



Improved Performance of Model Fit Indices with Small Sample Sizes in Cognitive Diagnostic Models

Hueying Tzou ^{1*}, Ya-Huei Yang ¹

¹ Department of Education, National University of Tainan, Tainan, Taiwan

ARTICLE HISTORY

Received: 14 November 2018

Revised: 13 March 2019

Accepted: 15 March 2019

KEYWORDS

RSS,

ζ^2 index,

model fit indices,

cognitive diagnostic models,

Q-matrix

Abstract: Selecting an appropriate cognitive diagnostic model (CDM) for data analysis is always challenging. Studies have explored several model fit indices for CDMs. The common results of these studies indicate that Q-matrix misspecifications lead to poor performance of the model fit indices in the context of CDMs. Thus, this study explored whether model fit indices improve performance with a modified Q-matrix. The average class size has reduced to 23 students in Taiwan because of the low birth rate; therefore, the study sought the effect of sample size on the performance of model fit indices. The results showed that Akaike's information criterion (AIC) was an excellent model fit index in small samples. Model fit indices with the modified Q-matrix presented superior performance.

1. INTRODUCTION

Recently, cognitive diagnostic models (CDMs) (DiBello, Roussos, & Stout, 2007) have been extensively studied in educational research (Jiao, 2009). CDMs are psychological models that are used to examine whether a subject is proficient in a skill or possesses a particular character (Chen, de la Torre, & Zhang, 2013) in order to provide more precise information regarding the subject (Ma, Iaconangelo, & de la Torre, 2016). When applying CDMs to analyze testing data to obtain diagnostic information regarding a subject, one must select the analytical model and define the Q-matrix of the test (Tatsuoka, 1983). Recently, CDMs have been developed in accordance with their applicable circumstances for different cognitive situations, such as the deterministic inputs, noisy “and” gate model (DINA; Junker & Sijtsma, 2001); the deterministic inputs, noisy “or” gate model (DINO; Templin & Henson, 2006); and the generalized deterministic inputs, noisy “and” gate model (GDINA; de la Torre, 2011). To ensure that valid diagnostic information is obtained from the model analytics, the model–data fit must be considered. Researchers can directly adopt the saturation model for data analysis and routinely have a high degree of fit; however, the complexity of the saturation model requires larger samples to produce accurate estimates (de la Torre & Lee, 2013). Practitioners are often unable to obtain sufficient samples; therefore, the application of CDMs to small sample sizes is critical.

CONTACT: Hueying Tzou ✉ tzou@mail.nutn.edu.tw 📧 Department of Education, National University of Tainan, 33, Sec. 2, Shu-Lin St., Tainan 700, Taiwan

ISSN-e: 2148-7456 /© IJATE 2019

For example, the average class size has reduced to 23 students in Taiwan because of the low birth rate.

In addition to selecting the correct CDM, the correct Q-matrix is equally critical in CDM analysis. Studies have confirmed that a misspecified Q-matrix negatively affects the recovery of parameters and the classification of subjects (Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008). Kunina-Habenicht (2012) indicated that with 30% misspecification of the Q-matrix at a sample size of 1000, the accurate classification rate was only 64%, even if the number was increased to 10,000 (10 times) under the same conditions.

Model fit indices were developed to select the appropriate model for data analysis. The indices are mainly divided into two types: absolute and relative fit indices. Common absolute fit indices are pair proportion correct, pair transformed correlation, and pair log-odds. According to Chen (2013), the pair proportion correct rate is a single-variable absolute fit index, and the performance of this indicator is poor; thus, it is not employed in this study. This study explored the performance of pair transformed correlation (hereinafter referred to as r) and pair log-odds indices (I) to correctly reject wrong models with the modified Q-matrix.

Relative fit indices are another type of model fit indices. More than two models can fit the same data set. To select the most appropriate models, relative model fit indices are required. Relative model fit indices are used for model comparisons and maximum log-likelihood (for example, $-2 \log$ -likelihood or $-2LL$). The two most commonly used are the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1976). This study explored the performance of AIC and BIC with the modified Q-matrix.

2. BACKGROUND

Studies have noted that Q-matrix misspecification affects model parameters (Rupp & Templin, 2008; de la Torre, 2008) and the accuracy of examinees' classifications (Chiu & Douglas, 2013). If a q-vector of an item was misspecified, the estimated item parameters and the examinees' classifications were significantly biased.

For this reason, researchers have focused on the development of Q-matrix correction methods. de la Torre (2008) developed the sequential δ method for the DINA model to perform item-by-attribute Q-matrix modification. According to de la Torre, if an item must be included in a particular attribute, the difference in the correct answering probability of the group with and without the particular attribute is maximized (de la Torre indicated the difference value as δ). Therefore, under the item level, we first assume the q-vector of the item as a zero vector and compare the δ values of each attribute (or a combination of attributes) to include the attribute with the maximum δ value into the q-vector.

de la Torre simulated 5000 examinees with uniform attribute distribution to explore the performance of the modification method (δ method) with different types of Q-matrix misspecification: overspecified (an attribute that is originally not measured but included in the q-vector), underspecified (an attribute that is originally measured but excluded in the q-vector), and mixed misspecification (both overspecified and underspecified in the same q-vector). The results showed that an appropriate cutting value would lead to an excellent modified Q-matrix (same as the original Q-matrix) regardless of the Q-matrix misspecification. However, there are numerous restrictions to this application; the method is only for the DINA model, and the fitting model must be known prior.

