ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);391-403

The Impact of Q-matrix Misspecification and Model Misuse on Classification Accuracy in the Generalized DINA Model*

Miao GAO** M. David MILLER*** Ren LIU****

Abstract
This simulation study explored the impact of Q-matrix misspecification and model misuse on examinees' classification accuracy within the generalized deterministic input, noisy "and" gate (G-DINA) model framework under the different conditions. The data was generated by saturated G-DINA model. Along with the generating model, two reduced models were used to fit the data: the additive CDM (A-CDM) and DINA model. The manipulated conditions included number of respondents, attribute correlations and test length. Two types of classification accuracy were examined: the overall classification accuracy and the class-specific classification accuracy. Results showed that the Q-matrix misspecification influenced classification accuracy more ominously than model misuse. The proportion of examinees classified correctly for each latent class was related to the types of Q-matrix misspecification. More test items had greater positive impact on classification accuracy than more respondents taking the test.

Key Words: Classification, cognitive diagnostic assessment, the generalized DINA model, Q-matrix misspecification

INTRODUCTION

Researchers and educational stakeholders have increasingly demanded more formative test information (Mislevy, 2006; Robets & Gierl, 2010; Rupp & Templin, 2008). They often wish to obtain the classification of respondents with respect to their skills. Teachers, students and parents often want to know the individual's level of skill mastery to facilitate an individual's development. Cognitive diagnosis models (CDMs) are used to measure the respondents' knowledge structures and the multiple attributes for the purpose of making classification-based decisions (Rupp, Templin, & Henson, 2010).

Despite the diversity of parametric models, general DCMs have gained increasing attention in recent years because they do not have idiosyncratic hypotheses about the impact of attribute relationship among items. They subsumed many popular models that were developed earlier such as the deterministic inputs, noisy "and" gate (DINA; Junker & Sijtsma, 2001) models, the deterministic inputs, noisy "or" gate (DINO; Templin & Henson, 2006) models, and the reduced reparameterized unified (R-RUM; Hartz, 2002) model. The three most common general DCMs are the general diagnostic model (GDM; von Davier, 2008, 2010), the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), and the generalized DINA model (G-DINA; de la Torre, 2011). Among the three models, the G-DINA extends the logit link function of the other two models to multiple link functions including identity and log links.

One of the most important steps before specifying CDMs is to identify the attributes measured by each items. This item by attribute specification is usually constructed by the content experts and is called the Q-matrix. In practice, if the item-attribute alignment we specified a priori is not supported by the

* This study is derived from a part of Ph.D. dissertation by Miao GAO entitled Assessing the Model Fit and Classification Accuracy in Cognitive Diagnosis Models, submitted to University of Florida Graduate School under the supervision of Professor M. David MILLER.

** Assistant professor, Nanjing Normal University, College of Education, Nanjing, China, e-mail: miaogaonj@163.com, ORCID ID: 0000-0001-6842-5468

*** Professor, University of Florida, Research and Evaluation Methodology Program, Gainesville-Florida, US, e-mail: dmiller@coe.ufl.edu, ORCID ID: 0000-0002-3506-1167.

**** Ph.D candidate, University of Florida, Research and Evaluation Methodology Program, Gainesville-Florida, US, e-mail: liurenking@ufl.edu, ORCID ID: 0000-0002-6708-4996.

Received: 04.08.2017
DOI: 10.21031/epod.332712
Accepted: 06.10.2017

data, the Q-matrix may be misspecified. Previous research has shown that parameter estimates and classification accuracy were affected by the misspecification of Q-matrix (e.g., Rupp & Templin, 2008; Kunina-Habenicht, Rupp, & Wilhelm, 2012). Specifically, Rupp and Templin (2008) used the different types of Q-matrix misspecification under the DINA model. They found the Q-matrix misspecification had caused biased parameter estimates and lower classification accuracy corresponding to the examinees' latent class. However, questions such as whether the results may be generalizable to more general contexts. The purpose of this study is to estimate the effects of specific types of Q-misspecification on examinee classification accuracy under the generalized G-DINA model.

The rest of the manuscript is structured as follows: In the theoretical framework, we first provide an overview of the Q-matrix, types of Q-matrix misspecifications and the generalized DINA model. In the method, the simulation design, the model estimation and the outcome assessment are described. Next the findings of this study are described. Lastly this manuscript is closed with a discussion of the findings.

