Comparison of Relative Fit Indices for Diagnostic Model Selection

Applied Psychological Measurement
2017, Vol. 41(6) 422–438
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617695521
journals.sagepub.com/home/apm

Sedat Sen¹ and Laine Bradshaw²

Abstract

The purpose of this study was to thoroughly examine the performance of three information-based fit indices—Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and sample-size-adjusted BIC (SABIC)—using the log-linear cognitive diagnosis model and a set of well-known item response theory (IRT) models. Two simulation studies were conducted to examine the extent to which relative fit indices can identify the generating model under a variety of data conditions and model misspecifications. Generally, indices performed better when item quality was stronger. When the IRT was the generating model, all three indices correctly selected the IRT model for all replications. When the true model was a diagnostic classification model, for all three fit indices, the multidimensional IRT model was incorrectly selected as frequently as 70% of the replications. The results of this study identify situations for researchers where commonly used—and typically well-performing—fit indices may not be appropriate to compare models for selection.

Keywords

diagnostic classification models, log-linear cognitive diagnosis model, model selection, relative fit

Over the last two decades, psychometricians have refined tools that enable educational assessment routines to change from reporting a single overall test score to a more detailed, fine-grained evaluation that shows students' performances on specific skills. A number of cognitive diagnosis models (CDMs), also known as diagnostic classification models (DCMs), have been developed for skill profile reporting in recent years (see DiBello, Roussos, & Stout, 2007; Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). As DCMs are newly developed models, some key methodological features of the models have not yet been well studied. Therefore, some general recommendations are currently lacking for practitioners and researchers utilizing DCMs. One key area that is not yet well studied is how to evaluate model–data fit in DCMs. Model–data fit plays an important role in the use of DCMs. As with all statistical models, the results of models with poor fit cannot be meaningfully interpreted. Accurately estimating DCM parameters, and consequently making valid inferences from DCM results, requires adequate

¹Harran University, Sanliurfa, Turkey

²University of Georgia, Athens, USA

Corresponding Author:

Sedat Sen, Faculty of Education, Harran University, Sanliurfa, 63300, Turkey.

Email: sedatsen@harran.edu.tr

model–data fit. Furthermore, attractive properties of DCMs (see Bradshaw & Madison, 2016) are enabled by item and person parameter invariance, which is present only when the model fits the data. Therefore, models should be vetted to determine whether they fit the data and preferentially selected with respect to how well they fit the data.

As with other psychometric models, evaluation of DCMs requires the examination of model–data fit with an eye toward both statistical and nonstatistical considerations, which can be examined at an item or test level and in an absolute or relative sense. Absolute fit is examined to see whether the model adequately represents the data and is required before examining relative fit. Relative fit compares fit among two or more models that have reasonably adequate absolute fit. A number of methods have been proposed and studied for examining absolute fit for DCMs, including limited information goodness-of-fit statistics (Hansen, Cai, Monroe, & Li, 2016; Jurich, 2014), resampling methods (Templin & Henson, 2006), and posterior predictive model checking (Sinharay & Almond, 2007). Studies have also used relative fit indices, frequently used in statistics and psychometrics, to make comparisons among DCMs (e.g., Kunina-Habenicht, Rupp, & Wilhelm, 2012); however, a comprehensive study on the performance of relative fit indices for candidate diagnostic model comparisons has not been conducted, leaving a need for empirically based guidance on the use of relative fit indices in the context of diagnostic models.

This article aims to address the need for empirical research on evaluating the use of information indices to make relative model–data fit comparisons involving DCMs. Research in this area is limited in two main ways: First, the performance of the indices has not been used to select among models that include general DCMs. General DCMs, such as the log-linear cognitive diagnostic model (LCDM; R. Henson, Templin, & Willse, 2009; see also de la Torre, 2011 for reparameterization of the LCDM), are more highly parameterized than their submodels, which include the deterministic input, noisy “and” gate model (DINA; de la Torre, 2009; Junker & Sijtsma, 2001); the deterministic input, noisy “or” gate (DINO; Templin, 2006) model; the noisy input, deterministic “or” gate (NIDO; Templin, Henson, & Douglas, 2006) model; and the compensatory reparameterized unified model (C-RUM; Hartz, 2002). While the submodels appear much more frequently in the literature, rarely are model–data fit comparisons made with general DCMs and corresponding submodels prior to model–data fit selection. Such comparison is warranted to justify the use of the simpler submodels over the more general parent models. Second, the performance of the indices has not been used to compare DCMs with models in other psychometric families. DCMs differ from the most commonly applied family of psychometric models in educational assessment, item response theory (IRT) models, in that they estimate categorical instead of continuous latent variables to represent latent examinee characteristics. The latent variable distribution is an assumption that can be vetted by model comparisons of DCMs and IRT models. In this study, the utility of relative model–data fit indices was investigated to select the best-fitting models among DCMs and IRT models.

Model–Data Fit in DCMs

DCMs have some commonalities and differences with other latent variable models such as IRT and factor analytic (FA) models. The primary distinction between these models is the latent variable distribution: Latent variables in DCMs are assumed to be categorical while they are assumed to be continuous in IRT and FA. DCMs are typically confirmatory, multidimensional latent variable models that utilize items that measure one or more latent variables. The Q-matrix is used to show hypothesized relationships between items and categorical latent variables termed attributes. These relationships are established prior to analyses by domain experts, making the

model confirmatory. Items may measure more than one attribute (within-item dimensionality), and across items multiple attributes can be measured (between-item dimensionality).

Like in IRT, model–data fit for DCMs can be examined in terms of either absolute fit or relative fit. As stated in Chen, de la Torre, and Zhang (2013), several model–data fit indices have been proposed for absolute fit assessment of DCMs, including mean absolute differences (R. Henson et al., 2009), chi-square and G statistics, item discrimination statistics (de la Torre, 2008), and root mean square error of approximation (Kunina-Habenicht et al., 2012). As in other modeling approaches, relative fit comparison among models can be made using the deviance statistic (i.e., $-2\log\text{-likelihood} (-2LL)$), information criteria indices, and the Bayes factor (e.g., de la Torre & Douglas, 2004).

