TI 2010-044/4
Tinbergen Institute Discussion Paper

Some Exact Tests for Manifest Properties of Latent Trait Models

Version 23 April 2010

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

[1] University of Amsterdam, and Tinbergen Institute, the Netherlands;
[2] Howard University, Washington DEC, USA.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Some Exact Tests for Manifest Properties of Latent Trait Models[I]

Jan G. De Gooijer[*,a], Ao Yuan[b]

[a] *Department of Quantitative Economics and Tinbergen Institute, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands*

[b] *Statistical Genetics and Bioinformatics Unit, National Human Genome Center, Howard University, Washington DC, USA*

Abstract

Item response theory is one of the modern test theories with applications in educational and psychological testing. Recent developments made it possible to characterize some desired properties in terms of a collection of manifest ones, so that hypothesis tests on these 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. To overcome this problem, a class of tests is proposed for making exact statistical inference about four manifest properties: covariances given the sum are non-positive (CSN), manifest monotonicity (MM), conditional association (CA), and vanishing conditional dependence (VCD). One major advantage is that these exact tests do not require large sample sizes. As a result, tests for CSN and MM can be routinely performed in empirical studies. For testing CA and VCD, the exact methods are still impractical in most applications, due to the unusually large number of parameters to be tested. However, exact methods are still derived for them as an exploration toward practicality. Some numerical examples with applications of the exact tests for CSN and MM are provided.

Key words: Conditional distribution, Exact test, Monte Carlo, Markov chain Monte Carlo

[I] The work of Ao Yuan is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048.

[*] Corresponding author at: Department of Quantitative Economics University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands, Tel.: +31-20-5254244; fax: +31-20-5254349.

Email addresses: `j.g.degooijer@uva.nl` (Jan G. De Gooijer), `ayuan@howard.edu` (Ao Yuan)

Accepted for publication in Computational Statistics & Data Analysis

1. Introduction

Item response theory (IRT), as opposed to the classical test theory, is a modern theory of standardized tests that are commonly used in educational and psychological measurement settings. In psychometrics, it describes the application of mathematical models to data from questionnaires and tests as a basis for measuring abilities, attitudes, or other variables. Items may be questions that have incorrect and correct responses, statements to indicate level of agreement, or patient symptoms scores, etc. IRT makes it possible in principle to analyse a collection of test items assigned to many subjects or examinees. Using various (non)parametric methods, the goal is to estimate a property (parameter) such as an examinee's ability, attitude, intelligence or strength of some traits. The properties are not directly observable. Once the parameter estimates are obtained, statistical tests are usually conducted to assess the extent to which the parameters predict item responses given the model used. Such tests provide information about the psychometric properties of assessment and the quality of estimates. The pioneering work of IRT occurred during the 1950s and 1960s, including the studies of the Educational Testing Service psychometrician Frederic Lord, the Danish mathematician George Rasch, and the Austrian sociologist Paul Lazarsfeld. Although the mathematical ground work was laid earlier, IRT gained popular application from the late 1970s and 1980s when the advent of computers provided the power for extensive evaluations. Compared to classical test theory, IRT generally has greater flexibility and provides more sophisticated information. It can perform many tasks that cannot be realized by using classical test theory. Some basic references to the historical literature in this field 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), Stout (1990), van der Linden and Hambleton (1997).

