An Evaluation of Methods for Assessing Model Fit for Bayesian Diagnostic Classification Models

W. Jake Thompson

Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas

Author Note

W. Jake Thompson  http://orcid.org/0000-0001-7339-0300

The author has no conflict of interest to declare. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210045 to the University of Kansas. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to W. Jake Thompson, Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas, 1122 West Campus Road, Lawrence, KS 66045-3101, Email: jakethompson@ku.edu

Abstract

Diagnostic classification models (DCMs) are psychometric models that can be used to estimate the presence or absence of psychological traits, or proficiency on fine-grained skills. Critical the use of any psychometric model in practice, including DCMs, is an evaluation of model fit. Traditionally, DCMs have been estimated with maximum likelihood methods and then evaluated with limited-information fit indices. However, recently, methodological and technological advancements have made Bayesian methods for estimating DCMs more accessible. When using a Bayesian estimation process, new methods for model evaluation are available to assess model fit. In the current study, we conduct a simulation study to compare the performance of the traditional measures of model fit to Bayesian methods. The results indicate that Bayesian measures of model fit generally outperform the more traditional limited-information indices. Notably, flags for model misfit were more likely to be true positives when using Bayesian methods. Additionally, Bayesian methods for model comparisons also showed better performance than has been reported for methods traditionally in conjunction with a maximum likelihood estimation. In summary, the findings suggest that Bayesian methods offer a better evaluation of model fit than more commonly used metrics.

Keywords: diagnostic assessment, model fit, classification

An Evaluation of Methods for Assessing Model Fit for Bayesian Diagnostic Classification Models

Diagnostic classification models (DCMs; also known as cognitive diagnostic models [CDMs]) are a class of psychometric models where the latent traits are a set of fine-grained, discrete variables (Bradshaw, 2016; de la Torre & Sorrel, 2023; Rupp et al., 2010). That is, rather than modeling a continuous scale score, DCMs estimate a respondent's proficiency or nonproficiency on a set of predefined skills. These estimates result in a profile of skills on which each respondent is proficient. These profiles offer stakeholders fine-grained results that are often more useful than an overall scale score. For example, in educational assessment, DCMs provide results that are more instructionally relevant (Lim et al., 2024; Thompson & Clark, 2024; S. Zhang et al., 2023). Similarly, applying DCMs to psychological assessment allows researchers to directly model the presence or absence of a trait, rather than attempting to apply a cut point to a continuous scale (R. Liu & Shi, 2020; J. Zhang et al., 2024).

As with all psychometric models, an evaluation of model fit is critical for users to have confidence in the inferences drawn from a DCM. Although many methods have been proposed for evaluating model fit for DCMs, most are constrained by the way DCMs are typically estimated. In practice, DCMs are most often estimated using a maximum likelihood procedure (e.g., Templin & Hoffman, 2013), as this is the only process historically offered by available software (George et al., 2016; Ma & de la Torre, 2020). The lack of accessible alternatives to maximum likelihood estimation has in turn limited the ways in which model fit can be evaluated. However, recent software advances have made Bayesian estimation of DCMs more accessible to practitioners (e.g., Templin, 2023; Thompson, 2023a), widening the scope of possible model-fit evaluations.

In this study, we examine the efficacy of model-fit indices that are available when a Bayesian estimation process is implemented. We first provide a high-level overview of DCMs and existing model-fit indices. We then conduct a simulation study to evaluate the performance of Bayesian model-fit indices and discuss the implications for the practice of diagnostic modeling.

Diagnostic Classification Models

DCMs are confirmatory latent class models, in which each class represents a particular profile of skill proficiency. Although the latent skills, known as attributes, can be polytomous, they are most often binary. For this discussion, we limit ourselves to binary attributes modeled with dichotomous item responses. Using binary attributes, the number of classes is 2^A, where A is the number of attributes measured by the assessment. For example, an assessment measuring three attributes would have $2^3 = 8$ classes: [0,0,0], [1,0,0], [0,1,0], [0,0,1], [1,1,0], [1,0,1], [0,1,1], and [1,1,1], where 1 indicates the presence of or proficiency on the given attribute and 0 indicates absence or nonproficiency.

In addition to the class definitions, we must also provide a Q-matrix (Tatsuoka, 1983) that defines which attributes are measured by each item. The Q-matrix has one row per item and one column per attribute. Each cell of the Q-matrix is either a 0 (the item does not measure the attribute) or 1 (the item does measure the attribute). Given the class definitions, the probability of respondent r providing a correct response is defined as

$$P(\mathrm{X}_r = x_r) = \sum_{c=1}^{C} \nu_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1-x_{ir}} \qquad (1)$$

where C is the number of classes, I is the number of items, and x_{ir} is the observed item response from respondent r on item i. For the parameters, π_{ic} is the probability of a respondent in

class *c* providing a correct response to item *i* and v_c is a mixing parameter that defines the base rate of membership in each class.