A number of researchers have examined the model–data fit issue for latent variable mixture models. Key references include J. M. Henson, Reise, and Kim (2007); Jedidi, Jagpal, and DeSarbo (1997); Li, Cohen, Kim, and Cho (2009); Lubke and Neale (2006); Nylund, Asparouhov, and Muthén (2007); Preinerstorfer and Formann (2012); Tofighi and Enders (2007); and Yang and Yang (2007). J. M. Henson et al. (2007) examined the accuracy of structural model parameter estimates in the latent variable mixture modeling context. Under different simulation conditions, these authors compared the efficacy of several relative model fit statistics, including the Akaike's Information Criterion (AIC; Akaike, 1974), the Bayesian Information Criterion (BIC; Schwarz, 1978), the sample-size adjusted BIC (SABIC), the consistent AIC (CAIC; Bozdogan, 1987), and the Vuong–Lo–Mendell–Rubin–adjusted (aVLMR) likelihood ratio (LR) test, in addition to comparing methods based on classification-based and distributional statistics. Their simulation study showed that the SABIC performed better than other relative fit indices. Lubke and Neale (2006) also found through a simulation study that the AIC and SABIC showed better performance than the BIC and CAIC indices in determining the correct mixture model. Tofighi and Enders (2007) examined the usefulness of several fit indices (e.g., AIC, BIC, SABIC, CAIC, SACAIC, Lo–Mendell–Rubin (LMR), MST, and MKT) for growth mixture models, and found that the SABIC and the LMR consistently extracted the correct number of latent classes. Yang and Yang (2007) evaluated performance of information criteria indices (AIC, BIC, CAIC, and SABIC) under different latent-class model conditions, and concluded that accuracy rates of AIC and BIC increased with sample size adjustment (CAIC and SABIC). Nylund et al. (2007) conducted a comprehensive simulation study to compare the performance of likelihood-based tests and information criteria indices (AIC, BIC, SABIC, and CAIC) using three mixture models (i.e., factor mixture models, latent-class models, and growth mixture models). Based on their simulation study, the BIC index performed better than other fit indices. The superiority of the BIC index over other information criteria indices was also shown by Li et al. (2009) and Preinerstorfer and Formann (2012) within mixture IRT models. Li et al. (2009) examined the performances of five fit indices (AIC, BIC, deviance information coefficient (DIC), pseudo-Bayes factor (PsBF), and posterior predictive model checks (PPMC)) using dichotomous mixture IRT models. Similarly, Preinerstorfer and Formann (2012) conducted a simulation study to examine two information criteria indices (AIC and BIC) in terms of model selection within mixture Rasch model context. Both simulation studies found that the BIC index is the most consistent index in determining the number of correct latent classes in mixture IRT models. Similarly, Jedidi et al. (1997) showed that the BIC is promising in determining the correct mixture structural equation model.

Despite the importance of model–data fit in terms of establishing a validity argument, it has been studied by only a few authors within the DCM context. DCMs are a special case of latent-class models, known as restricted latent-class models, where the number and nature of the latent classes are defined prior to analysis. Because they are members of the latent-class family, results from research on fit indices for latent-class models may generalize somewhat to DCMs,

though key differences may be uncovered through research. Three studies exploring these differences have been conducted by de la Torre and Douglas (2008), Kunina-Habenicht et al. (2012), and Chen et al. (2013). These studies are explained in detail below. In addition, a few researchers have examined the issue of item-level fit for DCMs (Lai, Gierl, & Cui, 2012; Sinharay & Almond, 2007; Wang, Shu, Shang, & Xu, 2015).

Relative fit indices have been commonly used to compare alternative models in the DCM context, the focus being on applying the indices for practical purposes of model selection not evaluating the theoretical performance of the indices. For example, de la Torre and Douglas (2008) evaluated model fit using four relative fit indices: the Bayes factor, the AIC, the BIC, and the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). They employed three DCMs that use a submodel form of the item response function (i.e., DINA; noisy-input, deterministic-and-gate model (NIDA), Junker & Sijtsma, 2001; and the multiple-strategy DINA model, de la Torre & Douglas, 2008) using a Markov Chain Monte Carlo (MCMC) Bayesian estimation algorithm. Although this study plays an important role in terms of demonstrating the application of model fit indices with Bayesian estimation techniques, the focus of that study was not on evaluating the performance of the model–data fit indices. To the authors’ knowledge, Kunina-Habenicht et al.’s (2012) study was the first that focused on comparing the performance of relative fit indices in a DCM context. They investigated the impact of model misspecifications on parameter estimation and model–data fit. The authors investigated the performance of two relative fit indices, the AIC and BIC, under different LCDM conditions with varying sample sizes, number of items, number of attributes, attribute correlations, and level of interactions between attributes. They found that the AIC and BIC indices showed good performance in determining the model with correct Q-matrix specification against the misspecified Q-matrices, in both overspecified and underspecified Q-matrix conditions. Chen et al. (2013) investigated the usefulness of various absolute and relative fit indices (i.e., $-2LL$, AIC, and BIC) under Q-matrix misspecifications and model misspecifications (e.g., generating model was DINA, and estimation model was DINO) using three constrained DCMs: the DINA, DINO, and additive CDM (A-CDM), each estimated using an identity link function. They found that the AIC and BIC showed good performance in determining the correct model under misspecifications of the model, Q-matrix, or both.

This study builds on previous studies of relative model–data fit to show the performance of three information-based fit indices (i.e., AIC, BIC, and SABIC; Sclove, 1987) in the context of a general DCM, the LCDM. As in Chen et al.’s (2013) study, the authors of the present study investigate model misspecification of attribute effects in the item parameterization, but they extend this work by examining model misspecifications that include general DCMs for the purpose of expanding the generalizability of their results beyond specific submodels and to the broader contexts of DCMs. The authors of the present study also extend previous work in the area by examining the performance of the indices to detect a type of model misspecification that has not yet been examined: misspecification of latent trait distributions. Two Monte Carlo simulation studies were conducted to answer the overarching research question:

Research Question 1: Do the relative fit indices show good performance in the context of DCMs under different model misspecification conditions?

