Improved Wald Statistics for Item-Level Model Comparison in Diagnostic Classification Models

Applied Psychological Measurement
2019, Vol. 43(5) 402–414
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618798664
journals.sagepub.com/home/apm

Yanlou Liu¹ , Björn Andersson², Tao Xin³,
Haiyan Zhang⁴ and Lingling Wang³

Abstract

Diagnostic classification models (DCMs) have been widely used in education, psychology, and many other disciplines. To select the most appropriate DCM for each item, the Wald test has been recommended. However, prior research has revealed that this test provides inflated Type I error rates. To address this problem, the authors propose to replace the asymptotic covariance matrix from the original version of the Wald statistic with a matrix obtained from improved computation methods. In this study, the Wald test based on the observed information matrix and the Wald test based on the sandwich-type matrix are proposed for item-level model comparisons and a simulation study is conducted to investigate their empirical behavior. Simulation results indicate that when the sample size is reasonably large ($N \geq 1,000$), the Type I error rates of the Wald test based on the sandwich-type matrix are accurate with adequate or excellent power under most of the simulation conditions.

Keywords

model comparison, diagnostic classification models, Wald test, asymptotic covariance matrix, likelihood ratio test

Introduction

Diagnostic classification models (DCMs) characterize relationships between responses of an individual obtained from questionnaire/test items and his or her latent attributes. DCMs have great promise to provide multidimensional discrete diagnostic information on a set of fine-grained attributes (such as cognitive processes, skills, abilities or knowledge representations) for instructional interventions (Rupp, Templin, & Henson, 2010). To date, DCMs have been widely used in educational, psychological, and psychiatric measurement, and also in many

¹Qufu Normal University, China

²University of Oslo, Norway

³Beijing Normal University, China

⁴Taishan University, Tai'an, China

Corresponding Author:

Tao Xin, Professor, Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, 19 Xinjiekouwai Street, Beijing 100875, China.

Email: xintao@bnu.edu.cn

other disciplines (e.g., Greeno, 1980; Leighton & Gierl, 2007; Sorrel et al., 2016; Templin & Henson, 2006).

As a class of statistical models, a variety of DCMs have been developed (Rupp et al., 2010) based on different assumptions about how the probability of a correct response conditional on the Q-matrix and an examinee's latent attribute mastery pattern is defined. The different DCMs reflect specific assumptions about how latent attributes influence test performance in a specific domain, where the Q-matrix specifies the relationship between items and latent attributes (Tatsuoka, 1990). For example, the deterministic inputs, noisy, "and" gate model (DINA; Haertel, 1989; Junker & Sijtsma, 2001) represents a noncompensatory assumption in that all required attributes are expected to have been mastered to obtain a correct response to an item. Hence, the absence of any of the required attributes cannot be compensated by the presence of other attributes. The deterministic inputs, noisy, "or" gate model (DINO; Templin & Henson, 2006) and the additive cognitive diagnostic model (A-CDM; de la Torre, 2011) are compensatory models, in which the absence of some attributes can be compensated by the presence of other attributes. The DINA, DINO, and A-CDM are specific DCMs, in other words these models can be shown to be constrained special cases of the general DCMs, or saturated DCMs, such as the general diagnostic model (von Davier, 2005, 2008), the log-linear cognitive diagnosis model (Henson, Templin, & Willse, 2009) and the generalized DINA model (GDINA; de la Torre, 2011).

As many DCMs are available, one critical step is to select the most appropriate DCM for response data at the item level. Compared with the saturated DCMs, the selected specific DCMs are more suitable for practical applications (de la Torre & Lee, 2013; Ma, Iaconangelo, & de la Torre, 2016; Rojas, de la Torre, & Olea, 2012; Sorrel, Abad, Olea, de la Torre, & Barrada, 2017; Sorrel, de la Torre, Abad, & Olea, 2017). In DCMs, given that the Q-matrix is correctly specified and that the saturated model provides the best model-data fit (Liu, Tian, & Xin, 2016), there are at least five different statistics to use for item-level model selection: the Wald test, the $S - X^2$ test, the likelihood ratio (LR) test, the Lagrange multiplier test, and the two-step likelihood ratio (2LR) test (Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017). De la Torre (2011) originally proposed that the Wald (W_d) test can be used to select the most interpretable reduced model from the saturated DCM. De la Torre and Lee (2013) investigated the empirical Type I error rate control and power of W_d by taking DINA, DINO, and A-CDM as examples. Furthermore, Ma et al. (2016) evaluated the performance of W_d in selecting the most appropriate model at the item level for additive models under the log and logit link functions (Hartz, 2002; Maris, 1999). Sorrel, Abad, et al. (2017) examined the performance of W_d , $S - X^2$, LR, and the Lagrange multiplier test and found that, compared with other statistics, only the empirical Type I error of $S - X^2$ was close to the nominal level. However, $S - X^2$ had low power to reject false models. Sorrel, de la Torre, et al. (2017) compared the performance of the 2LR, LR, and W_d in selecting the best reduced DCM. Although W_d has some advantages to other methods, such as ease of computation and slightly higher power, all of the previous studies consistently found that the W_d test that was computed using the item-wise information matrix proposed by de la Torre (2011) had inflated Type I error rates. This suggests that better computation methods of the covariance matrix or the item parameter standard errors (SEs) are needed (Ma et al., 2016) for the test to maintain the correct significance level.

