JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Summer 2013, Vol. 50, No. 2, pp. 123–140

Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling

Jinsong Chen
Sun Yat-Sen University
Jimmy de la Torre and Zao Zhang
Rutgers, The State University of New Jersey

As with any psychometric models, the validity of inferences from cognitive diagnosis models (CDMs) determines the extent to which these models can be useful. For inferences from CDMs to be valid, it is crucial that the fit of the model to the data is ascertained. Based on a simulation study, this study investigated the sensitivity of various fit statistics for absolute or relative fit under different CDM settings. The investigation covered various types of model–data misfit that can occur with the misspecifications of the Q-matrix, the CDM, or both. Six fit statistics were considered: –2 log likelihood (–2LL), Akaike's information criterion (AIC), Bayesian information criterion (BIC), and residuals based on the proportion correct of individual items (p), the correlations (r), and the log-odds ratio of item pairs (l). An empirical example involving real data was used to illustrate how the different fit statistics can be employed in conjunction with each other to identify different types of misspecifications. With these statistics and the saturated model serving as the basis, relative and absolute fit evaluation can be integrated to detect misspecification efficiently.

Cognitive diagnosis models (CDMs) are psychometric models developed primarily for assessing examinees' mastery and nonmastery of skills or attributes. In cognitive diagnostic assessment, CDMs are often used together with the Q-matrix (Tatsuoka, 1983) to provide diagnostic information about examinees. Such information can inform student learning and aid in the design of better instruction. More often than not, cognitive diagnosis modeling is used to refer only to the psychometric component of the process. However, a more holistic perspective should include the Q-matrix as a part of the cognitive diagnosis modeling process (e.g., de la Torre, 2008). The role of the Q-matrix as an integral part of the modeling process becomes even more critical when the validity of the inferences is of concern. This is so because both the CDM and Q-matrix can potentially contribute to the model–data misfit under the diagnostic modeling context. As such, for the purposes of this article, the modeling process is construed to involve both the CDM and Q-matrix.

Recently, different types of CDMs that can potentially be applied across a wide range of settings have been developed. Some of these CDMs are highly constrained models like the *deterministic inputs, noisy "and" gate* (DINA; Junker & Sijtsma, 2001) model and the *deterministic inputs, noisy "or" gate* (DINO; Templin & Henson, 2006) model; some are of additive nature like the additive CDM (*A*-CDM; de la Torre, 2011), the linear logistic model (LLM; Maris, 1999), and the reduced reparameterized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002); other CDMs might have a saturated formulation like the log-linear CDM

(Henson, Templin, & Willse, 2009) and the generalized DINA (G-DINA; de la Torre, 2011).

For inferences from different CDMs to be valid, it is important that the fit of the model to the data is ascertained (i.e., absolute fit evaluation). With the availability of various CDMs, choosing the most appropriate model for a particular application (i.e., relative fit evaluation) also is important. Under the CDM context, various fit statistics or methods have been developed or used. Some of these statistics include: those based on the residuals between the observed and predicted correlations and log-odds ratios of item pairs and the residuals between the observed and predicted proportion correct of individual items (de la Torre & Douglas, 2008; Sinharay & Almond, 2007); item discrimination indices (de la Torre, 2008; de la Torre & Chiu, 2010); χ^2 and G statistics based on the observed and predicted item-pair responses (Rupp, Templin, & Henson, 2010, pp. 276–277); and the mean absolute differences (Henson et al., 2009) between the observed and predicted item conditional probabilities of success and related root mean square error of approximation (Kunina-Habenicht, Rupp, & Wilhelm, 2012). Conventional fit statistics like the Akaike's information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarzer, 1976), the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), and the Bayes factor (Kass & Raftery, 1995) were adopted empirically for relative fit evaluation (e.g., de la Torre & Douglas, 2004, 2008; DeCarlo, 2011; Rupp et al., 2010, p. 278; Sinharay & Almond, 2007).

Only a handful of studies have been conducted to systematically evaluate the extent to which these statistics are sensitive to model–data misfit or useful for model selection. Kunina-Habenicht et al. (2012) found that AIC and BIC were useful in selecting the correctly specified Q-matrix against the misspecified Q-matrices when about 30% of the entries had been randomly permuted or the number of attributes was severely overspecified (from three to five) or underspecified (from five to three) under the log-linear CDM. In addition, they also found that the AIC was useful in selecting the correct model against the misspecified model when all interaction effects were omitted. However, the usefulness of the mean absolute differences and root mean square error of approximation were limited for absolute fit evaluation at the test level. Furthermore, most of these statistics or methods are primarily useful for relative fit evaluation. Because a z-score or p-value for these procedures generally is not available, absolute fit evaluation at fixed significance levels is difficult to undertake.

To address some of these concerns, this study investigated the usefulness or sensitivity of various fit statistics for absolute or relative model–data fit under different CDM settings. Diverse situations of misfit were covered, including the misspecifications of the CDM, the Q-matrix, and both. The investigation was carried out using a simulation study involving different factors. An empirical example with real data was used to illustrate how different fit statistics can be employed in concert to detect different model misspecifications. To simplify model presentation and provide an overarching framework, all models were formulated within the G-DINA model framework (de la Torre, 2011), which has been shown to subsume several specific models. However, because the G-DINA model in its saturated form is equivalent to other general models (e.g., Henson et al., 2009; von Davier, 2005) in their saturated form, identical results can be expected regardless of the general model involved.