The following section briefly presents details of the LCDM framework and relative fit indices used in this study. Following this, the simulation study design is described, and then results are reported that include the recovery of model parameters, classification accuracy percentages, and performance of the relative fit indices. Finally, recommendations are provided for the practical use of relative fit indices in the context of DCMs.

Log-Linear Cognitive Diagnostic Model

Although there are other general diagnostic models (generalized DINA (G-DINA) model, de la Torre, 2011; the general diagnostic model, von Davier, 2005), in this study, the authors focus on the LCDM framework. The LCDM subsumes most commonly used DCMs such as DINA, NIDA, DINO, NIDO, and C-RUM, and uses the canonical link function for binary response data. The LCDM specifies the conditional probability that examinee e with attribute pattern α_e provides a correct response to item i as

$$P(X_{ic} = 1 | \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i))}{1 + \exp(\lambda_{i,0} + \lambda_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i))}, \quad (1)$$

where $\mathbf{q}_i$ is a A -length row of the Q -matrix i , such that $q_{ia} = 1$ when item i measures attribute a and 0 otherwise. $\lambda_{i,0}$ is the intercept interpreted as the logit of a correct response for an examinee who has not mastered any of the attributes required by item i . The vector λ_i represents a vector of size $(2^A - 1) \times 1$, elements of which contain the main effect and interaction parameters for item i . In addition, $\mathbf{h}(\alpha_c, \mathbf{q}_i)$ is a vector of size $(2^A - 1) \times 1$ of linear combinations of α_c and $\mathbf{q}_i$, such that elements serve as an indicator for which effects should be present on each item for each attribute pattern. Other DCMs can be obtained by making constraints on the model parameters λ_i , some of which are demonstrated below.

To demonstrate the LCDM and the three most common submodels, assume a diagnostic assessment that measures four arithmetic skills—addition, subtraction, multiplication, and division—each defined as an attribute. Suppose Item 2 measures multiplication (Attribute 3) and division (Attribute 4; for example, $4 \times 7 / 2 = ?$). Based on the LCDM, the item response function is

$$P(X_{2c} = 1 | \alpha_{c3}, \alpha_{c4}) = \frac{\exp(\lambda_{2,0} + \lambda_{2,1(3)}\alpha_{c3} + \lambda_{2,1(4)}\alpha_{c4} + \lambda_{2,2,(3,4)}\alpha_{c3}\alpha_{c4})}{1 + \exp(\lambda_{2,0} + \lambda_{2,1(3)}\alpha_{c3} + \lambda_{2,1(4)}\alpha_{c4} + \lambda_{2,2,(3,4)}\alpha_{c3}\alpha_{c4})}, \quad (2)$$

where α_{c3} and α_{c4} are latent variables for Attributes 3 and 4, respectively. The logit of the correct response includes the intercept ($\lambda_{2,0}$), a main effect for Attribute 3 ($\lambda_{2,1,(3)}$), a main effect for Attribute 4 ($\lambda_{2,1,(4)}$), and a two-way interaction term between Attributes 3 and 4 ($\lambda_{2,2,(3,4)}$). The second subscript on the item parameters denotes the level of the effect, where 0 represents the intercept, 1 represents the main effect, and 2 represents the two-way interaction. The parenthetical subscripts for the main effects and interaction terms list the attribute(s) corresponding to each effect. If the authors of the present study used the DINA model specifications for their example Item 2, students would only be expected to answer the item correctly if they had mastered both attributes (division and multiplication). This assumption is parameterized by setting the main effects to equal 0 and estimating an interaction term that is positive. Based on these properties, the following LCDM equation can be used for DINA model parameter estimation:

$$P(X_{2c} = 1 | \alpha_{c3}, \alpha_{c4}) = \frac{\exp(\lambda_{2,0} + \lambda_{2,2,(3,4)}\alpha_{c3}\alpha_{c4})}{1 + \exp(\lambda_{2,0} + \lambda_{2,2,(3,4)}\alpha_{c3}\alpha_{c4})}. \quad (3)$$

Unlike the DINA model, for the DINO model, students are expected to get Item 2 correct if they master at least one of the two attributes required. In the DINO model, the probability of a correct response for Item 2 is equal to Equation 2 for the LCDM, with the constraint of $\lambda_{2,1,(3)} = \lambda_{2,1,(4)} = -\lambda_{2,2,(3,4)}$. This constraint results in the probability of a correct response for students who have mastered either one of the two attributes being equal to the probability of correct response for students who have mastered both attributes:

$$\begin{aligned}
 P(X_{2c} = 1 | \alpha_{c3} = 0, \alpha_{c4} = 1) &= P(X_{2c} = 1 | \alpha_{c3} = 1, \alpha_{c4} = 0) \\
 &= P(X_{2c} = 1 | \alpha_{c3} = 1, \alpha_{c4} = 1) = \frac{\exp(\lambda_{2,0} + \lambda_2)}{1 + \exp(\lambda_{2,0} + \lambda_2)},
 \end{aligned} \tag{4}$$

where $\lambda_2 = \lambda_{2,1,(3)} = \lambda_{2,1,(4)} = -\lambda_{2,2,(3,4)}$. Thus, like the DINA model, only two parameters are estimated per item in the DINO model. In the DINA model, those two parameters are the intercept and the highest-level interaction term, and in the DINO model, it is the intercept ($\lambda_{2,0}$) and a common parameter (λ_2). Finally, consider Item 2 for the C-RUM. The C-RUM does not include any interaction effects, resulting in the parameterization:

$$P(X_{2c} = 1 | \alpha_{c3}, \alpha_{c4}) = \frac{\exp(\lambda_{2,0} + \lambda_{2,1,(3)}\alpha_{c3} + \lambda_{2,1,(4)}\alpha_{c4})}{1 + \exp(\lambda_{2,0} + \lambda_{2,1,(3)}\alpha_{c3} + \lambda_{2,1,(4)}\alpha_{c4})}. \tag{5}$$

Estimated parameters in C-RUM include an intercept ($\lambda_{2,0}$) and the main effect parameters.