The definition of the π parameter is determined by the particular DCM subtype that is estimated. The choice of a DCM model defines certain assumptions about how attributes interact with each other on items that measure multiple attributes. For example, the deterministic-input, noisy "and" gate (DINA) model assumes that respondents should be proficient on all attributes measured by an item in order to provide a correct response (de la Torre & Douglas, 2004; Junker & Sijtsma, 2001). In contrast, the deterministic-input, noisy "or" gate (DINO) model assumes that respondents should provide a correct response if they are proficient on any of the attributes measured by the item (Templin & Henson, 2006).

In addition to models like the DINA and DINO that make strict assumptions about attribute interactions, there are general models that make fewer assumptions and subsume the more-restrictive models. One popular general DCM is the log-linear cognitive diagnostic model (LCDM), which parameterizes π similar to log-linear models (Henson et al., 2009; Henson & Templin, 2019). Consider an item measuring two attributes. Conditional on the attribute profile for class *c*, $\alpha_c = [\alpha_1, \alpha_2]$, the LCDM would define

$$\text{logit}\big(P(X_{ic} = 1|\alpha_c)\big) = \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_1 + \lambda_{i,1,(2)}\alpha_2 + \lambda_{i,2,(1,2)}\alpha_1\alpha_2 \qquad (2)$$

where $\lambda_{i,0}$ is an intercept and represents the log odds of providing a correct response to item *i* when neither attribute is present. We then have two main effects, $\lambda_{i,1,(1)}$ and $\lambda_{i,1,(2)}$, which represent the increase in the log odds when the first or second attribute is present, respectively. Finally, an interaction term, $\lambda_{i,2,(1,2)}$, represents the change in the log odds when both attributes are present.

By constraining the λ parameters in Equation 2, we can achieve models statistically equivalent to more-restrictive models such as the DINA and DINO models (Rupp et al., 2010). If both main effects are constrained to 0, the model is equivalent to the DINA model. That is, a respondent is proficient on either none ($\lambda_{i,0}$) or both ($\lambda_{i,2,(1,2)}$) attributes, and there is no increase in probability for being proficient on a subset of the required attributes. The DINO model can be achieved by constraining the two main effects to be equal (i.e., $\lambda_{i,1}$) and constraining the interaction to be $-1 \times \lambda_{i,1}$. With these constraints, the increase in log odds will be equal to $\lambda_{i,1}$ regardless of whether one or both attributes is present (i.e., the presence of any attribute is sufficient to provide a correct response).

Model Fit for DCMs

After choosing and estimating a DCM, one must evaluate the model's performance. In general, model fit can be evaluated in two ways. Measures of absolute fit describe how well an estimated model represents the observed data. Measures of relative fit directly compare the fit of two or more competing models. We will discuss different methods assessing absolute and relative fit for DCMs in the following sections.

Absolute Fit

For DCMs estimated with a maximum likelihood process, the most common model-fit indices are so called *limited-information indices* (Maydeu-Olivares & Joe, 2005, 2006). The most widely used of these indices is the M_2 statistic, which was originally developed for multidimensional item response theory models (Maydeu-Olivares & Joe, 2005) and later adapted for DCMs (Hansen et al., 2016; Y. Liu et al., 2016). Limited-information indices are required because of the sparse data tables created by categorical response data. For example, an assessment with 10 dichotomous items has $2^{10} = 1{,}024$ possible response patterns. With any

reasonable sample size, it is unlikely we would observe enough respondents at each response pattern to accurately and reliably compare the number of respondents the model expects at each response pattern to the observed number, and this problem become more pronounced as the number of items increases. We are therefore limited to lower-order summaries of the contingency tables. For example, the aforementioned M_2 statistic uses the first- and second-order marginal probabilities. Thus, these limited-information indices cannot capture higher-order aspects of the data.

When a Bayesian estimation process is used, we have options beyond sparse tables that necessitate the use limited-information indices. Rather, we can utilize posterior predictive model checks (PPMCs). Whereas maximum likelihood methods result in a point estimate for each parameter, Bayesian methods provide a posterior distribution of plausible values for each parameter. When using PPMCs, we simulate new data sets from the joint posterior distribution and then compare the simulated data sets to our observed data (Schad et al., 2021). The key decision, then, is to choose the features of the data in the simulated and observed data sets that we want to compare. For example, Sinharay et al. (2006) and Sinharay and Almond (2007) have used PPMCs to evaluate item-level fit for item response theory models and DCMs, respectively.