Chiu (2013) developed the minimum residual sum of squares (RSS) method to improve the limits of the δ method (the fitting model must be known prior). The RSS method is based on nonparametric classification (Chiu & Douglas, 2013) to obtain examinees' attribute patterns and the theoretical response (η_{ij}) of examinees' attribute patterns and the Q-matrix. The squared

value of the difference between the actual and theoretical responses is calculated with equation (1). Chiu argues that if the q -vector is correctly defined, theoretical responses are similar to actual responses, and the RSS value is minimized. The next step is to calculate the RSS value of each q -vector in the item level and choose the q -vector with the minimum RSS value as the new q -vector of the corresponding item.

$$RSS_i = \sum_{j=1}^N (X_{ij} - \eta_{ij})^2 \quad (1)$$

There were two data-generating models (DINA and noisy input, deterministic “and” gate [NIDA]), two attribute numbers ($K = 3$ and 5), three attribute pattern distributions (uniform, multivariate normal threshold, and higher order), three sample sizes (100, 500, and 1000), four item qualities ($s = g = 0.2, 0.3, 0.4, \text{ and } 0.5$), and two Q-matrix misspecification rates (random misspecification 10% or 20%). The criteria were to compare the recovery rates of the true and corrected Q-matrices. A higher recovery rate indicated the superior correction performance of the RSS method. The results showed that despite the small sample size, in the case of 0.3, the recovery rate was at least 88% if the Q-matrix misspecification was 10% and at least 75% if the Q-matrix misspecification was 20%.

Unlike the previous δ index which is only applied for the DINA model and required assuming about the fitting model, de la Torre and Chiu (2016) developed another more generalized Q-matrix modification method; they called this index ζ^2 (de la Torre & Chiu, 2016). The new modification method $-\zeta^2$, used the GDINA model to exceed the limits of the δ method, which was only applied with the DINA model. However, the GDINA model is a complex model because of the estimation of many parameters. In other words, a large sample size is required to obtain accurate estimates. The sample size in the study of de la Torre & Chiu (2016) was 2000; the performance of smaller samples has rarely been explored. Therefore, this study mainly focuses on small sample sizes and explores the performance of ζ^2 indicators.

As showed in Chiu’s (2013) study, the performance of the RSS method with the data generated from the DINA and NIDA models was excellent with small sample sizes. Nevertheless, the performance of the RSS method under the GDINA model was rarely discussed in literatures. In this case, we compare the performance of the RSS method and the ζ^2 method under the GDINA model.

3. METHOD

3.1. Research Purposes and Questions

The purposes of the study are as follows:

1. To explore the performance of the RSS method and the ζ^2 method with the setting sample sizes and Q-matrix misspecifications.
2. To explore the performance of model fit indices (AIC, BIC, r , l) with small sample sizes with the original and modified Q-matrix.
3. To compare the performance of the model fit indices with the original and modified Q-matrix.

3.2. Study I: Simulation design

This research was divided into two studies. In study I, both Q-matrix modification methods were compared, and the superior one would be used in the second study. Figure 1 presents the flow chart of study I.

Data generations and analyses were conducted using R software (R Core Team, 2017). The R package GDINA (Ma & de la Torre, 2018) was used to generate data sets. The item parameters were setting as $s=g=0.1$ for all items. We assumed the examinees’ attribute patterns were uniform. The number of attributes varied with the coverage of test; thus, we assumed that the smaller domain contained fewer attributes ($K = 3$), and the larger domain contained more

attributes ($K = 5$). Test length was set to 30 items. Meanwhile, Taiwan currently averages 23 students per class and 3.3 classes per grade in elementary schools. Therefore, we set the sample sizes to 50 (approximately two classes), 75 (three classes), 100, and 200.