Liu et al. (2016) have noted that since not only the item parameters but also the structural model parameters are estimated from the examinees' response data (Rupp et al., 2010), both the item parameters and structural model parameters should be considered in the computation of the information matrix. Philipp, Strobl, de la Torre, and Zeileis (2018) have analytically shown that the incomplete information matrix (de la Torre, 2009) or the item-wise information matrix (de la Torre, 2011) lead to underestimated item SEs. Furthermore, in DCMs two new

methods to directly estimate the item parameter covariance matrix, the inverse of the observed information matrix, and the sandwich-type estimator that take into account both the item and structural parameters have been proposed (Liu, Xin, Andersson, & Tian, 2018; Xin, Liu, & Tian, 2016). Liu et al. (2019) compared the performance of the two new methods, the empirical cross-product information matrix, and the two approaches proposed by de la Torre (2009, 2011) in providing consistent *SEs* of item parameter estimates and found that with correctly specified or slightly misspecified models the observed information matrix and the sandwich-type matrix had good performance. However, with substantial model misspecification only the *SEs* based on the sandwich-type covariance matrix were robust. Hence, in this study two new ways to calculate the Wald statistic for item-level model comparison will be introduced. That is, the Wald test based on the observed information matrix (W_{Obs}) or the Wald test based on the sandwich-type matrix (W_{SW}).

The main interest of this study is to investigate the empirical behavior of the two newly proposed methods together with the W_d and LR statistics for item-level model comparisons under the GDINA framework via simulations. The remainder of this article is organized as follows: First, useful background on the saturated GDINA and its reduced models are briefly described, followed by an introduction of the statistical tests for item-level model comparisons that will be investigated in this study, and a discussion of several methods that can be used to obtain the asymptotic covariance matrix of the item parameter estimates. Then, results from a simulation study that systematically evaluates the performance of W_d , W_{Obs} , W_{SW} , and LR were reported. The article concludes with discussion, concluding remarks, and some directions for further research.

The Saturated GDINA and the Reduced DCMs

The GDINA is a general framework and many commonly used DCMs such the DINA, DINO, or A-CDM can be viewed as special cases of the GDINA model (de la Torre, 2011). Suppose that K binary attributes are diagnosed in a DCM test with J dichotomous items. Define $\alpha = (\alpha'_1, \dots, \alpha'_j, \dots, \alpha'_L)'$ as the discrete attribute mastery profile matrix with a total number of possible attribute mastery patterns equal to $L = 2^K$. The structural parameter vector $\eta = (\eta_1, \dots, \eta_l, \dots, \eta_{L-1})'$ describes the probability $p(\alpha_l | \eta)$ of a randomly selected examinee being a member of the l th attribute mastery profile. The number of freely estimated structural parameters is thus $L - 1$ (Liu et al., 2016; Rupp et al., 2010). A $J \times K$ binary Q-matrix specifies which attributes are measured by each item, for example, $\mathbf{q}'_j = (q_{j1}, \dots, q_{jk}, \dots, q_{jK})$ is the j th row of the Q-matrix. In the Q-matrix, if attribute k is required for solving item j , $q_{jk} = 1$, and $q_{jk} = 0$ otherwise. Let $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N)'$ be the observed response matrix of N examinees, where $\mathbf{x}_i = (x_{i1}, \dots, x_{iJ})'$ is the i th examinee's response vector.

