Journal of Classification (2019) 36:295–305
https://doi.org/10.1007/s00357-018-9287-5

Conditional Independence and Dimensionality of Cognitive Diagnostic Models: a Test for Model Fit

Youn Seon Lim[1] · Fritz Drasgow[2]

Published online: 29 March 2019

Abstract

Nonparametric cognitive diagnosis methods are useful in cognitive diagnosis modeling for calibration efficiency, especially when sample size is small or large, or the latent attributes are more complex. This article proposes the Mantel-Haenszel chi-squared statistic as an index for detecting the misspecification of latent attributes as well as testlet effects in nonparametric cognitive diagnosis methods. The proposed theoretical considerations are augmented by simulation studies conducted to assess the performance of the Mantel-Haenszel statistic under various conditions within the nonparametric diagnosis framework, with a special focus on situations were the set of latent abilities assumed to underlie the data was underspecified.

Keywords Cognitive diagnosis model · Nonparametric approach · Local independence · Qmatrix validation

Cognitive diagnosis models are used to provide diagnostic feedback to examinees and stakeholders at a finer grain size than a single test score. Many different models have been proposed, but they all require a common feature, the $\boldsymbol{Q}$-matrix, that indicates the item J by latent attribute K relationship (Tatsuoka 1983). Each entry q_{jk} in the matrix indicates whether the kth attribute is necessary in the solution of the jth item. An examinee's performance with respect to what is measured is assumed to be influenced by a composite of the latent attributes such that different combinations define profiles of distinct proficiency classes, which are characterized by the K-dimensional latent attribute vectors $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, …, $\boldsymbol{\alpha}_C$, with $C = 2^K$.

✉ Youn Seon Lim
YounSeon.Lim@hofstra.edu

Fritz Drasgow
fdrasgow@illinois.edu

[1] Department of Science Education, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hofstra University, Hempstead, NY 11549, USA

[2] School of Labor & Employment Relations, Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, USA

The validity of a cognitive diagnosis model depends on whether the K-dimensional latent attribute vector entirely determines classes of examinees so that the conditional distributions of item scores are all independent of each other after adjusting for the effect of the attributes. This property is often called local independence (e.g., Rupp et al. 2010; Lord and Novick 1968). The assumption of local independence is equivalent to the assumption that the K attributes $\alpha_1, \alpha_2, \ldots, \alpha_K$ span the complete latent space—that is, no latent attributes have been missed or left out. Said differently, violations of local independence indicate the possible misspecification of attributes.

The testlet effect also calls into question the assumption of local independence. A testlet is a cluster of items that shares a common stimulus, such as a reading passage and measures something additional in common (Wainer and Kiely 1987). One way to account for the testlet effect is to incorporate specific dimensions in addition to the K-dimensional latent attributes the $\boldsymbol{Q}$-matrix specifies. Therefore, testing for local independence can be used as a diagnostic tool for detecting testlet effects as well as incorrect specifications of the latent attributes in cognitive diagnostic modeling.

In cognitive diagnosis modeling, evaluations of model-data fit provide information about the cognitive diagnosis model and data fit as well as the $\boldsymbol{Q}$-matrix and data fit (e.g., Chen et al. 2013). Various fit statistics and methods have been proposed for both types of evaluations. Some of them include conventional relative fit measures such as Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), log likelihood, and Bayes factor (e.g., Chen et al. 2013; Kunina-Habenicht et al. 2012; Rupp et al. 2010). Furthermore, absolute fit measures such as the residual between the observed and predicted Fisher-transformed correlation, the residual between the observed and predicted correct proportion, the residual between the observed and predicted log-odds ratios, and the G statistic have been proposed (e.g., Chen et al. 2013; Rupp et al. 2010). These statistics and methods are limited to use in parametric cognitive diagnosis models because most statistics are computed as a function of maximum likelihood estimates; the predicted item responses are generated based on the fitted model.

This article proposes the Mantel-Haenszel (MH) statistic as an index for detecting misspecification of latent attributes as well as testlet effects in nonparametric cognitive diagnosis methods. Obviously, under the nonparametric methods, evaluation of model-data fit is informative about the $\boldsymbol{Q}$-matrix and data fit only. The MH statistic is a well-researched tool for evaluating the conditional independence of binary variables that are stratified along the levels of a third random variable, for example, examining conditional independence of item pairs that are stratified along the levels of total test scores under an IRT model (Rosenbaum 1984).