Background

Q-matrix

A critical step in cognitive diagnostic model is to develop the Q-matrix because CDM and the Q-matrix are essential modeling process. Developing the Q-matrix defines the attribute structure measured by an assessment. An example of a JxK Q-matrix can be demonstrated as follows:

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ & & \\ 1 & 1 & 1 \end{bmatrix} \tag{1}$$

Where j indicates "item" and k indicates "attributes." The element, qjk, is specified as "1" if the jth item requires the kth attribute to answer this item correctly; otherwise, qjk is specified as "0".

In a Q-matrix, each element qjk indicates whether the item j measures the attribute k, where qjk =1 means item j measures the attribute k and qjk =0 means item j does not measure the attribute k. The Q-matrix reflects the loading structure of the multiple attributes on the items. The Q-matrix is specified by content experts and this specification process is a subjective activity (Rupp, Templin, & Henson, 2010). Hence, the quality of the Q-matrix determines the diagnostic information obtained from the CDM analysis.

The Generalized DINA Model Framework

The generalized DINA model, like all other CDMs, requires a *J x K* Q-matrix. The G-DINA discriminates latent classes into $2^{K_j^*}$ latent groups, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ represents the required attributes for item j. Each latent group is reduced to an attribute vector represented by α_{lj}^*. In this study, it would suffice to use the reduced vector $\alpha_{lj}^* = (\alpha_{lj1}^*, \dots, \alpha_{ljK_j^*}^*)$ instead of the full vector $\alpha_{lj} = (\alpha_{lj1}, \dots, \alpha_{ljk})$. Each latent group has the probability of answering correctly the item represented by P(α_{lj}^*). The item response function (IRF) for G-DINA could be written as:

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \dots + \delta_{j12\dots K_k^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{2}$$

where δ_{j0} is the intercept, δ_{jk} is the main effect by α_k , $\delta_{jkk'}$ is the interaction effect by α_k and $\alpha_{k'}$, and $\delta_{j12\dots K_j^*}$is the interaction effect by α_1 ,…, α_{k^*}.

The DINA model, that is the most commonly used reduced model, is a special case of the G-DINA model. By setting all the parameters, except δ_{j0} and $\delta_{j12\ldots K_j^*}$, to zero, the IRF for DINA model is as follows:

$$P(\alpha_{lj}^*) = \delta_{j0} + \delta_{j12\ldots K_k^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{3}$$

Another special case of the G-DINA model is the A-CDM, which contains only the intercept and the main effects. The IRF for A-CDM is defined as follows:

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_{lk} \tag{4}$$

This model contains only the intercept and the main effect of each attribute.

METHOD

Simulation Study

The simulation study was aimed to examine the effects of Q-matrix misspecification and CDM misuse on classification accuracy. All data generation and estimations were conducted using the software R (R Core team, 2016). The data was generated using the saturated model G-DINA. The number of respondents, the correlation between attributes, and the number of items measured in a test were manipulated and resulted in 12 data-generating conditions with 1000 replications for each condition. For each of the generated datasets, three CDMs within the G-DINA framework were applied for the data analysis: the G-DINA, A-CDM and DINA models. Six Q-matrices, including 1 correctly specified Q-matrix and 5 misspecified Q-matrices were examined. In total, there were 216 different settings for data analysis, which included 18 diverse estimations and 12 different data-generating conditions.

Number of respondents

Three levels of number of respondents reflecting small, moderate and large samples were investigated in this study: N = 500, 1000 and 5000. Previous research has shown this is a relevant factor that influences model fit, parameter estimates, and classification (Chen, de la Torre, & Zhang, 2013; Cui, Gierl, & Chang, 2012; de la Torre, 2009; de la Torre & Douglas, 2004; Shu, Henson, & Willse, 2013). Several studies have shown that number of respondents should be at least 500 in order to have an acceptable model fit and relatively accurate parameter estimates even when using the reduced model as the generated model (Chen et al., 2013; Cui et al., 2012; Shu et al., 2013). The pilot study indicated that when the sample size increased to 500, the model fit achieved an acceptable level.

Number of attributes

This study focused on one level of the number of attributes K =4. A review of the CDM simulation studies indicates that there are usually three to eight attributes being designed in an assessment, which also reflects the number of attributes in application examples (Cheng, 2009; Chen et al., 2013; DeCarlo, 2012; de la Torre, 2009; de la Torre & Douglas, 2004; Huebner & Wang, 2011; Kunina-Habenicht et al., 2012). Considering all the other factors being manipulated in the simulation and a fairly large estimation process, the attributes' number was fixed at four in this study.

Marginal attribute difficulty

A multivariate normal distribution for latent attributes with the mean vector and correlation matrix were used to generate respondents' true attribute patterns. In this study, the mean vector of(0, 0, 0, 0) was used for the four attributes test; this led to the same marginal mastery proportions for all attributes of .50. This mean vector is also called marginal attribute difficulty.

