(-j) = k) \\ &= P(X_j = 1 | X^+(-j) = r) \sum_{k=0}^{J-1} P(X^+(-j) = k) = P(X_j = 1 | X^+(-j) = r), \end{aligned}$$

for any $0 \leq r \leq J-1$. Since X_j is binary, this implies that $X_j \perp X^+(-j)$ $(1 \leq j \leq J)$ for all J. In particular, take $j = 1$ and $J = 2$, we have $X_1 \perp X_2$; take $J = 3$ we have $X_1 \perp (X_2 + X_3)$ which in turn from the independence between X_1 and X_2 implies that $X_1 \perp X_3$; ..., $X_1 \perp X_j$ $(j \neq 1)$. Similarly, take $j = 2$ and $J = 2, 3, \ldots$, we have $X_2 \perp X_j$ $(j \neq 2)$, and finally, $X_1, \ldots, X_J$ are independent of each other. The rest proofs are the same as in Proposition 1. $\square$

To perform the exact test for H_0 vs. H_1, the procedures are similar to those for the CSN. We first get the tables $\mathbf{t}(k, -j)$'s $(k = 0, \ldots, J-1; j = 1, \ldots, J)$ from the observed table $\mathbf{t}$, compute $\Delta^{(0)}$ by (7) and (8). Then draw Monte Carlo samples $\mathbf{t}^{(m)}(k, -j)$'s according to (5) as in the sampling for CSN, then compute the $\hat{\Delta}^{(m)}$'s by (7) or/and (8). The Monte Carlo sampling to compute the P-value is similar as before: Specify an integer M and a sequence $z_1, \ldots, z_M$ as that for CSN. For $m = 1, \ldots, M$ do the following: a) Steps (i) and (ii) are similar as before; b) If $\Delta^{(m)} \geq \Delta^{(0)}$, let $z_m = 1$ otherwise $z_m = 0$. The Monte Carlo P-value for H_0 vs. K is $\overline{z}$, its estimated standard error and confidence interval are the counter parts in the testing for CSN. Clearly, the Remark applies also to this case.

3.3 Test for CA

The principle for testing CA will be the same as that for CSN. In the following we refer to the notations and Proposition 4.4 in Yuan and Clarke (2001). Under these notations, CA is equivalent to H:

$$\Theta = \{\mathrm{Cov}(\chi_A(X(\omega(j))), \chi_B(X(\omega(j))) | X(\omega'(j')) \in D) : (j, j', \omega, \omega', \prec, \prec', A, B, D)\} \geq 0$$

vs. $\quad K : \theta < 0$, where the range of $(j, j', \omega, \omega', \prec, \prec', A, B, D)$ is

$$\begin{gathered} j + j' \leq J; \quad \omega(j), \omega'(j') \in \Omega; \quad \omega(j) \cap \omega'(j') = \phi; \\ \prec \in \Lambda(\omega(j)); \quad \prec' \in \Lambda(\omega'(j')); \quad A, B \in \mathcal{S}(\prec_\omega); \quad D \subset \mathcal{S}(\prec'_{\omega'}). \end{gathered}$$

The cardinality of Θ will usually be enormous even for $J \geq 3$.

Let Θ_0 be the subset of Θ consisting of all the components of Θ for which $\omega(j)$ be the $(1,\dots,J)$-complement of $\omega'(j')$ and $\omega(j)=\omega_1(j_1)\oplus\omega_2(j_2)$ for some $\omega_1(\cdot)$, $\omega_2(\cdot)$ and $j_1+j_2=j$. As before, for a level α test of H vs. K, we need only to construct a level α test for $H_0:\Theta_0=\mathbf{0}$ vs. $K_0:\Theta_0>\mathbf{0}$. By similar reasoning as before, this corresponds to independence of $\chi_A(X(\omega_1(j_1)))$ and $\chi_B(X(\omega_2(j_2)))$ pairs, for any $A\in\mathcal{S}(\prec_{\omega_1})$ and $B\in\mathcal{S}(\prec_{\omega_2})$, conditional on the event $X(\omega'(j'))\in D$. Now, for each fixed j' and $\omega'(j')$, let $\Gamma_D=\Gamma_D(\omega'(j'))$ be all the vectors X's with $X(\omega'(j'))\in D$. For fixed j_1, j_2, $A\in\mathcal{S}(\prec_{\omega_1})$ and $B\in\mathcal{S}(\prec_{\omega_2})$, let $y_{AB|D}$, $y_{AB^c|D}$, $y_{A^cB|D}$ and $y_{A^cB^c|D}$ be the cell counts of the events AB, AB^c, A^cB and A^cB^c in the set Γ_D. Define $y_{A|D}=y_{AB|D}+y_{AB^c|D}$, $y_{B|D}=y_{AB|D}+y_{A^cB|D}$ and $y_{++|D}=y_{A|D}+y_{B|D}$. Then under H_0, the two by two contingency table $y_D:=(y_{AB|D},y_{AB^c|D},y_{A^cB|D},y_{A^cB^c|D})$ are columnwise independent, and its conditional distribution given $(y_{A|D},y_{B|D})$ is standard (Agresti, 1990)

