Performance of Fit Indices in Choosing Correct Cognitive Diagnostic Models and Q-Matrices

Applied Psychological Measurement

2016, Vol. 40(6) 405–417

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621616647954

apm.sagepub.com

Pui-Wa Lei¹ and Hongli Li²

Abstract

In applications of cognitive diagnostic models (CDMs), practitioners usually face the difficulty of choosing appropriate CDMs and building accurate Q-matrices. However, functions of model-fit indices that are supposed to inform model and Q-matrix choices are not well understood. This study examines the performance of several promising model-fit indices in selecting model and Q-matrix under different sample size conditions. Relative performance between Akaike information criterion and Bayesian information criterion in model and Q-matrix selection appears to depend on the complexity of data generating models, Q-matrices, and sample sizes. Among the absolute fit indices, MX2 is least sensitive to sample size under correct model and Q-matrix specifications, and performs the best in power. Sample size is found to be the most influential factor on model-fit index values. Consequences of selecting inaccurate model and Q-matrix in classification accuracy of attribute mastery are also evaluated.

Keywords

model-fit indices, cognitive diagnostic models, Q-matrices, simulation

In applications of cognitive diagnostic models (CDMs; DiBello, Roussos, & Stout, 2007), one has to face the difficult decisions of choosing a model from a large number of models (Jiao, 2009) and a Q-matrix among multiple theoretically plausible ones. Model parsimony is an important consideration in CDM selection, because more complex CDMs require larger sample sizes to yield accurate estimates (e.g., de la Torre & Lee, 2013; Rojas, de la Torre, & Olea, 2012). It is also essential to choose the most appropriate Q-matrix for CDM analyses because previous studies of Q-matrix misspecifications have shown negative impact on both parameter recovery and classification accuracy (e.g., Choi, Templin, Cohen, & Atwood, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp & Templin, 2008). However, performances of model-fit indices that are supposed to inform model and Q-matrix choices under different conditions are not well understood. Although a number of fit indices have been developed in the

¹The Pennsylvania State University, University Park, USA

²Georgia State University, Atlanta, USA

Corresponding Author:

Pui-Wa Lei, Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, 140 CEDAR Building, University Park, PA 16802, USA.

Email: puiwa@psu.edu

general item response theory (IRT) modeling framework, few have been examined in CDM contexts as a means to inform model and Q-matrix selections. This study intends to examine several promising fit indices selected from the IRT literature in CDM contexts. It adds to the scarce literature in fit index performance for CDM and contributes to the understanding of fit index utility in model and Q-matrix selection.

Literature Review

The CDMs Used in the Present Study

This study uses the de la Torre (2011) generalized deterministic inputs, noisy “and” gate (GDINA), a general CDM. The findings about the fit indices in this study should generalize to other parameterizations of general CDMs (e.g., J. Chen, de la Torre, & Zhang, 2013; Kunina-Habenicht et al., 2012). The GDINA model classifies examinees into $2^{k_j^*}$ groups based on mastery of the required attributes for each item, where k_j^* is the number of required attributes for item j . Each group has its own associated probability of success as indicated by $P(\alpha_{ij}^*)$, where α_{ij}^* is a particular reduced attribute vector whose elements are indicators of mastery of the required attributes for item j . As shown in Equation 1 for the GDINA model (de la Torre), the probability of a correct response for examinees with an attribute pattern α_{ij}^* is a function of main effects, and all possible interaction effects among the k_j^* required attributes for item j .

$$P(\alpha_{ij}^*) = \delta_{j0} + \sum_{k=1}^{k_j^*} \delta_{jk} \alpha_{ik} + \sum_{k'=k+1}^{k_j^*} \sum_{k=1}^{k_j^*-1} \delta_{jkk'} \alpha_{ik} \alpha_{ik'} \dots + \delta_{j12 \dots k_j^*} \prod_{k=1}^{k_j^*} \alpha_{ik}, \quad (1)$$

where δ_{j0} is the intercept, δ_{jk} is the main effect due to a single attribute α_k , $\delta_{jkk'}$ is a first-order interaction effect between α_k and $\alpha_{k'}$, and $\delta_{j12 \dots k_j^*}$ is the highest-order interaction effect due to $\alpha_1, \dots, \alpha_{k_j^*}$. de la Torre also illustrated several more restrictive CDMs as special cases of the GDINA. Specifically, the deterministic inputs, noisy “and” gate (DINA) model can be obtained by setting all the parameters in Equation 1, except δ_{j0} and $\delta_{j12 \dots k_j^*}$ to 0. The guessing parameter of the DINA model is equivalent to δ_{j0} , and $\delta_{j0} + \delta_{j12 \dots k_j^*}$ is equal to the non-slipping or one minus the slipping parameter of the DINA. The deterministic input, noisy “or” gate (DINO) model can be obtained by constraining absolute values of all the main and interaction effect parameters in Equation 1 to be the same and alternating the signs of the parameters by the order of the interactions (i.e., negative for two-, four-, six- . . . way interactions; positive for three-, five-, seven- . . . way interactions). When all the interaction effects in Equation 1 are constrained to 0, the model becomes an additive CDM (additive cognitive diagnostic model [ACDM]). Using a log function on the ACDM will result in an equivalent model (i.e., model parameters need to be transformed) to the reduced reparameterized unified model (RRUM) (Hartz, 2002).