Correlation between attributes

Two levels of attribute correlation were set to values of .4 and .8 to represent moderate and high correlation(Henson, Roussos, Douglas & He, 2008), respectively. A range of .3 to .9 of the tetrachoric correlation is typical in educational assessment and CDM research (Cui et al., 2012; Henson, Templin & Douglas, 2007; Kunina-Habenicht et al., 2012). A weakly correlated attributes level could be included as a contrast, but I chose not to do this to keep the overall simulation and estimation manageable. The correlations were set to be equal across all attribute pairs in the correlation matrix.

Q-matrix specification

The number of items in a test was set to two levels in this study: J =14 and 28. The number of items and the number of attributes measured in a test are associated. For K =4, the number of all possible attribute patterns was 2^4=16, and there are 15 attribute patterns. Considering the computational time, we set the maximum number of attributes being assessed by an item to three. The item 1-14 in Table 1 showed the Q-matrix specification for generation when J=14. This simulation design also investigated the conditions where the test length is equal to and greater than the number of possible attribute patterns. Two levels of the item number were examined in this study: J = 14 and 28 for the number of attributes K = 4. The Q-matrix for J=28 was a duplicate of Q-matrix for J-14. The correctly identified Q-matrix for J = 28 is also shown in Table 1. The Q-matrix for J = 14 was embedded as a subset of this Q-matrix.

Table 1. Correct Q-Matrix of J = 14 and 28

	Attribute					Attribute			
Item	#1	#2	#3	#4	Item	#1	#2	#3	#4
1	1	0	0	0	15	1	0	0	0
2	0	1	0	0	16	0	1	0	0
3	0	0	1	0	17	0	0	1	0
4	0	0	0	1	18	0	0	0	1
5	1	1	0	0	19	1	1	0	0
6	1	0	1	0	20	1	0	1	0
7	1	0	0	1	21	1	0	0	1
8	0	1	1	0	22	0	1	1	0
9	0	1	0	1	23	0	1	0	1
10	0	0	1	1	24	0	0	1	1
11	1	1	1	0	25	1	1	1	0
12	1	1	0	1	26	1	1	0	1
13	1	0	1	1	27	1	0	1	1
14	0	1	1	1	28	0	1	1	1

Note. Items 1-14 are used when J = 14.

Different types of the Q-matrix misspecification were investigated: under-fitting the Q-matrix (defining 1 as 0), over-fitting the Q-matrix (defining 0 as 1), and a balanced misfit (exchanging 0 and 1). As shown in Table 2, taking the test with J=14 items as an example, *qt-14* was the true Q-matrix for data generation. Two under-specified Q-matrices *qu3-14* and *qu2-14* meant that qu3-14 Q-matrix changed all 3-attribute items into selected 2-attribute items, and this selection of the attribute deletion was random for each item; *qu2-14* Q-matrix changed all 2-attribute items into selected 1-attribute items, and this selection of the attribute deletion was random for each item. Similarly, two over-specifications *qo1-14* and *qo2-14* Q-matrices were created by randomly selecting the attribute being added. For creating the balanced misfit for the Q-matrix (*qm-14*), the items that needed to be altered were first randomly selected; then, the attributes that needed to be altered were selected randomly for each item.

Table 2. The Q-Matrix Misspecification and True Q-Matrix

K	J	Q-matrix	Alternations	Item Altered	Total # of changes (1 to 0)	Total # of changes (0 to 1)	Ave. # of attributes per item	Ave. # of items per attribute
4	14	*qt-14*	Data generating Q-matrix	0	0	0	2	7
		qu3-14	All 3-attribute items are changed into selected 2-attribute items.	I11 - I14	4	0	1.71	6
		qu2-14	All 2-attribute items are changed into selected 1-attribute items.	I5 - I10	6	0	1.57	5.5
		qo1-14	All 1-attribute items are changed into selected 2-attribute items.	I1-4	0	4	2.29	8
		qo2-14	All 2-attribute items are changed into selected 3-attribute items.	I5 - I10	0	6	2.43	8.5
		qm-14	Attributes are deleted and added to balance out the overall number of changes.	2 items randomly selected from I1-I4; 3 items randomly selected from I5-I10; 2 items randomly selected from I11-I14	7	7	2	7
4	28	*qt-28*	Data generating Q-matrix	0	0	0	2	14
		qu3-28	Half of the 3-attribute items are changed into selected 2-attribute items.	I11 - I14	4	0	1.86	13
		qu2-28	Half of the 2-attribute items are changed into selected 1-attribute items.	I5 - I10	6	0	1.79	12.5
		qo1-28	Half of the 1-attribute items are changed into selected 2-attribute items.	I1-4	0	4	2.14	15
		qo2-28	Half of the 2-attribute items are changed into selected 3-attribute items.	I5 - I10	0	6	2.21	15.5
		qm-28	Attributes are deleted and added to balance out the overall number of changes.	2 items randomly selected from I1-I4; 3 items randomly selected from I5-I10; 2 items randomly selected from I11-I14	7	7	2	14