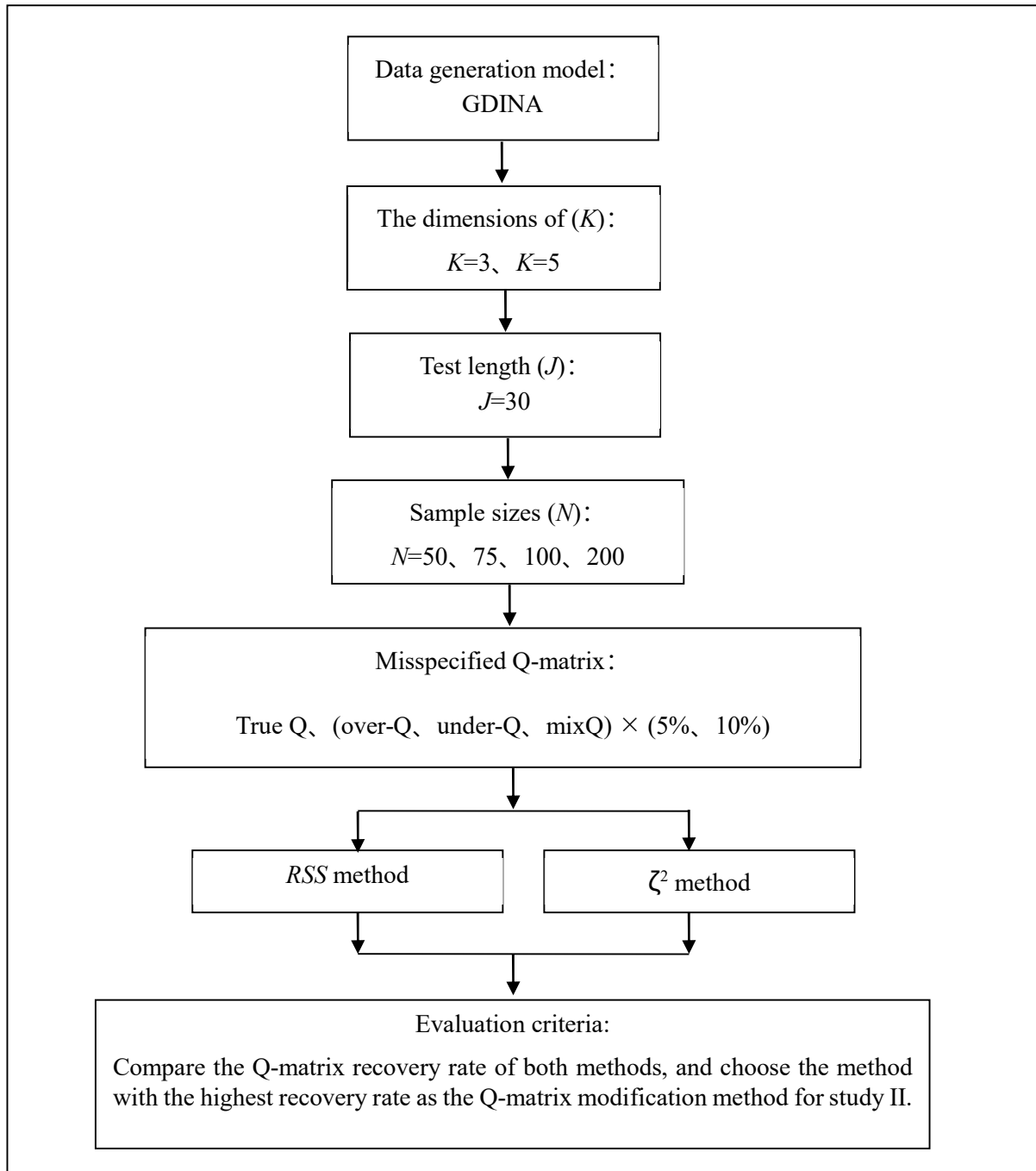


Figure1. Experimental flow chart for study I.

A total of six Q-matrix misspecification situations (three misspecification types \times two misspecification rates) were studied; the three misspecification types were overspecified (overQ), underspecified (underQ), and mix-specified (mixQ) Q-matrices. The misspecifications were randomly altered. OverQ meant that the item did not require the attribute, but the coding of the attribute was changed to 1 from 0 to become a requiring attribute; underQ meant that the item required the attribute, but the coding of the attribute was altered to 0; mixQ meant that in the same item, one required attribute was coded as 0 and another one not

required attribute was coded as 1. Misspecification rates were 5% and 10%. For example, 5%overQ meant 5% erroneous coding elements overspecified in the Q-matrix

In this study, we compared the Q-matrix recovery rates of both modification methods (RSS method and ζ^2 method) with the six Q-matrix misspecifications and eight simulation conditions. We simulated each of the 48 combinations in this study and replicated each conditions 30 times. The results were displayed with the mean Q-matrix recovery rate.

$$Q\text{-matrix recovery rates} = 1 - \left| \frac{\sum_{j=1}^J \sum_{k=1}^K q_{jk}^{\text{original}} - q_{jk}^{\text{corrected}}}{J \times K} \right| \tag{2}$$

J: test length

K: numbers of attributes

q_{jk}^{original} : original coding in item j and attribute k

$q_{jk}^{\text{corrected}}$: corrected coding in item j and attribute k

Many researches have shown that a misspecified Q-matrix affects the estimation of item parameters (de la Torre, 2008; Rupp & Templin, 2008; Kuninan-Habenicht et al., 2012). To prevent confounding effects on the study results caused by the structure of the Q-matrix, we made the Q-matrix as balanced as possible. The balanced design maintained the number of attributes measured by an item (mean item complexity) and the number of items measuring each attribute (attribute information) approximately the same.

Table 1 shows the correct Q-matrix of K = 5 (hereinafter referred to as True Q, TQ). The attribute information is the same (each attribute is measured by 12 items). There are 10 single-attribute items, 10 double-attribute items, and 10 triple-attribute items. Table 2 is the TQ of K = 3.

Table 1. True Q-matrix for K = 5

	$\alpha1$	$\alpha2$	$\alpha3$	$\alpha4$	$\alpha5$		$\alpha1$	$\alpha2$	$\alpha3$	$\alpha4$	$\alpha5$
Item01	1	0	0	0	0	Item16	0	1	0	1	0
Item02	0	1	0	0	0	Item17	0	1	0	0	1
Item03	0	0	1	0	0	Item18	0	0	1	1	0
Item04	0	0	0	1	0	Item19	0	0	1	0	1
Item05	0	0	0	0	1	Item20	0	0	0	1	1
Item06	1	0	0	0	0	Item21	1	1	1	0	0
Item07	0	1	0	0	0	Item22	1	1	0	1	0
Item08	0	0	1	0	0	Item23	1	1	0	0	1
Item09	0	0	0	1	0	Item24	1	0	1	1	0
Item10	0	0	0	0	1	Item25	1	0	1	0	1
Item11	1	1	0	0	0	Item26	1	0	0	1	1
Item12	1	0	1	0	0	Item27	0	1	1	1	0
Item13	1	0	0	1	0	Item28	0	1	1	0	1
Item14	1	0	0	0	1	Item29	0	1	0	1	1
Item15	0	1	1	0	0	Item30	0	0	1	1	1

Table 2. True Q-matrix for K = 3

	α_1	α_2	α_3		α_1	α_2	α_3
Item01	1	0	0	Item16	0	1	0
Item02	0	1	0	Item17	0	0	1
Item03	0	0	1	Item18	1	1	0
Item04	1	1	0	Item19	1	0	1
Item05	1	0	1	Item20	0	1	1
Item06	0	1	1	Item21	1	1	1
Item07	1	1	1	Item22	1	0	0
Item08	1	0	0	Item23	0	1	0
Item09	0	1	0	Item24	0	0	1
Item10	0	0	1	Item25	1	1	0
Item11	1	1	0	Item26	1	0	1
Item12	1	0	1	Item27	0	1	1
Item13	0	1	1	Item28	1	1	1
Item14	1	1	1	Item29	1	1	1
Item15	1	0	0	Item30	1	1	1

3.2.1. Q-matrix misspecification design

In the case of 5%overQ, there were 150 elements with 5% misspecifications. The researcher randomly selected 7 elements that originally coded as 0 and altered them to 1. In the case of 5%underQ, 8 elements that originally coded as 1 were altered to 0. In the case of 5%mixQ, 7 items were selected; the elements that originally coded as 0 were altered to 1, and the elements that originally coded as 1 were altered to 0 under the same item. In the case of K = 3, 4 elements were altered in 5%overQ, 5 elements were altered in 5%underQ, and 4 items were altered in 5%mixQ.

