Detecting Testlet Effects in Cognitive Diagnosis Models

Youn Seon Lim

Abstract A testlet is a cluster of items that shares a common stimulus (e.g., a set of questions all related to the same text passage). The testlet effect calls into question one of the key statistical assumptions of any tests: local independence of the test item responses. Local dependence among test item responses is typically induced by the under-specification of the latent ability dimensions supposed to underlie a test. Hence, evaluating whether local independence holds for the items of a given test can be used as a diagnostic tool for detecting testlet effects. This study studied and compared the MH statistic, the Chi-squared statistic and the absolute deviations of observed and predicted corrections in detecting testlet effects in cognitively diagnostic tests. Various simulation studies were conducted to evaluate their performance under a wide variety of conditions.

Keywords Cognitive diagnosis models · Testlet effects · Mantel-Haenszel statistic · Chi-squared statistic · Absolute deviations of observed and predicted corrections

1 Introduction

A testlet is "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (Wainer & Kiely, 1987, p. 190). A typical example is a reading comprehension test in which a reading passage is used as the stimulus for more than one item to measure examinees' ability to comprehend the reading passage. Another example refers to ordering sentences to make a complete passages, where the items (sentences) are embedded in the passage itself. Responses to items within a testlet calls into question of the key statistical assumptions of any test: local independence

Y. S. Lim (✉)
Quantitative and Mixed Methods Research Methodologies, Educational Studies, University of Cincinnati, Cincinnati, OH, USA
e-mail: limyo@ucmail.uc.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 393, https://doi.org/10.1007/978-3-031-04572-1_2

of the test item responses. Local dependence among test items is typically induced by the under-specification of the measured latent dimensions by a test (i.e., Lim & Drasgow, 2019a,b; Rupp et al., 2010).

Various methods have been suggested for examining local dependence in cognitive diagnosis models. For example, de la Torre and Douglas (2004) evaluated item pair dependence using bivariate information. Templin and Henson (2006) expanded their method by using a parametric bootstrap method to estimate the distribution of the item association measures to estimate a p-value for their test statistic. Chen et al. (2013) used comparing the residual between the observed and expected Fisher-transformed correlation, and the residual between the observed and expected between log-odds ratios for the measure of association for each pair of items. Lim and Drasgow (2019a,b) also modified the Mantel Haenszel (MH) statistic for the measure. Those statistics have been used for the model fit or Q-matrix fit evaluation. This study evaluates the performances of the MH statistic (Lim & Drasgow, 2019a,b), the Chi-squared statistic $x_{jj'}$ (Chen & Thissen, 1997), the absolute deviations of observed and predicted corrections $r_{jj'}$ (Chen et al., 2013) in detecting testlet effects in cognitively diagnostic test in various simulation conditions.

2 Cognitive Diagnosis Models

Three cognitive diagnosis models were considered in this study: Deterministic-Input, Noisy "And" gate (DINA) model, Generalized Deterministic Inputs, Noisy "And" gate (G-DINA) model (saturated model), and Additive Cognitive Diagnosis Model (A-CDM).

Let Y_{ij} denote the binary item response of the ith examinee to the jth item, $i = 1, \ldots, I$, $j = 1, \ldots, J$ with 1 = correct and 0 = incorrect. Cognitive diagnosis models formulate the conditional distribution of item responses Y_{ij} given examinee latent attributes $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$, for $k = 1, \ldots, K$. (e.g., de la Torre & Douglas, 2004). Each entry α_{ik} indicates whether the ith examinee has mastered the kth attribute with 1 = mastered and 0 = not mastered. The binary $J \times K$ Q-matrix is an essential component of cognitive diagnosis models. The Q-matrix has a row for each item, $j = 1, \ldots, J$, and a column for each attribute, $k = 1, \ldots, K$. Each entry q_{jk} in the matrix indicates whether the kth attribute is required for the solution of the jth item with 1 = required and 0 = not required.

A common cognitive diagnosis model is the DINA model (e.g., Junker & Sijtsma, 2001). In this model, an ideal response η_{ij} is used to indicate whether all required attributes for the jth item are mastered by the ith examinee. The item response function (IRF) for the DINA model is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})},$$

where $s_j = P(Y_j = 0 \mid \eta_j = 1)$ and $g_j = P(Y_j = 1 \mid \eta_j = 0)$.