The saturated GDINA model may produce the best absolute fit indices because it is highly parameterized. However, relative fit indices for the GDINA may not be good, and its parameter estimates may not be stable, especially when sample size is not large. In that case, a simpler model might be preferred. A simpler model also provides more straightforward interpretation of the attribute–item relationship (Rojas et al., 2012). In this study, the authors examine the fit indices with the GDINA model and the four simpler models (DINA, DINO, ACDM, and RRUM) within this general CDM.

Fit Indices for CDMs

Model-fit indices can be classified as either relative or absolute measures of fit. Relative fit indices are meant for model comparisons and are often functions of log-likelihood values of the

maximum likelihood (e.g., -2 log-likelihood or $-2LL$), some of which penalize more complicated models (e.g., Akaike information criterion $[AIC] = -2LL + 2P$, where P is the number of parameters) or take into account sample sizes as well (e.g., Bayesian information criterion $[BIC] = -2LL + P \ln[N]$, where $\ln[N]$ is the natural log of sample size). Models that yield the smallest value for each of these relative fit indices are preferred. In contrast, absolute fit indices aim to measure the magnitude of overall discrepancy between observed and model-predicted values and are usually functions of residuals. Although absolute fit indices should, in theory, be compared with established cutoff values to determine fit, comparing their values across different models could yield helpful information in practice (Kunina-Habenicht et al., 2012). This is especially true when cutoff values are not available or do not work well.

Findings about relative fit indices in CDM contexts have not been consistent due likely to different simulation designs. On one hand, J. Chen et al. (2013) analyzed three models (DINA, ACDM, GDINA) with data generated from two (DINA, ACDM) using codes written in OX (Doornik, 2003) and found BIC to perform better than AIC while $-2LL$ indiscriminately selected the saturated model in all conditions. On the other hand, Kunina-Habenicht et al. (2012) found AIC to perform better than BIC with a correct Q-matrix but incorrect model with omitted interaction effects using the Mplus code for log-linear CDM (Templin & Hoffman, 2013). The choices of absolute fit indices included in previous studies appeared to be somewhat ad hoc and did not overlap across studies. Kunina-Habenicht et al. included two univariate measures (the mean absolute difference [MAD] between observed and predicted proportion; and root mean square error of approximation [RMSEA] which compares the class-specific model-predicted and observed proportions for the correct response category, weighted by the proportion of respondents in the latent class) and found both to be insensitive to model misspecifications unless the Q-matrix was overspecified. Hence, the authors did not recommend relying solely on MAD or RMSEA for model-fit judgment. Among the three absolute fit indices that J. Chen et al. studied (absolute difference between observed and predicted values in proportion correct or p , Fisher transformation of item-pair correlation or r , and log-odds ratio of item pair or l), the bivariate measures (r and l) had better performance than the univariate one (p). However, more studies of more absolute fit indices for CDM are needed.

Research on IRT model assumptions suggested that measures of local dependence (LD) might yield valuable information about model misfit and could potentially be adapted to serve as overall model-fit indices. Misspecified CDM, Q-matrix, or both would affect their capability to accurately predict patterns of item responses, thereby resulting in unaccounted for systematic relationships among item residuals (i.e., LD). Hansen, Cai, Monroe, and Li (2014) found the χ^2 statistics (W. Chen & Thissen, 1997) to be slightly conservative but still useful in identifying potential misspecifications of CDMs. Yen's (1984) Q3 is another widely used LD index and has been found to perform better than χ^2 for posterior predictive model checking (PPMC) of IRT models when the assumed distribution is not relied on (Levy, Mislevy, & Sinharay, 2007). Residual item covariance (McDonald & Mok, 1995) has also been recommended in both the frequentist and PPMC frameworks (Levy et al., 2007), and a related measure of residual item correlation has been applied in CDM studies (DiBello et al., 2007) but not examined empirically. As such, these statistics appear promising to be adapted as overall absolute fit indices for CDM and Q-matrix selections. This study intends to examine relative performances of these adapted absolute fit indices in informing model and Q-matrix selections. Although RMSEA and χ^2 have been studied in CDM contexts as noted above, they were used as local item-level fit indices and examined in a different parameterization of the general CDM (log-linear). It would be interesting to see if previous findings at local levels generalize to the global test level.