Background

Q-Matrix or CDM Misspecification

When model–data misfit occurs in cognitive diagnosis modeling, the source of misfit could come from the nature of the attributes, the attribute structure, the Q-matrix, or the CDM. It would be difficult for one study to investigate every cause of model–data misfit. To narrow the focus of our study, we chose to examine two causes of misfit: Q-matrix and CDM misspecifications. The reasons for doing so are as follows: (1) as mentioned earlier, the Q-matrix and CDM are integral parts of the modeling process; (2) both misspecifications can be readily investigated compared to other sources of misfit such as misspecification of the attribute nature; and (3) both misspecifications can severely affect parameter estimation quality and classification accuracy and can even interact to cause deterioration in the estimation process.

One crucial step in developing cognitive diagnostic assessment is the incorporation of substantive knowledge by specifying the Q-matrix. Let q_{jk} denote the element in row j and column k of a $J \times K$ Q-matrix, where J and K represent the number of items and attributes, respectively. The entry, q_{jk}, is specified as 1 if mastery of attribute k is required to answer item j correctly and as 0 otherwise. In most if not all CDM applications, the process of establishing the Q-matrix through substantive knowledge tends to be subjective in nature and has raised serious validity concerns among researchers (de la Torre, 2008; Rupp & Templin, 2008). For this study, when Q-matrix misspecification is said to have occurred, it could mean that the attribute specified for an item has been under-specified (i.e., some 1s have been incorrectly specified as 0s), over-specified (i.e., some 0s have been incorrectly specified as 1s), or both under- and over-specified (i.e., some 1s have been incorrectly specified as 0s and some 0s have been incorrectly specified as 1s). In cognitive diagnosis modeling, the Q-matrix plays an important role of constraining the number of item parameters to be estimated. Such constraints can interact with different types of CDMs. For instance, whereas there are always two parameters per item for the DINA model, there are $K_j^* + 1$ parameters per item for the *A*-CDM, with K_j^* being the number of required attributes for item *j*. Consequently, different types of Q-matrix misspecifications intertwine with different types of CDMs to confound the source of the misfit.

CDM misspecification in this study refers to incorrect parameterization of the psychometric component of the modeling process. In choosing a CDM parameterization, researchers formalize their conceptualization of the hypothesized cognitive processes involved in answering test items. Given specific Q-matrix and attribute structures, many possible parameterizations that are different in nature can be found (e.g., DINA, DINO, *A*-CDM). Accordingly, the potential to incorrectly specify the CDM and the consequence of doing so cannot be ignored. Theoretically, saturated CDMs always fit the data better than any reduced CDMs because of their more complex parameterization. However, it is not clear that saturated models are always to be preferred. One reason is that saturated CDMs require larger sample sizes to be estimated precisely. Another reason is that reduced CDMs are simpler and easier to interpret, assuming that their fit is not substantially inferior. Model selection that involves the saturated and reduced CDMs can be addressed within the scope of CDM

misspecification by using fit statistics that can compensate for model parameter complexity (e.g., AIC, BIC).

Model Fit Evaluation

Two types of fit evaluation are considered in this article: relative and absolute. Relative fit evaluation refers to the process of selecting the best-fitting model among a set of competing models. Fit statistics that are useful for identifying misspecifications under relative fit evaluation should select the true model as the best-fitting model (given that it is among the competing models). In reality, however, it is difficult to tell if the true model is one of the competing models when misspecification is of concern. Absolute fit evaluation refers to the process of determining whether the model at hand fits the data adequately. Fit statistics that are sensitive to misspecifications under absolute fit evaluation should reject misspecified models with high probability. In practice, however, it is likely that more than one model can fit the data adequately. In this study, we investigated whether both relative and absolute fit evaluations can be used in concert to identify the true model or to provide useful information about model misspecification. The usefulness (for relative fit evaluation) and sensitivity (for absolute fit evaluation) of the fit statistics for identifying misspecifications of the Q-matrix and/or CDM were examined using a simulation study.

Fit Statistics

Six fit statistics were considered in this study: –2 log-likelihood (–2LL), AIC, BIC, the residual between the observed and predicted proportion correct of individual items, the residual between the observed and predicted Fisher-transformed correlation of item pairs (referred to as transformed correlation), and the residual between the observed and predicted log-odds ratios of item pairs. In this study, the first three statistics were used for relative fit evaluation, whereas the last three were for absolute fit evaluation. It is worth noting that the last three statistics are also called limited-information fit statistics; this is compared with full-information fit statistics like the χ^2 or G statistics based on the contingency table of the observed and predicted responses (Rupp et al., 2010, p. 274). Specifically, the proportion correct based on individual item provides only univariate information, whereas both the item-pair–based transformed correlation and the log-odds ratio provide bivariate information.

The –2LL, AIC, and BIC are computed as a function of the maximum likelihood (ML), which is based on the ML estimate of the item parameters (i.e., $\hat{\beta}$) with the attributes integrated out:

$$\mathrm{ML} = \prod_{i=1}^{N} \sum_{l=1}^{L} L(\mathbf{X}_i | \hat{\beta}, \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l), \tag{1}$$

where N is the sample size, L is the total number of attribute patterns, $\mathbf{X}_i$ is the response vector for examinee i, α_l is the lth attribute vector, $L(\mathbf{X}_i | \hat{\beta}, \alpha_l)$ is the likelihood of the response vector of examinee i given α_l, and $p(\alpha_l)$ is the prior probability of α_l. Based on (1), the three statistics are

$$-2LL = -2\ln(ML) \tag{2}$$

$$AIC = -2LL + 2P \tag{3}$$

$$BIC = -2LL + P\ln(N), \tag{4}$$

where P is the number of model parameters. For instance, P equals $(2J + 2^k - 1)$, $(\sum_{j=1}^{J} K_j^* + J + 2^k - 1)$, and $(\sum_{j=1}^{J} 2^{K_j^*} + 2^k - 1)$ for the DINA, *A*-CDM, and saturated models, respectively. For each of these three statistics, the fitted model with the smallest value is selected among the set of competing models.