Henson et al. (2009) proposed the Log-Linear Cognitive Diagnosis Model (LCDM). The LCDM can fit a full continuum of cognitive diagnosis models that range from fully compensatory models to fully conjunctive models. The DINA model can be written as a special case of the LCDM. In particular, if an item requires two attributes, the IRF can be written as

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{exp[\lambda_j\alpha_1 + \lambda_j\alpha_2 + (\lambda_j)\alpha_1\alpha_2 - \eta_1]}{1 + exp[\lambda_j\alpha_1 + \lambda_j\alpha_2 + (\lambda_j)\alpha_1\alpha_2 - \eta_1]},$$

where $\eta_j = -ln(g_j/1 - g_j)$ and $\lambda_j = \eta_j + ln(s_j - 1/s_j)$.

de la Torre's Generalized DINA (G-DINA) model is another example (de la Torre, 2011). Similar to the LCDM, the G-DINA model can be reduced to special cases of general cognitive diagnosis models with different link functions: identity, logit, and log. The framework of the G-DINA is based on the DINA model. However, the 2^K latent class memberships of the DINA model are partitioned into $2^{K_j^\star}$ latent groups, where $K_j^\star = \sum_{k=1}^{K} q_{jk}$ denotes the number of required attributes for item j. Let $\boldsymbol{\alpha}_{lj}^\star$ be the reduced attribute vector whose elements are the required attributes for item j. Then the probability that a test taker mastering the attribute pattern $\boldsymbol{\alpha}_{lj}^\star$ (i.e., all elements of $\alpha_{lj}^\star$would answer item j correctly is given by

$$\begin{aligned} P(\boldsymbol{\alpha}_{lj}^\star) &= P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_{lj}^\star) \\ &= \lambda_{j0} + \sum_{k=1}^{K_j^\star} \lambda_{jk}\alpha_{ljk} + \sum_{k'>k}^{K_j^\star} \sum_{k=1}^{K_j^\star - 1} \lambda_{jkk'}\alpha_{ljk}\alpha_{ljk'} \ldots + \lambda_{j1,\ldots K_j^\star} \sum_{k=1}^{K_j^\star} \alpha_{ljk}, \end{aligned}$$

where λ_{j0} is the intercept, λ_{jk} is the main effect, $\lambda_{jkk'}$ and $\lambda_{j1,\ldots K_j^\star}$ are interaction effects.

Without the interaction terms, the G-DINA model becomes the Additive-CDM (A-CDM). The A-CDM is one of several reduced models that can be derived from the saturated G-DINA model. The IRF of the additive model is given by

$$P(\boldsymbol{\alpha}_{lj}^\star) = P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_{lj}^\star) = \lambda_{j0} + \sum_{k=1}^{K_j^\star} \lambda_{jk}\alpha_{jk}.$$

Item j has $K_j^\star + 1$ parameters in this model. The mastery of an attribute has a constant and direct impact on the probability of a correct response.

3 Fit Statistics for Local Dependence

Three fit statistics were evaluated in this study: MH statistic (Lim & Drasgow, 2019a,b), the Chi-squared statistic denoted by $x_{jj'}$ (Chen & Thissen, 1997), the absolute deviations of observed and predicted corrections denoted by $r_{jj'}$ (Chen et al., 2013).

Lim and Drasgow (2019a,b) modified the MH chi-square statistic which were originally introduced by Mantel and Haenszel (1959) to test for conditional independence of two dichotomous or categorical item responses j and j' by forming the row-by-column contingency table, conditional on the levels of the control variable C, where $c = 1, 2, \ldots, 2^K = C$ proficiency class membership. Let $\{i_{j,j'c}\}$ denote the frequencies of examinees in the $2 \times 2 \times C$ contingency table, and then

$$\text{MH}\chi^2 = \frac{[\sum_c (i_{11c} - \sum_c E(i_{11c})]^2}{\sum_c \text{var}(i_{11c})},$$

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

The Chi-squared statistic $x_{jj'}$ (Chen & Thissen, 1997) is computed by forming the row-by-column contingency table,

$$\chi^2 = \sum_j \sum_{j'} \frac{(i_{jj'} - E(i_{jj'}))^2}{E(i_{jj'})},$$

where $E(i_{jj'}) = E_{pq} = N \int P_j(\theta)^p P_j(\theta)^q [1 - P_i(\theta)]^{(1-p)} [1 - P_j(\theta)]^{(1-q)} f(\theta) d\theta,$. where $P_i(\theta)$ is the trace link for item j, $f(\theta)$ is the population distribution. For cognitive diagnosis models, $E(i_{jj'})$ is estimated by an examinee's posterior distributions (Robitzsch et al., 2020).

The absolute deviations of observed and predicted corrections $r_{jj'}$ (Chen et al., 2013) is calculated by

$$r_{jj'} = |\mathbf{Z}[Corr(\mathbf{Y}_j, \mathbf{Y}_{j'})] - \mathbf{Z}[Corr(\mathbf{Y}_j, \mathbf{Y}_{j'})]|,$$

where $Corr(\cdot)$ is the Pearson's product-moment correlation, $\mathbf{Z}(\cdot)$ is the Fisher's transformation.