The assessment with number of items J = 28 had doubled the items as in the assessment J =14. The misspecification of Q-matrix in J =28 only occurred in items 1 to 14, and items 15 to 28 always remained the same as in true Q-matrix (qt-28). In this way, the number of misspecified items in J = 28 was the same as in J =14 when controlling the type of misspecification, which made the results comparable for different test length.

Item parameter specification for data generation

The parameter setting was referenced from an empirical study (Basokcu, Ogretmen, &Kelecioglus, 2013). The true item parameters (δ_{jk}) used in this simulation study were ranged from 0.12 to 0.68, and the detailed values were presented in Table 3. For simplicity, all the one-attribute items used the same parameter setting, and the same idea was followed for the two- and three-attribute items.

Table 3. Item Parameters for Data Generation (d_{jk})

	Attribute Pattern and Parameters							
1-attribute item	0	1						
	.21	.68						
2-attribute item	00	10	01	11				
	.18	.25	.15	.59				
3-attribute item	000	100	010	001	110	101	011	111
	.26	.12	.17	.18	.13	.27	.26	.51

Model selection

Each of the generated datasets was analyzed by three CDMs within the G-DINA framework. The true generating model was the G-DINA model. In addition to the true model, two misused CDMs were used to analyze the data. misusage of CDM refers to incorrect parameterization of the modeling process. As two comparison models, A-CDM contained only intercept and main effects for each item; and the DINA model contained only intercept and the highest order of interaction effect for each item.

Outcome Measures

Classification accuracy (CA) is defined as the degree to which the classification of examinees' latent classes analyzed by observed data agrees with examinees' true latent classes (Cui et al., 2012). The simulated examinee attribute patterns were used as the true examinees' latent classes; the attribute patterns estimated from the response data using MLE method were used as the estimated latent classes. The simulated and estimated latent class were then compared for each examinee. If they were consistent, a value of "1" was assigned to the examinee to represent being classified accurately; otherwise, a value of "0" was assigned for being classified inaccurately. By taking the average of 0/1 over all examinees and all replications, the overall correct classification rates were calculated for each condition, which refers to overall classification accuracy (OCA). By taking the average of 0/1 for the examinees by each latent class, the class-specific correct classification rates were calculated, which refers to class-specific classification accuracy (CCA). In order to simplify the interpretation of the findings, the CCA was calculated based on one generating condition (n = 5000, ρ = .4 and J = 14) and being fitted with the various CDMs and Q-matrices. The OCA and CCA were then compared for all the estimation settings.

RESULTS

In CDM estimations, the classification is usually of primary interest because the decisions about the examinees are made based on the classification (Rupp, Templin, & Henson, 2010). Two types of the classification accuracy were illustrated in this part: the OCA and CCA.

Overall Classification Accuracy (OCA)

Table 4 demonstrated the overall correct classification rates by the different levels of all factors. For the purpose of explaining the results more explicitly, the effects of N, J and on the OCA were the focus in Table 4. The impact of CDM misuse and Q-matrix misspecification on OCA is examined in Figure 1.