For the proportion correct, transformed correlation, and log-odds ratio, the model-predicted item responses are generated based on the fitted model. A large number of attribute patterns can be obtained by sampling from the posterior distribution of the attributes. The generated attribute patterns and estimated model parameters then can be used to generate the predicted item responses. Denote $\boldsymbol{X}_j = \{X_{1j}, \ldots, X_{ij}, \ldots, X_{Nj}\}'$ and $\tilde{\boldsymbol{X}}_j = \{\tilde{X}_{1j}, \ldots, \tilde{X}_{ij}, \ldots, \tilde{X}_{\tilde{N}j}\}'$ as the observed and predicted response vector for item j, respectively, where $\tilde{N}$ is the [generally large] sample size used to generate the predicted response patterns. For item j,

$$p_j = \left| \sum_{i=1}^{N} X_{ij}/N - \sum_{i}^{\tilde{N}} \tilde{X}_{ij}/\tilde{N} \right| \tag{5}$$

$$r_{jj'} = |Z[Corr(\boldsymbol{X}_j, \boldsymbol{X}_{j'})] - Z[Corr(\tilde{\boldsymbol{X}}_j, \tilde{\boldsymbol{X}}_{j'})]| \tag{6}$$

$$l_{jj'} = \left| \log\left(\frac{N_{11}N_{00}}{N_{01}N_{10}}\right) - \log\left(\frac{\tilde{N}_{11}\tilde{N}_{00}}{\tilde{N}_{01}\tilde{N}_{10}}\right) \right|, \tag{7}$$

where $j' \neq j$, $Corr(\cdot)$ is the Pearson's product-moment correlation, $Z[\cdot]$ is the Fisher transformation, and $N_{yy'}$ and $\tilde{N}_{yy'}$ are the number of observed and predicted examinees, respectively, who scored y on item j and y' on item j'. In other words, the proportion correct is based on the observed and predicted first moment of individual item responses, whereas both the transformed correlation and the log-odds ratio are based on the second moment. When the model fits the data, the statistics should be close to zero for all items.

To use the proportion correct, transformed correlation, and log-odds ratio for absolute fit evaluation, their standard errors (SEs) are needed. The approximate SEs of these statistics should be model based, and can be computed as follows:

$$\mathrm{SE}[p_j] = \left[\left(\sum_{i}^{\tilde{N}} \tilde{X}_{ij}/\tilde{N} \right) \left(1 - \sum_{i}^{\tilde{N}} \tilde{X}_{ij}/\tilde{N} \right) /N \right]^{1/2} \tag{8}$$

$$\mathrm{SE}[r_{jj'}] = [N - 3]^{1/2} \tag{9}$$

$$\mathrm{SE}[l_{jj'}] = \left[\tilde{N}(1/\tilde{N}_{11} + 1/\tilde{N}_{00} + 1/\tilde{N}_{01} + 1/\tilde{N}_{10})/N \right]^{1/2}. \tag{10}$$

With these SEs, the z-scores of the three statistics can be derived to test whether the residuals differ significantly from zero.

It should be noted that there are one proportion correct and J–1 transformed correlations and log-odds ratios for each item. For a test with J items, there are J proportions correct and $J(J–1)/2$ transformed correlations and log-odds ratios to examine. To evaluate the absolute fit of the model, we proposed to test the maximum z-score of each statistic. This will obviate the need to test the J items and $J(J–1)$ item-pair statistics. Rejection of any z-score is an indication that the model does not fit at least one item or item pair adequately. Implicitly, the above procedure involves multiple tests. To ensure that the multiple tests do not inflate the Type I error, an adjusted significance level α^* was used for each test. In particular, the significance level was adjusted using the Bonferroni correction: $\alpha^* = \alpha/J$ for the proportion correct and $\alpha^* = 2\alpha/J/(J–1)$ for the transformed correlation and log-odds ratio.

Simulation Study

Design

To systematically investigate the performance of the six fit statistics across various settings of misspecification, a simulation study was employed. Two types of reduced CDMs were used to generate the data: the DINA model and A-CDM. In addition to these two models, a saturated model (i.e., the G-DINA model) also was used to fit the data. Two test lengths were considered with $J = 15$ and $J = 30$. The correctly specified Q-matrix for $J = 30$ is presented in Table 1. The Q-matrix for $J = 15$ was imbedded as a subset of this Q-matrix.

Five attributes were considered, and each attribute was specified an equal number of times. The maximum number of required attributes $(K_j^*)_{\max} = 3$ in both Q-matrices. Different scenarios of Q-matrix misspecification were investigated as summarized in Table 2. These scenarios covered over-specification, under-specification, and both over- and under-specifications of single or multiple q-vectors. Among them, Q1 and Q6 were correctly specified; the rest were incorrectly specified.

Accordingly, there were four true CDM and Q-matrix combinations, each of which was fitted with 15 CDM and Q-matrix combinations (Table 3). The true item parameters can be found in Table 4, and the values were set so that $P(\boldsymbol{\alpha}_{lj}^*)_{\min} = 0.10$ and $P(\boldsymbol{\alpha}_{lj}^*)_{\max} = 0.90$ for all true CDMs, where $\boldsymbol{\alpha}_{lj}^*$ was the reduced attribute vector whose elements are the required attributes for item j (see de la Torre, 2011 for more details). Finally, two sample sizes were considered with $N = 500$ or 1,000, resulting in 120 ($2 \times 2 \times 2 \times 15$) simulation conditions. Each condition was replicated 500 times, and the estimation code was written in Ox (Doornik, 2003).