The adaptation of item-level fit indices to indicate overall test-level fit is straightforward, by simply averaging absolute values of all item-level statistics (see the online appendix for

computational details). Specifically, the authors included the average of χ^2 test statistics of independence for pairwise item response frequencies over all item pairs (W. Chen & Thissen, 1997; MX2); the mean of absolute values of the Q3 statistics that are pairwise correlations of item residuals (Yen, 1984; MADQ3); the average of absolute values of pairwise item covariance residuals (McDonald & Mok, 1995; MADres); the average of absolute deviations between observed and predicted pairwise item correlations (DiBello et al., 2007; MADcor); and the mean of item-level RMSEAs that compare the class-specific model-predicted and observed proportions for each response category, weighted by the proportion of respondents in the latent class, across classes and response categories. Relative fit indices (AIC and BIC) were also examined because previous findings about their performances were inconsistent.

Q-Matrices and Sample Sizes

Q-matrices represent different hypotheses about which attributes are required for correctly answering each item. Due to the lack of full understanding of the cognitive processes involved in taking a test, in reality, the true Q-matrix is unknown (DeCarlo, 2012). Establishing a Q-matrix through substantive knowledge sometimes may be subjective (de la Torre, 2008; Rupp & Templin, 2008), and empirical evidence from CDM analyses has been proven to be an important source to help researchers choose the most appropriate Q-matrix (e.g., de la Torre, 2008; Li & Suen, 2013; Wang & Gierl, 2011). Previous studies of Q-matrix misspecifications have shown negative impact on both parameter recovery and classification accuracy. For example, Rupp and Templin (2008) found that underspecified Q-matrices would overestimate slipping parameters of the DINA model, whereas overspecified ones would overestimate guessing parameters. Kunina-Habenicht et al. (2012) also demonstrated that Q-matrix misspecification (over-, underspecified, or random permutation) could seriously affect classification accuracy, parameter recovery of latent class distributions, correlations, and attribute proportions. However, Choi et al. (2010) found negligible effect with overspecified Q-matrices but negative effect with underspecified ones on both item parameter estimation and respondent classification, suggesting that the impact of Q-matrix misspecification may depend on the kind of misspecification and perhaps the specific CDM employed as well.

In addition, little is known about sample size requirement for fit indices to perform well in selecting CDMs or Q-matrices. Prior studies have examined the effect of sample size on parameter recovery within the log-linear CDM parameterization and found that sample sizes about 500 (Choi et al., 2010) or 1,000 (Kunina-Habenicht et al., 2012) would be sufficient to estimate intercept and main effects. Unfortunately, it was difficult to recover two- or higher-way interactions even with a sample size of 4,000 (Choi et al., 2010) or 10,000 sometimes (Kunina-Habenicht et al., 2012). Choi et al. also suggested that relative fit indices were able to select the correct model under the known Q-matrix condition with a high level of consistency when sample was 200 or more. However, more studies are needed to examine the effect of sample size on the performance of fit indices under different model and Q-matrix conditions.

The Present Study

In summary, performances of fit indices that are supposed to inform CDM model and Q-matrix choices are not well understood due to the dearth of systematic studies. Although LD statistics have been used to identify model misfit in the general IRT framework, their utilities in identifying model and Q-matrix misspecifications have not been thoroughly examined in CDM contexts. This study aims to evaluate the performance of several promising LD statistics adapted to indicate overall absolute fit, along with the relative fit indices whose relative performances have

been found to be inconsistent. Specifically, the authors examine the performance of these fit indices in selecting the correct model among a number of CDMs from the GDINA parameterization and the correct Q-matrix among a few plausible ones under different sample size conditions. Null distribution and empirical power of absolute fit indices are also evaluated, as are consequences of selecting inaccurate model and Q-matrix in classification accuracy.

Method

Manipulated Factors

The analysis model, Q-matrix, and sample size are believed to influence measures of model fit. The first factor manipulated was model. Data were simulated using both the GDINA and the ACDM, and analyzed by all five models (GDINA, DINA, DINO, ACDM, and RRUM) using the R package CDM version 1.5–10 (Robitzsch, Kiefer, George, & Uenlue, 2013).