Relative Fit Indices

Several relative fit statistics are available for comparing fit among two or more competing models. In general, information criterion indices or LR test statistics can be used within the latent variable modeling context. The former is appropriate for comparison of nested or non-nested models, while the latter can be used but only for nested models. The authors of the present study focus on information criterion indices in this study, specifically on the AIC, BIC, and SABIC. Multiple indices are often used to compare the model–data fit among a set of models for the same data when maximum-likelihood (ML) estimates of the parameters are obtained. Information criteria indices are based on some form of penalization of the likelihood function. Smaller values of information criterion indices indicate better fit. However, different indices may select different best-fitting models for the same data due to differences in the penalty function applied to the likelihood. The AIC is given by

$$\text{AIC} = -2\log L + 2p, \tag{6}$$

where p is the number of estimated parameters and $2p$ is used as a penalty for overparameterization, and L is the likelihood function. For DCMs, L represents the marginal likelihood function value of the model, and p includes the total number of item parameters and structural parameters. One problem with the AIC is that it tends to select more complex models. The lack of penalization for sample size leads to inconsistency in the performance of the AIC and a tendency for it to overestimate the correct number of classes (McLachlan & Peel, 2000). However, AIC is asymptotically efficient if the true model is not among the models being compared (Vrieze, 2012). To account for sample size, the BIC can be used and is given by

$$\text{BIC} = -2\log L + p \ln(N), \tag{7}$$

where L is the likelihood of the estimated model with p free parameters and $\ln(N)$ is the logarithmic function of the total sample size N . As can be seen in Equation 7, the penalty function for the BIC is based on the number of estimated parameters as well as the sample size. The BIC tends to apply more penalization to the likelihood function than the AIC when complex models are estimated. As a result, the BIC is more likely to select simpler models than the AIC due to the inclusion of the sample size in the penalty function: “The penalty of BIC with N makes statistical significance more and more difficult to achieve as the sample size increases” (Vrieze, 2012, p. 233). Namely, unlike the AIC, the BIC is said to be consistent asymptotically

(Shao, 1997), which means that as the sample size increases, the BIC tends to select the correct number of mixture classes consistently if the true model is among the models being compared (Vrieze, 2012). In contrast, as the sample size increases, the AIC tends to select a more complex model even when the true model is in the candidate model set. Shao's study on model-selection indices showed that the usefulness of the AIC and BIC may depend on the structure of the true model. The BIC is expected to perform better when the true model has a simple structure, and the AIC is expected to work better when the true model has a complex structure. In addition to the true model, the usefulness of the AIC and BIC may depend on several factors, including the loss function, study design, and research question (Vrieze, 2012). In latent variable mixture models, the complexity of the true model, class separation, and class proportions have been shown to affect the usefulness of these two fit indices (Lubke & Neale, 2006; Nylund et al., 2007; Vrieze, 2012). As mentioned above, the BIC performed better than the AIC within the latent variable model context (Jedidi et al., 1997; Li et al., 2009; Nylund et al., 2007; Preinerstorfer & Formann, 2012).

Adjustments for model complexity have been proposed for both the AIC and BIC in the literature. The adjusted version of the AIC (i.e., CAIC) was proposed by Bozdogan (1987), and the adjusted version of the BIC (i.e., SABIC) was proposed by Sclove (1987). The SABIC is similar to the BIC but uses $[N^* = (N + 2) / 24]$ instead of N to reduce the penalty on the sample size. As Yang (2006) stated, the SABIC is expected to perform well with large number of parameters or small sample sizes. Several simulation studies showed the superiority of the SABIC over other fit indices for model selection among latent variable mixture models (J. M. Henson et al., 2007; Tofighi & Enders, 2007; Yang, 2006).

Method

The performance of three relative fit indices was investigated for selecting appropriate models under various testing conditions that involved DCMs. The present study was divided into two simulation studies: one that focused on multidimensional models (Study I with 48 conditions) and one that focused on unidimensional models (Study II with 28 conditions). The primary results of the present study are the rankings of various estimation models according to the AIC, BIC, and SABIC, which was evaluated to determine how frequently the fit indices were able to select the true model underlying the data.

Simulation Study Design

Overall, seven different factors were manipulated with varying conditions. The following conditions were simulated: sample size (1,000), test length (12 and 24 items), number of attributes (1 and 3), base rate (0.25 and 0.50), tetrachoric correlation (.7), generating model type (saturated LCDM, reduced LCDM, Rasch model, two-parameter logistic (2-PL) model, and three-parameter logistic (3-PL) IRT model), and estimation model type (six to eight different models, described below in detail). One hundred replications were simulated for each condition.

Constant Factors

Study I used three-dimensional models in all conditions, and Study II used one-dimensional models in all conditions. For Study I, the correlation among dimensions was held constant at .70. This correlation is a pairwise tetrachoric correlation corresponding to the multivariate Bernoulli distributions specified as the true attribute profile distributions. Two multivariate

Table 1. Q-Matrix and Corresponding Item-Level Models for Conditions With Three Attributes.

Item	Attribute 1	Attribute 2	Attribute 3	LCDM-S	LCDM-R
1	0	0	1	LCDM	LCDM
2	0	0	1	LCDM	LCDM
3	0	0	1	LCDM	LCDM
4	0	0	1	LCDM	LCDM
5	0	1	0	LCDM	LCDM
6	0	1	0	LCDM	LCDM
7	0	1	0	LCDM	LCDM
8	0	1	0	LCDM	LCDM
9	1	0	0	LCDM	LCDM
10	1	0	0	LCDM	LCDM
11	1	0	0	LCDM	LCDM
12	1	0	0	LCDM	LCDM
13	1	0	1	LCDM	DINA
14	1	0	1	LCDM	DINO
15	1	0	1	LCDM	C-RUM
16	1	0	1	LCDM	LCDM
17	0	1	1	LCDM	DINA
18	0	1	1	LCDM	DINO
19	0	1	1	LCDM	C-RUM
20	0	1	1	LCDM	LCDM
21	1	1	0	LCDM	DINA
22	1	1	0	LCDM	DINO
23	1	1	0	LCDM	C-RUM
24	1	1	0	LCDM	LCDM

Note. LCDM = log-linear cognitive diagnostic model; LCDM-S = saturated LCDM; LCDM-R = reduced LCDM; DINA = deterministic input, noisy “And” gate; DINO = deterministic input, noisy “Or” gate; C-RUM = compensatory reparameterized unified model; MIRT = multidimensional IRT; IRT = item response theory.