The sensitivity of the six fit statistics to misspecifications of the CDM, the Q-matrix, and both were investigated. The –2LL, AIC, and BIC were used for relative fit evaluation, and the fitted model with the smallest value for each statistic was selected. The proportion of times each fitted model was selected out of the 500 iterations (i.e., selection rate) was reported and analyzed. The maximum z-scores of the proportion correct, transformed correlation, and log-odds ratio were used for absolute fit evaluation. In each iteration, the maximum z-scores were computed and

Table 1
Correctly Specified Q-matrix (J = 30)

	Attribute										
Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1*	1	0	0	0	0	16	1	0	0	0	0
2*	0	1	0	0	0	17	0	1	0	0	0
3*	0	0	1	0	0	18	0	0	1	0	0
4*	0	0	0	1	0	19	0	0	0	1	0
5*	0	0	0	0	1	20	0	0	0	0	1
6*	1	1	0	0	0	21	1	0	1	0	0
7*	1	0	0	0	1	22	1	0	0	1	0
8*	0	1	1	0	0	23	0	1	0	1	0
9*	0	0	1	1	0	24	0	1	0	0	1
10*	0	0	0	1	1	25	0	0	1	0	1
11*	1	1	1	0	0	26	1	0	1	1	0
12*	1	1	0	0	1	27	1	0	1	0	1
13*	1	0	0	1	1	28	1	0	0	1	1
14*	0	1	1	1	0	29	0	1	1	0	1
15*	0	0	1	1	1	30	0	1	0	1	1

Note. Items with * are used when $J = 15$.

Table 2
Summary of Q-matrix Misspecifications

J	Q-matrix	Item Altered	Alterations	Note
15	Q1	None	–	True
	Q2	1	α_2: 0 → 1	Over-specification
	Q3	6	α_1: 1 → 0	Under-specification
	Q4	11	α_1: 1 → 0, α_4: 0 → 1	Both
	Q5	1	α_2: 0 → 1	Over-specification
		6	α_1: 1 → 0	Under-specification
		11	α_1: 1 → 0, α_4: 0 → 1	Both
30	Q6	None	–	True
	Q7	1	α_2: 0 → 1	Over-specification
	Q8	6	α_1: 1 → 0	Under-specification
	Q9	11	α_1: 1 → 0, α_4: 0 → 1	Both
	Q10	1	α_2: 0 → 1	Over-specification
		6	α_1: 1 → 0	Under-specification
		11	α_1: 1 → 0, α_4: 0 → 1	Both

compared with the critical z-score z_c with $\alpha = .05$ and the Bonferroni correction to determine whether the fitted model was to be rejected or not. This time, the proportion of times each fitted model was rejected out of the 500 iterations (i.e., rejection rate) was analyzed and discussed.

Table 3
True and Fitted CDM and Q-matrix Combinations

	Q-matrix									
Model	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Saturated	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
DINA	**D1**	D2	D3	D4	D5	**D6**	D7	D8	D9	D10
A-CDM	**A1**	A2	A3	A4	A5	**A6**	A7	A8	A9	A10

Note. True combinations are in **boldface**; for D1 and A1, Q-matrices Q1 – Q5 are used (left panel); for D6 and A6, Q-matrices Q6 – Q10 are used (right panel).

Table 4
True Item Parameters ($P(\boldsymbol{\alpha}_{lj}^{*})$)

	Required Attribute Pattern							
	0	1						
True	00	10	01	11				
CDM	000	100	010	001	110	101	011	111
DINA	.10	.90						
	.10	.10	.10	.90				
	.10	.10	.10	.10	.10	.10	.10	.90
A-CDM	.10	.90						
	.10	.50	.50	.90				
	.10	.37	.37	.37	.63	.63	.63	.90

Results

Due to space considerations, only results that had immediate relevance for discussion purposes are presented in this article. The results in their entirety can be requested from the first author.

Relative Fit Evaluation

First, we evaluated the selection rates of the –2LL, AIC, and BIC when only CDM misspecification was of concern. For –2LL (not presented), the saturated model was always selected in all conditions. This was expected because the saturated model, which has a more complex parameterization, always has a higher ML than any reduced models. The selection rates for both AIC and BIC are presented in Table 5. For both statistics, either the true or saturated models were selected in all cases (i.e., the misspecified reduced CDMs were never selected). For the AIC, the true models had the highest selection rates in 29 out of 40 cases; in the remaining 11 cases, where the DINA model was the true CDM, the saturated model had the highest selection rates. For the BIC, the true models had the highest selection rates in 38 out of 40 cases. The only two exceptions for the BIC were associated with highly constrained reduced CDM (i.e., DINA), larger sample size (i.e., $N = 1{,}000$) and multiple misspecified q-vectors (i.e., Q5 or Q10). Although not presented, we