In the present study, we are interested in overall evaluations of model fit. Both Park et al. (2015) and Thompson (2019) describe a PPMC for the raw-score distribution of an assessment. Using the simulated data sets, we can calculate the expected number of respondents at each raw score point. We then can compare the counts from each individual data set to the expected count, creating a posterior distribution of a χ^2-like statistic. Finally, we can compare the counts of respondents at each score point in our observed data, calculate the χ^2-like statistic, and compare our observed value to the posterior distribution. The posterior distribution represents the

plausible values of the statistic if our estimated model were correct. If our observed value is outside of the posterior distribution (e.g., outside the middle 95% of the distribution), this indicates model misfit. This comparison is often summarized as the posterior predictive *p*-value (*ppp*), which is the proportion of the posterior distribution that is more extreme than our observed value.

The raw-score PPMC offers several theoretical advantages over limited-information methods. First, the raw-score distribution accounts for item dependencies that are excluded when looking only at first- and second-order probabilities, as in the M_2. Second, the joint posterior used to simulate the replicated data sets includes the estimated uncertainty in each of the parameters. Therefore, the summary statistics calculated for the PPMC reflect the plausible values given the uncertainty in the parameter estimates, rather than relying on point estimates from a maximum likelihood estimation. Third, because we are calculating an empirical distribution for the PPMC and comparing the observed value to prespecified quantiles of the distribution, we do not have to depend on asymptotic assumptions that may or may not be met. Thus, the raw-score PPMC χ^2 offers many potential benefits over methods that are more widely used. However, it should be noted that a Bayesian estimation does not preclude the calculation of limited-information indices. That is, when using a maximum likelihood estimation, we cannot calculate PPMCs, but when using a Bayesian estimation, we can calculate both PPMCs and limited-information indices.

Relative Fit

Relative-fit indices are used to compare competing models. These indices do not provide information on whether or not a model adequately fits the observed data. Rather, they compare

how well one model fits relative to another model. Therefore, relative-fit indices primarily useful after evaluating absolute fit (Sen & Bradshaw, 2017).

When using a maximum likelihood estimation, there are well documented methods for comparing competing models, such as the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978). These indices purport to estimate the predictive accuracy of models and can thus be used for comparing models to determine which model would offer better predictions, with some penalty for model complexity. Although commonly used, both the AIC and BIC have significant drawbacks when using a Bayesian estimation, making their use inappropriate. As noted by Hollenbach and Montgomery (2020), the AIC assumes that we have not placed priors on the parameters, which is common practice for Bayesian models. The AIC also assumes that the posterior is multivariate normal, which is not the case when using categorical classes. Additionally, the AIC has been shown to be unreliable when the data have a nested structure, such as items within respondents (Gelman et al., 2014). The BIC has similar weaknesses. The BIC has been shown to be inaccurate when using nonuniform prior distributions (Berger et al., 2003). Hollenbach and Montgomery (2020) and Gelman and Rubin (1995) went so far as to recommend avoiding the BIC altogether for Bayesian models, as, despite its name, the BIC cannot actually approximate any exact Bayesian solution for predictive accuracy and is undefined if specific prior distributions are not selected.

Therefore, when using Bayesian estimation, we must turn to other information criteria for comparing models, namely, leave-one-out cross validation (LOO), as described by Vehtari et al. (2017). A complete description of the LOO is beyond the scope of this paper, and we direct readers to Vehtari et al. (2017) for details. In short, the LOO uses the posterior density to estimate out-of-sample predictive fit for a model, known as the expected log predictive density

(ELPD). Then, just as with other information criteria such as the AIC and BIC, we can compare the ELPD for competing models. The model with the largest value is the preferred model (i.e., expected to have the highest predictive accuracy). As implemented in the loo package (Vehtari et al., 2023), we can also estimate the standard error of the difference. A difference between two models that is much larger than the standard error of the difference (e.g., 2.5; Bengio & Grandvalet, 2004) indicates a meaningful difference in the LOO estimates between the models.

The Current Study

Previous work has compared the efficacy of absolute (Hu et al., 2016) and relative (Lei & Li, 2016; Sen & Bradshaw, 2017) fit measures for DCMs. However, these studies were limited to model-fit indices that are possible when using maximum likelihood estimation. No research to date has compared the performance of Bayesian measures of absolute model fit to the maximum likelihood-based methods. Further, no study has yet examined the use of the LOO for DCMs, as all studies on relative fit have focused on the AIC, BIC, and other similar metrics. In this study, we conducted a simulation to evaluate how well Bayesian methods of model-fit performance compared to their maximum likelihood-based counterparts.