$$P(y_D|y_{A|D},y_{B|D})=\frac{\begin{pmatrix} y_{A|D} \\ y_{AB|D}\end{pmatrix}\begin{pmatrix} y_{B|D} \\ y_{A|D}-y_{AB|D}\end{pmatrix}}{\begin{pmatrix} y_{++|D} \\ y_{A|D}\end{pmatrix}}. \tag{9}$$

For given $A\in\mathcal{S}(\prec_{\omega_1(j_1)})$, $B\in\mathcal{S}(\prec_{\omega_2(j_2)})$ and $D\subset\mathcal{S}(\prec'_{\omega'}(j'))$, let n_{ABD} be the sample size for all the observations satisfying $X(\omega_1(j_1))\in A$, $X(\omega_2(j_2))\in B$ and $X(\omega'(j')\in D$. If $n_{ABD}>2$, an estimate $\hat{r}_{ABD}$ of $r_{ABD}=\mathrm{Cov}(\chi_A(X(\omega(j))),\chi_B(X(\omega(j)))|X(\omega'(j'))\in D)$ can be constructed by its empirical version.

Since the cardinality of Θ is huge, it seems impractical to construct a closed form testing statistic even for H_0. Instead, we use a random scan sampling method as follows.

Let $\mathcal{S}_0(\prec'_{\omega'}(j'))$ be the collection of all $\prec'_{\omega'}(j')$s for some $1\le j'\le J$ to which the observation $\mathbf{x}_i(\omega'(j'))$ belongs for at least two i's. Define $\mathcal{S}_0(\prec_{\omega_1(j_1)})$ and $\mathcal{S}_0(\prec_{\omega_2(j_2)})$ similarly. Define $\mathcal{N}_0$ be all the integer triples (j',j_1,j_2) with $j'+j_1+j_2=J$ and that there are $D\in\mathcal{S}_0(\prec'_{\omega'}(j'))$, $A\in\mathcal{S}_0(\prec_{\omega_1(j_1)})$ and $B\in\mathcal{S}_0(\prec_{\omega_2(j_2)})$. Let W_1, W_2 and W' be the vectors of proportions of the observed A, B and D's. For a collection $\mathcal{C}$ of sets, denote $U(\mathcal{C})$ as the uniform distribution over $\mathcal{C}$, and $\mathcal{D}(W,\mathcal{C})$ be the weighted distribution over $\mathcal{C}$ with weights W.

Set a prespecified sample size M, and a sequence $z_1,\dots,z_M$ to be specified. For $m=1,\dots,M$, go over the following steps:

(i) Draw (j',j_1,j_2) from $U(\mathcal{N}_0)$, A from $\mathcal{D}(W_1,\mathcal{S}_0(\prec_{\omega_1(j_1)}))$, B from $\mathcal{D}(W_2,\mathcal{S}_0(\prec_{\omega_2(j_2)}))$ and D from $\mathcal{D}(W_3,\mathcal{S}_0(\prec'_{\omega'}(j')))$.

(ii) Given the above A, B, D, compute n_{ABD}, $y_{AB|D}$, $y_{A|D}$, $y_{B|D}$, $y_{++|D}$ and $\hat{r}_{ABD}$ from the observed data table.