4 Simulation Studies

To investigate the performance of the MH statistic, the Chi-squared statistic $x_{jj'}$, the absolute deviations of observed and predicted corrections $r_{jj'}$, a variety of simulation conditions were studied by crossing the numbers of examinees I, items J, and examinees' latent attribute distributions ρ for three different cognitive diagnosis models.

For each simulation condition, a set of item response vectors was simulated for 100 replications. Item response data of sample sizes $I = 500$ (small), or 2000 (large) were drawn from a discretized multivariate normal distribution $MVN(0_K, \sum)$, where the covariance matrix $\sum$ has unit variance and common correlation $\rho =$ 0.3 (low) or 0.6 (high). Test lengths $J = 20$ (short) or 40 (long) were studied. A $Q-$matrix was generated randomly from a discrete uniform distribution on the

Table 1 Correctly specified Q (K = 3)

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

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

	Item																			
Testlet	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

maximum $2^K - 1$ possible q-vectors for each condition and fixed for replications. The correctly specified Q-matrix for $J = 20$ is presented in Table 1. The Q-matrix for $J = 40$ was obtained by duplicating this matrix two times to study the longer length of item under the same attribute specification conditions.

Data were generated using three different models: the DINA model, A-CDM, and a saturated model (i.e., the G-DINA model). For the DINA model, item parameters were drawn from Uniform (0, 0.3). For the A-CDM and the saturated model, like Chen et al. (2013), the parameters were restricted as $P(\alpha_{ij}^{\star})_{\min} = 0.10$ and $P(\alpha_{ij}^{\star})_{\max} = 0.90$, where $\alpha_{ij}^{\star}$ was the reduced attribute vector whose components are the required attributes for the jth item (see de la Torre, 2011, more details).

A fixed and pre-specified Item-by-testlet T-matrix was utilized to simulate testlet data. The entry t_{mj} of the T-matrix indicates whether the m_{th} testlet, for $m = 1, 2, \ldots, M$, includes the jth item. For each replication of 100 replications, the transpose of T-matrix shown in Table 2 was combined with Q-matrix ($K = 3$) in Table 1, to simulate item responses. A model was fitted only with the Q-matrix ($K = 3$).

The R Core Team (2020) was used for the estimation in this study (CDM package Robitzsch et al., 2020).

The MH statistic, Chi-squared statistic $x_{jj'}$ (Chen & Thissen, 1997), absolute deviations of observed and predicted corrections $r_{jj'}$ (Chen et al., 2013), and their corresponding p-values were computed for all $(J \times (J-1))/2$ item-pairs in an individual replication. Across 100 trials for each condition, the proportion of times the p-value of each item-pair was smaller than the significance level 0.05 was recorded and is summarized.

5 Results

Across 100 trials for each condition, the proportion of times the p-value of each item-pair was smaller than the significance level 0.05 was recorded and is summarized in the tables shown below. The type 1 errors and power rates of the three statistics are reasonable for detecting testlet effects in cognitive diagnosis models.

5.1 Type I Error Study

In this simulation study, the correctly specified Q-matrices ($K = 3$) were used to fit the data to examine type I error rates. The summarized rejection rates are reported in Table 3. The type I error rates of the $r_{jj'}$ became conservative when the numbers of items J and examinees I were increased. The Chi-squared test statistic $x_{jj'}$ was very conservative, with type I error rates below 0.024. The MH statistic got consistent under all conditions when item $J = 40$.

Table 3 Type I error study when $K = 3$

	$J = 20$						$J = 40$					
	α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
I	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
DINA model												
500	0.042	0.019	0.044	0.045	0.014	0.033	0.048	0.017	0.053	0.042	0.020	0.053
2000	0.046	0.023	0.052	0.045	0.015	0.033	0.049	0.019	0.052	0.048	0.019	0.045
A-CDM												
500	0.036	0.009	0.029	0.031	0.009	0.026	0.039	0.011	0.030	0.036	0.011	0.028
2000	0.048	0.013	0.030	0.049	0.010	0.026	0.048	0.010	0.029	0.047	0.010	0.026
Saturated model												
500	0.034	0.010	0.025	0.033	0.009	0.026	0.040	0.010	0.029	0.035	0.011	0.028
2000	0.047	0.010	0.028	0.045	0.010	0.025	0.046	0.010	0.029	0.047	0.009	0.026