Method

To evaluate the performance of Bayesian absolute- and relative-fit indices for DCMs, we conducted a simulation study. In this study, we manipulated the number of assessed attributes (two or three), the minimum number of items measuring each attribute (five or seven), the sample size (500 or 1,000). These factors were chosen to represent test designs that are commonly seen in applied research (e.g., Bradshaw et al., 2014; Templin & Hoffman, 2013). Using a full factorial design, these factors resulted in a total of eight test-design conditions. Within each test-design condition, we also manipulated the data-generating model (LCDM or

DINA) and the estimated model (LCDM or DINA) to evaluate the performance of model-fit metrics when the estimated model should and should not fit the data. With fully crossed data-generating and estimating models, there are four modeling conditions within each condition, resulting in 32 total conditions across all test designs. We conducted 50 replications per condition.

The simulation and subsequent analyses were conducted in R version 4.3.3 (R Core Team, 2024). All DCMs were estimated using *Stan* (version 2.32.2; Carpenter et al., 2017) via the measr package (Thompson, 2023a, 2023b), and replications were conducted on AWS EC2 instances using the portableParallelSeeds package (P. E. Johnson, 2024). All R code for the simulation and subsequent analyses is available in a public OSF project repository.[1]

Data Generation

The data generation followed the design of the data simulations used by M. S. Johnson and Sinharay (2018) and Thompson et al. (2023) in their evaluations of reliability indices for DCMs. In this study, the number of respondents was determined by the simulation condition. The true attribute profile for each respondent was determined by a random draw from all possible profiles. Additionally, each simulated assessment measured two or three attributes, with each attribute measured by at least five or seven items. The total number of items for each simulated assessment is therefore the product of the number of attributes and the minimum number of items for each attribute. In the simulation, the Q-matrix for each simulated assessment was specified so that the first three items measuring each attribute were single-attribute items.

[1] The project repository can be found at https://osf.io/t5v96/.

The remaining two or four items for each attribute (for the five-item and seven-item conditions, respectively) had a 50% chance of also measuring a second attribute.

Item-parameter generation depended on the data-generating model. In conditions where data were generated from the LCDM, item parameters included item intercepts, main effects, and interactions, all of which are on the log-odds scale. Item intercepts were drawn from a uniform distribution ranging from -3.0 to 0.6, and main effects were drawn from a uniform distribution ranging from 1.0 to 5.0. In the LCDM, interaction terms are constrained to be greater than -1 times the smallest main effect to ensure monotonicity of the model. Thus, the interaction parameters were drawn from a uniform distribution ranging from the calculated lower bound to 2.0. In conditions where data were generated from the DINA model, item parameters include the slipping and guessing parameters, which are both on the probability scale. For consistency with the simulation of LCDM data, parameters were generated on the log-odds scale and converted to probability values. Guessing parameters were drawn from a uniform distribution ranging from -3.0 to 0.6, consistent with the LCDM intercepts. The final guessing parameters were then calculated as the inverse logit of the generated parameter. Slipping parameters were generated from a uniform distribution ranging from 1.0 to 5.0, consistent with the main effects in the LCDM. Because the slipping parameter represents the probability of not providing a correct response when a respondent is proficient on the measured attributes, the final slipping parameter was calculated as 1 minus the inverse logit of the generated parameter value.

The generated attribute profiles, Q-matrix, and item parameters were then used to simulate a data set for each replication.

Simulation Process and Analysis

After the data were generated, both an LCDM and a DINA model were estimated on the simulated data set. We then calculated indices of absolute fit (i.e., M_2 and raw-score PPMC χ^2) and relative fit (i.e., LOO) for each model. All of these indices were calculated for both estimated models to evaluate how they performed under different conditions of known model specifications. When data were generated from the LCDM, we expected the LCDM to show adequate model fit and be the preferred model, as the DINA model was underspecified. On the other hand, when data were generated from the DINA model, we expected both models to show adequate absolute fit, as the LCDM subsumes the DINA model. However, because the LCDM was overspecified in this condition, we expected the DINA model to be preferred by the relative-fit indices.

For each absolute-fit index, an estimated model was flagged for misfit if the *p*-value or the *ppp* was less than .05 for the M_2 and PPMC χ^2, respectively. We then used the flags to calculate the positive and negative predictive values (Altman & Bland, 1994; Smith, 2012). The positive predictive value (PPV) is the proportion of positive results that are true positives, that is, the proportion of models where the fit index indicated poor model fit where we expected it. Similarly, the negative predictive value (NPV) is the proportion of models in which the fit index indicated adequate model fit where we expected it.

Finally, for relative fit, we determined the preferred model by calculating the difference between the LOOs for each of the estimated models, as well as the standard error of the difference. Using the criteria suggested by Bengio and Grandvalet (2004), when the difference between the criterion for the LCDM and the DINA model was greater than 2.5 times the standard error of the difference, we determined the preferred model to be the model with the lowest index

value. When the difference was less than 2.5 times the standard error, we determined the models to be equally fitting and therefore selected the more parsimonious model (i.e., the DINA model) as the preferred model. We then calculated the proportion of replications within each condition in which the LOO selected the correct model (i.e., the model that was used to generate the data).