Bernoulli distributions were generated to have a different base rate (as described below and in Table 2) but keeping the correlation constant at .70.

For both Studies I and II, the sample size was held constant at 1,000, and the simulation was based on an assessment that measured each dimension with 12 items. The Q-matrix for Study I is given in Table 1, which shows that the Q-matrix is balanced with respect to complexity of the items measuring each attribute and the number of items measuring each attribute: Each attribute is measured alone 4 times and in conjunction with one other attribute 8 times.

Manipulated Factors

Both Studies I and II contain four manipulated factors: generating model types (varies by Studies I and II), quality of test items (medium and high), base rate of attribute mastery (0.25 and 0.50), and the estimation model types (varies by Studies I and II).

Generating models. For Study I, two models were used to generate item response data: the saturated LCDM and the reduced LCDM. For the saturated LCDM, a fully factorial LCDM structure was simulated to underlie the complex items in the Q-matrix, meaning that Items 13 to 24 each had four nonzero parameters: an intercept, a main effect for each measured attribute, and a two-way interaction between the two attributes (see Table 1). For the reduced LCDM, the underlying DCM structure was generated to differ across the complex items with the three most

Table 2. Attribute Profile Distributions Under Generating DCMs.

Base rate = 0.25			Base rate = 0.50		
Profile	Pattern	Proportion	Profile	Pattern	Proportion
1	[000]	.593	1	[000]	.293
2	[001]	.069	2	[001]	.075
3	[010]	.069	3	[010]	.075
4	[011]	.027	4	[011]	.054
5	[100]	.069	5	[100]	.075
6	[101]	.027	6	[101]	.054
7	[110]	.027	7	[110]	.054
8	[111]	.119	8	[111]	.321

Note. Attribute patterns equal $[\alpha_1\alpha_2\alpha_3]$. DCMs = diagnostic classification models.

common submodels of the LCDM represented: Three complex items were generated to be like DINA items (Items 13, 17, and 21), three to be like DINO items (Items 14, 18, and 22), three like C-RUM items (15, 19, and 23), and three like the saturated LCDM (16, 20, and 24). The four generating models for Study II were the one-attribute DCM, the Rasch model, the 2-PL IRT model, and the 3-PL IRT model. With only one attribute, all DCMs are equivalent, so that no distinction is made in Study II with respect to whether the model was the LCDM, DINO, and so on.

Quality of test items. Both Studies I and II manipulated the quality of items at two levels, termed medium and high. The higher quality items have greater item discrimination, meaning that the items are better at separating masters and nonmasters of the measured attributes. For DCMs, item discrimination is defined as the difference in the probability of a correct response for two groups of students. As shown in Table 1A (see the online appendix), for simple structure items, the item discrimination was set to be 0.60 (i.e., 0.85-0.25) for the medium-quality condition and 0.64 (i.e., 0.82-0.18) for the high-quality condition. For the complex items, the item discrimination for masters of both attributes and nonmasters of either attribute remained at 0.60 (i.e., 0.85-0.25) for the medium-quality condition and increased to 0.83 (i.e., 0.95-0.12) for the higher quality condition. Table 2A (see the online appendix) provides the item parameter values that correspond to the item response probabilities in Table 1A. The values in the higher quality condition are mimicking the values used by Kunina-Habenicht et al. (2012) to allow for comparisons of results. Because item quality from DCM real data examples in the literature is typically lower than that simulated by Kunina-Habenicht et al. (2012), the medium-quality condition was added for contrast. It was anticipated that as the item quality improves, it will be easier to detect the true model as the better fitting model. For the IRT models, the item quality was not manipulated. Table 3A (see the online appendix) provides the generating item parameters for the IRT models.

Base rate of mastery. The base rate of mastery for an attribute is the proportion of examinees who have mastered the attribute in the population. Two levels for the base rate factor were used: 0.25 and 0.50. A base rate of 0.50 is commonly used in the literature. The 0.25 conditions were added to investigate model fit comparison under less optimal conditions. It was hypothesized that the less even distribution of masters and nonmasters of each attribute would increase estimation difficulty, and consequently influence the model fit results. Table 2 shows the attribute profile distributions that result from fixing the pairwise attribute correlations equal to .70 under each base rate condition.

Estimation models. In Study I, the data generated from both the saturated and reduced LCDMs were estimated with six models: the saturated LCDM, the reduced LCDM, the DINA model, the DINO model, the C-RUM, and a multidimensional IRT (MIRT) model. For Study II, each of the four generating models was used as estimation models for all conditions. For example, LCDM-generated data sets were estimated with four different models, including the LCDM, Rasch, 2-PL, and 3-PL IRT models. All estimations were carried out with marginal ML estimation using Mplus 7.0 (Muthén, & Muthén, 1998/2014).

Evaluation Criteria

Prior to investigating the performance of the fit indices, the estimation accuracy of item and examinee parameters was examined. Recovery of item parameters was assessed using the root mean square error (RMSE). Recovery of examinee parameters was assessed using the classification accuracy of the attribute profiles. Classification accuracy is the percentage of examinees whose generated (true) attribute profile was equal to the estimated attribute profile. To evaluate the performance of the relative model–data fit indices, the AIC, BIC, and SABIC values were calculated for all replications. For each replication, the first, second, and third model choice was recorded for each index in Study I, whereas first through the fourth model choices were recorded in Study II. Results show how often fit indices were able to determine the true generating model, as well as which models were incorrectly selected as the best-fitting model when the correct model was not identified.