The second factor manipulated was Q-matrix. Given that Q-matrix misspecification may confound model performance (Rupp & Templin, 2008), three Q-matrices (correct, over-, and underspecified) were used (see Table 1 in the online supplement). The correct Q-matrix came from an existing test, the 20-item Michigan English Language Assessment Battery (MELAB) reading test Form E. This Q-matrix with four attributes (vocabulary, syntax, extracting explicit information, and understanding implicit information) was constructed and validated by Li and Suen (2013). To construct the overspecified Q-matrix, Attribute 2 was added to Items 1, 5, and 7 and Attribute 3 to Items 2 and 12. These five Q-matrix entries were included in the initial Q-matrix examined by Li and Suen but were later removed because the RRUM model calibration showed that these five entries were not substantial. Because Attribute 3 (extracting explicit information) could conceivably be considered a prerequisite for Attribute 4 (connecting and synthesizing), Attribute 3 was also added as required for all items requiring Attribute 4 for the overspecified Q-matrix. For the underspecified Q-matrix, the authors trimmed from the correct Q-matrix. According to Hartz (2002), if the diagnostic discrimination parameter (r_{ik}) estimated by the RRUM model is greater than .9, it indicates that item i lacks diagnostic capacity for discriminating the masters from the non-masters on attribute k , and the corresponding Q-matrix entry can be dropped. To construct a theoretically plausible underspecified Q-matrix, the authors removed from the correct Q-matrix entries whose r_{ik} parameter estimate from the RRUM model was larger than .8. Overall, the average attribute per item was 1.70 for the correct Q-matrix, 2.45 for the overspecified Q-matrix, and 1.25 for the underspecified Q-matrix.

The third factor manipulated was sample size. Four sample size levels (500, 1,000, 2,000, and 4,000) were considered because they have been used in previous simulation studies (e.g., J. Chen et al., 2013; Choi et al., 2010; Kunina-Habenicht et al., 2012). In total, 120 (4 sample sizes $\times$ 3 Q-matrices $\times$ 2 generating models $\times$ 5 analysis models) conditions were examined with 200 replications each.

Simulation Procedures

Two hundred sample item response datasets for each of the four sample sizes were generated using the GDINA model and the ACDM with the SIM.GDINA package in R. Parameter values for the GDINA generating model (including attribute thresholds; correlations between attributes; coefficients for the intercept, main effect, and interaction terms among the required attributes based on the true Q-matrix and the GDINA) were estimated from an extant dataset with item responses to a reading test collected from 2,019 students (Li & Suen, 2013). Parameter values for the ACDM generating model were identical to the GDINA generating model, except

that coefficients for interaction terms were not applicable to ACDM and thereby fixed to 0. Therefore, this study simulated data with two models (GDINA and ACDM), a single true Q-matrix, and a set of true parameters.

Each of the sample datasets was analyzed with all 15 combinations of model and Q-matrix (i.e., 5 models $\times$ 3 Q-matrices). Regarding fit indices, the absolute value of LL (instead of $-2LL$ because the direction of interpretation remained the same), AIC, and BIC was recorded for relative fit. RMSEA, MADQ3, MADres, MADcor, and MX2 were also recorded for absolute fit. Replications that did not yield results for any analysis models were eliminated from further analyses. The GDINA generating model produced results for 95.5%, 99.5%, and 100% of the replications for sample sizes 500, 1,000, and $\geq 2,000$, respectively. Similarly, the ACDM generating model produced results for 98.5%, 99%, and 100% of the replications for sample sizes 500, 1,000, and $\geq 2,000$, respectively. Estimation problems occurred when the analysis used the GDINA with the overspecified Q-matrix for some small samples, which resulted in singular matrices that could not be solved by the system.

Results

Contributing Factors to Variance in the Fit Indices

For each of the data generating models, proportions of variance due to different study factors and their interactions were examined for each fit index to identify the key factors that influenced their magnitudes using three-way (Sample size $\times$ Q-matrix $\times$ Model) ANOVA. Main and interaction effects with moderate effect sizes or larger (eta-square $> .06$) were further examined. For both data generating models, sample size was the key factor that determined the magnitude of the fit indices (accounted for $\geq 29\%$ of total variance). Values of the likelihood-based relative fit indices increased as sample sizes increased. With large values of the likelihood-based fit indices, differences due to different Q-matrices or models appeared relatively minor (eta-square $< .001$). In contrast, values of most absolute fit indices tended to decrease as sample size increased, except for MX2.

In addition, the sample size by Q-matrix interaction accounted for a fair amount of variance (6%-21%) in RMSEA, MADcor, MADres, and MX2 for the GDINA generating model. When data were generated by the GDINA, RMSEA consistently favored the overspecified Q-matrix over the true Q-matrix across sample sizes; RMSEA also preferred the underspecified Q-matrix over the true Q-matrix at a sample size of 500 but not at $N \geq 2,000$. Average MADcor, MADres, and MX2 values were nearly the same between the correctly specified and overspecified Q-matrices, and they were lower than the underspecified Q-matrix counterparts. This indicated that MADcor, MADres, and MX2 were sensitive to the underspecified but not the overspecified Q-matrix. Larger differences between matrices were observed for these fit indices when sample size was larger. Although the pattern of sample size by Q-matrix interaction on these indices appeared largely similar for the ACDM generating model, their effect sizes were small ($< 3\%$). Moreover, the sample size by model interaction explained a non-negligible amount of variance in MX2 (7% - 8%) for both data generating models. MX2 tended to favor the GDINA model across sample sizes. As sample size increased, larger between-model differences in MX2 were observed.