IRT models can be divided into two families: unidimensional and multidimensional. The unidimensional model assumes that the response data are unidimensional in the reference population, i.e. the item response probabilities are a function of a single underlying property. However, because of the greatly increased complexity, the majority of IRT research and applications utilize a unidimensional model. Another commonly used condition is monotonicity, i.e. the item response characteristic curves are nondecreasing functions. In this context the works by Junker (1991), Junker (1993), and Junker and Ellis (1997) are worth mentioning. Their main results include an asymptotic characterization of monotone unidimensional property 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: covariances given the sum are non-positive (CSN), manifest monotonicity (MM), conditional association (CA), and vanishing conditional dependence (VCD) (see Section 2 for a brief introduction). An IRT model can have none or some of the manifest properties mentioned above. However, since the desired properties are characterized by a (usually large) collection of statistics, the asymptotic validity requires unrealistically huge sample sizes. As a rule of thumb, asymptotic methods will be valid if the data sample size $n \geq 30m^2$, where m is the number of unknown parameters in the problem. In practice, data sample sizes are often much smaller than that required by valid asymptotic methods. For example, in investigating some of the manifest properties based on performances of students of some grade in a given region, the data size is often in the low hundreds or less. We will see latter (Sections 2 and 4) that the required sample sizes, using asymptotic methods, are in the thousands, tens of thousands and more for the above four manifest properties. So the asymptotic tests are apparently impractical. The objective of the current paper is to construct exact tests of certain properties, for datasets with relatively small sample sizes, so that many realistic studies can be carried out in practice such as the above mentioned example.

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 (MCMC) 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 hypotheses, testing can be performed on those specified on the boundary of the parameter set. As a result some kind of conditional independence can be achieved which gives rise to parameter-free exact tests. These simpler hypothesis tests are also tests for the original ones with the same significance level. We elaborate on the form of the exact tests in subsequent sections. In particular, exact tests for CSN and MM can be routinely performed in empirical studies. For CA and VCD, the exact methods are still impractical in most applications, due to the unusually large number of parameters to be tested. But we still derive exact computational methods for them as an exploration toward practicality. First, in Section 2, we provide key definitions, and notations. Next, in Section 3, we give four exact tests for four different manifest conditions. The finite-sample performance of two exact test statistics is considered in Section 4 by simulation for several unidimensional IRT models. This is followed by a small illustration of the tests to an empirical item response dataset. Finally, we provide some concluding remarks in Section 5.

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 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 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, 1987). Usually direct computation of (1) is difficult, instead, various sampling methods can be used. That is, sample $\mathbf{t}^{(n)}$ $(n = 1, \dots, N)$ 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 properties (characterized by their corresponding 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 properties we want to test, and in Section 3 we discuss the corresponding test statistics $h(\cdot)$, the conditioning statistic $\mathbf{S}$, the conditional distributions, and the sampling.

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,

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

Note CSN is an intuitive property since for a fixed total, increasing some component, the other components tend to decrease. But this property is not automatically true for all IRT models, as it requires all the components vary in a coordinated way. An IRT model should have some special dependence nature for this to be true. Thus in practice we expect some of the commonly used IRT models or the corresponding data tables possess this property.

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

Definition (MM): Manifest monotonicity holds if

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

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

The following concept, conditional association (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 nondecreasing 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.

An IRT model can have none, some, or even all of the four manifest properties defined above, and generally having one or some of the properties do not necessarily imply one or some of the other properties; see, e.g., Junker (1993), and Junker and Ellis (1997) for a characterization of the relationships among CSN, CA, MM, VCD and some other properties.

From the definitions of these properties, asymptotic methods will be valid for CSN provided the sample size $n \geq 30[J(J-1)/2]^2$; for MM if $n > 30[J(J-1)]^2$. Sample sizes for CA and VCD are much larger. In practice, tests are often made-up of $J > 3$ items. If $J = 6$, the valid sample size using asymptotic methods for CSN is $n \geq 6,750$; for MM we have $n > 27,000$; let alone sample sizes for the properties CA and VCD. We see that the sample size required for valid asymptotic methods are unrealistic in many applications. To perform exact tests of the above properties, the key is to derive the conditional distributions for each of the properties and the corresponding sampling methods. We will see that we only require the conditional models at the boundary of each assumption. This makes the corresponding models very simple, otherwise exact methods will be infeasible. In Section 3, we consider these issues one by one.

3. Construction of the tests

The exact tests derived below are based on the condition that the level α test is determined by the boundary condition under which all J items are independent. Without this condition, the conditional distributions and related samplings will be difficult to handle. Testing for CSN, MM, CA and VCD, denoted by the null hypothesis H, will be non-standard, and often the number of parameters involved will be huge. To overcome this problem, we first simplify the conditions to be tested on the boundary of the parameter set, giving rise to a simpler null hypothesis H_0. This will be done such that any level α test for H_0 is also a level α test for H, although these two hypotheses are not equivalent.