Results

Recovery of Item Parameters

Mean RMSE values of generated diagnostic models for Studies I and II are presented in Table 4A (see the online appendix). Results are in line with previous studies (e.g., Bradshaw & Cohen, 2010). RMSE values for the intercept parameter ranged from 0.09 to 0.172, indicating that parameter estimates for the intercept were recovered well for most of the models. As expected, recovery of interaction parameters (0.48–0.95) was found to be worse than the recovery of the intercept (0.09–0.17) and main effect (0.25–0.44) parameters. As expected, greater accuracy was observed with higher item quality conditions, and parameters were recovered more accurately under the simpler reduced LCDM when compared with the more complex saturated LCDM. Diagnostic models with one attribute, as well as the IRT models, had more accurate parameter recovery compared with models with three attributes. The mean RMSE values of the item difficulty parameter were 0.103, 0.188, and 0.249 for the Rasch model, 2-PL, and 3-PL IRT models, respectively. Item discrimination parameters produced mean RMSE values of 0.136 and 0.145 for 2-PL and 3-PL IRT models, respectively. The mean RMSE value for the pseudo-guessing parameter was 0.009.

Classification Accuracy

Table 4A provides the attribute profile classification accuracy rates in the last column. Classification accuracy rates for the three-attribute conditions were similar across the base rate conditions, ranging from 74.40% to 97.40% for conditions with a base rate of 0.25, and from 70.05% to 95.40% for conditions with a base rate of 0.50. Classification accuracy rates for the one-attribute conditions ranged from 83.13% to 94.70% for conditions with a base rate of 0.25, and from 78.70% to 94.20% for conditions with a base rate of 0.50. As with item parameter

recovery, classification accuracy was stronger under conditions with higher item quality and the simpler model (i.e., the reduced LCDM).

Relative Fit Evaluation

Table 3 provides results on the performance of the AIC, BIC, and SABIC in Study I. Under the different conditions, the values in Table 3 show the proportions of replications for which each fit index ranked each estimation model. For example in Row 1, under the conditions with a base rate of 0.25, medium item quality, and the true model being the saturated LCDM, the AIC correctly selected the estimated saturated LCDM as the best-fitting model in 19% of the replications, but incorrectly selected the estimated saturated LCDM as the second best-fitting model in 47% of the replications and as the third best-fitting model in 34% of the replications.

Performance of the AIC in Study I. Table 3 shows that the AIC index showed almost perfect performance (96% accurate) in selecting true saturated LCDM when the item quality was high and the base rate was 0.50. Under the lower base rate conditions of 0.25, this accuracy reduced to 87%. When item quality was medium, however, the AIC was not able to select the correct model. The AIC correctly selected the saturated LCDM only 29% and 19% under the 0.50 and 0.25 base rates, respectively. Under these conditions, the AIC was most likely to incorrectly select the MIRT model at rates of 70% and 63%, respectively, and the AIC was often likely to select the C-RUM as the second best-fitting model behind the MIRT model. The AIC showed poor performance in selecting the true reduced LCDM for most of the conditions. The MIRT model was incorrectly selected over the true reduced LCDM 66% to 69% of the time under medium item quality, and the saturated LCDM was incorrectly selected as the true model almost as frequently as the reduced LCDM under high item quality with a 0.50 base rate. Item quality affected the performance of AIC index for data conditions generated with reduced LCDM, with the true model being selected more frequently under the conditions with the higher item quality.

Performance of the BIC in Study I. When the saturated LCDM was the true model, the BIC index showed poor performance in selecting the true generating model in all conditions, with accuracy near 0% in three conditions and 39% in the high-quality item with 0.50 base rate condition (see Table 3). The BIC index incorrectly selected the C-RUM (89% and 60%) over the true generating model under high item quality conditions and incorrectly selected the MIRT (63% and 70%) and DINA (32% and 25%) models under medium item quality conditions. As shown in Table 3, the BIC index showed good performance in selecting the true reduced LCDM under high item quality conditions, with perfect accuracy for the 0.50 base rates and an 80% accuracy under the 0.25 base rate. However, performance of the BIC index was low for selecting the reduced LCDM under medium item quality conditions, ranging from 18% to 20% accuracy.

Performance of the SABIC in Study I. As shown in Table 3, the SABIC showed good performance under high item quality conditions with the saturated LCDM as the true model when the base rate was 0.50 (90% accuracy) and much lower performance when the base rate was 0.25 (65% accuracy). Under high item quality, the C-RUM was most often incorrectly selected as the true model (8%-27% of the replications). Under medium item quality, the MIRT model was selected over the true saturated LCDM 63%-70% of the time. Consistent results were observed with the reduced LCDM conditions where the SABIC showed good performance under high item quality conditions with a 0.50 base rate and much weaker performance under the other conditions.

Performance of fit indices in Study II. Results for the performance of the three fit indices under conditions where the true model was the one-attribute LCDM are presented in Table 4. Under high item quality conditions and 0.50 base rates when the one-dimensional LCDM was the true

Table 3. Study I Results Based on Three Fit Indices (AIC, BIC, and SABIC).