Null Distribution and Empirical Power for Absolute Indices

After further examining distributions of the absolute fit indices at different sample sizes under the true model and Q-matrix conditions, the authors found it difficult, if not impossible, to find

a common cutoff value for each of these indices except perhaps for MX2. Using the 95% empirical critical values (.71 - .75 for the GDINA generating model and .85 - .88 for the ACDM generating model) as cutoffs, MX2 largely demonstrated high empirical power to detect most misspecifications with the following exceptions: (a) the incorrect combination of the overspecified Q-matrix with the true generating model or the overparameterized model with the true or overspecified Q-matrix across all sample sizes, (b) the combination of the ACDM with the true or overspecified Q-matrix at $N \leq 1,000$ when data were generated from the GDINA. Although the other absolute fit indices were much more sensitive to sample sizes than MX2 under correct specifications, empirical power was similarly tabulated using corresponding empirical critical values. Power for MADcor and MADres was high except in the same low-power conditions as MX2. In addition, MADcor and MADres had low power for the combination of the RRUM with the true or overspecified Q-matrix at smaller sample sizes. MADQ3 had extremely low power across all conditions. Power for RMSEA was consistently low in the same low-power conditions as MX2 or at smaller sample sizes ($N \leq 1,000$ for data generated from the GDINA and $N \leq 2,000$ from the ACDM) in other conditions.

Model and Q-Matrix Selection

The Q-matrix (among the three for each model), model (among the five for each Q-matrix), and the combination of Q-matrix and model (among the 15) selected by each of the fit indices were noted for each replication and sample size by data generating model. Proportions of selection out of the successful replications for each of the Q-matrices, models, or their combination by each of the fit indices were tabulated. Choices with the largest proportions were examined to find out what model, Q-matrix, or their combination each of the fit indices favored when used in a relative sense. Table 2 in the online supplement displays the highest proportion of model by Q-matrix combination choices among the 15 possible combinations by the fit indices for each of the data generating models. LL consistently favored the overspecified Q-matrix and the most complicated GDINA model regardless of the data generating model. AIC penalized complicated models with a larger number of parameters, and was able to identify the true combination of Q-matrix and model majority of the time ($\geq 51\%$) under the study conditions except when sample size was very large and data were generated from a less highly parameterized model. When GDINA was the generating model, AIC tended to favor the true combination of Q-matrix and model, with higher proportion of correct choices as sample size increased. When ACDM was the generating model, AIC also favored the true combination of Q-matrix and model until sample size was increased to 4,000, in which case AIC favored the true Q-matrix and GDINA combination. In contrast, BIC tended to favor the less complicated Q-matrix and/or model (relative to the true) when sample size was small. Because the penalty component of BIC increased with sample size, this pattern of sample size effect implied that $-2LL$ rose faster than the penalty with sample size, and that the penalty had a larger impact compared with $-2LL$ (or was too severe) at small samples. When the generating model was GDINA, BIC favored the underspecified Q-matrix and the less highly parameterized ACDM at sample size 500, or the true Q-matrix with ACDM at larger sample sizes. BIC chose the correct combination of Q-matrix and model with a high proportion only at sample size 4,000 when the generating model was GDINA. When the generating model was the simpler ACDM, BIC favored the incorrect combination of the true Q-matrix and DINO at sample size 500 and started to choose the correct combination of Q-matrix and model with higher proportions at larger sample sizes ($N \geq 1,000$).

Unfortunately, none of the absolute fit indices included in this study were able to identify the correct combination of model and Q-matrix with high percentages. Regardless of the data

generating model, MADcor, MADres, and MX2 were similar in their choice of the GDINA model with the overspecified Q-matrix, and the percentage of this combination selection increased with sample size. RMSEA tended to favor the DINO model with the overspecified Q-matrix when the generating model was GDINA, or the RRUM model with the underspecified Q-matrix when the generating model was ACDM, until sample size reached 4,000. In contrast, MADQ3 tended to favor the DINA model with the overspecified Q-matrix regardless of the generating model unless sample size was small (500). Overall, the bivariate measures of absolute fit (MADcor, MADres, and MX2) appeared to behave more consistently than the univariate measures (RMSEA and MADQ3) and tended to suggest that the saturated model and the overspecified Q-matrix fit best.

Q-Matrix Selection by Fixed Model

When one would like to choose a Q-matrix based on model-fit indices given a preselected model, then results for the highest proportion of Q-matrix choices (among 3) by the different fit indices for a given model might be informative. Table 3 in the online supplement provides the Q-matrix choices with the highest proportions for both the GDINA and the ACDM generating models. When the GDINA or the correct model was analyzed, Table 3 shows that (a) AIC was able to identify the true Q-matrix with high percentages ($\geq .87$), regardless of sample size; (b) BIC tended to choose the true Q-matrix when sample size was large, or the underspecified Q-matrix when sample size was smaller; and (c) the absolute fit indices tended to favor the overspecified Q-matrix with a few exceptions.

