Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Evaluating Methods for Assessing Model Fit in Diagnostic Classification Models

W. Jake Thompson *

ATLAS, University of Kansas, Lawrence, 66045, Kansas, United States
*Corresponding author. 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 to 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, recent methodological and technological advances 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 conducted a simulation study to compare the performance of traditional measures of model fit with 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

Diagnostic classification models (DCMs; Bradshaw, 2016; de la Torre & Sorrel, 2023; Rupp et al., 2010) are confirmatory latent class models, in which each class represents a particular profile of skill (also called "attribute") proficiency. Under the large umbrella of DCMs, there are different models that make their own 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), in which proficiency on each attribute or set of attributes provides a unique increase to probability of providing a correct response (Henson & Templin, 2019; Henson et al., 2009).

Given the range of available DCMs, it is important to evaluate the performance of the chosen model. In general, model fit can be evaluated in two ways. Measures of absolute fit describe how well an estimated model represents the observed data. Relative fit metrics directly compare the fit of two or more competing models. In the following sections, we will discuss different methods for assessing the absolute and relative fit of DCMs.

1. Model Fit for DCMs

1.1 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), due to their use of lower-order summaries of the contingency tables. The most widely used of these indices is the M_2 statistic, which has been adapted for DCMs by Hansen et al. (2016) and Liu et al. (2016).

When using a Bayesian estimation process, we can utilize posterior predictive model checks (PPMC). To use 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). For example, both Park et al. (2015) and Thompson (2019) describe a PPMC for the raw-score distribution of an assessment, which 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. Finally, because we are calculating an empirical distribution for the PPMC, we do not have to depend on asymptotic assumptions that may or may not be met.

1.2 Relative Fit

There are well documented methods for comparing competing models when using a maximum likelihood estimation, such as the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978). Although commonly used, both the AIC and BIC have significant drawbacks when using a Bayesian estimation, making their use inappropriate (for a full discussion of limiations, see Berger et al., 2003; Gelman & Rubin, 1995; Hollenbach & Montgomery, 2020). Therefore, we must turn to other information criteria for comparing models, namely, leave-one-out cross validation (LOO), as described by Vehtari et al. (2017). 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). We can then 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).

1.3 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.

2. Method

To evaluate the performance of Bayesian absolute- and relative-fit indices for DCMs, we conducted a simulation study where 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). 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, 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 (Carpenter et al., 2017) via the measr package (Thompson, 2023a, 2023b). All R code for the simulation and subsequent analyses is available in a public OSF project repository.[1]

2.1 Data Generation

Within each 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. 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. 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. Slipping parameters were generated from a uniform distribution ranging from 1.0 to 5.0, consistent with the main effects in the LCDM.

2.2 Simulation Process and Analysis

Both an LCDM and a DINA model were estimated using 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. 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. 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 and standard error of the difference between the LOOs for each of the estimated models. 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 that the preferred model was the model with the largest ELPD. 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 (i.e., data-generating) model.

The expected 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	

3. Results

3.1 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.

3.2 Relative Model Fit

Evaluations of relative fit are only meaningful when the competing models have been found to have adequate absolute model fit (Sen & Bradshaw, 2017). Accordingly, for the relative-fit results, we filtered the simulation output to include only 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.

Across all conditions, the LOO determined the correct model in 82% of replications. Figure 1 shows the percentage of replications in which the LOO selected the correct model by test-design condition and data-generating model. 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. Thus, even when