Table 5
Selection Rates of AIC and BIC for CDM Misspecification

			$N = 500$						$N = 1{,}000$					
			AIC			BIC			AIC			BIC		
J	T	Q	S	D	A	S	D	A	S	D	A	S	D	A
15	D	Q1		**1.00**			**1.00**			**1.00**			**1.00**	
		Q2	1.00	**.00**			**1.00**		1.00	**.00**		.09	**.91**	
		Q3	.03	**.97**			**1.00**		.02	**.98**			**1.00**	
		Q4	.67	**.33**			**1.00**		1.00	**.00**			**1.00**	
		Q5	1.00	**.00**		.05	**.95**		1.00	**.00**		.88	**.12**	
	A	Q1	.06		**.94**			**1.00**	.02		**.98**			**1.00**
		Q2	.07		**.93**			**1.00**	.02		**.98**			**1.00**
		Q3	.10		**.90**			**1.00**	.03		**.97**			**1.00**
		Q4	.12		**.88**			**1.00**	.03		**.97**			**1.00**
		Q5	.28		**.72**			**1.00**	.11		**.89**			**1.00**
30	D	Q6		**1.00**			**1.00**			**1.00**			**1.00**	
		Q7	1.00	**.00**			**1.00**		1.00	**.00**		.06	**.94**	
		Q8		**1.00**			**1.00**			**1.00**			**1.00**	
		Q9	.11	**.89**			**1.00**		.77	**.23**			**1.00**	
		Q10	1.00	**.00**			**1.00**		1.00	**.00**		.84	**.16**	
	A	Q6			**1.00**			**1.00**			**1.00**			**1.00**
		Q7			**1.00**			**1.00**			**1.00**			**1.00**
		Q8			**1.00**			**1.00**			**1.00**			**1.00**
		Q9			**1.00**			**1.00**			**1.00**			**1.00**
		Q10			**1.00**			**1.00**			**1.00**			**1.00**

Note. N = sample size; J = items; T = true model; S = saturated model; D = DINA; A = A-CDM; rates for the true models are in **boldface**; blank cells are never selected (i.e., 0).

verified that when the selection was to be made between the saturated and misspecified reduced models, the AIC and BIC always selected the saturated model regardless of what Q-matrix was used.

In summary, except for a few conditions, the BIC almost always selected the true model primarily; the saturated model was selected only second, whereas the misspecified CDMs were never selected at all. In comparison, using the AIC only resulted in selecting the saturated model in a few conditions (i.e., DINA is the true CDM). In other words, given the same Q-matrix, if the saturated model is selected against a reduced model using the BIC, it can be an indication that the reduced model was misspecified. But the same cannot be said of AIC.

Second, we evaluated the selection rates of the –2LL, AIC, and BIC when only Q-matrix misspecification was of concern (Table 6). As shown, the –2LL selected the true Q-matrices (i.e., Q1 or Q6) in most cases (i.e., 81% or above) when DINA was the fitted model in all conditions. When the saturated model was fitted, the –2LL selected the over-specified Q-matrices (i.e., Q2 or Q7) in most cases (i.e., 78% or above) in all conditions. This suggests that over-specified Q-matrices can

Table 6
Selection Rate of –2LL, AIC, and BIC for Q-matrix Misspecification

				−2LL				AIC			BIC		
J	*N*	*T*	*F*	Q1/6	Q2/7	Q3/8	Q5/10	Q1/6	Q2/7	Q5/10	Q1/6	Q3/8	Q5/10
15	500	D	S	**.01**	.99			**.78**	.22		**1.00**		
			D	**1.00**				**1.00**			**1.00**		
			A	.20	.20	.06	.50	.25	.13	.50	.34	.10	.49
		A	S	**.04**	.96			**.81**	.18		**1.00**		
			D	.81		.10		.81			.82	.10	
			A	**.55**	.44			**.87**	.13		**.98**		
	1,000	D	S	**.16**	.84			**.87**	.13		**1.00**		
			D	**1.00**				**1.00**			**1.00**		
			A	.18	.21	.01	.60	.24	.14	.60	.31	.02	.60
		A	S	**.17**	.83			**.85**	.15		**1.00**		
			D	.95		.04		.95			.95	.04	
			A	**.60**	.40			**.90**	.10		**.99**		
30	500	D	S	**.22**	.78			**.89**	.11		**1.00**		
			D	**1.00**				**1.00**			**1.00**		
			A	.56	.41			.86	.10		.95		
		A	S	**.20**	.80			**.88**	.12		**.99**		
			D	.97		.02		.97			.97	.02	
			A	**.50**	.50			**.88**	.12		**.99**		
	1,000	D	S	**.21**	.79			**.91**	.09		**1.00**		
			D	**1.00**				**1.00**			**1.00**		
			A	.62	.38			.89	.11		.98		
		A	S	**.25**	.75			**.90**	.10		**1.00**		
			D	1.00				1.00			1.00		
			A	**.51**	.49			**.90**	.10		**.99**		

Note. J = items; N = sample size; T = true model; F = fitted model, S = saturated model; D = DINA; A = *A*-CDM; rates for the true and saturated models are in **boldface**; blank cells are never selected (i.e., 0); Q-matrices before and after "/" are used for $J = 15$ and 30, respectively; Q-matrices with rates <.1 in any condition are omitted.

produce higher ML than the true Q-matrices. This is reasonable because models with over-specified Q-matrices have more parameters and accordingly can fit the data better than true models. For both the AIC and BIC, the true Q-matrices were mostly selected (i.e., at least 78% of the time) when either the saturated or true models were fitted in all conditions. The performance of the BIC was more remarkable, with selection rates of 98% or above in all conditions. This suggests that, together with the BIC, the saturated model can be used as the true model to compare different Q-matrices when the true model is not known.