Table 4. Overall Classification Accuracy (OCA) in All Conditions

				Q-matrix Specification					
Model	*ρ*	*J*	*N*	*qt*	*qu3*	*qu2*	*qo1*	*qo2*	*qm*
G-DINA	0.4	14	500	0.711	0.689	0.525	0.706	0.695	0.467
			1000	0.719	0.698	0.533	0.717	0.713	0.477
			5000	0.727	0.705	0.542	0.726	0.725	0.481
		28	500	0.886	0.879	0.837	0.885	0.883	0.815
			1000	0.889	0.884	0.843	0.889	0.888	0.826
			5000	0.893	0.888	0.850	0.893	0.893	0.834
	0.8	14	500	0.720	0.731	0.647	0.714	0.709	0.629
			1000	0.723	0.734	0.651	0.720	0.720	0.639
			5000	0.726	0.733	0.650	0.725	0.724	0.644
		28	500	0.873	0.876	0.846	0.871	0.870	0.824
			1000	0.875	0.879	0.852	0.874	0.873	0.831
			5000	0.877	0.881	0.858	0.877	0.876	0.837
A-CDM	0.4	14	500	0.669	0.645	0.536	0.657	0.641	0.460
			1000	0.675	0.652	0.540	0.666	0.647	0.462
			5000	0.689	0.653	0.543	0.678	0.654	0.453
		28	500	0.838	0.833	0.810	0.833	0.828	0.776
			1000	0.850	0.846	0.816	0.847	0.845	0.792
			5000	0.859	0.851	0.820	0.860	0.859	0.800
	0.8	14	500	0.738	0.726	0.641	0.730	0.717	0.617
			1000	0.750	0.736	0.642	0.751	0.743	0.619
			5000	0.755	0.742	0.643	0.760	0.758	0.609
		28	500	0.887	0.868	0.845	0.887	0.886	0.827
			1000	0.888	0.868	0.846	0.889	0.888	0.831
			5000	0.889	0.870	0.848	0.890	0.889	0.835
DINA	0.4	14	500	0.643	0.648	0.512	0.448	0.526	0.374
			1000	0.647	0.652	0.513	0.454	0.530	0.375
			5000	0.648	0.652	0.515	0.458	0.532	0.379
		28	500	0.851	0.855	0.767	0.736	0.773	0.737
			1000	0.858	0.861	0.771	0.740	0.785	0.744
			5000	0.865	0.867	0.781	0.746	0.794	0.753
	0.8	14	500	0.673	0.664	0.652	0.505	0.613	0.485
			1000	0.678	0.666	0.653	0.509	0.613	0.488
			5000	0.682	0.668	0.653	0.511	0.613	0.490
		28	500	0.870	0.872	0.828	0.712	0.811	0.750
			1000	0.878	0.879	0.832	0.715	0.821	0.753
			5000	0.882	0.883	0.836	0.717	0.825	0.759

As shown in Table 4, when test length increased, the correct overall classification rates were much higher. For example, in G-DINA model with qt matrix, the correct overall classification rates went up from .711 to .886 as test length increased from J=14 to 28,controlling ρ= .4 and N = 500. This is expected because more pieces of information provided by the items for each dimension can be used to detect the classification. Second, as the sample size increased, the overall classification rates slightly increased for all conditions. For example, again in G-DINA model with qt matrix, the overall classification accuracy increased from .886 to .893 as sample size increased from 500 to

5000,controlling ρ= .4 and J=28. Comparing the effects of J and N on classification accuracy, we can see that more items in a test are more critical than more examinees to get a better classification accuracy. Third, the increase in attribute correlation slightly increased the overall classification accuracy with few exceptional conditions.

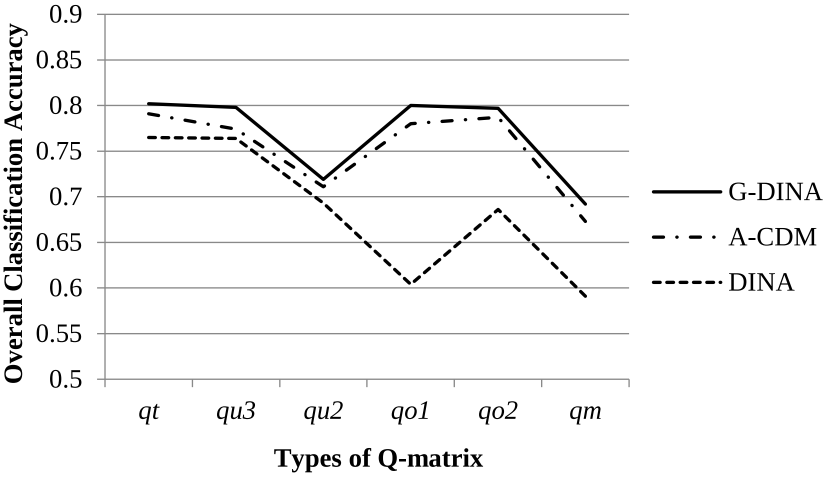

Figure 1. Overall Classification Accuracy (OCA) by CDM and Q-matrices

To investigate the effects of the misspecification of CDM and Q-matrix, the correct overall classification rates were shown in Figure 1. The classification rates used in this figure were collapsed over the other factors N, J and for the simpler illustration.

For CDM misuse, Figure 1 showed that the overall classification accuracy was highest in G-DINA no matter which specified Q-matrix was used. This makes sense because G-DINA was the generating model. Comparing the other two CDMs, A-CDM has higher classification rates than DINA. The A-CDM yielded very similar overall classification rates with the true model G-DINA where A-CDM contained only main effects of the attributes and omitted all the interactions. The DINA model showed the lowest classification rates among three CDMs where DINA contained only the highest order of interactions among attributes.