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), we can reformulate CSN as follows

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

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 $n_k > 0$)

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

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

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

Note that (4) tends to have small values under H_0 and big values under the alternative. This makes $h(\cdot)$ to be a valid test statistic. Clearly, $\hat{r}(i,j|0) = \hat{r}(i,j|J) \equiv 0, \ \ \forall i,j$. Let

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

be the collection of all $r(i,j|k)$'s. Then the null hypothesis for testing 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 satisfying

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

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. $K : \sup_{\theta \in \Theta} > \mathbf{0}$.

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 we condition 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

$$\begin{aligned} 0 &= Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) \\ &= P(X_i = 1, X_j = 1) - P(X_i = 1)P(X_j = 1). \end{aligned} \tag{6}$$

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

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

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

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

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 $r^{(m)} \leq r_0$, 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 variance sd^2 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 $\overline{z} = \frac{1}{M}\sum_{m=1}^{M} z_m$. The corresponding $100(1-\alpha)\%$ 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 $100(1-\alpha/2)\%$ quantile of the standard normal distribution. Since $sd^2 \approx \frac{1}{M-1}\hat{\alpha}(1-\hat{\alpha}) \leq 1/4$, to estimate $\hat{\alpha}$ within accuracy β, one should choose $M \geq \Phi^{-2}(1-\alpha/2)/(4\beta^2)$. For $\alpha = 0.05$, $\beta = 0.01$, we have $M \geq (\frac{2.576}{2\times 0.01})^2 \approx 17,000$. If $\hat{\alpha}$ is smaller than some prespecified level α, H_0 and hence CSN is rejected.

Remark: The sampling scheme 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 computational 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 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 $s_2 \neq s_1$, 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

$s_2 \neq s_1$, 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 ergodicity of the Monte Carlo samples, and the convergence of the corresponding P-value. Note that P-values for other properties, expressed in terms of covariances $\leq (\geq) 0$, can be computed in the same way as above.

3.2. Test for MM

Using the notation of Yuan and Clarke (2001), consider the total score of the i-th examinee over the J items, but subtract the term for the j-th item. Denote this by $x_i^+(-j) = \sum_{r=1, r\neq j}^J x_{ir}$. 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 $\Theta = \{\Delta_k(-j) : k = 0, \ldots, J-1; j = 1, \ldots, J\}$. So, the null hypothesis H: MM is equivalent to $H_0 : \Theta \geq \mathbf{0}$ vs. $K : \Theta < \mathbf{0}$. We first get natural estimators of $\Delta_k(-j)$'s, and so a test statistic for MM. To this end we partition the collection of examinees' binary response vectors based on the values of $x_i^+(-j)$. Let $\mathbf{t}(k, -j) = \{\mathbf{x}_i : \quad x_i^+(-j) = k\}$ ($k = 0, 1, \ldots, J-1$; $j = 1, \ldots, J$), and $t(k, -j) = |\mathbf{t}(k, -j)|$ is its cardinality. Now, 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 $h(\cdot)$ 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, given the independence between X_1 and X_2, implies that $X_1 \perp X_3, \ldots, X_1 \perp X_j$ $(j \neq 1)$. 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 of the proof is the same as in Proposition 1. $\square$

To perform the exact test for H_0 vs. H_1, the Monte Carlo procedure is similar to the one used for testing CSN. In particular, get the tables $\mathbf{t}(k, -j)$'s $(k = 0, \ldots, J - 1; j = 1, \ldots, J)$ from the observed table $\mathbf{t}$. Then compute $\Delta^{(0)}$ by (7) and (8). Next, draw Monte Carlo samples $\mathbf{t}^{(m)}(k, -j)$'s according to (5) as in the sampling setup for testing CSN. Then compute the $\hat{\Delta}^{(m)}$'s by (7) or/and (8). Further, the Monte Carlo sampling to compute P-values is similar as before. Specify an integer M and a sequence $z_1, \ldots, z_M$ similar 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 of these quantities corresponding to testing for CSN. Clearly, the Remark given in Section 3.1 applies also to this case.