Third, we evaluated the selection rates of the –2LL, AIC, and BIC when both CDM and Q-matrix misspecifications were of concern (Table 7). As shown, the −2LL cannot select the true CDM-Q-matrix combinations (i.e., zero selection rate) in any condition. In contrast, both the AIC and BIC selected the true combinations

Table 7
Selection Rate of –2LL, AIC, and BIC for Both CDM and Q-matrix Misspecifications

			−2LL		AIC				BIC		
J	*N*	*T*	S1/S6	S2/S7	S1/S6	D1/D6	A1/A6	A2/A7	D1/D6	A1/A6	A2/A7
15	500	D	.01	.99		**1.00**			**1.00**		
		A	.04	.96	.04		**.82**	.12		**.98**	.02
	1,000	D	.16	.84		**1.00**			**1.00**		
		A	.17	.83	.01		**.89**	.09		**.99**	.01
30	500	D	.22	.78		**1.00**			**1.00**		
		A	.20	.80			**.88**	.12		**.99**	.01
	1,000	D	.21	.79		**1.00**			**1.00**		
		A	.25	.75			**.90**	.10		**.99**	.01

Note. J = *items;* N = sample size; T = true model; F = fitted model, S = saturated model; D = DINA; A = *A*-CDM; rates for the true CDM and Q-matrix combinations are in **boldface**; blank cells are never selected (i.e., 0); Q-matrices before and after "/" are used for $J = 15$ and 30, respectively; fitted models with rates no more than .01 in any condition are omitted.

predominantly (82% or higher) across all conditions. The performance of the BIC was more remarkable, with correct selection rates of 98% or above in all conditions. This suggests that when both CDM and Q-matrix misspecifications were of concern, BIC was the best choice among the three statistics to compare across different CDM-Q-matrix combinations because it was most likely to select the true combination (assuming such combination is part of the consideration).

Absolute Fit Evaluation

The rejection rates of the proportion correct, transformed correlation, and log-odds ratio were used for absolute fit evaluation. We found that the proportion correct statistic was either unreliable or had low rejection rates (close to 0%) in most cases and thus was insensitive to either CDM or Q-matrix misspecifications. Accordingly, it was omitted in subsequent analyses.

Table 8 presents the rejection rates for the transformed correlation and log-odds ratio. The following results can be culled from the table: (1) the rejection rates for the transformed correlation and log-odds ratio were essentially the same in all cases, suggesting that the sensitivity of the two statistics was almost identical. Therefore, unless stated otherwise, the two statistics were not distinguished in subsequent analysis; (2) for both statistics, the rejection rates for the true model–Q-matrix combination were always extremely low (i.e., 1% or below), suggesting a low Type I error rate; (3) when the true model was *A*-CDM, the rejection rates for the true model with over-specified Q-matrices were always extremely low (i.e., 1% or below); (4) the rejection rates for the saturated model with true or over-specified Q-matrices were always extremely low (i.e., 1% or below); (5) when the true model was *A*-CDM, the fitted combination was *A*-CDM or the saturated model with Q4, and $N = 500$, the rates were moderately high (i.e., 71% or 83%); (6) in all other incorrect CDM and Q-matrix combinations

Table 8
Rejection Rate of ρ *and l With* α = *.05 and Bonferroni Correction*

			$N = 500$						$N = 1{,}000$					
			ρ			*l*			ρ			*l*		
J	*T*	*Q*	S	D	A	S	D	A	S	D	A	S	D	A
15	D	Q1	**.00**	**.01**		**.00**	**.00**		**.00**	**.00**		**.00**	**.00**	
		Q2	.00			.00	.99		.00			.00		
		Q3	.98	.98		.98	.98							
		Q4	.94			.91								
		Q5	.99			.98								
	A	Q1	**.01**	.99	**.01**	**.01**	.99	**.01**	**.01**		**.01**	**.01**		**.01**
		Q2	.01		.01	.01		.01	.00		.01	.00		.01
		Q3	.98		.98	.98		.98						
		Q4	.71		.83	.71		.83	.99			.99		
		Q5	.96		.98	.96		.98						
30	D	Q6	**.00**	**.01**		**.00**	**.01**		**.00**	**.00**		**.00**	**.01**	
		Q7	.00			.00			.00			.00		
		Q8												
		Q9	.99			.97								
		Q10												
	A	Q6	**.01**		**.01**	**.00**		**.01**	**.00**		**.01**	**.00**		**.01**
		Q7	.01		.01	.01		.01	.00		.01	.00		.01
		Q8												
		Q9	.91		.94	.91		.94						
		Q10												

Note. N = sample size; J = items; T = true model; Q = Q-matrix, S = saturated model; D = DINA; A = *A*-CDM; rates for the true or saturated models with the true Q-matrices are in **boldface**; blank cells are always rejected (i.e., 1).

(i.e., either the CDM or the Q-matrix was misspecified), the rejection rates were always very high (i.e., 91% or above), suggesting a high statistical power. The above findings can be interpreted as follows. First, taking Results 2, 5, and 6 into account, both statistics were sensitive to either CDM or Q-matrix misspecifications, with a few exceptions. Under the Bonferroni correction, the statistics were too conservative, with Type I error rates much lower than the nominal level. Despite this, both statistics still have very high power. Second, considering Results 4 and 6, both statistics treated the saturated model as the true model in most conditions. This implies that the saturated model can be used to evaluate Q-matrix misspecification when the true model is unknown, which is a common occurrence in practice. Third, based on Result 4, both statistics were insensitive to over-specified Q-matrices when the saturated model was fitted or when a highly constrained model (i.e., DINA) was not involved. This implies that over-specified Q-matrices can still produce adequate model–data fit such that they could not be rejected by the two statistics.