For investigating Q-matrix misspecification, the condition qt was the correct Q-matrix and could be used as baseline rates in each CDM. Figure 1 showed that the OCA in qt was higher than the other misspecified Q-matrices in three CDMs. The effects of the misspecified Q-matrices on classification accuracy were then compared with the true Q-matrix in different CDMs. The classification rates in G-DINA and A-CDM showed similar patterns for the Q-matrix misspecification. Within these two models, the OCA for the condition qu3, qo1 and qo2 was close to the rates in qt. The misspecified qu2 had lower overall classification rates, and the misspecified qm showed the lowest overall classification rates. This is not surprising because the qm included all types of misspecification. To compare the effects of different Q-matrices in the DINA model, the OCA was highest in qt; the condition qu3 yielded almost the same results with qt; while the lowest classification still occurred in qm among all the conditions. The Q-matrices qo2 and qu2 in DINA model yielded the moderate classification rates.

Table 5. Correct Overall and Class-specific Classification Rates for Misspecifications of CDM and Q-matrix

Model	Q-matrix	Overall	Attribute Classes															
			0000	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1101	1011	0111	1111
G-DINA	qt	.802	.587	.653	.666	.666	.673	.810	.816	.830	.822	.836	.847	.922	.934	.934	.940	1
	qu3	.798	.688	.637	.609	.609	.578	.756	.728	.726	.758	.756	.815	.840	.890	.890	.912	1
	qu2	.719	.792	.532	.568	.568	.305	.502	.745	.500	.532	.406	.492	.849	.807	.807	.672	.944
	qo1	.800	.586	.648	.661	.661	.668	.806	.812	.827	.818	.832	.843	.923	.933	.933	.939	1
	qo2	.797	.592	.631	.648	.648	.654	.792	.801	.816	.804	.817	.829	.921	.932	.932	.936	1
	qm	.692	.741	.290	.353	.353	.526	.438	.414	.556	.518	.471	.624	.673	.557	.557	.574	.994
A-CDM	qt	.791	.647	.628	.623	.623	.599	.731	.753	.759	.767	.771	.776	.864	.872	.872	.877	.995
	qu3	.774	.740	.612	.592	.592	.541	.636	.652	.637	.688	.680	.782	.712	.809	.809	.830	.991
	qu2	.711	.781	.545	.593	.593	.307	.488	.728	.485	.524	.375	.481	.826	.800	.800	.664	.936
	qo1	.780	.612	.608	.603	.603	.583	.723	.733	.746	.759	.764	.764	.869	.879	.879	.882	.994
	qo2	.787	.638	.608	.611	.611	.589	.726	.747	.754	.753	.762	.768	.867	.877	.877	.880	.997
	qm	.673	.735	.334	.362	.362	.481	.390	.355	.549	.434	.429	.516	.654	.553	.553	.531	.994
DINA	qt	.765	.597	.551	.593	.593	.608	.684	.690	.698	.710	.715	.733	.922	.908	.908	.925	1
	qu3	.764	.601	.505	.547	.547	.571	.715	.705	.727	.765	.749	.766	.887	.906	.906	.925	1
	qu2	.693	.735	.449	.482	.482	.253	.506	.705	.462	.463	.330	.384	.843	.726	.726	.675	1
	qo1	.604	.246	.242	.240	.240	.286	.664	.629	.668	.548	.724	.627	.892	.870	.870	.848	1
	qo2	.686	.555	.439	.455	.455	.467	.568	.557	.570	.579	.579	.573	.864	.820	.820	.610	1
	qm	.591	.396	.221	.376	.376	.419	.384	.433	.492	.462	.445	.626	.725	.517	.517	.635	1

Class-specific Classification Accuracy (CCA)

When we examined the respondents' classification at each latent class level, it was worthwhile to note that classes with more attributes tended to have generally higher classification accuracy in various CDMs and Q-matrices (Table 5). For example, in the G-DINA and qt condition, the CCA ranged from 0.587 to 1 for the class with no attribute to the class with all attributes. The attribute class in which all attributes were mastered (attribute pattern 1111) maintained very high correct classification rates no matter which CDM and Q-matrix were used. Especially in the DINA model, the misclassification of examinees in this attribute class never occurred.

Comparing the different CDMs, the G-DINA model yielded the highest CCA in almost all the latent classes with few exceptions. When using qt in G-DINA model, the correct classification rates for one-attribute mastery classes were at least 65%; and these rates reached at least 80% and 90% for two- and three-attribute mastery classes, respectively. A-CDM performs better than DINA in the classes with zero-, one- and two-attributes. The CCA by using qt and A-CDM were approximately .6 for one-attribute mastery classes, .75 for two-attribute mastery classes, .87 for three-attribute mastery classes.