In the saturated GDINA, the probability of a correct response on item j for examinee i with attribute mastery pattern $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iK})'$ is given by de la Torre (2011).

$$P_j(\alpha_i) = \lambda_{j,0} + \sum_{k=1}^K \lambda_{j,1,(k)} \alpha_{ik} q_{jk} + \sum_{k=1}^{K-1} \sum_{k'=k+1}^K \lambda_{j,2,(k,k')} \alpha_{ik} \alpha_{ik'} q_{jk} q_{jk'} + \dots, \quad (1)$$

where $\lambda_{j,0}$ is the intercept parameter, $\lambda_{j,1,(k)}$ is the main effect parameter associated with attribute k , and $\lambda_{j,2,(k,k')}$ is the two-way interaction effect parameter associated with attribute k and attribute k' . For example, if the first, the second, and the third attributes are required by item j , $\mathbf{q}'_j = (1, 1, 1)$ and the i th examinee has mastered all the required attributes by the item $\alpha'_i = (1, 1, 1)$, then according to Equation 1 the saturated GDINA can be expressed as follows,

$$P_j(\alpha_i) = \lambda_{j,0} + \lambda_{j,1,(1)} + \lambda_{j,1,(2)} + \lambda_{j,1,(3)} + \lambda_{j,2,(1,2)} + \lambda_{j,2,(1,3)} + \lambda_{j,2,(2,3)} + \lambda_{j,3,(1,2,3)}. \quad (2)$$

By setting some of the parameters of the saturated model to fixed values, the commonly used reduced DCMs can be obtained, see de la Torre (2011). In this example, the DINA model can be obtained from GDINA by setting the main and the lower order interaction effect terms to zero and thus the model becomes

$$P_j(\alpha_i) = \lambda_{j,0} + \lambda_{j,3,(1,2,3)}. \quad (3)$$

To obtain the A-CDM, all the interaction effect parameters in the saturated GDINA are fixed to zero and the item response function of the A-CDM can be written as

$$P_j(\alpha_i) = \lambda_{j,0} + \lambda_{j,1,(1)} + \lambda_{j,1,(2)} + \lambda_{j,1,(3)}. \quad (4)$$

The mathematical function for the DINO model is given by

$$P_j(\alpha_i) = \lambda_{j,0} + \lambda_{j,1,(1)}, \quad (5)$$

by setting $\lambda_{j,1,(1)} = \lambda_{j,1,(2)} = \lambda_{j,1,(3)} = -\lambda_{j,2,(1,2)} = -\lambda_{j,2,(1,3)} = -\lambda_{j,2,(2,3)} = \lambda_{j,3,(1,2,3)}$.

Information Matrix Estimation Procedures for DCMs

In DCMs, the item parameters λ and the structural parameters η are simultaneously estimated from the observed data, typically by using the MMLE/EM (marginal maximum likelihood estimation via expectation–maximization) algorithm. Denote $\hat{\gamma} = (\hat{\lambda}', \hat{\eta}')$ as the MLEs of the model parameters that maximizes the log likelihood function of the observed response data $\ell(\gamma|\mathbf{x})$.

To directly estimate the asymptotic covariance matrix of the item parameter estimates in DCMs, there are at least five procedures available, namely, the incomplete information matrix in which only the item parameters are considered, the item-wise information matrix, the empirical cross-product information matrix, the observed information matrix, and the sandwich-type covariance matrix (Liu et al., 2018; Xin et al., 2016).

Philipp et al. (2018) have proved that the incomplete information matrix (de la Torre, 2009) and the item-wise information matrix (de la Torre, 2011) underestimates item *SEs*. They also studied the property of the empirical cross-product information matrix in providing consistent item parameter *SEs*, in which the item and structural parameters were simultaneously incorporated. The empirical cross-product information matrix is the cross-product of the derivative of the log likelihood of the observed data with respect to the model parameters,

$$\mathcal{I}_{\text{XPD}} = \frac{\partial \ell(\gamma|\mathbf{x})}{\partial \gamma} \frac{\partial \ell(\gamma|\mathbf{x})}{\partial \gamma'}. \quad (6)$$

They found that compared with the incomplete or the item-wise information matrix, the *SEs* based on the empirical cross-product information matrix showed smaller bias.

The observed information matrix and the sandwich-type estimator that included both the item and structural parameters for the estimation of the asymptotic covariance matrix were proposed by Liu et al. (2019). The observed information matrix is the negative of the second derivatives of the log likelihood of the observed data matrix with respect to the model parameters

$$\mathcal{I}_{\text{Obs}} = -\frac{\partial^2 \ell(\gamma|\mathbf{x})}{\partial \gamma \partial \gamma'}, \quad (7)$$