3.3. Test for CA

In principle, testing CA will be the same as testing 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. $K : \theta < 0$, where the range of $(j, j', \omega, \omega', \prec, \prec', A, B, D)$ is

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

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 $(1, \ldots, J)$-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 $1 \leq j' \leq J$ to which observation $\mathbf{x}_i(\omega'(j'))$ belongs to 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 $z_1, \ldots, z_M$ to be specified. For $m = 1, \ldots, M$, 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 for CA. Also, sampling from (11) parallels sampling of CA given in Section 3.3.

Denote $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,J}, x_{i,J+1}, \dots, x_{i,J+k})$ where $i = 1, \dots, n$. 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 $z_1, \dots, z_M$ to be specified, and set $m = 0$. Then do the following:

(i) Draw J_0 from $\{2, \dots, J-1\}$, j from $\{1, \dots, J_0 - 1\}$, k from $\{J_0 + 1, \dots, J\}$, $\omega(j)$ from $\{1, \dots, J_0\}$, 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 $y_{i,1} = \chi_A(X_i(\omega(j))$, $y_{i,2} = \chi_B(X_i(\omega^c(j))$, $y_{i,3} = \chi_{A^c \cap B^c}(X_i(\omega(j))$, $(i = 1, \dots, n_G)$, y_{+1}, y_{+2}, y_{+3} and $\hat{\theta}(\mathbf{y})$ by (12).

(ii) Sample $Y^{(m)}$ from (11). 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. Finite-sample performance

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. This section presents three sets of Monte Carlo experiments illustrating the finite-sample performance of the exact tests for CSN and MM. In all experiments the number of replicates is set at 1,000, with $M = 30,000$. Although this setup allows for meaningful power results, the actual number of replicates may be considered low. But the costs in computing the tests statistics for CSN and MM with $M = 30,000$ was a limitation for considering a larger number of replicates.

4.1. First experiment

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 2PLM, defined via the conditional probability of an item response, is given by

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

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 1PLM is a special case of (13) when $a_j = 1$ $(j = 1, \ldots, J)$; see, e.g., Patz and Junker (1999) and van der Linden and Hambleton (1997) for more details on these models.

Using the computer program *WinGen2* (Han and Hambleton, 2007) we simulate item and person parameters, item responses for a set of $J = 10, \ 20$ items, and $n = 25$ and 50 examinees. For the 1PLM, b_j is sampled randomly from a $U[0.6, 1.9]$ distribution. This range is selected because estimated discrimination parameters for real data often fall within these values. θ_i is

Table 1: Empirical quartiles Q_1, Q_2, and Q_3 of 1,000 computed P-values for testing CSN and MM, and number of P-values < 0.05; Experiment 1.

Model	n	J	CSN		MM		No. P-values < 0.05	
			Q_2	(Q_1, Q_3)	Q_2	(Q_1, Q_3)	CSN	MM
1PLM	25	10	0.466	(0.233, 0.683)	0.234	(0.157, 0.342)	59	14
		20	0.237	(0.060, 0.512)	0.332	(0.225, 0.471)	76	13
	50	10	0.640	(0.385, 0.830)	0.192	(0.148, 0.249)	23	2
		20	0.327	(0.165, 0.530)	0.323	(0.192, 0.403)	76	0
2PLM	25	10	0.437	(0.191, 0.728)	0.358	(0.290, 0.448)	72	2
		20	0.360	(0.110, 0.678)	0.391	(0.263, 0.528)	20	17
	50	10	0.300	(0.131, 0.558)	0.259	(0.216, 0.308)	106	0
		20	0.789	(0.569, 0.917)	0.366	(0.317, 0.432)	6	1

sampled randomly from a $N(0, 1)$ distribution. For the 2PLM the item discrimination parameters a_j are drawn from a log-normal distribution with mean 0 and standard deviation 0.25. The item difficulty parameters b_j are sampled from a $N(0, 1)$ distribution. These parameter distributions can be considered realistic in practice.