(iii) Sample n_{ABD} of y_Ds from (9), and compute the estimate $\tilde{r}_{ABD}$, using the sampled data, of r_{ABD} by the same formula for $\hat{r}_{ABD}$.

(iv) If $\tilde{r}_{ABD} > \hat{r}_{ABD}$, set $z_m = 1$, else $z_m = 0$.

The estimated P-value and its estimated standard error are computed in the same way as before.

3.4 Test for VCD

Using the same notation as in the previous subsection. Proposition 5.1 in Yuan and Clarke (2001) says that VCD is equivalent to the condition that for each k there is an $\epsilon = \epsilon(k)$, with $\epsilon(k)$ going to zero, so that

$$\max_{j,\omega(j),A,B,D} |Cov(\chi_A(X(\omega(j)), \chi_B(X(\omega^c(j))|X_{J,k} \in D)| \leq \epsilon(k), \tag{10}$$

in which the operation $\max_{j,\omega(j),A,B,D}$ denotes the maximum over

$$1 \leq j < J; \quad \omega(j) \in \Omega; \quad A \in \mathcal{S}(\omega(j)); \quad B \in \mathcal{S}(\omega^c(j); \quad \text{and} \quad D \in \mathcal{S}_{J,k}.$$

Let $\theta = |Cov(\chi_A(X(\omega(j)), \chi_B(X(\omega^c(j))|X_{J,k} \in D)|$, $\Theta = \{\theta : j, \omega(j), J, k, A, B, D\}$, $\overline{\theta} = \max \theta \in \Theta$. Then CVD can be formulated as $H : \overline{\theta} < \epsilon$ vs. $K : \overline{\theta} \geq \epsilon$, for some ϵ. As before, for a level α test for H vs. K, if we use the testing statistic $\hat{\overline{\theta}}$ with rejection rule of the form: $\hat{\overline{\theta}} > \theta_0$, where θ_0 is θ evaluated at the observation (x_{ij}), then we only need to get a level α test for $H_0 : \overline{\theta} = 0$ vs. K. For fixed $\omega(j)$, J, k and D, let $G = G_D = \{i : \mathbf{x}_{J,k} = D\}$, $n_G = |G|$ be the cardinality of G, $Y_{i,1} = \chi_A(X_i(\omega(j))$, $Y_{i,2} = \chi_B(X_i(\omega^c(j))$, $Y_{i,3} = \chi_{A^c \cap B^c}(X_i(\omega(j))$, the Y_{ij}'s are binary and under H_0, they are independent conditional on $X_{J,k}$, and conditional on the Y's total will eliminate the nuisance parameters. Thus we have

$$P(Y|Y_{+1}, Y_{+2}, Y_{+3}) = \prod_{j=1}^{3} \frac{Y_{+j}!(n_G - Y_{+j})!}{n_G!}, \tag{11}$$

so the test will be similar to that of before. Also, sampling from (11) is the same as before.

Denote $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,J}, x_{i,J+1}, \ldots, x_{i,J+k})$ where $i = 1, \ldots, n$. The averages of examinees' scores over G are

$$\overline{\chi}_A(D) = (1/n_G) \sum_{i \in G} \chi_A(\mathbf{x}_i(\omega(j))) \quad \text{and} \quad \overline{\chi}_B(D) = (1/n_G) \sum_{i \in G} \chi_B(\mathbf{x}_i(\omega^c(j))).$$

So,

$$\hat{\theta} = \frac{1}{n_G} \left| \sum_{i \in G} (\chi_A(\mathbf{x}_i(\omega(j))) - \overline{\chi}_A(D))(\chi_B(\mathbf{x}_i(\omega^c(j))) - \overline{\chi}_B(D)) \right| \tag{12}$$

is an estimator of θ.

In principle, to test H_0 vs. K, we still need to go through all the combinations $\{j, \omega(j), J, k, A, B, D\}$ to find the maximum, which is impractical. Instead, we use random scan as in the previous section, in which, at each Monte Carlo iteration m, we randomly select a $\theta \in \Theta$, draw a sample $(\mathbf{x}_{ij}^{(m)})$, and compute $\hat{\theta}(\mathbf{x}^{(m)})$ and $\hat{\theta}(\mathbf{x})$. Any occurrence of $\hat{\theta}(\mathbf{x}^{(m)}) \geq \hat{\theta}(\mathbf{x})$ is evidence against H_0. Specifically, the sampling is as follows.