The next section describes the assumption of conditional independence underlying cognitive diagnosis models and provides a brief review of nonparametric cognitive diagnosis methods. Then, the MH test of model fit is presented. Next, simulation studies are described with a wide range of conditions. Then, an analysis of real data is described. In the final section, applications and implications of the method are discussed.

1 Conditional Independence and Its Violations

Let Y_{ij} denote the binary item response of the ith examinee to the jth item, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Cognitive diagnosis models describe the joint distribution of item response vector $\boldsymbol{Y}_i$ conditional on binary attribute vector $\boldsymbol{\alpha}_{ic} = \{\alpha_{ick}\}$, for $c = 1, 2, \ldots, 2^K$, and for $k = 1, \ldots, K$. Each entry α_{ick} indicates whether the ith examinee has mastered the kth attribute. Each binary entry q_{jk} in the $\boldsymbol{Q}$-matrix indicates whether the kth attribute is relevant for the jth item, with 1 meaning the attribute is relevant and 0 indicating it is irrelevant. The joint probability of a cognitive diagnosis model for the ith examinee is

$$P\left(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_i\right) = \prod_{j=1}^{J} P\left(Y_{ij} \mid \boldsymbol{\alpha}_i\right). \tag{1}$$

Therefore, most models are required to satisfy the assumption of conditional independence among item responses Y_{ij} given the attribute vector $\boldsymbol{\alpha}_{ic}$ (e.g., Rupp et al. 2010) because the assumption makes it possible to assess the joint probability or likelihood of the models.

The assumption of conditional independence is violated when the dimensionality of the $\boldsymbol{Q}$-matrix is incorrectly specified. More specifically, a necessary attribute may be omitted. The assumption may be also a concern when the response to an item is based on the responses to the previous items, or when items are grouped by sharing a common stimulus such as a reading passage or a common scenario. Such a grouping of items is referred to the testlet effect, and an additional dimension may be required to adequately model the data, but would be considered as a nuisance dimension because it is not substantively meaningful (e.g., Wainer and Kiely 1987). Most cognitive diagnosis models ignore the testlet effect, and it may result in underspecified dimensions. Therefore, the existence of testlets calls into question the assumption of local independence. A misfit item j may indicate that the item is problematic or $\boldsymbol{q}_j$ is underspecified; a few misfit items indicate the dimensions of $\boldsymbol{Q}$-matrix may be underspecified; misfit items sharing a common stimulus may indicate a testlet effect.

2 Nonparametric Cognitive Diagnosis Methods

Nonparametric cognitive diagnostic methods assess examinees' mastery and nonmastery of attributes without regard to parametric form. These methods are useful in cognitive diagnosis modeling, especially when parametric model fitting is inefficient because of too small or large sample sizes, or more complex sets of latent attributes (Junker 2011).

One approach to nonparametric cognitive diagnosis methods is to apply cluster analysis to identify groups of examinees with similar pattern of latent attributes given the assumption of a conjunctive relationship among attributes and a valid $\boldsymbol{Q}$-matrix. Chiu et al. (2009) clustered the sum score vectors $W_i = (W_{i1}, \ldots, W_{iK})$ using hierarchical agglomerative and K-means clustering to produce the 2^K latent classes. Ayers et al. (2008) utilized the capability score vectors $B_i = (B_{i1}, \ldots, B_{iK})$, where $B_{ik} = \sum_j Y_{ij} q_{jk} / \sum_j q_{jk}$ instead of W_i. Park and Lee (2011) mapped item responses to an attribute matrix and then conducted K-means and hierarchical agglomerative clustering.