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (11)$$

Note that for the A-CDM, $\mathbf{R}\hat{\boldsymbol{\lambda}}_j = (\hat{\lambda}_{j,2,(1,2)}, \hat{\lambda}_{j,2,(1,3)}, \hat{\lambda}_{j,2,(2,3)}, \hat{\lambda}_{j,3,(1,2,3)})'$. To impose restriction on the saturated GDINA to obtain the DINO model and to test the hypothesis that $\hat{\lambda}_{j,1,(1)} = \hat{\lambda}_{j,1,(2)} = \hat{\lambda}_{j,1,(3)} = -\hat{\lambda}_{j,2,(1,2)} = -\hat{\lambda}_{j,2,(1,3)} = -\hat{\lambda}_{j,2,(2,3)} = \hat{\lambda}_{j,3,(1,2,3)}$, the following restriction matrix is defined,

$$\mathbf{R} = \begin{pmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}. \quad (12)$$

Close inspection of Equation 9 reveals that the Wald statistic is dependent on the accuracy of the item parameter estimates $\hat{\boldsymbol{\lambda}}_j$ and the corresponding asymptotic covariance matrix $\hat{\boldsymbol{\Sigma}}_j$. However, in the original implementation of the Wald statistic W_d for selecting the most suitable model, the item parameter covariance matrix was computed based on the item-wise information matrix, which leads to underestimated item parameter *SEs* (Philipp et al., 2018).

In this study, the authors propose that the Wald test for item-level model comparison is computed using $\mathcal{I}_{\text{Obs}}$ or $\mathcal{I}_{\text{SW}}$. As previous studies (Liu et al., 2018; Xin et al., 2016) have shown that the covariance matrices $\hat{\boldsymbol{\Sigma}}_j$ of the item parameter estimates based on the inverse of $\mathcal{I}_{\text{Obs}}$, or $\mathcal{I}_{\text{SW}}$ are more accurate than those based on the empirical cross-product, the incomplete, or the item-wise information matrix. In this study, the Wald statistic in which the computation of the item parameter covariance matrix $\hat{\boldsymbol{\Sigma}}_j$ is based on $\mathcal{I}_{\text{Obs}}$ or $\mathcal{I}_{\text{SW}}$, will be denoted by W_{Obs} or W_{SW} . In theory, it is reasonable to expect that the performance of W_{Obs} or W_{SW} would be better than the original version W_d .

The LR Test

In DCMs, another way to address the item-level model comparison question is through use of the LR test. The LR statistic is computed as

$$\text{LR} = -2 \log \left(\frac{L_r}{L_s} \right) = 2(\ell_s - \ell_r), \quad (13)$$

where L_r is the likelihood for the reduced DCM and L_s is the likelihood for the saturated DCM. Suppose that the model parameter estimates $\hat{\boldsymbol{\gamma}}$ lie in the interior of the parameter space. Then, under the null hypothesis, this statistic is also asymptotically chi-square distributed with degrees of freedom equal to the number of item parameters that are constrained in the reduced DCM at the item level. By comparing the log likelihoods of the saturated and the reduced models $\ell_s - \ell_r$, the LR can be used to test whether the difference between these two models is statistically significant. The calculation of the LR is straightforward and does not require the estimation of the information matrix.

Sorrel, Abad, et al. (2017) found that the Type I error control of the LR test was more robust than W_d . Hence, it is of interest to empirically compare the behavior of the W_{Obs} , W_{SW} , and LR under the same conditions.

Simulation Study

The primary interest of the current study was to examine the empirical performance of the W_{Obs} , W_{SW} , W_d , or LR via Monte Carlo simulations under conditions varying by sample size, attribute

correlation and fitted DCM. The purpose of the first part of the simulation was to investigate the Type I error rate control of these statistics. The binary item response data matrix was generated from DINA, A-CDM, or DINO separately. The second part of the simulation was conducted to illustrate the empirical power of these statistics when the fitted model was misspecified. The saturated GDINA model was used to generate the response data.

Three factors were manipulated in a $5 \times 3 \times 3$ factorial design: (a) To systemically investigate the impact of sample size on the use of asymptotic sampling distributions, five sample sizes were considered, 500, 1,000, 2,000, 3,000, and 5,000. (b) To better reflect the reality of the correlations between attributes in applied research (Kunina-Habenicht, Rupp, & Wilhelm, 2012), the correlation coefficients between the attributes were $\rho = 0$, $\rho = .5$, and $\rho = .8$. (c) The fitted DCMs were the DINA, A-CDM, and DINO models. This gave a total of 45 simulation conditions for each method. To further assess the effect of the manipulated between-subjects factors, the fitted DCM, sample size, and attribute correlation on the performance of the W_{Obs} , W_{SW} , W_d , or LR method with respect to the Type I error, the marginal means of the Type I error rates for each methods were analyzed using ANOVAs, and computing the effect size, omega-squared index (ω^2). Kirk (1996) offered the following guidelines for interpreting the magnitude of effect size ω^2 : small (0.010), medium (0.059), and large (0.138), consequently, only those factors whose ω^2 index is greater than 0.138 were focused.