Table 1 shows empirical quartiles Q_1, Q_2 (median), and Q_3 of the 1,000 computed P-values. It is quite obvious from the values of Q_2 that in a large number of 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 $(Q_3 - Q_1)$ is low. The last two columns of Table 1 show the number of P-values less than 0.05 out of 1,000 replications. Recall from Section 3.1 that the nominal level α is established on the boundary of the null parameter space of $\Theta = \mathbf{0}$. When the actual case is $\Theta < \mathbf{0}$, the observed nominal levels can be significantly smaller than α. So when we observe 0 rejections out of 1,000 replications in Table 1, this does not mean the test for MM is of level $\alpha = 0$. Rather it means that the actual case is more likely $\Theta < \mathbf{0}$. It should be noted here that for any specified α, a size-α critical value $h(\alpha)$ can only be obtained via Monte-Carlo sampling under H_0, as the $(1-\alpha)$-th sample quantile. For instance, for CSN this implies following steps (i)–(iii) in Section 3.1. Then a size-α test for CSN is given by the rejection rule: reject the null, if $h(\text{observed}) > h(t(\alpha))$, where $t(\alpha)$ corresponds to the α-th quantile of the null distribution. Hence, the size of the tests is not related to the number of P-values less than 0.05.

Given the above 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) showed 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 scale, i.e. the items constitute a unidimensional ordered series such that an answer to a given item predicts the answers to all previous items in the series. Indeed, by constructing such a scale, we are 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 issue here further. Experiment 2 below presents a counterexample in which the CSN property is rejected.

4.2. Second experiment

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 $n = 25,\ 50$, $J = 10$, $r = 0.5,\ 0.6,\ 0.7$, and $p_i = 0.5$. Table 2 shows empirical quantiles of 1,000 computed P-values. We see that, when $n = 25$ and $r = 0.5$, the CSN property is rejected in quite a few cases, with 38% of the P-values lying between 0 and 0.05. When $n = 25$ and $r = 0.6,\ 0.7$ these percentages are 61.3% and 78.6% respectively. Thus, as the correlation increases the null hypothesis of CSN is more strongly rejected. This result is typical for other sample sizes and values of J.

4.3. Third experiment

For the last experiment, we compute the exact tests for CSN and MM using data taken from the 1992 Trial State Assessment Program in Reading at Grade 4 of the US National Assessment of Educational Progress (NAEP). In fact, the dataset under study concerns a random sub-sample of size $n = 3,000$ drawn from the population of fourth-grade students in the US; see Patz and

Table 2: Empirical quantiles of 1,000 P-values for testing the CSN property ($J = 10$); Experiment 2.

		Empirical quantiles								
n	r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
25	0.5	0.002	0.015	0.033	0.063	0.109	0.165	0.260	0.377	0.565
	0.6	0.000	0.000	0.005	0.011	0.025	0.051	0.091	0.174	0.344
	0.7	0.000	0.000	0.000	0.000	0.004	0.011	0.026	0.058	0.146
50	0.5	0.014	0.038	0.082	0.130	0.182	0.256	0.343	0.451	0.597
	0.6	0.002	0.009	0.018	0.037	0.059	0.091	0.158	0.233	0.383
	0.7	0.000	0.001	0.002	0.006	0.013	0.024	0.041	0.077	0.144

Junker (1999), Table 1. The responses concern $J = 6$ items from each student, for each item a response of 1 represents a correct answer and 0 for incorrect. The questions themselves and the associated reading passage have not been publicly released by NAEP. Patz and Junker (1999) analysed the complete dataset (3,000 examinees) using MCMC sampling methods for 2PLM item calibration. Here we assume that the dataset has the nature of a population, and 1,000 random samples of sizes $n = 25,\ 50,\ 100$ and 200 are drawn without replacement from the full dataset. Recall that these sample sizes are far less than the minimal sample sizes required for using asymptotic methods for CSN (No. of parameters =15, $n \geq 6,750$) and MM (No. of parameters =30, $n > 27,000$).