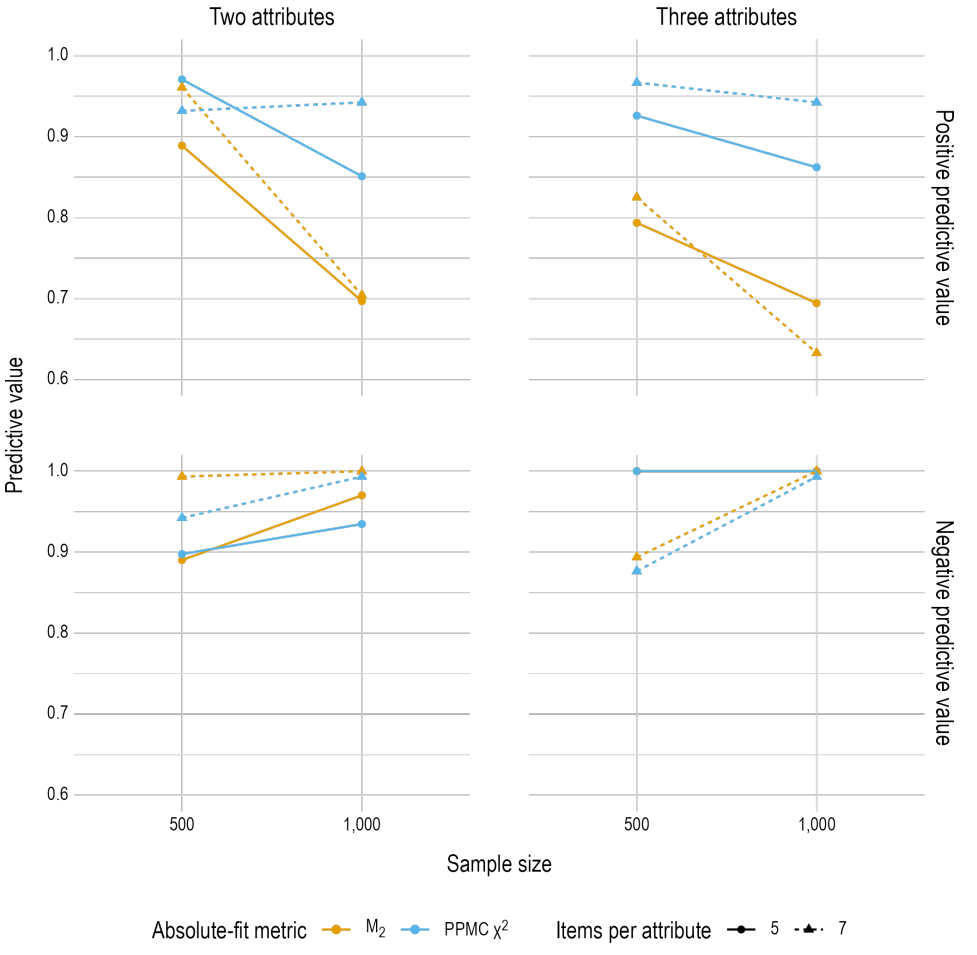

Figure 1. Positive and Negative Predictive Values, by Test-Design Condition

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

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.

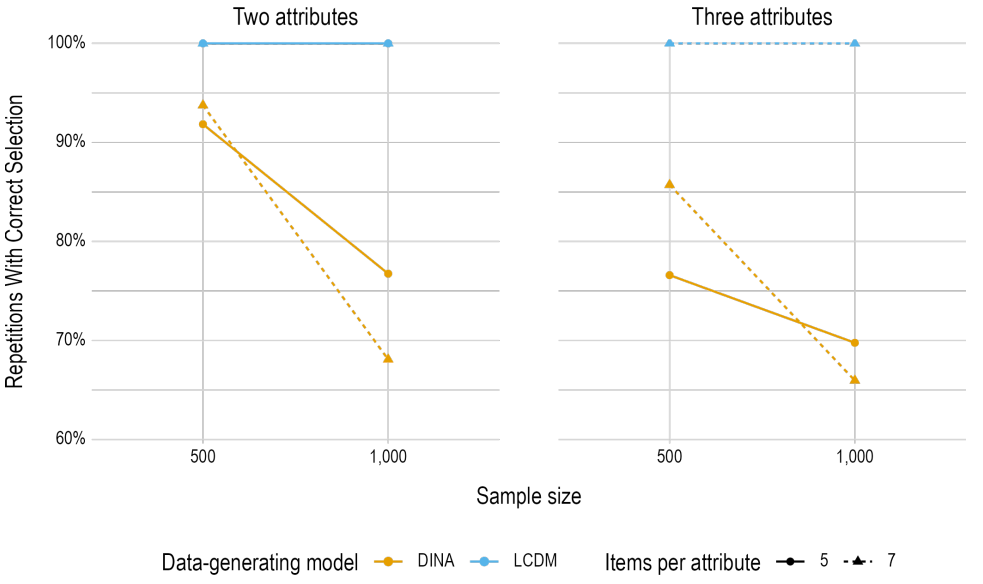

Figure 2. Positive and Negative Predictive Values, by Test-Design Condition

4. 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.

Across all conditions, the M_2 statistic performed well, with results consistent with previous research evaluating the efficacy of the method (e.g., Liu et al., 2016). However, the PPMC χ^2 statistic showed comparable or improved performance in all conditions. 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. This is likely because the M_2 captures only two moments of the data, whereas the PPMC χ^2 is able to capture higher-order moments and reflect discrepancies that may be missed by the M_2.

When evaluting relative fit, 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.

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 under a limited set of test designs. Future work should examine performance of these metrics under more complex designs (e.g., more attributes, more-complex item structures). 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.

Funding Statement This 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 of the U.S. Department of Education.

Competing Interests None.

Notes

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

References

Akaike, H. (1973, September). 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, 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., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, *25*, 165–173. https://doi.org/10.2307/271064

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. (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

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., & 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

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

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*. Computer software. Version Version 4.3.3. 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

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

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 No. 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'*. Computer software. Version R package version 0.3.1. The Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.measr

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

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