Table 9
Q-Matrix Q11 for the Fraction Subtraction Data

Item	Text	Attributes							
		α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
1	5/3 – 3/4	0	0	0	1	0	1	1	0
2	3/4 – 3/8	0	0	0	1	0	0	1	0
3	5/6 – 1/9	0	0	0	1	0	0	1	0
4	3 1/2 – 2 3/2	0	1	1	0	1	0	1	0
5	4 3/5 – 3 4/10	0	1	0	1	0	0	1	1
6	6/7 – 4/7	0	0	0	0	0	0	1	0
7	3 – 2 1/5	1	1	0	0	0	0	1	0
8	2/3 – 2/3	0	0	0	0	0	0	1	0
9	3 7/8 – 2	0	1	0	0	0	0	0	0
10	4 4/12 – 2 7/12	0	1	0	0	1	0	1	1
11	4 1/3 –2 4/3	0	1	0	0	1	0	1	0
12	11/8 – 1/8	0	0	0	0	0	0	1	1
13	3 3/8 – 2 5/6	0	1	0	1	1	0	1	0
14	3 4/5 – 3 2/5	0	1	0	0	0	0	1	0
15	2 – 1/3	1	0	0	0	0	0	1	0
16	4 5/7 – 1 4/7	0	1	0	0	0	0	1	0
17	7 3/5 – 4/5	0	1	0	0	1	0	1	0
18	4 1/10 – 2 8/10	0	1	0	0	1	1	1	0
19	4 – 1 4/3	1	1	1	0	1	0	1	0
20	4 1/3 – 1 5/3	0	1	1	0	1	0	1	0

Note. α_1 = Convert a whole number to a fraction; α_2 = Separate a whole number from a fraction; α_3 = Simplify before subtracting; α_4 = Find a common denominator; α_5 = Borrow from whole number part; α_6 = Column borrow to subtract the second numerator from the first; α_7 = Subtract numerators; α_8 = Reduce answers to their simplest form.

An Empirical Example

In this example we analyzed empirical data to illustrate how the above findings for the fit statistics can be applied when evaluating model–data misfit in a real setting. The data used were a subset of the data that originally were described and analyzed by Tatsuoka (1990) and more recently were used by Tatsuoka (2002), de la Torre and Douglas (2004), and DeCarlo (2011), among others. The data consisted of responses of 536 middle school students to 20 fraction subtraction items measuring eight attributes. The Q-matrix, which is referred as Q11, is shown in Table 9. Using Q11, we compared six CDMs: the saturated model, the DINA model, the DINO model, the additive CDM (*A*-CDM), the LLM, and the R-RUM.

Table 10 gives the results of the model–data fit using –2LL, AIC, BIC, transformed correlation, and log-odds ratio. The saturated model had the smallest –2LL, but the LLM was the best model based on either the AIC or the BIC. For the BIC, three models (i.e., DINA, LLM, and R-RUM) performed better than the saturated model. Based on the findings of the simulation study, these results indicate that all three reduced CDMs are acceptable (i.e., cannot be ruled out as the true model). A possible

Table 10
Fit Results for the Fraction Subtraction Data Using Q11

CDM	–2LL	AIC	BIC	Max. z(r)	Max. z(l)
Saturated	8,529	9,419	1,1326	10.239	11.669
DINA	9,170	9,760	11,024	10.142	11.673
DINO	11,775	12,365	13,629	20.524	22.362
A-CDM	11,430	12,092	13,511	16.649	10.329
LLM	8,685	9,347	10,765	9.754	11.098
R-RUM	8,707	9,369	10,787	8.891	10.493

Note. Max. z(r) = maximum z score for r; Max. z(l) = maximum z score for l; critical z score $z_c = 3.467$, 3.649, 4.044 for $\alpha = .1, .05, .01$, respectively (with the Bonferroni Correction).

explanation is that the underlying model is made up of the three reduced CDMs so that each of them is partially correct. This finding suggests that the data, which often were fitted with the DINA model in the literature, additionally may require other less-constrained models.

For absolute fit evaluation, however, all CDMs were rejected at the lowest significance level (i.e., .01) with large z-scores based on transformed correlation or log-odds ratio. Because even the saturated model—which should perform similarly as the true model under both statistics—was rejected, it is likely that there were misspecifications in the Q-matrix.

De la Torre and Douglas (2004) noted that Item 8 (i.e., 2/3 – 2/3) does not necessarily require mastery of the only prescribed attribute (i.e., subtract numerators); students who have not mastered the attribute but are familiar with the inverse property of addition can still answer the item correctly. Without this attribute, the item does not measure any of the eight attributes. Hence, it can be dropped from the test. With this item removed, the data were re-analyzed with the same six CDMs and Q-matrix Q12. The results are presented in Table 11. The LLM was still the best-fitting CDM, with the R-RUM being slightly worse. Similar to the above findings, the three models (i.e., DINA, LLM, and R-RUM) still performed better than the saturated model based on the BIC and accordingly could be partially correct. Based on the z-scores of either the transformed correlation or log-odds ratio, the p-value of the saturated model was between .05 and .01, which implies that the Q12 might need additional adjustment if a significance level of .05 is the target. In Table 11, it is clear that the z-scores of the LLM and R-RUM were much closer to the critical z-score than the z-scores of other reduced CDMs; this suggests that the true model might be closer to a mixture of the LLM and R-RUM (i.e., some items are best fitted with the LLM whereas others are best fitted with the R-RUM). However, item-level fit evaluation is necessary to further adjust the Q-matrix or to figure out how different CDMs can be mixed at the item level—this is beyond the scope of this study.