		True model																	
		LCDM-S						LCDM-R						LCDM-S					
		Model fit rank based on AIC						Model fit rank based on BIC						Model fit rank based on SABIC					
BR	Item Q	Est. model	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2
0.25	Medium	LCDM-S	.34	.47	.19	.31	.37	.18	.00	.00	.00	.25	.24	.12	.39	.10	.00	.31	.34
		LCDM-R	.00	.00	.00	.19	.48	.15	.00	.00	.00	.12	.53	.20	.01	.00	.00	.18	.50
		DINA	.29	.08	.02	.05	.01	.00	.4	.63	.32	.15	.02	.00	.27	.31	.21	.04	.02
		DINO	.03	.00	.00	.15	.02	.00	.46	.21	.03	.26	.07	.00	.07	.03	.00	.18	.03
		C-RUM	.34	.45	.16	.30	.10	.01	.50	.16	.02	.22	.12	.02	.26	.56	.16	.29	.09
	High	MIRT	.00	.00	.63	.00	.02	.66	.00	.00	.63	.00	.02	.66	.00	.00	.63	.00	.02
		LCDM-S	.01	.12	.87	.15	.76	.09	.06	.90	.03	.23	.77	.00	.03	.32	.65	.20	.76
		LCDM-R	.15	.00	.00	.00	.22	.78	.73	.00	.00	.00	.20	.00	.42	.00	.00	.00	.19
		DINA*	.00	.00	.00	.00	.00	.00	.05	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		DINO*	.00	.00	.00	.10	.02	.12	.11	.00	.00	.06	.02	.19	.00	.00	.00	.05	.05
0.5	Medium	C-RUM*	.06	.87	.05	.75	.00	.00	.01	.09	.89	.71	.01	.00	.04	.67	.27	.75	.00
		MIRT	.78	.00	.00	.00	.00	.00	.04	.00	.00	.00	.00	.00	.51	.00	.00	.00	.00
		LCDM-S	.13	.58	.29	.34	.31	.18	.20	.00	.00	.36	.28	.10	.43	.23	.07	.32	.29
		LCDM-R	.00	.00	.00	.26	.64	.10	.00	.00	.00	.16	.66	.18	.00	.00	.00	.23	.66
		DINA	.38	.01	.00	.00	.00	.00	.05	.70	.25	.04	.00	.00	.35	.30	.06	.01	.00
	High	DINO	.00	.00	.00	.00	.00	.00	.10	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00
		C-RUM	.49	.41	.01	.40	.03	.00	.65	.30	.05	.43	.04	.00	.22	.47	.17	.44	.03
		MIRT	.00	.00	.70	.00	.02	.69	.00	.00	.70	.00	.02	.69	.00	.00	.70	.00	.69
		LCDM-S	.00	.04	.96	.00	.53	.47	.01	.60	.39	.00	1.00	.00	.00	.10	.90	.00	.86
		LCDM-R	.00	.00	.00	.00	.47	.53	.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00	.14
		DINA	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		DINO	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		C-RUM	.02	.96	.02	.99	.00	.00	.00	.40	.60	1.00	.00	.00	.02	.90	.08	1.00	.00
		MIRT	.98	.00	.02	.01	.00	.00	.99	.00	.01	.00	.00	.00	.98	.00	.02	.00	.00

Note. Values represent the proportion of replications for which the model was selected for each of the top three ranks, among the set of estimated models. The bold font cells correspond to the true model for each condition. The conditions with asterisks lacked convergence for some of the analyses. AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SABIC = sample-size-adjusted BIC; LCDM = log-linear cognitive diagnostic model; LCDM-S = saturated LCDM; LCDM-R = reduced LCDM; BR = base rate; Item Q = item quality; Est. model = estimation model; DINA = deterministic input, noisy "And" gate; DINO = deterministic input, noisy "Or" gate; C-RUM = compensatory reparameterized unified model; MIRT = multidimensional IRT; IRT = item response theory.

Table 4. Study II Results: Selection Rates of AIC, BIC, and SABIC for One-Dimensional Models.

BR	Item Q.	Est. model	AIC				BIC				SABIC			
			4	3	2	1	4	3	2	1	4	3	2	1
0.25	Medium	LCDM-I	.00	.11	.56	.33	.00	.44	.49	.07	.00	.20	.63	.17
		Rasch	.17	.05	.13	.65	.00	.08	.02	.90	.09	.07	.01	.83
		2-PL	.04	.63	.31	.02	.00	.48	.49	.03	.00	.64	.36	.00
		3-PL	.79	.21	.00	.00	1.00	.00	.00	.00	.91	.09	.00	.00
	High	LCDM-I	.00	.00	.00	1.00	.00	.00	.62	.38	.00	.00	.01	.99
		Rasch	.02	.11	.87	.00	.00	.00	.38	.62	.00	.00	.99	.01
		2-PL	.23	.70	.07	.00	.00	1.00	.00	.00	.00	1.00	.00	.00
		3-PL	.75	.19	.06	.00	1.00	.00	.00	.00	1.00	.00	.00	.00
0.5	Medium	LCDM-I	.00	.11	.45	.44	.00	.51	.45	.04	.00	.19	.58	.23
		Rasch	.24	.17	.09	.50	.00	.10	.01	.89	.08	.18	.05	.69
		2-PL	.00	.48	.46	.06	.00	.39	.54	.07	.00	.55	.37	.08
		3-PL	.76	.24	.00	.00	1.00	.00	.00	.00	.92	.08	.00	.00
	High	LCDM-I	.00	.00	.00	1.00	.00	.00	.01	.99	.00	.00	.00	1.00
		Rasch	.00	.01	.99	.00	.00	.00	.99	.01	.00	.00	1.00	.00
		2-PL	.00	.99	.01	.00	.00	1.00	.00	.00	.00	1.00	.00	.00
		3-PL	1.00	.00	.00	.00	1.00	.00	.00	.00	1.00	.00	.00	.00

Note. Values represent the proportion of replications for which the model was selected for each of the four ranks, among the set of four estimated models. The bold font cells correspond to the true model for each condition. AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SABIC = sample-size-adjusted BIC; BR = base rate; Item Q = item quality; LCDM = log-linear cognitive diagnostic model; LCDM-I = LCDM with one attribute; 2-PL = two-parameter logistic; 3-PL = three-parameter logistic.

model, all three indices had perfect (or nearly perfect) performance in selecting the LCDM as the best-fitting model. The AIC and SABIC were also perfect for the high item quality and lower base rate conditions. However, in all of the remaining conditions, all three indices performed poorly, each selecting the Rasch model as the correct model much more frequently (up to 65%, 90%, and 83% of replications for the AIC, BIC, and SABIC, respectively) than the LCDM. The remaining conditions for Study II used one of three IRT models as the generating model. These data sets were estimated using the one-attribute LCDM and the three unidimensional IRT models. When data were generated with an IRT model, all fit indices performed perfectly, selecting the true IRT model rather than the one-attribute LCDM for every replication, regardless of whether the generating model was the Rasch model, 2-PL, or 3-PL IRT model.

Discussion

Model–data fit is a critical issue for any psychometric model. As model–data fit improves, the argument for the validity of inferences based upon results is strengthened. Comparing the relative fit of competing models is a common practice to improve model–data fit. In this study, two simulation studies were used to examine the performance of three relative fit indices in determining the true diagnostic model over misspecified DCMs and over IRT models. The designs in this study reflected two common situations for comparing models in the context of diagnostic modeling. The first is to figure out which DCM parameterization to use among the many available options. Researchers face the task of figuring out if reduced versions of the LCDM are more parsimonious than the saturated LCDM; as in other statistical modeling contexts, overfitting a model may lead to estimation complexities and/or inaccuracies. The second is to figure out whether the assumption of binary latent traits is appropriate, or whether it is more appropriate to use a continuous latent variable distribution as in an IRT model.