3.3. Results of Study I: Performance of Q-matrix Modification Methods

The performance of the ζ^2 index under the condition of K = 3 is shown in Table 3. The lowest recovery rate (0.815) was shown in N = 50; this indicates 18.5% type I error. The highest recovery rate (0.956) was shown in N = 200; this indicates 4.4% type I error. The lowest recovery rate under the condition of K = 5 (0.516) was shown in N = 50; this indicates 48.4% type I error. The highest recovery rate (0.620) was shown in N = 200; this indicates 38% type I error. De la Torre and Chiu (2016) used the ζ^2 index to modify the data generated from the GDINA model with fixed sample size (N = 2000), test length (J = 30), attribute numbers (K = 5), and random Q-matrix misspecification rates (5%). The result showed that type I error was 2%, and the Q-matrix recovery rate was 0.971. By comparison, the results of the current study showed much higher type I error and a much lower Q-matrix recovery rate. The effect of sample size on the performance of the ζ^2 method was notable. The results of the current study were quite different from those of de la Torre and Chiu (2016) under the same K = 5 simulation conditions. This may be caused by the sample size. The ζ^2 method requires item parameters and examinees' attribute patterns for Q-matrix modification. The largest sample size in this study (N = 200) was only one tenth of that in the study of de la Torre and Chiu; therefore, the estimated parameters were less accurate and led to poor modification results. This can be partially supported by the simulation result of K = 3. Because of the decreased number of attributes (K = 3), the estimated parameters were decreased, and the accuracy of parameter estimation was improved. Therefore, the modification performance of the ζ^2 index method was improved. Despite the sample size of only 200, type I error was reduced to below 5%, and the Q-matrix recovery rate was increased to 0.95.

The results of the RSS method in the case of $K = 3$ are shown in Table 4. The recovery rate of RSS for TQ was approximately 0.77, and the type I error was approximately 23%. Under $K = 5$, the lowest recovery rate (0.812) was shown in $N = 50$; this indicates that the type I error was 18.8%. The highest recovery rate (0.835) was shown in $N = 200$; this indicates that the type I error was 16.5%. The ζ^2 method exhibited lower type I error in $K = 3$; the RSS method exhibited lower type I error in $K = 5$.

For $K = 3$, 5% Q-matrix misspecification, and $N = 200$, the recovery rate of the ζ^2 method exceeded 0.95; that is, the misspecification rates of the modified Q-matrix were lower than 5%. In the case of 10% Q-matrix misspecification and $N = 100$, the recovery rate of the ζ^2 method exceeded 0.9. According to the results of $K = 3$ and $N = 200$, the misspecification rate of the modified Q-matrix was lower than that of the original Q-matrix. This indicates that the ζ^2 method is an effective modification method. However, in the case of $K = 5$, the performance of the ζ^2 method was not acceptable. For $K = 5$, the Q-matrix recovery rates were all lower than 0.7; that is, the misspecification rate of the modified Q-matrix was higher than the settings. Therefore, the ζ^2 method is not suitable for $K = 5$ and $N < 200$.

Table 3. Q-Matrix Recovery Rates of the ζ^2 Method

K=3	TQ	overQ		underQ		mixQ	
		5%	10%	5%	10%	5%	10%
N							
50	0.815	0.817	0.813	0.816	0.820	0.810	0.801
75	0.865	0.864	0.860	0.864	0.866	0.865	0.852
100	0.902	0.899	0.900	0.902	0.900	0.897	0.901
200	0.956	0.953	0.952	0.954	0.952	0.954	0.952
K=5	TQ	overQ		underQ		mixQ	
N		5%	10%	5%	10%	5%	10%
50	0.516	0.511	0.504	0.531	0.546	0.525	0.538
75	0.512	0.499	0.498	0.522	0.549	0.522	0.542
100	0.516	0.510	0.506	0.536	0.550	0.541	0.550
200	0.620	0.608	0.596	0.638	0.653	0.631	0.637

Note: N = sample size; TQ = True Q-matrix, the Q-matrix used for data generation; overQ = overspecified Q-matrix; underQ = underspecified Q-matrix; mixQ = mix-misspecified Q-matrix; 5% = 5% of entries of the Q-matrix were changed; 10% = 10% of entries of the Q-matrix were changed.

Table 4. Q-Matrix Recovery Rates of the RSS Method

K=3	TQ	overQ		underQ		mixQ	
		5%	10%	5%	10%	5%	10%
N							
50	0.778	0.776	0.776	0.777	0.778	0.778	0.769
75	0.773	0.771	0.769	0.772	0.774	0.774	0.773
100	0.773	0.773	0.773	0.773	0.774	0.774	0.771
200	0.776	0.777	0.775	0.774	0.774	0.776	0.774
K=5	TQ	overQ		underQ		mixQ	
N		5%	10%	5%	10%	5%	10%
50	0.812	0.805	0.788	0.816	0.809	0.797	0.757
75	0.822	0.819	0.793	0.822	0.824	0.812	0.777
100	0.829	0.826	0.821	0.823	0.825	0.816	0.782
200	0.835	0.836	0.833	0.833	0.831	0.833	0.816

3.3.2. The effect of sample size

In the case of $K = 3$, the Q-matrix recovery rates of the ζ^2 method increased with the sample size; by contrast, the recovery rates of the RSS method were fixed at approximately 0.77, and no increasing trend was observed. Additionally, under the condition of $K = 3$, the Q-matrix recovery rates of the ζ^2 method were higher than those of the RSS method; and the difference in recovery rates between the ζ^2 and RSS methods increased with sample size.