Another approach utilizes the Hamming distance technique that was originally proposed by Barnes (2003) with a valid $\boldsymbol{Q}$-matrix. In this technique, examinees' latent attribute vectors are obtained by minimizing the Hamming distance between the observed item responses $\boldsymbol{Y}_i$ and all possible ideal responses $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_C$, $C = 2^K$,

$$D(\boldsymbol{Y}_i, \boldsymbol{\alpha}_c) = \sum_{j=1}^{J} |Y_{ij} - \eta_{cj}|. \tag{2}$$

Like Barnes (2003), Chiu and Douglas (2013) posited a conjunctive relationship among the attributes. Lim and Drasgow (2017) proposed an algorithm given the assumption of conjunctive, disjunctive, or compensatory relationships among attributes. The theoretical justification of this approach is that the true attribute pattern minimizes the expected distance between $\boldsymbol{Y}_i$ and $\boldsymbol{\eta}_c$ regardless of what the true model is, under some regularity conditions (Lim and Drasgow 2017; Wang and Douglas 2015).

3 Mantel and Haenszel Test of Model Fit

The MH statistic χ^2 introduced by Mantel and Haenszel (1959) is generally used to test for conditional independence—of two dichotomous or categorical variables j and j' by forming the row-by-column contingency tables, conditional on the levels of the control variable C. For IRT models, the MH statistic has been commonly used to detect differential item functioning, items that function differently for two groups of examinees called focal and reference groups with different experiences or backgrounds (Holland and Thayer 1988). In the procedure, the sample is stratified into C classes according to their observed total test score.

In this study, the latent attribute vector $\boldsymbol{\alpha}_c = (\alpha_{c1}, \alpha_{c2}, \cdots, \alpha_{cK})'$, for $c = 1, 2, \ldots 2^K = C$ is proposed as the stratification variable. As discussed above, in cognitive diagnosis models, item responses are assumed to be independent given the correct $\boldsymbol{\alpha}_c$, and a higher value of $\boldsymbol{\alpha}_c$ implies a higher probability that $Y_j = 1$ for each $j = 1, 2, \ldots, J$ (e.g., Holland and Rosenbaum 1985). Then, any pair of vectors of monotonic nondecreasing functions $g_j(\boldsymbol{Y})$ and $g_{j'}(\boldsymbol{Y})$ of a vector of dichotomous responses $\boldsymbol{Y}$ to item j and j', given any monotonic nondecreasing function $h(\boldsymbol{\alpha}_c)$, has a nonnegative conditional covariance, a result of Rosenbaum (1984).

Let $\left\{ i_{jj'_c} \right\}$ denote the frequencies of examinees in the $2 \times 2 \times C$ contingency table. The marginal frequencies are the row totals $\{ i_{1+_c} \}$ and the column totals $\{ i_{+1'_c} \}$, and i_{++_c} represents the total sample size in the cth stratum. Strata having a minimum total sample size i_{++_c} equal or larger than 1 are included. If any cell count in a table is 0, then the Haldane correction is applied to each cell in the table to obtain a more accurate significance level of the MH test (e.g., Li et al. 1979). Under the null hypothesis of conditional independence between j and j', the following statistic is proposed:

$$\text{MH } \chi^2 = \frac{\left(\left| \sum_c i_{11c} - \sum_c E(i_{11c}) \right| - 1/2 \right)^2}{\sum_c \text{var}(i_{11c})}, \tag{3}$$

where $E(i_{11c}) = i_{1+c} i_{+1c} / i_{++c}$ and $\text{var}(i_{11c}) = i_{0+c}\ i_{1+c}\ i_{+0c}\ i_{+1c} / i^2_{++c}\ (i_{++c} - 1)$.

Under the null hypothesis, the test statistic has approximately a chi-squared distribution with degrees of freedom equal to 1 when sample sizes in each contingency table become large, and in cognitive diagnosis models, if each examinee's true latent attribute vector $\boldsymbol{\alpha}_i$ is known. Mantel and Haenszel (1959) indicate that this summary chi-square reference distribution is suitable even when some of the strata have small counts. This statistic would be suitable for the analysis of sparse contingency tables, provided the overall counts for each cell in the combined table obtained by collapsing across all C contingency tables are sufficiently large. The null hypothesis of independence is equivalent to the odds ratio equal to 1.

$$\text{Odds ratio}_{MHj,j'} = \frac{\sum_{c=1}^{C} \left(i_{11_c}\ i_{00_c} \right) / i_c}{\sum_{c=1}^{C} \left(i_{10_c} i_{01_c} \right) / i_c}, \tag{4}$$

where $i_c = i_{11_c} + i_{00_c} + i_{10_c} + i_{01_c}$.