However, the DINA model had higher than expected classification accuracy in three- and four-attribute mastery classes, even with misspecified Q-matrices. More specifically, focusing on the three-attribute mastery classes, the CCA of the DINA model using qt were .922, .908, .908 and .925, while the G-DINA model using qt has almost the same classification accuracy. In qu2, qo1 and qo2, the CCA of the DINA model was slightly lower than G-DINA and higher than A-CDM in three-attribute classes' estimations. In qu3, DINA even performed best among three CDMs in the classification accuracy of three-attribute latent classes (.887, .906, .906 and .925).

Considering the Q-matrix misspecification, the class-specific classification rates are related to the different types of misspecified Q-matrices (under-, over- or mixed misspecification). G-DINA and A-CDM showed a similar pattern: The over-specified Q-matrices (qo1 and qo2) did not have much impact on the class-specific classification accuracy. The under-specified Q-matrices, especially qu2, had much lower CCA in these two models. While in the DINA model, the misspecified qu2 and qo2 seemed to have a more severe impact on CCA; the qo1 mainly affected the correct classification rates on the classes with fewer attributes. The misspecified qm, for all three fitting models, showed the lowest classification rates, and the low class-specific classification rates occurred in almost all attribute classes.

Furthermore, we noticed that the low class-specific classification rates corresponded to the attribute patterns that matched the manipulated attribute classes. For example, in the misspecified qu2 where two-attribute items were changed into one-attribute items, the correct classification rates of two-attribute mastery classes (e.g. attribute class [1100]) dropped a great deal when compared with qt condition. The correct classification rates of one-attribute mastery classes (e.g. attribute class [0001]) decreased as well in all three CDMs. Unlike G-DINA and A-CDM, in the condition qo1 where the one-attribute items were changed to two-attribute items, the classification rates for having one attribute in DINA were very low which matched the manipulated items. In the condition qo2, the CCA of two-attribute mastery classes were low as well in the DINA model.

DISCUSSION and CONCLUSION

The G-DINA model offers a flexible framework to investigate the issues in examinees' diagnostic classification. The specification of Q-matrix and the choice of CDM play a critical role for achieving better classification accuracy. This study helps to understand better of the effects of CDM misuse and Q-matrix misspecification on classification accuracy under various conditions. The different factors, such as number of test items, number of examinees and attribute correlation, all have certain impacts on examinees' classification. The outcome of CDMs provides meaningful formative test information about the multiple proficiencies of the attributes measured in each examinee. Although this study is sufficiently complex, it clearly can be extended by using a broader range of design.

This simulation study contributed in the following four aspects. First, the G-DINA model was used as a framework that aligned with the trend in CDM development. The simulation was conducted in the saturated model and fit the data with two reduced models as well as the saturated model, which better aligns to the practice of real data analysis. Second, both the Q-matrix misspecification and CDM misuse were investigated separately and conjunctively. Third, the under-, over- and mixed misspecified Q-matrices allow us to detect the more specific effects of Q-matrix misspecification under various conditions in a generalized CDM framework. Fourth, the overall classification accuracy and the class-specified classification rates (often the primary interest in CDM analysis) were investigated under different conditions in this study.

Both the number of respondents and test length illustrated clear positive effects on classification accuracy. Despite the model selection and Q-matrix specification, the increase of the number of respondents and/or the test items always demonstrated the growth in the correct classification rates. One noticeable finding is that the increase in test length improved the classification accuracy more dramatically than the increase in sample size. It provides an insightful direction to the practitioners, to assist in making the decision of which factors will be manipulated, in order to effectively improve the examinees' classification accuracy.

Our results also demonstrated that model misuse does not noticeably affect the overall classification accuracy, even though the G-DINA model still maintained the highest level of classification accuracy. We simulated data in the saturated G-DINA model by mimicking the complex empirical situation. When estimating the data with various CDMs, we found the models performed differently under the consideration of examinees' latent classes. For the examinees who have fewer attributes (e.g. one- or two-attribute), G-DINA and A-CDM models yield more accurate classification rates than the DINA model. A-CDM showed a better classification accuracy in the non-attribute mastery class. This may due to the structure of G-DINA and A-CDM models that they contains the main effects. For the examinees who have more attributes (e.g. three- or four-attribute), the DINA model that contained only the highest order of interaction had higher than expected classification accuracy even with the Q-matrix misspecification. Given these, although A-CDM is easier to interpret in practice, if we have large number of attributes, it may be worth considering having higher order interaction effects.