It was interesting to note that when the simpler DINA model was analyzed, all three relative fit indices, MADcor, and MADres were able to identify the true Q-matrix with relative high percentages ($\geq .46$) regardless of the generating model, and their accuracy rates were higher with larger sample sizes. In contrast, when the DINO model was analyzed, these fit indices were able to identify the correct Q-matrix with high percentages only when ACDM was the generating model. The fit indices did not perform consistently in choosing the correct Q-matrix when other models were analyzed.

Overall, AIC appeared to be the most helpful in selecting Q-matrix when one does not know the true model. AIC had high accuracy rates in most model conditions with a few exceptions when GDINA was the generating model. Because the fit indices (except RMSEA, MADQ3, and MX2) tended to agree with each other more often about the correct Q-matrix when the DINA model was analyzed, the DINA model might be used to first choose a Q-matrix.

Model Selection by Fixed Q-Matrix

If a Q-matrix has been identified, then results on the highest proportion of model choices (among 5) by the different fit indices for a given Q-matrix might be helpful. Table 4 in the online supplement shows the model choices by AIC, BIC, RMSEA, and MADQ3 with the highest proportions for both the GDINA and the ACDM generating models. LL, MADcor, MADres, and MX2 consistently favored the most highly parameterized GDINA model regardless of the generating model across Q-matrices and sample sizes. As such, their model selection results are not presented. When the generating model was GDINA, AIC appeared to be the most helpful in choosing the correct model across Q-matrices, except at the smallest sample size with the overspecified Q-matrix. BIC tended to favor the less highly parameterized models (ACDM or DINA), except at the largest sample size with the true or underspecified Q-matrix. When the generating model was ACDM, AIC was able to identify the correct model better than

BIC with the correct or overspecified Q-matrix at smaller sample sizes ($N \leq 1,000$). However, BIC performed better at larger sample sizes ($N > 1,000$) or with the underspecified Q-matrix.

Unfortunately, the other fit indices did not work well in choosing the correct model irrespective of the generating model. With few exceptions, RMSEA tended to favor the RRUM with the underspecified Q-matrix, or the DINO with the true or overspecified Q-matrix. MADQ3 tended to favor the DINO with the underspecified Q-matrix, or the DINA with the true or overspecified Q-matrix.

Classification Accuracy

True mastery status of each of the attributes for each simulee was used to tabulate classification accuracy by the study models and Q-matrices. Proportions of correct mastery decisions across the four attributes (i.e., correct identification of the entire attribute profile) were tabulated for each simulee. These proportions were averaged across simulees, and proportions of their variance due to the study factors and interactions were calculated to identify factors that affected classification accuracy. Q-matrix, model, and their interaction accounted for the most variance in classification accuracy (around 80% for the GDINA generating model and 65% for the ACDM generating model). As expected, the correct combination of model and Q-matrix yielded the highest correct classification rate (about 83% for the GDINA generating model and 81% for the ACDM generating model). The combinations of correct Q-matrix and more highly parameterized models, such as the RRUM, the ACDM, and the GDINA (when they were not the generating model), were close seconds in classification accuracy (around 80%). DINA appeared to be the least sensitive to Q-matrix choices on classification accuracy rate (hovering about 76%-77% for the GDINA generating model and 74% for the ACDM generating model). In contrast, classification accuracy rate for RRUM dropped dramatically (by about 60% for the GDINA generating model and 50% for the ACDM generating model) when the underspecified Q-matrix was analyzed rather than the other two Q-matrices. The effect of Q-matrix choice on classification accuracy rate for the other models was not large ($< 10\%$ difference regardless of the generating model), though accuracy rates were somewhat higher when the correct Q-matrix was analyzed. Excluding the Q-matrix-sensitive RRUM, the DINO was the least accurate (59% - 68% regardless of the generating model) among the models in identifying mastery of attributes. It is important to note that the above classification accuracy is only a quantitative measure. The actual meanings of the attributes may change when different Q-matrices and/or models are estimated.