4 Heuristic Justification of the Large Sample Chi-square Reference Distribution

The estimated test statistic MH $\hat{\chi}^2$ would have an asymptotic chi-square distribution with one degree of freedom as the true MH statistic MH χ^2 would, if the true attribute vector $\boldsymbol{\alpha}$ were known. Mantel and Haenszel (1959) asserted that under the null hypothesis, the MH χ^2 has an asymptotic chi-squared distribution with one degree of freedom, under some general conditions.

It is assumed here that the number of items J is sufficiently large so that $P[\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}]$ is close to 1, a result of previous theoretical studies (Lim and Drasgow 2017; Wang and Douglas 2015). A rigorous argument requires that the number of items J grows sufficiently fast with the sample size N. Note that

$$\mathrm{MH}\ \hat{\chi}^2 = \mathrm{MH}\ \chi^2 + \left(\mathrm{MH}\ \hat{\chi}^2 - \mathrm{MH}\ \chi^2\right), \tag{5}$$

where $\left(MH\hat{\chi}^2 - MH\chi^2\right)$ represents error, due to using $\hat{\boldsymbol{\alpha}}$ rather than $\boldsymbol{\alpha}$. If in (5), we have convergence in probability to zero in the second term on the right, we see that our approximate M-H test statistic $MH\hat{\chi}^2$ has the same asymptotic distribution as the desired MH statistic $MH\chi^2$. Specifically,

$$\left(MH\hat{\chi}^2 - MH\chi^2\right) \rightarrow^P 0 \Rightarrow MH\hat{\chi}^2 \rightarrow^D \chi_1^2. \tag{6}$$

The result in (6) is obtained if J is sufficiently large, so that under the null hypothesis the overwhelming majority of estimated attribute patterns are identical to the true attribute patterns. Finite test length and sample size properties are studied in the following simulation studies, and type I error rate power rates are summarized.

Table 1 Correctly specified $\boldsymbol{Q}$-matrix ($K = 5$)

Item	Attribute				
	k_1*	k_2*	k_3*	k_4	k_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4**	0	0	0	1	0
5**	0	0	0	0	1
6	1	1	0	1	1
7	1	0	0	1	0
8	0	1	0	1	0
9	0	0	1	1	0
10	0	1	0	1	0
11	1	1	0	0	0
12	1	1	0	0	1
13	1	0	0	1	1
14	0	1	0	1	1
15	0	0	1	1	0
16	1	0	0	1	0
17	0	1	0	0	1
18	0	0	1	0	0
19	1	0	0	1	0
20	1	0	0	1	1

*Attributes used for the three-dimensional $\boldsymbol{Q}$-matrix. **For the three-dimensional $\boldsymbol{Q}$-matrix, the attribute pattern for item 4 was 101; for item 5, the attribute pattern was 011

Table 2 Type I error study: correctly specified $\boldsymbol{Q}$-matrix

I	$K=3$				$K=5$			
	$\boldsymbol{\alpha}$ with $\rho=.3$		$\boldsymbol{\alpha}$ with $\rho=.6$		$\boldsymbol{\alpha}$ with $\rho=.3$		$\boldsymbol{\alpha}$ with $\rho=.6$	
	$J=20$	$J=40$	$J=20$	$J=40$	$J=20$	$J=40$	$J=20$	$J=40$
With estimated latent attribute profiles								
500	.032	.044	.044	.046	.035	.046	.030	.046
2000	.057	.044	.052	.047	.051	.049	.053	.051
With true latent attribute profiles								
500	.048	.049	.049	.049	.049	.049	.050	.049
2000	.049	.049	.050	.049	.051	.049	.050	.049

5 Simulation Study

To investigate the performance of the MH statistic, a variety of simulation conditions were studied by crossing the number of examinees, the length of tests, the number of attributes, and the distribution of $\boldsymbol{\alpha}$ under the nonparametric cognitive diagnosis model.