One important finding in this study is that the misspecification of Q-matrix affected the overall classification accuracy in a more obvious way than model misuse. In practical application, the true Q-matrix is unknown and there is a possibility that Q-matrix could be misspecified in the designing process. As expected, the true Q-matrix yielded the most accurate classification. In general, the under-misspecified Q-matrices had more severe impact on CA than over-misspecified Q-matrices especially in the models with main effects. The misspecified Q-matrix qm was most problematic because the correct classification rates were low in almost all the conditions. Although the number of attributes held constant in qm, a large number of misspecification occurred. The qm contained all types of the misspecification and represented the most severe misspecification. Thus, it is not only the number of misspecified items that matters but also the types of misspecification. The attribute structure, rather than the number of attribute by item, is a much more important component in the diagnosis process. In practice, we may face a situation where there is an uncertainty in determining whether one item measures the attribute. We suggest that over-specification may be better than under-specification.

Besides the effect on overall classification rates, the different types of misspecified Q-matrices also showed the effects on the corresponding latent class. When a certain attribute combination is not represented in the Q-matrix, the respondents mastering the same attribute combination are more likely to be misspecified. A typical example in all three CDMs is the misspecified qu2, where two-attribute items were changed into one-attribute items. The classification rates decreased noticeably in the corresponding two-attribute mastery classes in all three CDMs. Thus inferences for the examinees in the associated classes should be more cautious.

Moreover, the effects of differently specified Q-matrices on classification accuracy varied in three CDMs. For example, the over-specified Q-matrix (qo1and qo2) influenced the DINA model more severely, but not in the G-DINA and A-CDM. The balanced misfit Q-matrix qm had shown more dramatic negative effect on the classification rates in DINA than the other two models. This may be due to the different features of three CDMs. The saturated G-DINA model contains the main effects and all the ways of interactions, the A-CDM contains the main effects only, and the DINA includes only the highest order of interaction. In sum, the G-DINA model had a more stable performance in all latent classes when considering Q-matrix misspecification, although A-CDM performed well in zero- and one-attribute mastery classes and DINA showed high classification accuracy in three- and four-attribute mastery classes.

Regardless of the different types of CDMs and Q-matrices, it was noteworthy that the examinees in the latent classes with more attributes had higher classification accuracy, and the examinees in the latent classes with fewer attributes could not be classified accurately. This becomes considerable in practice when applying these CDMs to identify the mastery and non-mastery of multiple attributes, especially for the examinees at the lower end. The attribute class mastering all attributes almost never showed any misspecification rates; while the attribute class with no attributes had low correct classification rates. For addressing the possible reasons of this phenomenon, future research may examine the impact of item difficulty and the distribution of attribute patterns.

In practice, the importance of diagnostic test development framework and Q-matrix validation methods should be emphasized. After the Q-matrix is designed, we recommend validating the Q-matrix using the method proposed in de la Torre (2008) and de la Torre and Chiu (2016) to check the possibility of misspecification. Yet it is not easy to evaluate the correctness of the Q-matrix due to its subjective nature and the complexity when applied to the model. When there is an uncertainly in determining if one item measures the attribute, over-specification may be better than under-specification. In order to classify the examinees into latent groups, the selection of the CDMs may relate to which group of examinees are more concerned with. The saturated model usually yields more stable classification accuracy across all the latent classes. The model with higher order interactions should be considered when there are a number of attributes, although the model with only main effects is easier to interpret. Hopefully the findings of this study will provide some insights for practitioners and researchers in determining the Q-matrix and cognitive diagnostic models when facing various situations.

REFERENCES

Basokcu, T. O., Ogretmen, T., & Kelecioglu, H. (2013). Model data fit comparison between DINA and G-DINA in cognitive diagnostic models. *Education Journal, 2*(6), 256-262.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123-14.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619-632.

Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19-38.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447-468.

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. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115-13.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

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 at Urbana-Champaign, Urbana-Champaign, IL.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275-288.

Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement*, *44*(4), 361-376.

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*(2), 191-210. Doi: 10.1007/S11336-008-9089-5

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*(2), 407-419.

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.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-305). Washington, DC: American Council on Education.

R Core Team (2016). R (Version 3.3) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Robets, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29*(3), 25-38.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

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

Shu, Z., Henson, R., & Willse, J. (2013). Using neural network analysis to define methods of DINA model estimation for small sample sizes. *Journal of Classification, 30*(2), 173-194.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*(3), 287-305. Doi: 10.1037/1082-989X.11.3.287

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychology Science Quarterly*, *52*(1), 8-28.