However, under the condition of $K = 5$, the performance of the methods was considerably different. The Q-matrix recovery rates of the RSS method were significantly higher than those of the ζ^2 method. The difference in both methods was the largest at $N = 50$ and the smallest at $N = 200$. In other words, with larger sample sizes, the recovery rates became more similar.

3.3.3. The effect of Q-matrix misspecification rates

Under the condition of $K = 3$, the Q-matrix recovery rates of both methods did not reduce with the increase in misspecification rates. For example, the recovery rate of the ζ^2 method was 0.817 at 5%overQ and 0.813 at 10%overQ. The recovery rates were almost the same even though the misspecification rate increased from 5% to 10%. Furthermore, the recovery rate of the RSS method was 0.776 at 5%overQ and 10%overQ with no difference between the misspecification rates. In the case of $K = 5$, the Q-matrix recovery rates of the ζ^2 and RSS methods decreased slightly due to the increase in misspecification rates, but it is was not significant

3.3.4. The effect of Q-matrix misspecification types

The difference in the recovery rates was not distinct among the three Q-matrix misspecification types for the ζ^2 and RSS methods. For example, in the case of $K = 3$, $N = 200$, and 10% misspecification, the recovery rates of the ζ^2 method were 0.952 for overQ, underQ, and mixQ; meanwhile, the recovery rates of the RSS method were 0.775, 0.774, and 0.774, respectively. The results imply that the Q-matrix misspecification type has a minor effect on both methods.

For $K = 3$, the ζ^2 method exhibited superior modification performance; by contrast, the RSS method exhibited superior modification performance for $K = 5$. Given these results, the ζ^2 method was applied to Q-matrix modification for $K = 3$, and the RSS method was applied for $K = 5$ in study II.

3.4. Study II: Simulation design

The simulation data are the same as those used for study I. Study II compared the performance of model fit indices with the true Q-matrix, misspecification Q-matrix, and the corresponding modified Q-matrix. Figure 2 presents the flow chart of study II.

3.5. Results of Study II

For the readability, the results of $K=5$ were shown in the appendix page.

3.5.1. The performance of relative indices with original Q-matrix

As shown in Table 5, under the condition of $K = 3$, AIC always correctly selected the GDINA (the correct data-generating model) as the fitting model; the performance of BIC in correct model selection varied with the misspecification type. In the cases of overQ and underQ, the selection rate of GDINA was much higher than that of the other two models in BIC, except for underQ at $N = 50$; in the case of 10%mixQ, the selection rate of GDINA was much higher than that of the other two models at $N = 100$ and $N = 200$. The results implied that in the case of $K = 3$, BIC was considerably affected by Q-matrix misspecification types.

Under the condition of $K = 5$, the selection rate of GDINA in AIC was much higher than that of the other two models for all simulation conditions. Different from that in the $K = 3$ scenario,

the performance of model selection in BIC was affected by sample size for $K = 5$. Only for $N = 200$ was the selection rate of GDINA in BIC much higher than that of the other two models. The results showed that the correct model selection rate of AIC was high under various conditions. BIC was affected by the Q-matrix misspecification type under the condition of $K = 3$ and by sample size under the condition of $K = 5$. These results were similar to the results of Hu et al. (2015).