Specify a sample size M, a sequence $z_1, \ldots, z_M$ to be specified, and set $m = 0$. Then do the following

(i) Draw J_0 from $\{2, \ldots, J-1\}$, j from $\{1, \ldots, J_0 - 1\}$, k from $\{J_0 + 1, \ldots, J\}$, $\omega(j)$ from $\{1, \ldots, J_0\}$, A from $\mathcal{S}(\omega(j))$, B from $\mathcal{S}(\omega(j)^c)$, and D from $\mathcal{S}_{J,k}$.

(ii) For the above D, get the set G_D for the observation $\mathbf{x}$, if G_D is empty, go back to (i), else increase m by 1, compute $y_{i,1} = \chi_A(X_i(\omega(j))$, $y_{i,2} = \chi_B(X_i(\omega^c(j))$, $y_{i,3} = \chi_{A^c \cap B^c}(X_i(\omega(j))$, $(i = 1, \ldots, n_G)$, y_{+1}, y_{+2}, y_{+3} and $\hat{\theta}(\mathbf{y})$ by (11).

(ii) Sample $Y^{(m)}$ from (10), compute $\hat{\theta}(Y^{(m)})$, if $\hat{\theta}(Y^{(m)}) \geq \hat{\theta}(\mathbf{y})$, set $z_m = 1$, else $z_m = 0$. If $m < M$, go to (i); else, stop.

The Monte Carlo P-value and its estimated standard error are computed in the same way as before.

4 Simulation and illustration

The tests for CA and VCD above, although feasible here as compared to their theoretical versions, are still not convenient to use. They need unrealistic huge sample sizes to perform the formal tests. So here we concentrate on illustrating the exact tests, using Monte Carlo simulations for CSN and MM, and via an illustration.

4.1 Simulated data

Example 1: Two known unidimensional parametric IRT models for binary response are used: the one-parameter logistic model (1PLM, also called the Rasch model), and the two-parameter logistic model (2PLM). The 2PML, defined via the associated item response function, is given by

$$P(X_j = 1|\theta_i) = \frac{1}{1 + \exp(-a_j(\theta_i - b_j))}, \ (i = 1, \dots, n; j = 1, \dots, J) \tag{13}$$

where θ_i represents the ability of examinee i, a_j, and b_j are item parameters. a_j is the item discrimination parameter, and b_j represents the item difficulty parameter. The 1PML is a special case of (13) when $a_j = 1$ $(j = 1, \dots, J)$.

Using the computer program *WinGen2* (Han and Hambleton, 2007) we simulated item and person parameters, item responses for a set of $J = 10, 20$ items, and $n = 100, 200$ examinees. For the 1PLM b_j was sampled randomly from a $U[0.6, 1.9]$ distribution. This range was selected because estimated discrimination parameters for real data often fall within these values. θ_i was sampled randomly from a $N(0, 1)$ distribution. For the 2PLM the item discrimination parameters a_j were drawn from a log-normal distribution with mean 0 and standard deviation 0.25. The item difficulty parameters b_j were sampled from a $N(0, 1)$ distribution. These parameter distributions can be considered realistic in practice. The number of replications in all experiments was set at 100, with $M = 30,000$.

Table 1 shows empirical quartiles Q_1, Q_2 (median), and Q_3 of the computed P-values. It is quite obvious from the values of Q_2 that in all cases there is no indication to reject the null hypotheses, i.e. there is no violation of the CSN and MM properties. Moreover, the variability in the P-values as measured by the sample interquartile range $(Q_3 - Q_1)$ also indicates that there is no evidence against CSN and MM.

Given these results, it seems that CSN and MM are rather general properties of multivariate binary data. In fact, by reviewing the theory underlying monotonicity, Junker and Sijtsma (2000) show that MM holds for the 1PLM. For the 2PLM these authors construct three theoretical counterexamples in which MM fails. Two counterexamples give rise to a nearly perfect (deterministic) Guttman pattern. Indeed, by constructing such a pattern, we were able to reject MM using the sampling process discussed in Section 3.2. But, since the ideal of a Guttman scale is difficult to achieve in real testing, we do not explore this set-up here further. An example violating the CSN property is provided below.