Overall, results showed a range of conditions where conventional relative fit indices cannot be trusted to select among competing parameterizations of DCMs, nor to compare IRT and DCM models. Generally, indices performed better when item quality was stronger. The AIC, BIC, and SABIC were particularly poor in selecting the correct model when item quality was only medium, with the most alarming result being that when item quality was medium and the true model was the saturated LCDM: The BIC never selected the true model, and the AIC and SABIC only selected the true model with 29% and 7% accuracy, respectively. Results showed, however, that strong item quality did not ensure accurate results from the fit indices. Under high-quality items, the AIC, BIC, and SABIC still dipped down to 53%, 3%, and 65% accuracy.

Complexity of the true underlying model also played a role in the performance of the fit indices. The reduced LCDM was more frequently correctly identified by all three fit indices as the correct model than was the saturated LCDM. This result is due in part to the difficulty in estimating interaction terms in the saturated LCDM; as shown by the RMSE results, the recovery of the interaction terms is difficult, particularly when item quality is medium. Often under the conditions where the saturated LCDM was the true model and the item quality was medium, the C-RUM, which is equivalent to the saturated LCDM without any interaction terms, was selected as a better fitting model. Results, however, showed that the poor performance of the fit indices was not entirely due to the model complexity; the indices did not perform consistently well when the true generating model was the one-dimensional LCDM. Although the indices performed almost perfectly when item quality was high in the one-attribute models, accuracy drastically dropped to as low as 4% under medium-quality items.

Across all conditions, the fit indices tended toward selecting the true model to be an IRT model. When the IRT was the generating model, all three indices correctly selected the IRT model for all replications and all IRT model specifications (i.e., Rasch, 2-PL, and 3-PL). The indices, however,

also frequently selected the IRT models when they were not the generating models. When the true model was a DCM, for all three fit indices, the MIRT model was incorrectly selected as frequently as 70% of the replications under the three-attribute conditions, and the Rasch model was selected up to 65%, 90%, and 83% of the replications by the AIC, BIC, and SABIC, respectively, under the one-attribute conditions. These results demonstrated that holding other factors constant, the indices did not perform well in detecting the true underlying distribution of the dimension. Further research in this area may include an analytical explanation for why the fit indices more commonly select the continuous underlying dimension when the true model is a binary underlying dimension. The authors hypothesize that the penalty parameters do not account for variable *type*, and continuous variables are able to explain more variance than categorical variables, though they have the same number of item parameters in the DCM and (M)IRT forms.

Together, findings in this study illustrated that under a variety of plausible testing conditions, there exists a sizable chance of choosing the wrong model when relative fit indices are used to select among DCM parameterizations, or used to compare DCMs with IRT models. While a single study cannot simulate all testing conditions, a wide range of conditions were explored to demonstrate a concern with the accuracy of commonly used fit indices. The authors did not explore important factors expected to affect fit indices but rather demonstrated issues with performance within a set of practical conditions. Based on these results, they recommend in similar practical testing conditions that model–data comparisons involving non-nested DCMs rely on multiple sources of evidence and avoid overweighting the importance of the relative fit indices explored in this research. Future work may examine how additional factors that were held constant (e.g., sample size, Q-matrix features) may influence the performance of fit indices in the context of model comparisons involving DCMs. Results may find that, in other conditions, the relative performance of the different fit indices changes, that the performance is also a concern, or that the indices perform adequately. In the absence of this future work, researchers seeking to use these indices to select a best-fitting model involving one or more DCMs may need to examine the performance of the indices under the researcher's conditions via simulation prior to using the indices to make model-selection decisions.

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.

Supplemental Material

The online appendix is available at <http://journals.sagepub.com/doi/suppl/10.1177/0146621617695521>.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.
- Bradshaw, L., & Cohen, A. (2010, May). Accuracy of multidimensional item response model parameters estimated under small sample sizes. In A. Izsák (Chair), *Using cognitive attributes to develop*

- mathematics assessments, opportunities, and challenges*. Symposium conducted at the annual American Educational Research Association conference in Denver, CO.
- Bradshaw, L., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, 16, 99-118.
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50, 123-140.
- 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. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115-130.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73, 595-624.
- 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.
- Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response theory models. *British Journal of Mathematical and Statistical Psychology*, 69, 225-252.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana, IL.
- Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 202-226.
- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16, 39-59.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- Jurich, D. P. (2014). *Assessing model fit of multidimensional item response theory and diagnostic classification models using limited-information statistics* (Unpublished doctoral dissertation). James Madison University, Harrisonburg, VA.
- 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.
- Lai, H., Gierl, M. J., & Cui, Y. (2012, April). *Item consistency index: An item-fit index for Cognitive Diagnostic Assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.
- Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33, 353-373.
- Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, 41, 499-532.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
- Muthén, L. K., & Muthén, B. O. (1998/2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.
- Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, 65, 251-262.

- Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic models: A review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6, 219-262.
- Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.
- Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7, 221-242.
- 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. Statistical Methodology*, 64, 583-639.
- Templin, J. (2006). CDM: Cognitive diagnosis modeling with Mplus [Computer software]. Retrieved from <http://jtemplin.coe.uga.edu/research/>
- Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.
- Templin, J., Henson, R., & Douglas, J. (2006, April). *General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates*. Paper presented at the 2007 National Council on Measurement in Education training session, Chicago, IL.
- Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 317-341). Greenwich, CT: Information Age.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Technical Report No. RR-05-16). Princeton NJ: Educational Testing Service.
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17, 228-243.
- Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item-level fit for the DINA model. *Applied Psychological Measurement*, 39, 525-538.
- Yang, C. C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50, 1090-1104.
- Yang, C. C., & Yang, C. C. (2007). Separating latent classes by information criteria. *Journal of Classification*, 24, 183-203.