Discussion

Performances of relative fit indices in model and/or Q-matrix selection appear to depend on the true data generating model, Q-matrix, sample size, or a combination of these. In this study, AIC was found to perform slightly better than BIC in selecting the correct combination of model and Q-matrix as well as in selecting the correct model, given a true or overspecified Q-matrix. This advantage of AIC was observed across sample sizes when the generating model was GDINA but only at smaller sample sizes, $N \leq 1,000$, when the generating model was simpler. In addition, AIC was found to perform better than BIC in choosing Q-matrix with the saturated GDINA or the correct analysis model. Using the general log-linear model parameterization, Kunina-Habenicht et al. (2012) also demonstrated that AIC had higher accurate model selection rate than BIC with a correct Q-matrix when the generating models were more complicated (with two- and three-way interactions) than the analysis model (main effects model). However, J. Chen et al. (2013) found the opposite with the less complicated DINA and ACDM as their generating models. Interestingly, this discrepancy was observed under the same ACDM generating

model and sample size conditions between J. Chen et al. and the current study. It was perhaps due to the difference in Q-matrix misspecifications. J. Chen et al.'s misspecified Q-matrices involved fewer items and misspecifications than the ones employed in this study.

BIC's performance appears to improve with a simpler generating model at larger sample sizes. BIC generally imposes larger penalty than AIC for more highly parameterized models and more so for larger samples. Therefore, it is reasonable to expect the relative performance between AIC and BIC to be different depending on the complexity of the data generating model, Q-matrix, and sample size. This indicates that one cannot always rely on one index to make model and Q-matrix selection when the true model and Q-matrix are unknown in practice. As AIC tends to err in favor of the more complex models and BIC the less complex ones, agreement between them should lend more support for the decision.

Unfortunately, none of the absolute fit indices examined in this study identified the correct combination of model and Q-matrix with high percentage. The three bivariate absolute fit indices (i.e., MADcor, MADres, and MX2) tend to favor the saturated GDINA model and the overspecified Q-matrix, whereas the behaviors of RMSEA and MADQ3 are not consistent. Using a different set of bivariate absolute fit indices, J. Chen et al. (2013) also found that they consistently selected the GDINA model with true or overspecified Q-matrices even though GDINA was the wrong model in their study.

Absolute fit indices do not appear to be proper for model or Q-matrix selection when used in a relative sense and should better be used in an absolute sense, as they are supposed to. Bivariate measures of absolute fit (MADcor, MADres, and MX2) tended to perform better than univariate measures (RMSEA and MADQ3) in empirical power. This finding was consistent with J. Chen et al. (2013) though different fit indices were used in the two studies. When used in an absolute sense, bivariate fit measures had high empirical power to detect model and Q-matrix misspecifications with few exceptions (e.g., when the model and Q-matrix are more complex relative to the true, or slightly less complicated than the true at small sample sizes). Among the bivariate absolute fit indices, MX2 performed the best for being least sensitive to sample size under the true model and Q-matrix combination and for being consistently powerful in detecting model and Q-matrix misspecifications with the fewest exceptions. MADQ3 had extremely low power across all conditions, and the power of RMSEA was consistently low in most conditions.

The present study's finding that sample size only explained a negligible proportion of variance in classification accuracy is consistent with that of Kunina-Habenicht et al. (2012). As expected, misspecification of Q-matrix or model would affect classification accuracy. When the Q-matrix was correctly specified, the more complicated models (RRUM, GDINA, and ACDM when they were not the generating model) had slightly lower classification accuracy rate than the true model, but the DINO had considerably lower accuracy rate. The misspecified Q-matrix also lowered classification accuracy for all study models. The Q-matrix misspecification effect on classification accuracy was found less serious in this study than that in Kunina-Habenicht et al. (2012). This is perhaps because their misspecification of Q-matrix by permutation was more severe than the theoretically plausible ones of this study. However, the RRUM's classification accuracy was greatly reduced by the underspecified Q-matrix. Fortunately, only the less well performing univariate measures of absolute fit (RMSEA and MADQ3) had selected the DINO or RRUM model with high frequency in certain conditions. RMSEA and MADQ3 are, therefore, not recommended to be used alone for model-fit decisions.

In application studies, researchers usually compare competing plausible Q-matrices (e.g., Wang & Gierl, 2011) because the true Q-matrix is typically unknown. J. Chen et al. (2013) suggested that using the saturated model to evaluate Q-matrix when the true Q-matrix is unknown is a common practice, and they recommended doing so with BIC. However, results of the

current study suggest that AIC may work better than BIC in selecting Q-matrix with the saturated model. As discussed above, the discrepancy in relative performance between AIC and BIC could be due to different Q-matrix misspecifications, different generating models, or a combination of both. Moreover, both J. Chen et al.'s study and the current study found that $-2LL$ and some other fit indices tended to select the true Q-matrix when the DINA model was analyzed regardless of the data generating models and sample sizes. Perhaps, the possibility of using the DINA model to choose Q-matrix should be further investigated. Successful replications of the present study's findings would suggest that selection of Q-matrix could be based on agreement among $-2LL$, AIC, BIC, MADcor, and MADres with the DINA model.