The expected model-fit results for each combination of generating and estimating models are shown in Table 1.

Table 1

Expected Model-Fit Results

Generating model	Estimated model	Absolute-fit flag	Relative-fit preference
DINA	DINA	No	DINA
	LCDM	No	
LCDM	DINA	Yes	LCDM
	LCDM	No	

Note. LCDM = log-linear cognitive diagnostic model; DINA = diagnostic input, noisy "and" gate.

Results

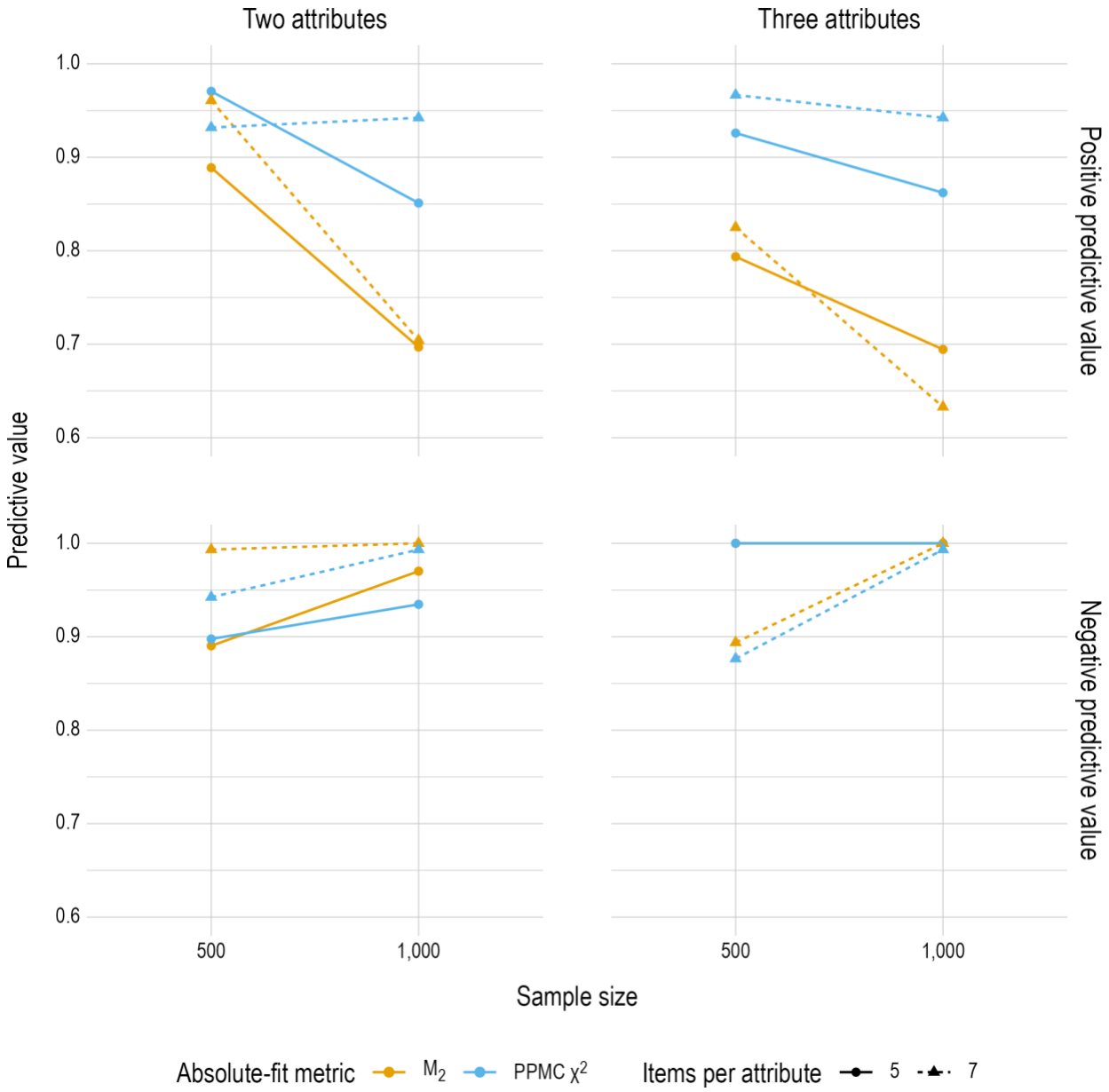
Absolute Model Fit

Across all conditions, the M_2 statistic had a PPV of .753 and an NPV of .964. In contrast, the PPMC χ^2 had a PPV of .919 and an NPV of .952. The NPVs indicate that negative test values for both metrics (i.e., a nonsignificant result) were usually true negatives. On the other hand, the PPVs indicate that positive test results (i.e., a significant result that indicates model misfit) were a true positive only 75% of the time for the M_2 statistic, compared to 92% of the time for the PPMC χ^2 statistic.