Table 4 Simulation study: testlet dependent data

		$J = 20$						$J = 40$					
		α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
I	S	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
DINA model													
500	T	0.928	0.925	0.946	0.803	0.869	0.925	0.815	0.853	0.900	0.896	0.903	0.937
	E	0.052	0.050	0.113	0.041	0.080	0.168	0.042	0.103	0.188	0.044	0.107	0.198
2000	T	0.999	0.998	0.999	0.968	0.995	1.000	0.983	0.993	0.996	0.977	0.996	0.998
	E	0.131	0.058	0.124	0.073	0.174	0.267	0.052	0.271	0.372	0.056	0.271	0.366
A-CDM													
500	T	0.713	0.770	0.848	0.672	0.779	0.862	0.733	0.758	0.844	0.683	0.769	0.858
	E	0.041	0.041	0.081	0.037	0.065	0.117	0.043	0.047	0.094	0.042	0.076	0.137
2000	T	0.998	0.999	1.000	0.987	0.995	0.999	0.996	0.995	0.998	0.989	0.993	0.996
	E	0.075	0.105	0.172	0.063	0.197	0.281	0.049	0.140	0.217	0.048	0.272	0.362
Saturated model													
500	T	0.535	0.606	0.708	0.448	0.562	0.702	0.567	0.608	0.714	0.456	0.577	0.691
	E	0.039	0.032	0.071	0.040	0.039	0.084	0.040	0.043	0.087	0.041	0.060	0.960
2000	T	0.922	0.963	0.978	0.845	0.928	0.955	0.945	0.953	0.979	0.874	0.929	0.960
	E	0.070	0.079	0.140	0.064	0.149	0.234	0.050	0.106	0.178	0.049	0.199	0.287

5.2 Power Study: Testlet Model

As shown in Table 4, high rejection rates for testlet dependent item pairs T were obtained for the three statistics (i.e., 0.803 or above in the DINA model, 0.672 or above in the A-CDM, and 0.448 or above in the saturated model). The power rates were moderately consistent under all conditions. Unlike the Chi-squared statistic $x_{jj'}$, and transformed correction statistic $r_{jj'}$, the rejection rates of the MH statistic for the item pairs in which only one item of a pair was testlet-dependent E were low (i.e., 0.075 or below). This implies that the MH test can play an important role in detecting only testlet dependent items. Not surprisingly, the performance of the test tends to slightly deteriorate in the saturated model.

6 Discussion

The simulation studies investigated the usefulness and sensitivity of the MH statistic, the Chi-squared statistic $x_{jj'}$, the absolute deviations of observed and predicted corrections $r_{jj'}$ in a variety of cognitive diagnosis modeling settings with testlet dependent items. The primary findings are that most type I error rates of the three different statistics were around the nominal significance level of 0.05. Furthermore the statistics perform reasonably well in detecting testlet dependent items. Nonetheless, the statistics are somewhat conservative and less sensitive to

different model settings. In summary, the statistics might be a promising tool for detecting testlet effects in cognitive diagnostic modeling.

For the popularity of testlets in large-scale assessments, it is necessary to investigate the issues related to testlet effects in cognitive diagnosis models. Ignoring testlet effects leads to inaccurate estimates of item parameters and misclassifications of examinees depending on the strengths of testlet effects with minimal influences of other properties of test constructions and administration (Lim et al., 2022). A few (unpublished) dissertations and two or three papers (e.g., Hansen, 2013) study testlet effects—but mainly in terms of how to model testlet effects. Till now, few testlet-effect detection procedures for cognitive diagnosis model have been investigated. Therefore, the significance of this study lies in investigating test statistics to detect testlet effects.

This study is not without limitations. One limitation is that the performance of the statistics was not evaluated with an empirical data. Another limitation is that the statistics were investigated with simple cognitive diagnosis models with testelts. With those limitations, researcher recommends further studies to be conducted with more complex cognitive diagnosis models and real datasets. Furthermore, the findings show that a cognitive diagnosis model that accounts for testlet effects is necessary.

References

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

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

Hansen, M. (2013). *Hierarchical Item Response Models for Cognitive Diagnosis*. Doctoral Dissertation, University of California, LA.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika*, *74*, 191–210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Lim, Y. S., & Drasgow, F. (2019a). Conditional independence and dimensionality of nonparametric cognitive diagnostic models: A test for model fit, *Journal of Classification, 36*, 295–305.

Lim, Y. S., & Drasgow, F. (2019b). Assessing the dimensionality of the latent attribute space in cognitive diagnosis through testing for conditional independence. In Wiberg, M., Culpepper, S., Janssen, R., González, J., & Molenaar, D. (Eds.). *Quantitative psychology research* (pp. 183–194). New York, NY: Springer.

Lim, Y. S., Fangxing, & B., Kelcey B. (2022). Sensitivity of Cognitive Diagnosis Models to Local Item Dependence has been selected for presentation. In: *Annual Conference of the National Council on Measurement in Education (NCME) 2022*, San Diego, CA.

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.

R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2020). *CDM: Cognitive diagnostic modeling*. R package version 3.4–21.

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

Templin, J. L., & Henson, R. A. (2006), Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.

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