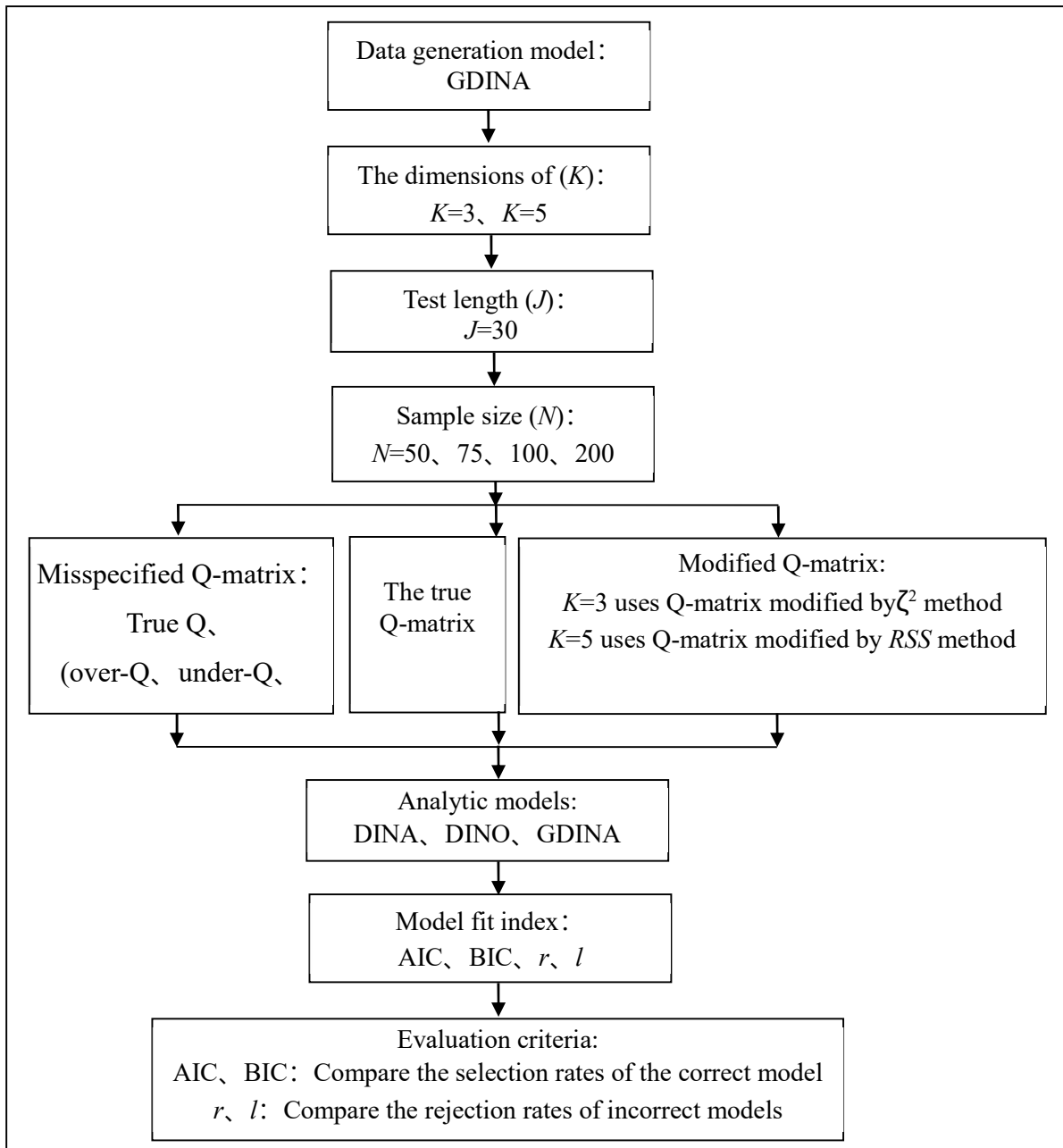


Figure 2. Experimental flow chart for study II.

Table 5. Selection Rates of the Relative Indices Under Various Simulation Conditions.

K=3		TQ		overQ				underQ				mixQ			
				5%		10%		5%		10%		5%		10%	
N	M	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
50	DINA	0	0.167	0	0.033	0	0	0	0.267	0	0.267	0	0.333	0	0.333
	DINO	0	0.333	0	0.100	0	0.033	0	0.300	0	0.267	0	0.400	0	0.600
	GDINA	1	0.500	1	0.867	1	0.967	1	0.433	1	0.467	1	0.267	1	0.067
75	DINA	0	0	0	0	0	0	0	0.033	0	0.067	0	0.067	0	0.167
	DINO	0	0.033	0	0	0	0	0	0.133	0	0.067	0	0.167	0	0.467
	GDINA	1	0.967	1	1	1	1	1	0.833	1	0.867	1	0.767	1	0.367
100	DINA	0	0	0	0	0	0	0	0	0	0	0	0	0	0.033
	DINO	0	0	0	0	0	0	0	0	0	0	0	0.067	0	0.200
	GDINA	1	1	1	1	1	1	1	1	1	1	1	0.933	1	0.767
200	DINA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	DINO	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	GDINA	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: M = analytic model

3.5.2. The performance of relative indices with the modified Q-matrix

As shown in Table 6, under the condition of K = 3, AIC still correctly selected the GDINA model as the fitting model. The correct model selection rates of BIC with the modified Q-matrix were higher than those with the original Q-matrix. For example, in the case of the original underQ, a perfect correct selection rate of the GDINA in BIC was observed at N = 100 and N = 200 but observed at N = 75, 100, and 200 in the corresponding modified underQ. Under the situation of K = 5, AIC showed better correct selection rates at sample sizes equal to or greater than 75; while BIC showed better correct selection rates at sample sizes equal to 200.

Table 6. Selection Rates for Relative Indices with the Modified Q-Matrix

K=3, ZETA		overQ				underQ				mixQ			
		5%		10%		5%		10%		5%		10%	
N	M	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
50	DINA	0	0.133	0	0.133	0	0.167	0	0.133	0	0.200	0	0.200
	DINO	0	0.067	0	0.067	0	0.067	0	0.067	0	0.067	0	0.067
	GDINA	1	0.800	1	0.800	1	0.767	1	0.800	1	0.733	1	0.733
75	DINA	0	0	0	0	0	0	0	0	0	0	0	0
	DINO	0	0	0	0	0	0	0	0	0	0	0	0
	GDINA	1	1	1	1	1	1	1	1	1	1	1	1
100	DINA	0	0	0	0	0	0	0	0	0	0	0	0
	DINO	0	0	0	0	0	0	0	0	0	0	0	0
	GDINA	1	1	1	1	1	1	1	1	1	1	1	1
200	DINA	0	0	0	0	0	0	0	0	0	0	0	0
	DINO	0	0	0	0	0	0	0	0	0	0	0	0
	GDINA	1	1	1	1	1	1	1	1	1	1	1	1

Note: M = analytic model

3.5.3. The performance of absolute indices with the original Q-matrix

According to Table 7, under the condition of $K = 3$, the rejection rates of three models (DINA, DINO, and GDINA) for mixQ were all 1 in both absolute model fit indices (r, l); the rejection rates of the three models increased with sample sizes in both absolute model fit indices under the situation of underQ. In the cases of TQ and overQ, the rejection rates of the DINA and DINO models increased with sample size, but the GDINA model decreased with sample size. These results implied that r and l tended to reject correct models for underQ and mixQ but tended to fail to reject correct models in the cases of TQ and overQ. These results were consistent with those of Chen et al., (2013) and Hu et al., (2015). Overall, when applying r and l , they accepted the correct model under the conditions of TQ and overQ. The results of $K = 5$ were roughly similar to that of $K = 3$.