The true probabilities of a correct response for examinees who mastered none and all of the required attributes were fixed at .2 and .9. The main and/or interaction effect parameters were fixed to $(.9 - .2)/n_j$, where n_j is the number of the main and/or interaction effect parameters for item j . The Q-matrix with five attributes and 30 items is provided in the online appendix Table A1. For each simulation condition, 1,000 replications were performed to provide stable results. The 95% confidence interval $p_{epmical} \pm 1.96 \sqrt{p_{epmical}(1 - p_{epmical})/R}$ of the empirical Type I error rate of these statistics which cover .05 is considered to be accurate (Bradley, 1978), $p_{epmical}$ is the empirical Type I error rate, $R = 1,000$ is the number of replications in each condition. As with de la Torre and Lee (2013) a power higher than of equal to .8 is considered adequate, at least .9 is considered excellent. The code was written in R (R Core Team, 2017) where the R functions for model parameter estimation, Wald test calculation and item parameter covariance matrix estimation were modified from the R packages CDM (Robitzsch, Kiefer, George, & Uenlue, 2017) and dcminfo (Liu & Xin, 2017). The R codes in the current study are available upon request from the corresponding author.

Results

Type I Error Rate Control

The empirical Type I error rate control results of the W_d , W_{Obs} , W_{SW} , and LR statistics across simulation conditions for DINA, A-CDM, and DINO at the 5% significance level are summarized in Tables 1 to 3, respectively. Results were averaged over items with the same number of required attributes (K_j). The 95% confidence interval of the empirical Type I error rate which do not cover .05 is marked in boldface.

As shown in Table 1, for the DINA model among the four statistics under investigation, the performance of the W_d or W_{Obs} to select the most appropriate DCM tended to inflate the Type I error rates, and although the W_{Obs} was slightly better than that of W_d , the W_{SW} or LR test had much better Type I error rates under the same conditions. The empirical Type I error rates presented in Table 2 tell a different story. For the A-CDM, the performance of the W_{Obs} and W_{SW} was better than W_d and LR in almost all cases. The Type I error rates of the LR were inflated

Table 1. Type I Error Rates for DINA ($\alpha = .5$).

ρ	N	W_d		W_{Obs}		W_{SW}		LR	
		$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$
0	500	0.066	0.110	0.060	0.096	0.044	0.070	0.055	0.063
	1,000	0.065	0.072	0.061	0.068	0.053	0.058	0.054	0.053
	2,000	0.055	0.065	0.053	0.062	0.050	0.058	0.053	0.053
	3,000	0.054	0.055	0.052	0.054	0.050	0.051	0.050	0.049
	5,000	0.048	0.051	0.047	0.050	0.046	0.049	0.053	0.049
0.5	500	0.082	0.139	0.071	0.111	0.046	0.069	0.061	0.069
	1,000	0.064	0.089	0.060	0.081	0.051	0.065	0.057	0.057
	2,000	0.054	0.069	0.052	0.066	0.047	0.059	0.053	0.055
	3,000	0.056	0.061	0.053	0.059	0.051	0.055	0.047	0.055
	5,000	0.054	0.061	0.053	0.060	0.051	0.057	0.050	0.048
0.8	500	0.112	0.131	0.102	0.121	0.037	0.050	0.072	0.097
	1,000	0.076	0.135	0.068	0.111	0.048	0.074	0.058	0.070
	2,000	0.063	0.090	0.060	0.084	0.052	0.069	0.054	0.055
	3,000	0.056	0.072	0.052	0.067	0.047	0.060	0.052	0.057
	5,000	0.053	0.064	0.051	0.061	0.049	0.057	0.054	0.051

Note. W_d is the Wald test based on the item-wise information matrix; W_{Obs} is Wald test based on the observed information matrix; W_{SW} is Wald test based on the sandwich-type covariance matrix; LR is the likelihood ratio test; K_j is the number of required attributes by the j th item. The 95% confidence interval of the value which do not cover .05 is marked in boldface. DINA = deterministic inputs, noisy, “and” gate model; ρ = the attribute correlation; N = the sample size.

Table 2. Type I Error Rates for A-CDM ($\alpha = .5$).