Example 2: Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be an i.i.d. sample from $\mathbf{X} = (X_1, \dots, X_J)$. Further, let $\mathbf{Y} \sim N(\mathbf{0}, \Omega)$, where all the off-diagonal elements of the $J \times J$ covariance matrix Ω are positive and equal to r. If $Y_i < \Phi^{-1}(p_i)$, set $x_{ij} = 1$ otherwise $x_{ij} = 0$ $(i = 1, \dots, n; j = 1, \dots, J)$. Given this general set-up we consider testing for CSN with $n = 100$, $J = 10, 20$, $r = 0.6, 0.7, 0.8$,

Table 1: Empirical quartiles Q_1, Q_2, and Q_3 of P-values for testing the CSN and MM properties. Simulation results are based on 100 replications using Monte Carlo samples with $M = 30,000$.

Model	n	J	CSN		MM	
			Q_2	(Q_1, Q_3)	Q_2	(Q_1, Q_3)
1PLM	100	10	0.8732	(0.7121,0.9435)	0.6581	(0.5970,0.7009)
		20	0.6295	(0.4230,0.7450)	0.2611	(0.2283,0.3033)
	200	10	0.9814	(0.9478,0.9961)	0.9537	(0.9453,0.9619)
		20	0.9257	(0.7969,0.9852)	0.2284	(0.1821,0.2772)
2PLM	100	10	0.2930	(0.1054,0.6185)	0.4127	(0.3790,0.4594)
		20	0.9810	(0.9362,0.9944)	0.5769	(0.5573,0.5914)
	200	10	0.2338	(0.0901,0.5298)	0.3982	(0.3644,0.4296)
		20	0.9700	(0.9188,0.9959)	0.4831	(0.4647,0.5064)

Table 2: Empirical quantiles of P-values for testing the CSN condition. Simulation results are based on 100 replications of sample sizes $n = 100$ using Monte Carlo samples with $M = 30,000$.

		Empirical quantiles										
J	r	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
10	0.6	0.0028	0.0071	0.0105	0.0184	0.0418	0.0708	0.1051	0.1683	0.1858	0.2620	0.4851
	0.7	0.0002	0.0015	0.0035	0.0054	0.0130	0.0205	0.0323	0.0520	0.0687	0.1101	0.1939
	0.8	0.0000	0.0000	0.0000	0.0002	0.0006	0.0023	0.0029	0.0067	0.0113	0.0163	0.0455
20	0.6	0.0073	0.0232	0.0444	0.0457	0.0966	0.1487	0.2349	0.3629	0.4407	0.5371	0.6649
	0.7	0.0000	0.0000	0.0001	0.0001	0.0008	0.0019	0.0032	0.0076	0.0162	0.0286	0.0649
	0.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

$p_i = 0.5$. Table 2 shows empirical quantiles of P-values based on 100 replications using Monte Carlo samples with $M = 30,000$. We see that, when $r = 0.6$, the CSN condition is not rejected for $J = 10$ and $J = 20$ in quite a few cases. However, when $r = 0.7$ and $r = 0.8$ the null hypothesis of CSN is strongly rejected for both values of J. These results are typical. Thus the CSN condition can only be rejected when there is a strong form of positive dependence between the items in the dataset.

4.2 Illustration

As an illustration, we apply the exact tests for CSN and MM to a dataset taken from the 1992 Trial State Assessment Program in Reading at Grade 4 of the US National Assessment of Educational Progress; see Patz and Junker (1999, Table 1). The data concerns the responses of $n = 3,000$ students to a test made-up of $J = 6$ items. The number of Monte Carlo samples was fixed at $M = 30,000$. For CSN, the empirical median value (Q_2) of the $r^{(m)}$'s is -0.0206, the lower empirical quartile (Q_1) is -0.0220, and the upper quartile (Q_3) is -0.0191. With $r_0 = -0.009469$, the Monte Carlo P-value equals 0.999967 ($sd = 3.316 \times 10^{-5}$). Hence, on the basis of this result, CSN is not rejected. For MM, the median value of the $\Delta^{(m)}$'s is 1.1449, $Q_1 = 0.5389$, and $Q_3 = 1.7826$. With $\Delta^{(0)} = 1.123382 \times 10^{-9}$, the Monte Carlo P-value is 0.904633 ($sd = 1.696 \times 10^{-3}$). Hence, on the basis of this result, MM is not rejected.