Table 3 shows empirical quartiles Q_1, Q_2, and Q_3 computed on the basis of 1,000 P-values. Clearly, for all values of n the empirical quartiles do not show evidence to reject the CSN property, with less evidence against violation of the null hypothesis as n increases from 25 to 200. Interestingly, the opposite occurs when testing for MM. That is, evidence to reject the MM property increases as n increases. The next step would be to fit 2PLMs to the 1,000 data subsets for each n. Then, following Junker and Sijtsma (2000), estimates of $P(X_j = 1|X^+(-j))$ may well reveal violations of monotonicity at certain locations of its empirical distribution.

Table 3: Empirical quartiles Q_1, Q_2, and Q_3 of 1,000 P-values for testing the CSN and MM properties; Experiment 3.

n	CSN		MM	
	Q_2	(Q_1, Q_3)	Q_2	(Q_1, Q_3)
25	0.3236	(0.1419, 0.5852)	0.2571	(0.1742, 0.3863)
50	0.2796	(0.1111, 0.5191)	0.1760	(0.1012, 0.2815)
100	0.3966	(0.2242, 0.5925)	0.0855	(0.0474, 0.1836)
200	0.5742	(0.4037, 0.7458)	0.0288	(0.0173, 0.0499)

5. Some concluding remarks

We propose exact hypothesis tests for CSN, MM, CA, and VCD. In particular, tests for CSN and MM are now computationally feasible and practical, with Monte Carlo P-values computed under H_0. For CA and VCD to be practical, it is still open to further research. Moreover, the Monte Carlo method may extend to some more properties. 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 MCMC method is to update a sub-table per iteration, so it can be used in practice without actual size limitation. Yuan and Yang (2001) 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. For data with really large table size, this method can be considered to refine our method.

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 properties CSN, MM, CA and VCD have a common feature: columnwise independence, although on the corresponding H, these properties are not the same.

Acknowledgements: The authors thank three anonymous referees for detailed and helpful comments.

References

Agresti, A., 1990. Categorical Data Analysis. 2nd Edition. Wiley, New York.

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, Reading, MA.

Cox, D.R., 1972. The analysis of multivariate binary data. Appl. Statist. 21, 113–120.

Cressie, N., Holland, P.W., 1983. Characterizing the manifest probabilities of latent trait models. Psychometrika 48, 129–141.

Ellis, J., Junker, B.W., 1997. Tail measurability in monotone latent variable models. Psychometrika 62, 495–523.

Fisher, R.A., 1935. The logic of inductive inference. J. Roy. Statist. Soc. Ser. B 98, 39–54.

Fisher, G., 1974. Einf*ü*hrung in die Theorie Psychologischer Tests: Grundlagen und Anwendungen. Bern: Huber.

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

Han, K.T., 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., Rosenbaum, P.R., 1986. Conditional association and unidimensionality in monotone latent trait models. Ann. Statist. 14, 1523–1543.

Joag-Dev, K., Proschan, F., 1983. Negative association of random variables, with applications. Ann. Statist. 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. Ann. Statist. 21, 1359–1378.

Junker, B.W., Ellis, J., 1997. A characterization of monotone unidimensional latent variable models. Ann. Statist. 25, 1327–1343.

Junker, B.W., Sijtsma, K., 2000. Latent and manifest monotonicity in item response models. Appl. Psychological Measurement 24, 65–81.

Lehmann, E.L., 1986. Testing Statistical Hypothesis. 2nd Edition. Wiley, New York.

Lord, F.M., Norvick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Mehta, C.R., Patel, N.R., Senchaudhuri, P., 1988. Importance sampling for estimating exact probabilities in permutational inference. J. Amer. Statist. Assoc. 83, 999–1005.

Patz, R.J., Junker, B.W., 1999. A straightforward approach to Markov chain Monte Carlo methods for item response models. J. 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., Hambleton, R.K. (Eds.), 1997. Handbook of Modern Item Response Theory. Springer, New York.

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

Yuan, A., Yang, Y., 2005. A Markov chain sampler for contingency table exact inference. Comput. Statist. 20, 63–80.

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