ρ	N	W_d		W_{Obs}		W_{SW}		LR	
		$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$
0	500	0.068	0.098	0.057	0.075	0.034	0.035	0.094	0.094
	1,000	0.056	0.067	0.051	0.059	0.040	0.042	0.083	0.069
	2,000	0.053	0.062	0.050	0.058	0.047	0.050	0.075	0.065
	3,000	0.052	0.059	0.050	0.056	0.047	0.050	0.074	0.064
	5,000	0.052	0.055	0.051	0.053	0.048	0.050	0.067	0.062
0.5	500	0.073	0.116	0.063	0.090	0.030	0.036	0.078	0.085
	1,000	0.058	0.077	0.054	0.068	0.045	0.047	0.071	0.072
	2,000	0.055	0.060	0.052	0.056	0.048	0.047	0.068	0.064
	3,000	0.056	0.060	0.053	0.056	0.049	0.051	0.066	0.061
	5,000	0.049	0.057	0.046	0.055	0.044	0.051	0.062	0.059
0.8	500	0.087	0.176	0.083	0.162	0.018	0.036	0.084	0.111
	1,000	0.065	0.098	0.056	0.078	0.038	0.045	0.066	0.072
	2,000	0.062	0.072	0.056	0.064	0.050	0.049	0.062	0.065
	3,000	0.059	0.063	0.054	0.059	0.050	0.051	0.058	0.060
	5,000	0.056	0.053	0.051	0.049	0.049	0.043	0.059	0.058

Note. W_d is the Wald test based on the item-wise information matrix; W_{Obs} is Wald test based on the observed information matrix; W_{SW} is Wald test based on the sandwich-type covariance matrix; LR is the likelihood ratio test; K_j is the number of required attributes by the j th item. The 95% confidence interval of the value which do not cover .05 is marked in boldface. A-CDM = additive cognitive diagnostic model; ρ = the attribute correlation; N = the sample size.

Table 3. Type I Error Rates for DINO ($\alpha = .5$).

ρ	N	W_d		W_{Obs}		W_{SW}		LR	
		$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$
0	500	0.079	0.108	0.071	0.087	0.044	0.059	0.061	0.075
	1,000	0.062	0.100	0.058	0.094	0.050	0.080	0.058	0.057
	2,000	0.057	0.072	0.055	0.070	0.051	0.066	0.052	0.051
	3,000	0.054	0.064	0.052	0.063	0.049	0.059	0.053	0.056
	5,000	0.053	0.058	0.052	0.057	0.051	0.055	0.048	0.048
0.5	500	0.091	0.071	0.077	0.063	0.039	0.035	0.081	0.083
	1,000	0.065	0.128	0.061	0.115	0.052	0.094	0.061	0.057
	2,000	0.062	0.090	0.060	0.084	0.054	0.076	0.055	0.054
	3,000	0.055	0.072	0.053	0.071	0.050	0.066	0.053	0.051
	5,000	0.055	0.064	0.054	0.063	0.052	0.060	0.052	0.057
0.8	500	0.091	0.035	0.089	0.084	0.029	0.040	0.227	0.209
	1,000	0.093	0.136	0.081	0.095	0.055	0.067	0.195	0.179
	2,000	0.072	0.138	0.069	0.128	0.060	0.101	0.146	0.142
	3,000	0.065	0.109	0.062	0.105	0.057	0.093	0.119	0.108
	5,000	0.057	0.083	0.056	0.081	0.053	0.076	0.080	0.093

Note. W_d is the Wald test based on the item-wise information matrix; W_{Obs} is Wald test based on the observed information matrix; W_{SW} is Wald test based on the sandwich-type covariance matrix; LR is the likelihood ratio test; K_j is the number of required attributes by the j th item. The 95% confidence interval of the value which do not cover .05 is marked in boldface. DINO = deterministic inputs, noisy, “or” gate model; ρ = the attribute correlation; N = the sample size.

under most conditions. The Type I error rates for the DINO model are reported in Table 3. It can be observed that only the W_{SW} demonstrated reasonably robust Type I error behavior across all conditions. Under the $\rho = .8$ condition the LR rejected too many well-fitting items (as high as .227).