5 Some concluding remarks

We proposed exact hypothesis tests for CSN, MM, CA, and VCD. The tests are computationally feasible and practical, with Monte Carlo P-values computed under H_0. Moreover, the Monte Carlo method may extend to some more latent traits. Nevertheless, the tests considered here may not be best ones in some sense and admit rooms for improvements. However, based on permutation, the amount of computation grows factorially (faster than exponential growth) along with the datatable size. So for collections with large table sizes, the simple Monte Carlo method may again becomes computationally impractical. For this, the Markov chain Monte Carlo method is to update a sub-table per iteration, so it can be used in practice without actual size limitation. Yuan and Yang (2005) proposed a Markov chain method for contingency table exact inference, in which a sub-table of user specified size is sampled at each iteration. This chain has high sampling efficiency and can be modified to the present case.

Finally, it is worth mentioning that the null hypotheses of CSN, MM, CA, and VCD considered here are not of the simple Pearson type. Hence tests with some optimality such as UMP tests, generally do not exist. Thus we have only dealt with level α tests for these hypothesis. We find level α tests on the corresponding H_0, which are also level α tests on the corresponding H. On each H_0, all the traits CSN, MM, CA and VCD have a common feature: columnwise independence, although on the corresponding H, these traits are not the same.

References

Agresti, A. (1990). *Categorical Data Analysis.* Wiley.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability (Part 5). In F. Lord and M. Novick (Eds.), *Statistical Theorems of Mental Test Scores*, (pp. 397-479). Addison-Wesley.

Cox, D.R. (1972). The analysis of multivariate binary data, *Applied Statistics*, **21**, 113-120.

Cressie, N. and Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, **48**, 129-141.

Ellis, J. and Junker, B.W. (1997). Tail measurability in monotone latent variable models. *Psychometrika*, **62**, 495-523.

Fisher, R.A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, **B98**, 39-54.

Fisher, G. (1974). *Einführung in die Theorie Psychologischer Tests: Grundlagen und Anwendungen.* Bern: Huber.

Fitzmaurice, G. and Laird, N.M. (1993). A likelihood-based method for analyzing longitudinal binary responses, *Biometrika*, **80**, 141-151.

Han, K.T. and Hambleton, R.K. (2007). *User's Manual for WinGen: Windows Software that Generates IRT Model Parameters and Item Responses.* Center for Educational Assessment Research Report No. 642, University of Massachusetts.

Holland, P.W. and Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent trait models. *Annals of Statistics*, **14**, 1523-1543.

Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables, with applications. *Annals of Statistics*, **10**, 286-295.

Junker, B.W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, **56**, 255-278.

Junker, B.W. (1993). Conditional association, essential independence, and monotone unidimensional item response models. *Annals of Statistics*, **21**, 1359-1378.

Junker, B.W. and Ellis, J. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics*, **25**, 1327-1343.

Junker, B.W. and Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, **24**, 65-81.

Lehmann, E.L. (1986). *Testing Statistical Hypothesis.* Wiley.

Lord, F.M. and Norvick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Mehta, C.R., Patel, N.R. and Senchaudhuri, P. (1988). Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association*, **83**, 999-1005.

Patz, R.J. and Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, **24**, 146-178.

Rosenbaum, P.R. (1987). Comparing item characteristic curves. *Psychometrika*, **52**, 217-233.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, **52**, 293-325.

Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika.*, **55**, 293-325.

van der Linden, W.J. and Hambleton, R.K. (1997, Eds.). *Handbook of Modern Item Response Theory.* Springer-Verlag.

Yuan, A. and Clarke, B. (2001). Manifest characterization and testing for certain latent properties. *Annals of Statistics*, **29**, 876-898.

Yuan, A. and Yang, Y. (2005). A Markov chain sampler for contingency table exact inference, *Computational Statistics*, **20**, 63-80.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika*, **77**, 642-648.