Discussion

Validity of inferences from cognitive diagnosis modeling is of increasing concern. It is important to evaluate the fit of the model to the data for inferences from various

Table 11
Fit Results for the Fraction Subtraction Data Using Q12 (With Item 8 Removed)

CDM	–2LL	AIC	BIC	Max. z(r)	Max. z(l)
Saturated	7,968	8,854	10,752	3.751	3.831
DINA	8,620	9,206	10,461	8.935	9.251
DINO	11,236	11,822	13,077	20.639	22.453
A-CDM	10,873	11,531	12,940	16.777	9.626
LLM	8,126	8,784	10,194	4.345	4.194
R-RUM	8,147	8,805	10,214	3.776	4.187

Note. Max. z(r) = maximum z score for r; Max. z(l) = maximum z score for l; critical z score $z_c = 3.467$, 3.649, 4.044 for $\alpha = .1, .05, .01$, respectively (with the Bonferroni Correction).

CDMs to be valid. Based on the simulation study, this research investigated the usefulness or sensitivity of various fit statistics for absolute or relative fit under different CDM settings. The investigation covered the misspecifications of the Q-matrix, the CDM, and both. The simulation study showed that for relative fit evaluation, the BIC, and to some extent, the AIC, can be useful to detect misspecification of the CDM, the Q-matrix, or both. The saturated model can play an important role in detecting CDM or Q-matrix misspecifications. For CDM misspecification, it can be used to distinguish between possibly true and misspecified CDMs. For Q-matrix misspecification, it can be used as the true model to compare across Q-matrices. For absolute fit evaluation, the residual between the observed and predicted correlation of item pair with the Fisher transformation and the residual between the observed and predicted log-odds ratios of pairwise item responses had similar performance and were sensitive to different misspecifications in most conditions, although both statistics tended to be conservative. For these two statistics, the saturated model can be used as the true model in most cases. However, both were insensitive to over-specified Q-matrices unless highly constrained CDMs were involved.

Through the empirical example, we illustrated how the AIC or BIC and transformed correlation or log-odds ratio can be used jointly to provide useful fit information from both the relative and absolute perspectives: with the saturated model, the AIC or BIC can distinguish between incorrect and partially correct CDMs for further analysis whereas the transformed correlation or log-odds ratio can be used with the saturated model to inform misspecification of the Q-matrix and to reject incorrect CDMs.

Although the results of this work are encouraging, additional work is needed to further understand model–data fit evaluation under the CDM context and to broaden the generalizability of the current findings. First, this study only covered two types of misspecifications. Other possible causes of misfit–such as misspecification of the number and structure of the attributes–need to be investigated. Due to the important role the Q-matrix played in cognitive diagnosis modeling, future research should systematically examine the impact of not only the type but also the degree of Q-matrix misspecifications on the different fit statistics. Specifically, it would be interesting to find out the extent to which the reliability of the fit indices deteriorates with

increasing number of misspecified Q-matrix entries. Second, investigation of fit statistics at the item level is needed to fine-tune detection of CDM or Q-matrix misspecifications. Although many fit statistics (e.g., AIC, BIC, root mean square error of approximation, mean absolute differences) can be used or adapted at the item level, their performance has not been systematically documented. In addition to fit statistics, specific procedures have been developed to detect and correct for misspecifications in the Q-matrix or CDM at the item level (e.g., de la Torre, 2008; de la Torre & Chiu, 2010; Lee & de la Torre, 2010). Investigating how these item-level fit statistics and procedures can be used jointly can improve the detection of different types of misfit across a wide range of settings. Third, because CDMs can be used with polytomous data or attributes, we also need to investigate how different fit statistics and procedures can be extended and their performance can be examined when polytomous responses or attributes are involved.

Fit evaluation in cognitive diagnosis modeling can be complicated and challenging due to the many possible causes of misfit: it often requires simultaneous consideration of the impacts of different components of the model. However, as this article has shown, absolute and relative fit evaluation in conjunction with a saturated model can provide a viable means of detecting misspecifications. With additional developments, item-level and test-level fit statistics can be integrated to fine-tune the process of evaluating model–data fit in cognitive diagnosis modeling.

Acknowledgments

This research was supported by National Science Foundation Grant DRL-0744486.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, *19*, 716–723.

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*, 343–362.

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

de la Torre, J., & Chiu, C.-Y. (2010, April). *A general method of empirical Q-matrix validation using the G-DINA model discrimination index*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

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

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

DiBello, L., Roussos, L., & Stout, W. (2007). Cognitive diagnosis Part I. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics*. Amsterdam, The Netherlands: Elsevier.

Doornik, J. A. (2003). Object-oriented matrix programming using Ox (Version 3.1) [Computer software]. London, UK: Timberlake Consultants Press.

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

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

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.

Lee, Y.-S., & de la Torre, J. (2010, April). *Item-level comparison of saturated and reduced cognitive diagnosis models*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, *68*, 78–98.

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

Schwarzer, G. (1976). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*, 239–257.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, *64*, 583–639.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *51*, 337–350.

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

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.), *Diagnostic monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum.

Templin, J., & 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* (Research Report: RR-05–16). Princeton, NJ: Educational Testing Service.

Authors

JINSONG CHEN is Associate Professor in the Department of Psychology at Sun Yat-Sen University, Guangzhou 510275, China; e-mail: jinsong.chen@live.com. His research interests include psychometrics, assessment, and latent variable modeling.

JIMMY DE LA TORRE is Associate Professor in the Department of Educational Psychology at Rutgers, the State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901; e-mail: j.delatorre@rutgers.edu. His research interests are psychological and educational testing and measurement, with specific emphasis on IRT and cognitive

diagnosis modeling, and how assessments can be used to inform classroom instruction and learning.

ZAO ZHANG is a Graduate Student, Department of Educational Psychology at Rutgers, the State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901; zhangzao@eden.rutgers.edu. Her research interests include educational measurement and language assessment.