Another important finding of this study is the effect of sample size on fit indices. Among the three factors manipulated (model, Q-matrix, and sample size), sample size explained the largest proportion of variance in all the fit indices. As sample size increased, relative fit indices increased but most absolute fit indices (except for MX2) decreased. The sensitivity of fit indices to sample size might not be an issue for relative fit indices because they are typically used for comparison among several models or Q-matrices fit to the same data with the same sample size. Absolute fit indices are more commonly compared with some rule-of-thumb criteria to determine the level of acceptable fit. Using Henson, Templin, and Willse's (2009) cut point for MADcor (around .06) and Kunina-Habenicht, Rupp, and Wilhelm's (2009) criterion for RMSEA ($< .05$), most of the models and Q-matrices examined in this study would be deemed acceptable in reproducing the generated data. The absolute fit indices might be more sensitive to more severe misspecifications of model or Q-matrix. However, the levels of model and Q-matrix misspecifications examined in this study are considered realistic, and they did result in meaningfully lower classification accuracy rates. Because absolute fit indices are meant to quantify goodness of model fit, they should not be sensitive to sample sizes. Future research might investigate the possibility of incorporating a sample size adjustment or establishing a proper cutoff criterion for each of the absolute fit indices.

The true and misspecified Q-matrices in this study are based on real examples and plausible theoretical considerations. While Q-matrices based on real examples may be more realistic, it should be realized that this approach may limit the generalization of the results. For example, the test length was not long, and the strategies used to produce the misspecified Q-matrices were rather limited. It is necessary to manipulate more types and levels of Q-matrix misspecifications, such as different number of attributes, different percentages of Q-matrix misspecifications, and different interactions among the attributes in future studies. Different population skill distributions, test diagnosticity, and test lengths might also affect performance of the fit indices (e.g., Huebner & Wang, 2011) and should be investigated in the future. In addition, the authors assumed that the same model was applicable for all items in the test and focused on test-level fit. However, it is possible that individual items may fit different models to different extents (de la Torre & Lee, 2013). Future extension can include item-level fit and allow the selection of different models for different items.

Finally, the true models and Q-matrices are known in simulation studies but they are never known in applications. Whether true models or Q-matrices exist in reality is debatable, and such discussion is beyond the scope of this study. However, there are reasonable chances that they may not even be among the competing models or Q-matrices that researchers seek to compare in practice. The present study suggests that fit indices could inform users about which competing Q-matrix or model is more appropriate than others in terms of goodness of fit with data, even if all competing models and Q-matrices are false. As such, it is important to choose model and Q-matrix based on not only goodness of fit with data but also interpretability of the attributes as defined by the model and Q-matrix. The chosen model and Q-matrix should be replicated with independent samples and subjected to further examinations, the same way as factor

structures in factor analytic models should. Furthermore, attribute scores should be validated for the purposes of their uses, the same way as test scores should in general.

In conclusion, CDMs hold promise for deriving rich multidimensional subscale information to aid learning; however, their applications are limited due partly to the lack of guidance for model and Q-matrix selections. Model-fit indices that can inform such choices, especially absolute fit indices, have not been systematically studied in the field. The present study examines the performance of several promising fit indices adapted from the LD literature in choosing correct CDMs and Q-matrices as well as in detecting misspecifications under different sample sizes. It provides some preliminary evidence for the utility of the adapted LD model-fit indices.

Authors' Note

A similar version of this article was presented at the 2014 Annual Meeting of the National Council on Measurement in Education, Philadelphia, Pennsylvania.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- 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.
- Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.
- Choi, H.-J., Templin, J. L., Cohen, A. S., & Atwood, C. H. (2010, April). *The impact of model misspecification on estimation accuracy in diagnostic classification models*. Paper presented at the meeting of the National Council on Measurement in Education (NCME), Denver, CO.
- DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, 36, 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, 343-362.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50, 353-373.
- 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, The Netherlands: Elsevier.
- Doornik, J. A. (2003). *Object-Oriented Matrix Programming Using Ox* (Version 3.1) [Computer software]. London, England: Timberlake Consultants Press.
- Hansen, M., Cai, L., Monroe, S., & Li, Z. (2014). *Limited-information goodness-of-fit testing of diagnostic classification item response theory models* (CRESST Report 840). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California.
- 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.

- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, 71, 407-419.
- Jiao, H. (2009). Diagnostic classification models: Which one should I use? *Measurement: Interdisciplinary Research & Perspective*, 7(1), 65-67.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, 35, 64-70.
- 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.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2007, April). *Posterior predictive model checking for conjunctive multidimensionality in item response theory*. Paper presented at the 2007 Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, 18, 1-25.
- McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30, 23-40.
- Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2013). *Cognitive Diagnosis Modeling* (R package version 1.5-10). Retrieved from <https://sites.google.com/site/alexanderrobitzsch/software>
- Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.
- Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, 68, 78-98.
- Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, 32(2), 37-50.
- Wang, C., & Gierl, M. J. (2011). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, 48, 165-187.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.