When looking at the PPVs and NPVs by test-design condition, as shown in Figure 1, we see that the NPVs for each metric are similar for all test designs. The condition-specific results are consistent with the overall results, which showed similar NPVs for the M_2 and the PPMC χ^2. However, the PPVs for the M_2 are consistently lower than the PPVs for the PPMC χ^2. This difference becomes more pronounced as the data set becomes larger (i.e., larger samples, more attributes). Thus, as the sample gets larger and the test design gets more complex, the M_2 becomes more likely to result in a false positive, indicating model misfit when there is in fact none. On the other hand, the PPMC χ^2 demonstrated consistently high PPVs across all simulation conditions.

Figure 1

Positive and Negative Predictive Values, by Test-Design Condition



Note. PPMC = Posterior predictive model check.

Relative Model Fit

As previously discussed, evaluations of relative fit are only meaningful only when the competing models have been found to have adequate absolute model fit. Accordingly, for the

relative-fit results, we filtered the simulation output to include only include replications in which both the LCDM and the DINA model showed adequate absolute model fit. Using the absolute-fit findings in the previous section, we used the PPMC χ^2 statistic to determine absolute fit. Table 2 shows the number of replications by test-design condition and data-generating model in which both estimated models demonstrated adequate absolute fit. As expected, both the estimated DINA model and the LCDM often had adequate absolute fit when data were generated from the DINA model, as the LCDM subsumes the DINA model. In contrast, it was much less likely that both models would show adequate absolute fit when data were generated from the LCDM.

Table 2

Number of Replications in Which Both Models Demonstrated Absolute Fit

Test designs			Data-generating model	
Attributes	Items	Sample size	DINA	LCDM
2	5	500	49	17
2	5	1,000	43	10
2	7	500	48	9
2	7	1,000	47	1
3	5	500	47	0
3	5	1,000	43	0
3	7	500	49	21
3	7	1,000	47	1

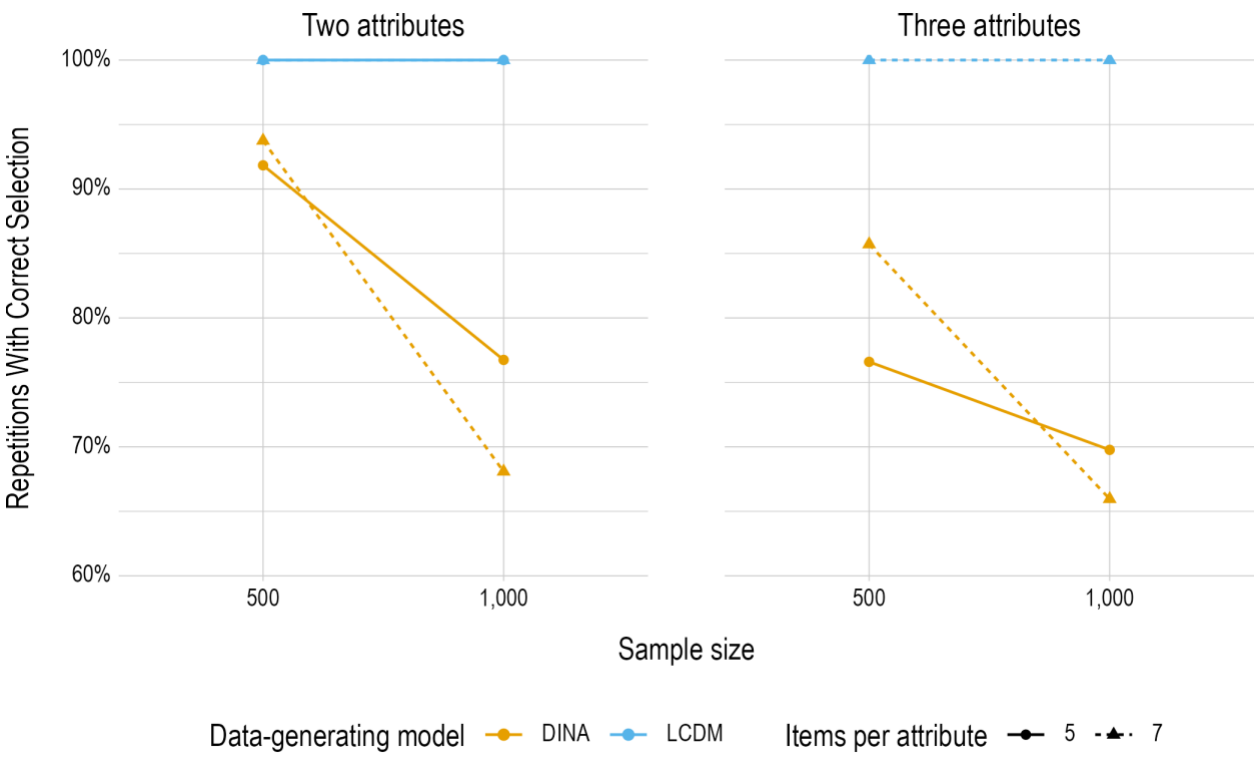
Note. LCDM = log-linear cognitive diagnostic model; DINA = diagnostic input, noisy "and" gate.