6 Simulation Design

For each condition, item response data of sample sizes $I=500$, or 2000, were drawn from a discretized multivariate normal distribution MVN(0_K, Σ), where the covariance matrix Σ has unit variance and common correlation $\rho=.3$, or .6 (e.g., Chiu et al. 2009). The K-dimensional continuous vector $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\cdots,\theta_{iK})'$ were dichotomized by

$$\boldsymbol{a}_{ik}=\begin{cases}1, & \text{if}\,\theta_{ik}\geq\Phi^{-1}\left(\frac{k}{(K+1)}\right);\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

Test lengths $J=20$ or 40 were studied with attribute vectors of length $K=3$ or 5. The correctly specified $\boldsymbol{Q}$-matrix for $J=20$ is presented in Table 1. The $\boldsymbol{Q}$-matrix for $J=40$ was obtained by duplicating the matrix. Item response data sets were generated from the DINA model and its item parameters were drawn from the uniform (0, .3) distribution. The Hamming distance–based nonparametric cognitive diagnosis model (Lim and Drasgow 2017) was used for the estimation of latent attributes. The main advantage with this proposed method is that it can be applied to parametric models because only class information is necessary.

Table 3 Power study: underspecified $\boldsymbol{Q}$-matrices with true $K=5$ and fitted $K=3$

I	$\boldsymbol{q}_j$ for both items underspecified				$\boldsymbol{q}_j$ for one item underspecified			
	$\boldsymbol{\alpha}$ with $\rho=.3$		$\boldsymbol{\alpha}$ with $\rho=.6$		$\boldsymbol{\alpha}$ with $\rho=.3$		$\boldsymbol{\alpha}$ with $\rho=.6$	
	$J=20$	$J=40$	$J=20$	$J=40$	$J=20$	$J=40$	$J=20$	$J=40$
500	.481	.519	.512	.513	.271	.200	.236	.186
2000	.653	.659	.754	.729	.430	.367	.472	.336

Table 4 T-matrix: testlet specification ($M = 2$)

Testlet	Item																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
M_1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0

6.1 Results

For each condition, sets of item response vectors were simulated for 100 replications. The proposed MH statistics and their corresponding p values were computed for all $J \times (J-1)/2$ item pairs in an individual replication. Of 100 trials, the proportion of times the p value of each item pair was smaller than the significance level .05, which was recorded and summarized in the tables.

Type I Error Study In this simulation study, the correctly specified $\boldsymbol{Q}$-matrices ($K = 5$, or $K =$ 3) were used to fit the data to examine type I error rates. Table 2 shows that most type I error rates were around the nominal significance level .05. The MH statistic appears consistent under all conditions when $J = 40$, confirming asymptotic consistency. In the condition $K = 5$, $J = 20$, and $I = 2000$, the type I error rate was slightly increased.

The MH statistic with known true class membership $\boldsymbol{\alpha}$ was also examined because it is not confounded by possible estimation errors due to the specific algorithm used to estimate latent attributes. The rejection rates were very close to the nominal significance level .05 for all conditions.

Power Study with Underspecified $\boldsymbol{Q}$-matrices A data set was generated with the $\boldsymbol{Q}$-matrix ($K = 5$) in Table 1. The data was fitted with the embedded $\boldsymbol{Q}$-matrix ($K = 3$) in each replication (Table 3). One dimension (a total of 9 items) or two dimensions (4 items) were underspecified. The average power rate of the item pairs where both items were underspecified in the same dimension was .572 with power relatively consistent across all conditions. The average rejection rate across item pairs where either item was underspecified was .124. Taking this finding into account, like the other statistics, the MH test is sensitive to $\boldsymbol{Q}$-underspecification and has moderately high power, particularly for the larger sample size.

Table 5 Testlet-dependent data with $\boldsymbol{Q}$-matrix ($K = 3$)

I	Both items were in a testlet				One item was in a testlet			
	$\boldsymbol{\alpha}$ with $\rho = .3$		$\boldsymbol{\alpha}$ with $\rho = .6$		$\boldsymbol{\alpha}$ with $\rho = .3$		$\boldsymbol{\alpha}$ with $\rho = .6$	
	$J = 20$	$J = 40$	$J = 20$	$J = 40$	$J = 20$	$J = 40$	$J = 20$	$J = 40$
500	.922	.959	.941	.975	.051	.049	.048	.047
2000	.997	.998	.994	.998	.088	.051	.062	.053

