Article

Is the Area Under Curve Appropriate for Evaluating the Fit of Psychometric Models?

Educational and Psychological Measurement
2023, Vol. 83(3) 586–608

DOI: 10.1177/00131644221098182
journals.sagepub.com/home/epm
SAGE

Yuting Han[1], Jihong Zhang[2], Zhehan Jiang[1] and Dexin Shi[3]

Abstract

In the literature of modern psychometric modeling, mostly related to item response theory (IRT), the fit of model is evaluated through known indices, such as χ^2, M2, and root mean square error of approximation (RMSEA) for absolute assessments as well as Akaike information criterion (AIC), consistent AIC (CAIC), and Bayesian information criterion (BIC) for relative comparisons. Recent developments show a merging trend of psychometric and machine learnings, yet there remains a gap in the model fit evaluation, specifically the use of the area under curve (AUC). This study focuses on the behaviors of AUC in fitting IRT models. Rounds of simulations were conducted to investigate AUC's appropriateness (e.g., power and Type I error rate) under various conditions. The results show that AUC possessed certain advantages under certain conditions such as high-dimensional structure with two-parameter logistic (2PL) and some three-parameter logistic (3PL) models, while disadvantages were also obvious when the true model is unidimensional. It cautions researchers about the dangers of using AUC solely in evaluating psychometric models.

Keywords

goodness of fit, absolute fit indices, AUC, item response theory, multidimensional structures

[1]Peking University, Beijing, China
[2]The University of Iowa, Iowa City, USA
[3]University of South Carolina, Columbia, USA

Corresponding Author:
Zhehan Jiang, Institute of Medical Education and National Center for Health Professions Education Development, Peking University, Beijing 100191, China.
Email: jiangzhehan@gmail.com

Introduction

Item response theory (IRT) models refer to a family of widely used psychometric models to predict response performance according to the characteristics of items and examinees (Birnbaum, 1968; Embretson & Reise, 2000/2013; Rasch, 1960; Thissen & Steinberg, 2009). The effective use of IRT models depends on the degree of goodness of fit (GOF) of the chosen model to the actual data. That said, the benefits and capabilities of IRT can only be truly realized if the model used shows a good fit (Orlando & Thissen, 2000), else errors in parameter estimation, test equating, and the detection of differential item functioning (DIF) can occur (Kang et al., 2009). Therefore, when using IRT for analysis, it is essential to thoroughly examine the chosen model's fit to the actual data (McKinley & Mills, 1985). Generally, the GOF statistics used in the IRT can be divided into both item and test levels. Item-level statistics can be used to screen individual items (Bock, 1972; Chalmers & Ng, 2017; Drasgow et al., 1985; Kang and Chen, 2008; McKinley & Mills, 1985; Orlando and Thissen, 2000, 2003; Wright & Masters, 1982; Yen, 1981), while test-level statistics are used to assess the degree of GOF between the model chosen and the actual data at the overall level (e.g., Cai et al., 2006; Cai & Hansen, 2013; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2006).

The GOF statistic for test-level testing based on χ^2 statistic assumes that the model chosen is the appropriate model, dividing the examinees into groups according to certain criteria (e.g., their ability or observed scores on the test) and calculating the difference between the observed and expected frequencies in each group to find the value of the χ^2 statistic. These GOF statistics can be categorized into two main types: full-information test statistics, which are computed from all possible response patterns (full contingency table), and limited information test statistics, which use the summary characteristics of the full contingency table. The two most commonly used full-information GOF statistics are Pearson's test statistic χ^2 and the likelihood ratio test statistic G2 (Koehler & Larntz, 1980; McKinley & Mills, 1985). The χ^2 test requires a sufficiently large sample size resulting in the expected frequencies in each group should be no less than five; otherwise, the statistic may violate the χ^2 distribution and result in questionable reliability of the test. Specifically, in sparse tables, the empirical Type I error rates of full-information GOF statistics are inflated. For the statistic G2, it no longer follows the approximate χ^2 distribution when the more complex model of the nest models does not fit the data under a large sample (Maydeu-Olivares & Joe, 2006).