Across all conditions, the LOO determined the correct model in 82% of replications. Figure 2 shows the percentage of replications in which the LOO selected the correct model by

test-design condition and data-generating model. Across all test-design conditions, when the LCDM was used to generate the data and both the LCDM and the DINA model showed adequate absolute fit, the LOO always selected the correct model (i.e., the LCDM). On the other hand, when the DINA model was used to generate the data, the LOO selected the LCDM as the preferred model in up to 34% of replications (the three-attribute, seven-item, 1,000-sample-size condition). Thus, even when the DINA model was used to generate the data and the estimated DINA model showed adequate model fit, the LOO still preferred the more-complex model in some situations. However, even with a slight preference for the more-complex model, the LOO still identified the correct model in more than 65% of replications in all conditions.

Figure 2

Correct Model Selections, by Test-Design Condition



Note. LCDM = log-linear cognitive diagnostic model; DINA = diagnostic input, noisy "and" gate.

Discussion

In this study, we examined the performance absolute and relative model-fit indices for Bayesian DCMs. Overall, the findings support the use of Bayesian estimation for DCMs to facilitate the use of Bayesian methods for model evaluation.

For evaluating absolute model fit, we examined the M_2 and PPMC χ^2 statistics. Across all conditions, the M_2 statistic performed well, with results consistent with previous research evaluating the efficacy of the method (e.g., Y. Liu et al., 2016). However, the PPMC χ^2 statistic showed comparable or improved performance in all conditions. This improvement was

particularly true when examining the PPVs for each statistic (i.e., the probability that a flag actually indicates model misfit). Although both the M_2 and PPMC χ^2 had similar negative predictive values, the PPMC χ^2 had consistently higher positive predictive values. Thus, when using the PPMC χ^2, practitioners can be more confident that a positive test result truly indicates model misfit.

Bayesian methods offer different methods for evaluating relative model fit than are appropriate when using a maximum likelihood estimation. Whereas indices such as the AIC and BIC can be used when models are estimated using maximum likelihood estimation, the LOO is used for Bayesian models. In this study, the LOO showed good performance, selecting the correct model in 82% of replications. In contrast, the AIC and BIC have been found to identify the correct model in as few as 30% of replications (Sen & Bradshaw, 2017). The performance of the LOO in this study compared to reported performance of the AIC and BIC means that using a Bayesian estimation process to access the LOO for model comparisons offers a marked improvement over methods that are used with a maximum likelihood estimation.

Limitations and Future Directions

There are some limitations to the present study. For example, we investigated a relatively limited set of test designs. Future research should confirm the efficacy of the PPMC χ^2 and LOO under more-complex designs (e.g., more attributes, more-complex item structures). Additionally, the present study assumed that a DCM was always desired, and therefore model comparisons focused on choosing between different DCM subtypes. Future research could also consider how the LOO behaves when practitioners are comparing different psychometric models. For example, we may compare a Bayesian DCM to an item response theory model that was also estimated with a Bayesian procedure.

Conclusion

Model fit is a crucial evaluation of model performance for any psychometric model, including DCMs. It is important not only to evaluate model fit, but also to ensure that we are using the best methods possible for the evaluation. The present study offers evidence that the PPMC χ^2 and LOO, which can only be utilized when a Bayesian estimation is used, offer improvements over existing maximum likelihood measures of model fit. By using improved methods for model evaluation, we can have greater confidence that the inferences of respondent proficiency we draw from DCMs are valid indications of the respondents' knowledge and skills.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Akadémiai Kiadó.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *British Medical Journal*, *309*(6947), 102. https://doi.org/10.1136/bmj.309.6947.102

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning*, *5*, 1089–1105. http://www.jmlr.org/papers/v5/grandvalet04a.html

Berger, J. O., Ghosh, J. K., & Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, *112*(1), 241–258. https://doi.org/10.1016/S0378-3758(02)00336-1