Table 6 ***Q***-matrix for fraction subtraction data

Item	$K=8$								Item	$K=8$							
1	0	0	0	1	0	1	1	0	11	0	1	0	0	1	0	1	0
2	0	0	0	1	0	0	1	0	12	0	0	0	0	0	0	1	1
3	0	0	0	1	0	0	1	0	13	0	1	0	1	1	0	1	0
4	0	1	1	0	1	0	1	0	14	0	1	0	0	0	0	1	0
5	0	1	0	1	0	0	1	1	15	1	0	0	0	0	0	1	0
6	0	0	0	0	0	0	1	0	16	0	1	0	0	0	0	1	0
7	1	1	0	0	0	0	1	0	17	0	1	0	0	1	0	1	0
8	0	0	0	0	0	0	1	0	18	0	1	0	0	1	1	1	0
9	0	1	0	0	0	0	0	0	19	1	1	1	0	1	0	1	0
10	0	1	0	0	1	0	1	1	20	0	1	1	0	1	0	1	0

Power Study with Testlet-Dependent Data For this simulation study, the fixed *T*-matrix in Table 4 was utilized to generate the testlet-dependent data. The entry t_{mj} of the *T*-matrix indicates whether the *m*th testlet, for $m = 1, 2, \ldots M$, includes the *j*th item. For each replication, the transpose of *T*-matrix was combined with the ***Q***-matrix ($K = 3$) embedded in Table 1, to simulate item responses. A model was fitted only with the ***Q***-matrix ($K = 3$). The *T*-matrix for $J = 40$ was obtained by duplicating the matrix.

As shown in Table 5, high rejection rates for testlet-dependent item pairs were obtained (i.e., .922 or above). The power rates were moderately consistent across conditions. The rejection rates of the MH statistic for item pairs in which only either item was testlet dependent were low (i.e., .088 or below). This implies that the MH test can play an important role only in detecting testlet-dependent items.

7 Fraction Subtraction Data

Fraction subtraction data (e.g., Tatsuoka 1983) were analyzed to investigate the performance of the MH statistic in practice. The data include the item responses to 20 items with 8 necessary attributes from 536 examinees. In this study, the ***Q***-matrix (see Table 6) that appeared originally in de la Torre and Douglas (2004) was used. The specified attributes are interpreted as (1) convert a whole number to a fraction, (2) separate a whole number from fraction, (3) simplify before subtracting, (4) find a common denominator, (5) borrow from whole number part, (6) column borrow to subtract the second numerator from the first, (7) subtract numerators, and (8) reduce answers to simplest form.

Table 7 Proportion of conditionally dependent item pairs

	Model						
	Non-P	DINA	DINO	A-CDM	Saturated	LLM	R-RUM
MH	.084	.147	.174	.079	.137	.200	.194
$x_{jj'}$	–	.321	.489	.221	.068	.426	.687
$r_{jj'}$	–	.426	.584	.253	.111	.500	.573

Table 8 Most frequently rejected items

Frequently rejected items	Agreement rate**	Frequencies		
		MH	$x_{jj'}$	$r_{jj'}$
5	0.750	25*	47	45
8	1.00	8	75*	76*
9	0.875	15	52*	61*
13	0.875	26*	50*	50
15	0.875	13	46	54*
16	1.00	20	51*	58*
19	0.875	28*	45	54*
20	0.875	26*	33	41

Items with single asterisk were ranked as top 1 to 4 most frequently rejected items over all models. ***Agreement rates between the implemented ***Q*** and data-driven ***Q*** (Lim and Drasgow 2017)

7.1 Results

The data were analyzed with seven different cognitive diagnosis models: the nonparametric model, the DINA model, the DINO model, the A-CDM, the saturated model, the log linear model, and the R-RUM. Additional fit statistics, the chi-squared statistic $x_{jj'}$ (Chen and Thissen 1997) and absolute deviations of observed and predicted corrections $r_{jj'}$ (Chen et al. 2013), were used for the evaluation of model-data fit. The average rejection rates of 190 item pairs are summarized and reported in Table 7. Interestingly, the MH statistic indicates substantially fewer model violations than the other two fit measures.