ANOVAs were conducted to further assess the effect of the manipulated between-subjects factors on the performance of the W_{Obs} , W_{SW} , W_d , or LR methods. For each method, the mean empirical Type I error was used as the ANOVA dependent variable. The $\hat{\omega}^2$ values for W_d , W_{Obs} , W_{SW} , and LR presented in Table 4 can be summarized as follows: (a) The Type I error of W_d was largely affected by sample size; (b) The main effects of sample size and attribute correlation for W_{Obs} were large; (c) The performance of W_{SW} was substantially affected by the fitted DCM, and sample size; (d) The main effects of the fitted DCM and attribute correlation factors, and their interaction for LR were large. In considering the $\hat{\omega}^2$ values for W_d , W_{Obs} , W_{SW} , and LR in Table 4 and the empirical Type I error rates in Tables 1 to 3, the authors can note that (a) as the sample size increased, the performance of all statistics were considerably better; (b) the attribute correlation strongly affected the performance of W_{Obs} and increasing the attribute correlation leads to inflated Type I error rates; (c) the Type I error rates of the W_{SW} were well controlled for the DINA and A-CDM under most of the simulation conditions, but in general slightly inflated for the DINO model; and (d) for A-CDM, the Type I error of LR was closer to the nominal level as the attribute correlation increased, however, for the DINO model, the Type I error was inflated when the attribute correlation was high ($\rho = .8$).

Close inspection of the simulation results in Tables 1 to 4 reveals that, as expected, the Wald test W_d was improved by using the observed information matrix or the sandwich-type covariance matrix, and among the four tests, the newly proposed W_{SW} had the best performance in providing reliable Type I error control across most of the simulation conditions.

Table 4. $\hat{\omega}^2$ Values for W_d , W_{Obs} , W_{SW} , and LR.

	DCM	N	ρ	DCM $\times$ N	DCM $\times\rho$	N $\times\rho$
W_d	0.015	0.467	0.125	0.092	−0.031	−0.041
W_{Obs}	0.066	0.429	0.179	0.058	−0.022	0.010
W_{SW}	0.343	0.260	−0.002	0.117	0.004	0.075
LR	0.142	0.127	0.198	0.006	0.334	0.027

Note. W_d is the Wald test based on the item-wise information matrix; W_{Obs} is Wald test based on the observed information matrix; W_{SW} is Wald test based on the sandwich-type covariance matrix; LR is the likelihood ratio test; DCM is the fitted diagnostic classification model; N = the sample size; ρ = the attribute correlation; The $\hat{\omega}^2$ value that is greater than 0.138 is marked in boldface.

Table 5. The Correct Detection Rates for A-CDM ($\alpha = .5$).

ρ	N	W_d		W_{Obs}		W_{SW}		LR	
		$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$	$K_j = 2$	$K_j = 3$
0	500	0.699	0.728	.673	0.683	0.582	0.560	0.714	0.733
	1,000	0.928	0.950	0.928	0.950	0.916	0.936	0.942	0.964
	2,000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000
	3,000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	5,000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	500	0.624	0.833	0.592	0.800	0.456	0.697	0.634	0.823
	1,000	0.884	0.985	0.877	0.983	0.863	0.978	0.893	0.989
	2,000	0.994	1.000	0.994	1.000	0.994	1.000	0.995	1.000
	3,000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	5,000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.8	500	0.495	0.760	0.453	0.682	0.179	0.440	0.469	0.721
	1,000	0.754	0.955	0.731	0.946	0.667	0.922	0.747	0.953
	2,000	0.958	0.999	0.955	0.999	0.951	0.999	0.965	1.000
	3,000	0.994	1.000	0.994	1.000	0.994	1.000	0.995	1.000
	5,000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. W_d is the Wald test based on the item-wise information matrix; W_{Obs} is Wald test based on the observed information matrix; W_{SW} is Wald test based on the sandwich-type covariance matrix; LR is the likelihood ratio test; ρ = the attribute correlation; N = the sample size; K_j is the number of required attributes by the j th item. The value that is smaller than .80 in each condition is in boldface. The condition with inflated or deflated type I error is in strikeout font.

Power

For the DINA and DINO models, the correct detection rates of the W_d , W_{Obs} , W_{SW} , and LR tests were all above .831 under all of the simulation conditions, with one exception: The empirical power of W_{SW} was .524 ($K_j = 2$) and .729 ($K_j = 3$) for DINA model, when $\rho = .8$ and $N = 500$. Thus, only the correct detection rates for A-CDM are provided in Table 5. The correct detection rates of W_d , W_{Obs} , W_{SW} , and LR in Table 5 are shown in strikeout font for conditions with inflated or deflated Type I errors. As shown in Table 5, for A-CDM, when $N > 500$, among the four tests under investigation in this study, only the W_{SW} had accurate Type I error rate control and adequate or excellent power (higher than .863) under all conditions with one exception ($\rho = .8$ and $N = 1,000$). Sample size and attribute correlation influenced the power of W_{SW} , the A-CDM tended to have less power to correctly reject the null hypothesis when the sample size

was small ($N = 500$), and attribute correlation was high ($\rho = .8$). This is not surprising, given that the previous study showed that the omission of interaction effect terms had little impact on classification accuracy (Kunina-Habenicht et al., 2012). Although the correct detection rates of the W_d , W_{Obs} , and LR tests are similar to the patterns of the power results of the W_{SW} , it should be pointed out that due to the inaccurate Type I error rates in the simulation, these results are not necessarily indicative of good performance in real applications.