In contrast to full-information test statistics, limited information test statistics (e.g., Bartholomew & Leung, 2002; Cai et al., 2006; Cai & Hansen, 2013; Cai & Monroe, 2014; Maydeu-Olivares & Joe, 2006) do not use the full contingency table but lower-order marginal tables only. Maydeu-Olivares and Joe's M2 family of statistics using residuals based on lower-order margins of the contingency table are practically useful in testing the overall GOF of IRT models. However, as the sample size increases, it becomes too sensitive to reject the fitted model which may have a tolerable or negligible degree of misfit (Xu et al., 2017). Therefore, root mean square error

approximation (RMSEA; Steiger & Lind, 1980) calculated based on M2 was used to measure effect size for evaluating the degree of the model-data misfit (Maydeu-Olivares & Joe, 2014). Other limited information test statistics such as TLI (Tucker–Lewis index; Bentler, 1990) and CFI (comparative fit index; Tucker & Lewis, 1973), initially introduced in structural equation modeling literature to report incremental fit measure, have also been applied to the GOF assessment in IRT. As a rule of thumb, a CFI or a TLI $\geq$ 0.90 is considered an acceptable fit, while a CFI or a $\geq$ 0.95 is considered an excellent fit. TLI has a higher penalty for adding parameters than CFI. They both only applicable when items all have a suitable null model and the data is not overly sparse.

When several models fit the data, the relative model fit statistics can help select a more proper model. There is no absolute cutoff point for relative model fit statistics but a comparison between the values produced from the ''fitting.'' χ^2 statistics cannot satisfy this need, while statistics based on information formula are generally adopted, for instance, 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 likelihood ratio (LR) test. However, Kang and Cohen (2007) found that AIC and BIC were accurate when the data were generated by the one-parameter logistic (1PL) model or the two-parameter logistic (2PL) model, but they tended to select 2PL models when the true models were three-parameter logistic (3PL) models.

Recent developments show a merging trend of psychometric and machine learnings (ML), for example, Bergner and colleagues (2012) conducted collaborative filtering analysis through an ML-based IRT; Pliakos and colleagues (2019) adopted ML approaches to assist the interpretability of IRT parameters; Jiang and colleagues (2019b) adopted AdaBoost algorithm from ML family to supplement cognitive diagnosis modeling; Silva and colleagues (2020) integrated clustering approaches to IRT-based computerized adaptive testing. It can be found that the trend takes place mostly in modeling strategies; however, model fit evaluation engrafting from one field to the other is paid less attention. In this study, the area under curve (AUC), a mainstream fit index in the ML-relevant literature, is the focus since it has been gradually cited in psychometric modeling studies (e.g., Gonzalez, 2021; Lee, 2019; Park et al., 2019; Su et al., 2018; R. Wu et al., 2017, Z. Wu et al., 2020).

Drawn from signal detection theory, AUC is a core ML measure prevalently used to evaluate the accuracy of classifiers. The superior performance of AUC has been discussed (e.g., Huang & Ling, 2005), and its applications have been substantially seen in fields such as bioinformatics, computer science, physics, business, and others. Recently, AUC has been applied to more educational and psychological research to examine the accuracy of certain computerized mechanisms, for instance, early warning systems and corresponding indicators empower AUC to evaluate classifiers (e.g., Bowers & Zhou, 2019; Carlson, 2018; Jiang et al., 2019a; Johnson & Semmelroth, 2010; Nicholls et al., 2010; Stuit et al., 2016), and intelligent tutoring systems devise AUC to compare the performance of ML predictors (Khajah et al., 2016; Le et al.,

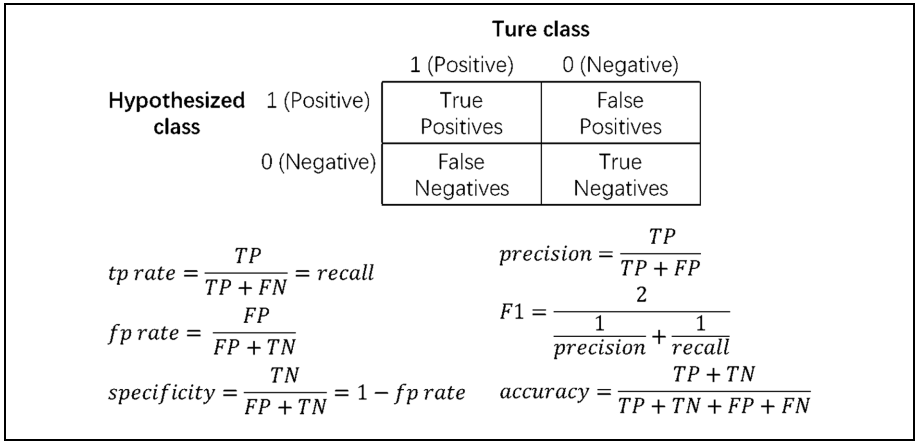


Figure 1. Confusion Matrix and Common Performance Metrics.