Table 8 reports the most frequently rejected four items for each of the statistics over all model settings. The results of statistics were consistent with those of Lim and Drasgow (2017). In their data-driven ***Q***-matrix estimation study, the component-wise agreement rates between the implemented ***Q***-matrix in this study and a data-driven ***Q***-matrix were obtained as shown in Table 8. The items for which the *q*-vectors may have been incorrectly specified were the most frequently rejected by the MH statistic. The disagreement across methods is especially noticeable for item 8. This result may imply that this item was overspecified based on the results of previous studies (e.g., Chen et al. 2013).

8 Discussion

The significance of this study lies in proposing a test of model fit for detecting ***Q***-matrix misspecifications and identifying testlet effects. The only requirement for this method is the availability of an estimate of the latent attributes, which serves as the stratification variable in

Table 9 Mean of absolute difference of estimated and true DINA item parameters (α with $\rho = .3$)

I	$J = 40$, $K = 5$*		$J = 40$, $K = 3$**		$J = 20$, $K = 5$*		$J = 20$, $K = 3$**	
	Guess	Slip	Guess	Slip	Guess	Slip	Guess	Slip
500	.015	.023	.073	.120	.021	.027	.081	.075
2000	.015	.020	.067	.069	.008	.012	.075	.131

*Correctly specified ***Q***-matrix. **Underspecified ***Q***-matrix the MH statistic. Several simulation studies investigated the usefulness and sensitivity of the MH statistic in a variety of conditions. The primary findings were that the MH test could play an important role in identifying underspecified q-vectors when the true model is unknown. It performs reasonably well in detecting testlet-dependent items. These results are important because ignoring such dependencies could possibly lead to inaccurate estimates of model parameters as shown in Table 9 as well as misclassifications of examinees (e.g., Chen et al. 2015; Rupp et al. 2010).

The real data analysis illustrated how the MH test can be used with different cognitive diagnosis models along with other model fit test statistics. The MH test found less misfit and was less sensitive to the use of different models. For q-vector misspecifications, it can be effective to identify problematic items. When it is used with the other test statistics, the results can provide more detail—whether an item may be underspecified, or a different model is needed for the data.

Whether the fit evaluation is to detect the $\boldsymbol{Q}$-matrix underspecification, or testlet effects, the MH test is simple, easy to implement, and theoretically supported. The results of the simulations suggest that the MH is a reasonably efficient test of model fit. Nevertheless, some consideration of other tests of model fit will always be desirable. Future research might include more attributes as well as more complex models. At the present time, however, the MH test appears to be a promising statistic for the detection of local dependence in cognitive diagnosis models.

References

Ayers, E., Nugent, R., & Dean, N. (2008). Skill set profile clustering based on student capability vectors computed from online tutoring data. In R. S. J. de Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining 2008: proceedings of the 1st International Conference on Educational Data Mining* (pp. 210–217). Montréal: International Data Mining Society.

Barnes, T. (2003). *The Q-matrix method of fault-tolerant teaching in knowledge assessment and data mining*. North Carolina State University, USA. http://www.lib.ncsu.edu/resolver/1840.16/4612. Accessed March 27, 2011

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

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123–140.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*(510), 850–866.

Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*(2), 225–250.

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika, 74*, 633–655.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

Holland, W. P., & Rosenbaum, P. R. (1985). *Conditional association and unidimensionality in monotone latent variable models (Research Report No. 85 – 47)*. Princeton: Educational Testing Service.

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

Junker, B. W. (2011). The role of nonparametric analysis in assessment modeling: then and now. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: proceedings of a conference in honor of Paul W. Holland* (pp. 67–85). New York: Springer-Verlag.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59–81.

Li, S., Simon, R. M., & Gart, J. J. (1979). Small sample properties of the Mantel-Haenszel test. *Biometrika, 66*, 181–183.

Lim, Y. S. & Drasgow, F. (2017 – accepeted). Nonparametric calibration of item-by-attribute matrix. *Multivariate Behavioral Research*.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute, 22*, 719–748.

Park, Y., & Lee, Y. (2011). Diagnostic cluster analysis of mathematics skills. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4*, 75–107.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumption of item response theory. *Psychometrika, 49*, 425–436.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic assessment: theory, methods, and applications*. New York: Guilford.

Tatsuoka, K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: a case study for testlets. *Journal of Educational Measurement, 24*, 195–201.

Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika, 80*, 85–100.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.