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (1st ed., pp. 297–327). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch13

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*(1), 2–14. https://doi.org/10.1111/emip.12020

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Sorrel, M. A. (2023). Cognitive diagnosis models. In E. N. Dzhafarov, F. G. Ashby, & H. Colonius (Eds.), *New handbook of mathematical psychology: Vol. 3. Perceptual and cognitive processes* (pp. 385–420). Cambridge University Press. https://doi.org/10.1017/9781108902724.010

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, *25*, 165–173. https://doi.org/10.2307/271064

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1–24. https://doi.org/10.18637/jss.v074.i02

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252. https://doi.org/10.1111/bmsp.12074

Henson, R., & Templin, J. (2019). Loglinear cognitive diagnostic model (LCDM). In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 171–185). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_8

Henson, R., Templin, J., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Hollenbach, F. M., & Montgomery, J. M. (2020). Bayesian model selection, model comparison, and model averaging. In L. Curini & R. Franzese (Eds.), *The SAGE handbook of research methods in political science and international relations* (pp. 937–960). SAGE. https://doi.org/10.4135/9781526486387

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, *16*(2), 119–141. https://doi.org/10.1080/15305058.2015.1133627

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664. https://doi.org/10.1111/jedm.12196

Johnson, P. E. (2024). *portableParallelSeeds: Allow replication of simulations on parallel and serial computers* (R package version 0.97) [Computer software]. The Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.portableParallelSeeds

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement*, *40*(6), 405–417. https://doi.org/10.1177/0146621616647954

Lim, Y. S., Willey, J. M., & Bangeranye, C. (2024). An exploration of cognitive diagnosis in medical education: Constructing comprehensive feedback for enhanced student learning. *Medical Science Educator*. Advance online publication. https://doi.org/10.1007/s40670-024-02064-2

Liu, R., & Shi, D. (2020). Using diagnostic classification models in psychological rating scales. *The Quantitative Methods for Psychology*, *16*(5), 442–456. https://doi.org/10.20982/tqmp.16.5.p442

Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*(1), 3–26. https://doi.org/10.3102/1076998615621293

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26. https://doi.org/10.18637/jss.v093.i14

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Park, J. Y., Johnson, M. S., & Lee, Y.-S. (2015). Posterior predictive model checks for cognitive diagnostic models. *International Journal of Quantitative Research in Education*, *2*(3–4), 244–264. https://doi.org/10.1504/IJQRE.2015.071738

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.3.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

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

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/10.1037/met0000275

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement*, *41*(6), 422–438. https://doi.org/10.1177/0146621617695521

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models. *Educational and Psychological Measurement*, *67*(2), 239–257. https://doi.org/10.1177/0013164406292025

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Smith, C. J. (2012). Diagnostic tests (2) – positive and negative predictive values. *Phlebology*, *27*(6), 305–306. https://doi.org/10.1258/phleb.2012.012J06

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Templin, J. (2023). *blatent: Bayesian latent variable models* (R package version 0.1.2) [Computer software]. The Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.blatent

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. https://doi.org/10.1111/emip.12010

Thompson, W. J. (2019). *Bayesian psychometrics for diagnostic assessments: A proof of concept* (Research Report 19-01). University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://doi.org/10.35542/osf.io/jzqs8

Thompson, W. J. (2023a). measr: Bayesian psychometric measurement using Stan. *Journal of Open Source Software*, *8*(91), 5742. https://doi.org/10.21105/joss.05742

Thompson, W. J. (2023b). *measr: Bayesian psychometric measurement using 'stan'* (R package version 0.3.1) [Computer software]. The Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.measr

Thompson, W. J., & Clark, A. K. (2024). Improving instructional decision-making using diagnostic classification models. *Educational Measurement: Issues and Practice*. Advance online publication. https://doi.org/10.1111/emip.12619

Thompson, W. J., Nash, B., Clark, A. K., & Hoover, J. C. (2023). Using simulated retests to estimate the reliability of diagnostic assessment systems. *Journal of Educational Measurement*, *60*(3). https://doi.org/10.1111/jedm.12359

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2023). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* (R package version 2.6.0) [Computer software]. The Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.loo

Zhang, J., Cui, S., Xu, Y., Cui, T., Barnhart, W. R., Ji, F., Nagata, J. M., & He, J. (2024). Introducing diagnostic classification modeling as an unsupervised method for screening probable eating disorders. *Assessment*. Advance online publication. https://doi.org/10.1177/10731911241247483

Zhang, S., Liu, J., & Ying, Z. (2023). Statistical applications to cognitive diagnostic testing. *Annual Review of Statistics and Its Application*, *10*(1), 651–675. https://doi.org/10.1146/annurev-statistics-033021-111803