Discussion

Many DCMs with different assumptions or theories about how an examinee's latent attributes influence test performance are now available in education, psychology, and many other disciplines. Selecting the most appropriate DCM for each item is crucial in theory and actual applications, since no particular DCM suits all the diagnostic test items in many, if not all, diagnostic assessments. For item-level model selection, a Wald test based on a simplified statistic W_d has been commonly used. However, according to several studies the W_d statistic yielded inflated Type I error rates (de la Torre & Lee, 2013; Ma et al., 2016; Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017).

As an alternative, the item parameter asymptotic covariance matrix $\hat{\Sigma}_j$ from the original version of the Wald statistic W_d was replaced with a matrix obtained from the observed information $\mathcal{I}_{Obs}$ or the sandwich-type matrix $\mathcal{I}_{SW}$ to address the inflation of Type I error in this study, building on work by Liu et al. (2016), Xin et al. (2016), Philip et al. (2018), and Liu et al. (2019). Evidence from the Monte Carlo simulation suggests that when the sample size was reasonably large ($N \geq 1,000$) for all the three fitted models, the DINA, DINO, and A-CDM, the Type I error rate control of W_{SW} was accurate or robust. When $N \geq 1,000$, for the DINA, DINO, or A-CDM, the power of W_{SW} was adequate or excellent under all the simulation conditions with only one exception. Although the Type I error rate control of W_{Obs} was better than that of W_d , the W_{SW} had the best performance under all simulation conditions. In this study, the LR statistic demonstrated good Type I error behavior under some conditions, however, its good behavior is not guaranteed. Similar to previous studies (Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017), the authors also found that the computation of the LR was time consuming. For example, with the DINO model the LR and Wald took approximately 1 minute and less than 1 second to compute in one replication, respectively, when $N = 5,000$ and $\rho = .8$. According to the simulation results in this study, W_{SW} is the recommended tool for item-level model comparisons in empirical applications.

Although, as expected, W_{Obs} or W_{SW} had Type I error rates close to the nominal level and had good power under most of simulation conditions, and the simulation results indicate that the W_{SW} may serve as a practical tool for empirically selecting the most appropriate model at item level in DCMs, further research is needed. Many other factors that could potentially affect the behavior of these statistics were not manipulated in this study. For example, previous studies (Ma et al., 2016; Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017) have found that item quality, attribute distribution, the fitted model and test length affected the performance of W_d , and the authors believe it would be instructive to investigate how these factors affect the statistical properties of W_{Obs} or W_{SW} .

Acknowledgments

The authors would like to thank Dr. Hua-Hua Chang, Dr. John Donoghue, and the two anonymous reviewers for their valuable comments on this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author's research was supported by a grant from Social Science Foundation of Shandong Province, China (Grant No. 18CJYJ16).

ORCID iD

Yanlou Liu <https://orcid.org/0000-0002-4229-7573>

Supplemental Material

Supplemental material is available for this article online.

References

- Bradley, J. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115-130.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50, 355-373.
- Greeno, J. G. (1980). Trends in the theory of knowledge for problem solving. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 9-23). Hillsdale, NJ: Lawrence Erlbaum.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301-323.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation), University of Illinois, Urbana-Champaign.
- Henson, R. A., Templin, J. L., & Willse, J. T. (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.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.
- 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.
- Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.
- Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, 41, 3-26.
- Liu, Y., & Xin, T. (2017). dcminfo: Information matrix for diagnostic classification models (R package version 0.1.6). Retrieved from <https://CRAN.R-project.org/package=dcminfo>
- Liu, Y., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology*, 72, 18-37.

- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*, 200-217.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.
- Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics, 43*, 88-115.
- R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2017). CDM: Cognitive diagnosis modeling (R package version 5.9-27). Retrieved from <http://CRAN.R-project.org/package=CDM>
- Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*, 614-631.
- Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 13*, 39-47.
- Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*, 506-532.
- Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton, NJ: Educational Testing Service.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307.
- Xin, T., Liu, Y., & Tian, W. (2016, April). *Information matrix estimation procedures for cognitive diagnostic model*. Paper presented at the meeting of the National Council on Measurement in Education, Washington, DC.