Table 7. Rejection Rates of the Absolute Indices Under Various Simulation Conditions

K=3		TQ		overQ				underQ				mixQ			
				5%		10%		5%		10%		5%		10%	
N	M	r	l	r	l	r	l	r	l	r	l	r	l	r	l
50	DINA	0.900	0.900	1	1	1	1	0.867	0.933	0.967	1	1	1	1	1
	DINO	0.800	0.900	1	1	1	1	0.800	0.900	1	1	1	1	1	1
	GDINA	0.067	0	0.067	0.033	0.067	0	0.533	0.567	0.867	1	1	1	1	1
75	DINA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0	0.033	0	0.033	0	0.033	0.667	0.667	1	1	1	1	1	1
100	DINA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0	0	0	0	0	0	0.767	0.833	1	1	1	1	1	1
200	DINA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0	0	0	0	0	0	0.967	0.967	1	1	1	1	1	1

3.5.4. The performance of absolute indices with the modified Q-matrix

As shown in Table 8, under the condition of $K = 3$, r and l tended to fail to reject the GDINA (rejection rates were less than 0.1), and the rejection rates of the DINA and the DINO were all 1. These results showed that the rejection rates of the correct model could be effectively reduced after the Q-matrix was modified. However, there was no such finding for $K = 5$. It meant that the rejection rates of the correct model weren't affected by the modified Q-matrix as $K = 5$.

Table 8. Rejection Rates for Absolute Indices with the Modified Q-Matrix

K=3, ZETA		overQ				underQ				mixQ			
		5%		10%		5%		10%		5%		10%	
N	M	r	l	r	l	r	l	r	l	r	l	r	l
50	DINA	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0.033	0	0.033	0	0.067	0	0.067	0	0.033	0	0.067	0
75	DINA	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0	0	0	0	0.033	0.033	0.033	0.033	0	0	0	0
100	DINA	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
200	DINA	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033

4. DISCUSSION

Before applying CDM results, one must ensure the CDM fits the data. Enhancing the fitness of the data and model strengthens the validity and inferences of the results. Therefore, selecting an appropriate model–fit index is crucial. The decreasing birth rate in Taiwan causes the number of students to decrease each year; the teaching and learning style are demanding to provide more individualized information. Conventional single scores have been inappropriate to help teachers, parents, and students to understand learning results. CDMs can exactly meet the needs of the current education and give individual students feedback on learning strengths and weaknesses. Previous studies have shown that misspecified Q-matrices and model selection affect the performance and applicability of CDMs, but no study has designed to explore whether the effect of CDMs applications can be improved as the Q-matrix has been modified in advance. The current study not only explore the effect of CDMs applications with the modified Q-matrix, but also explore the effect of CDMs applications with the small sample sizes to meet the education field needs.

The ζ^2 index and the RSS methods of Q-matrix modification were explored in this study. According to the modification results, in the case of $K = 3$, the ζ^2 method can effectively correct the misspecification of the Q-matrix. However, in the case of $K = 5$, the performance of both methods was not as good as expected. We also found that the performance RSS method was affected by the data generation models. In this study, for $K = 3$, the Q-matrix recovery rates of the RSS method were in the interval of 0.771 to 0.778 with the GDINA model generating data; while the Q-matrix recovery rates were more than 0.9 (as high as 1.0) in the case of $K = 3$ with the DINA model generating data. Even though under the same conditions (10% Q-matrix misspecification rate, $K = 3$ or $K = 5$), the Q-matrix recovery rate of this study underperformed Chiu’s. It implied that the RSS method might not be suitable for the GDINA-generated data.

The relative index, AIC, showed excellent performance with small samples; therefore, AIC was an appropriate model fit index for small samples. Conversely, BIC was sensitive to Q-matrix misspecification type and sample size; BIC was only suitable for overQ and $N \geq 200$. The

absolute indices of the study were sensitive to Q-matrix misspecification type and only displayed excellent performance in the cases of TQ and overQ.

The results showed that in the case of $K = 3$, all relative and absolute model fit indices improved model selection with the modified Q-matrix. This implied that Q-matrix modification could improve the performance of model fit indices as few attributes or small domain measured. However, in the case of $K = 5$, both modification methods exhibited poor performance. It might be resulted from the complexity of more attribute or the generating GDINA model since the recovery rate of the modified Q-matrix in the case of $K = 3$ performed better. Meanwhile, we also found similar pattern on the performance of the model fit indices by using the modified Q-matrix. Therefore, these limitations should be taken into consideration in future studies to expand the application of CDMs in practice.

ORCID

Hueying Tzou  <https://orcid.org/0000-0002-6740-6852>

Ya-Huei Yang  <https://orcid.org/0000-0002-4109-2381>

5. REFERENCES

- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, *19*, 716-723.
- DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979-1030). Amsterdam, Netherlands: Elsevier.
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnostic modeling. *Journal of Educational Measurement*, *50*, 123-140.
- Chiu, C.Y. (2013). Statistical refinement of the Q-Matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598-618.
- 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.
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343-362.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 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*(4), 355-373.
- de la Torre, J., & Chiu, C. Y. (2016). A General Method of Empirical Q-matrix Validation. *Psychometrika*, *81*(2), 253-273.
- Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. (2016). Evaluation of Model Fit in Cognitive Diagnosis Models. *International Journal of Testing*, *16*(2), 119-141.
- Jiao, H. (2009). Diagnostic classification models: Which one should I use? *Measurement: Interdisciplinary Research & Perspective*, *7*(1), 65-67.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.
- 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.
- Lei, P.-W., & Li, H. (2016). Choosing Correct Cognitive Diagnostic Models and Q-Matrices. *Applied Psychological Measurement*, *40*(6), 1-12.

- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model Similarity, Model Selection, and Attribute Classification. *Applied Psychological Measurement, 40*(3), 200-217.
- Ma, W., & de la Torre, J. (2018). GDINA: The generalized DINA model framework. R package version 2.3. Retrived from <https://CRAN.R-project.org/package=GDINA>
- R Core Team (2017). R: A language and environment for statistical computing (Version 3.4.3) [Computing software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org>
- Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification nonparameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.
- Schwarzer, G. (1976). Estimating the dimension of a model. *Annals of Statistics, 6*,461–464.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

6. APPENDIX

Table A1. Selection Rates of the Relative Indices Under Various Simulation Conditions

K=5		TQ		overQ				underQ				mixQ			
				5%		10%		5%		10%		5%		10%	
N	M	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
50	DINA	0	0.133	0	0.333	0	0.4	0	0.167	0	0.167	0	0.233	0	0.267
	DINO	0	0.867	0	0.667	0	0.6	0	0.833	0.067	0.833	0	0.767	0.067	0.733
	GDINA	1	0	1	0	1	0	1	0	0.933	0	1	0	0.933	0
75	DINA	0	0.033	0	0.133	0	0.3	0	0.167	0	0.067	0	0.2	0	0.167
	DINO	0	0.567	0	0.767	0	0.7	0	0.7	0.033	0.833	0	0.733	0	0.833
	GDINA	1	0.4	1	0.1	1	0	1	0.133	0.967	0.1	1	0.067	1	0
100	DINA	0	0	0	0.067	0	0.133	0	0	0	0.033	0	0.033	0	0.267
	DINO	0	0.2	0	0.4	0	0.667	0	0.333	0	0.567	0	0.6	0	0.733
	GDINA	1	0.8	1	0.533	1	0.2	1	0.667	1	0.4	1	0.367	1	0
200	DINA	0	0	0	0	0	0	0	0	0	0	0	0	0	0.067
	DINO	0	0	0	0	0	0	0	0	0	0.033	0	0	0	0.333
	GDINA	1	1	1	1	1	1	1	1	1	0.967	1	1	1	0.6

Note: M = analytic model

Table A2. Selection Rates for Relative Indices with the Modified Q-Matrix

K=5, RSS		overQ				underQ				mixQ			
		5%		10%		5%		10%		5%		10%	
N	M	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
50	DINA	0.600	0.933	0.600	0.967	0.433	0.967	0.633	0.933	0.600	0.967	0.567	1
	DINO	0.033	0.033	0.033	0.033	0	0.033	0.033	0.067	0	0	0	0
	GDINA	0.367	0.033	0.367	0	0.567	0	0.400	0.067	0.400	0.033	0.433	0
75	DINA	0.133	0.833	0.267	0.867	0.200	0.900	0.167	0.900	0.167	0.800	0.100	0.900
	DINO	0	0.033	0	0	0	0	0	0	0	0	0	0
	GDINA	0.867	0.133	0.733	0.133	0.800	0.100	0.833	0.100	0.833	0.200	0.900	0.100
100	DINA	0.100	0.767	0.133	0.733	0.133	0.800	0.167	0.800	0.100	0.800	0.133	0.700
	DINO	0	0	0	0	0	0	0	0	0	0	0	0
	GDINA	0.900	0.233	0.867	0.267	0.867	0.200	0.833	0.200	0.900	0.200	0.867	0.300
200	DINA	0.033	0.267	0.033	0.300	0.067	0.333	0	0.333	0.033	0.200	0.033	0.200
	DINO	0.033	0.033	0.033	0.033	0.033	0.033	0	0	0.033	0.033	0	0
	GDINA	1	0.767	1	0.733	0.967	0.700	1	0.667	1	0.833	0.967	0.800

Table A3. Rejection Rates of the Absolute Indices Under Various Simulation Conditions

K=5		TQ		overQ				underQ				mixQ			
				5%		10%		5%		10%		5%		10%	
N	M	r	l	r	l	r	l	r	l	r	l	r	l	r	l
50	DINA	0.967	0.967	1	0.967	1	1	0.933	1	1	1	1	1	1	1
	DINO	0.800	0.900	0.967	1	1	1	0.867	0.967	0.967	1	1	1	1	1
	GDINA	0.033	0	0	0	0.033	0	0.667	0.800	0.967	1	1	1	1	1
75	DINA	0.967	1	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0	0.033	0	0.033	0	0	0.967	0.933	1	1	1	1	1	1
100	DINA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	0.967	1	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0	0.033	0	0	0	0	0.933	0.967	0.967	0.967	1	1	1	1
200	DINA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	0	0	0	0	0	0	1	1	1	1	1	1	1	1

Table A4. Rejection Rates for Absolute Indices with the Modified Q-Matrix

K=5, RSS		overQ				underQ				mixQ			
		5%		10%		5%		10%		5%		10%	
N	M	r	l	r	l	r	l	r	l	r	l	r	l
50	DINA	0.467	0.600	0.467	0.633	0.333	0.467	0.400	0.600	0.500	0.600	0.700	0.767
	DINO	0.733	0.867	0.767	0.900	0.500	0.700	0.533	0.700	0.733	0.900	0.733	0.833
	GDINA	0.233	0.367	0.300	0.533	0.300	0.400	0.267	0.467	0.400	0.500	0.367	0.467
75	DINA	0.533	0.633	0.733	0.767	0.400	0.667	0.367	0.533	0.633	0.800	0.667	0.800
	DINO	0.733	0.800	0.833	0.933	0.600	0.700	0.600	0.767	0.800	0.800	0.833	0.900
	GDINA	0.433	0.500	0.600	0.633	0.367	0.500	0.333	0.500	0.467	0.600	0.567	0.733
100	DINA	0.733	0.833	0.867	0.900	0.767	0.900	0.733	0.833	0.700	0.867	0.833	0.900
	DINO	0.833	0.867	0.900	0.900	0.833	0.900	0.833	0.900	0.833	0.867	0.867	0.900
	GDINA	0.700	0.767	0.800	0.833	0.700	0.833	0.767	0.900	0.767	0.867	0.767	0.867
200	DINA	1	1	1	1	1	1	1	1	1	1	1	1
	DINO	1	1	1	1	1	1	1	1	1	1	1	1
	GDINA	1	1	1	1	1	1	1	1	1	1	1	1