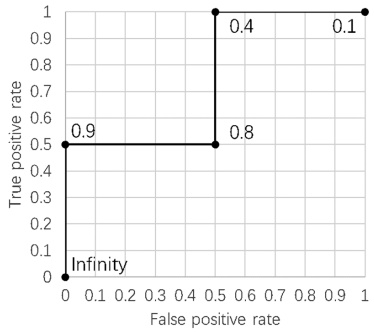
2018; Piech et al., 2015; Walker & Jiang, 2019). In terms of its application to psychometric modeling, AUC is a rarely investigated from a systemic view, although it has been quoted directly in a few studies under the context of IRT models (e.g., Cheng et al., 2019; Niemeijer et al., 2020; Svicher et al., 2019; Windle & Windle, 2017).

To summarize, AUC's performance in evaluating the fit of psychometric models remains unknown, yet researchers using IRT and other similar diagnostic analytics make critical inferences based on its estimates. Therefore, it is necessary to assess the appropriateness of adopting AUC to evaluate the targeted models from two perspectives: (a) selecting the best-fitting one and (b) serving as an alternative to absolute fit indices and its plausible recommendation.

Method

AUC is an extension of the receiver operating characteristics (ROC) curve which summarizes label-assignment performance by combining a confusion matrix (i.e., 2 $\times$ 2 table including true/false positive and true/false negative counts) at all threshold levels, which of the changes would alter classification accuracies. The confusion matrix and equations of several common metrics that can be calculated from it are shown in Figure 1. For a binary classifier, if the predicted category of the instance is 1, it is labeled as positive. If the predicted category of the instance is 0, it is labeled as negative. And label true for a correct prediction, label false for an incorrect prediction. Thus, each instance can be mapped to one cell of the confusion matrix, which aggregates the counts of instances for each of the four categories.

The ROC curve is drawn on a two-dimensional plane, the horizontal coordinate of which is the false positive rate (FPR) and the vertical coordinate is the true positive rate (TPR). A discrete classifier produces one confusion matrix and corresponding to



Scores and classifications of 4 instances

Instance No.	True class	Predict probability of being positive	Threshold	TP	FP	FN	TN	FPR	TPR
			Infinity	0	0	2	2	0	0
1	p	0.9	0.9	1	0	1	2	0	0.50
2	n	0.8	0.8	1	1	1	1	0.5	0.50
3	p	0.4	0.4	2	1	0	1	0.5	1
4	n	0.1	0.1	2	2	0	0	1	1

Figure 2. Scores and Classifications of Four Instances, and the Resulting ROC Curve.
Note. ROC = receiver operating characteristics.

a single point in ROC space. For probability or scoring classifier, various thresholds can be set to get the confusion matrixes, and each threshold value produces a different point in ROC curve. Tom Fawcett (2006) described the algorithm for efficient generation of ROC curves is in detail.

AUC turns the ROC curve into a numeric representation of performance for a binary classifier. Essentially, AUC aggregates the performance of a model across all threshold values. AUC is the area under ROC curve and ranges between 0 and 1: the best possible value indicates a perfect classifier, while zero if all the predictions are wrong. The AUC from a finite set of instances can be calculated as follows: first sort the instances by their predicted probability of being positive, then set these predicted probabilities as threshold values in descending order and calculate TPR and FPR accordingly. This would draw a series of upward and rightward points on the ROC plane and form the ROC curve. Finally, the AUC can be calculated as the sum of successive areas of trapezoids enclosed by instance points i and the latter point $i+1$ and the horizontal axis (FPR), as described in the following equation:

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(FPR_{i+1} - FPR_i)(TPR_{i+1} + TPR_i) \ , \tag{1}$$

where χ^2 represents the number of instances. Figure 2 shows an example of an ROC curve on a set of four instances. The instances, two positive and two negative, and their corresponding coordinates by setting each score as threshold are shown in the table beside the graph. The AUC for this example can then be calculated as[1]

$$AUC = \frac{1}{2}[(0.5-0)(0.5+0.5)+(0.5-0.5)(1+0.5)+(1-0.5)(1+1)] = 0.75.$$

To conclude, AUC is not dependent on classification threshold (i.e., classification-threshold invariant) and therefore can be used without additional subjective

judgments. AUC, however, is for one binary outcome variable but not multivariate scenarios. This study proposed using AUC to construct a statistic for evaluating the IRT model fit. From the similar vein that test-level M2 is constructed from its item-level statistics, AUC statistic in this study measures the average classification accuracy across all items with the following form:

$$avgAUC = \sum_{j=1}^{J} AUC_j \; / \; J, \tag{2}$$

where AUC_j denotes the area under ROC curve of the item j classifier under the examined model and J represents the total number of items. AUC_j quantified the tradeoff between the sensitivity and specificity of the prediction of binary item responses. One interpretation of AUC_j is the probability that the IRT model ranks a random case with a correct response for item j higher than a random case with an incorrect response. Thus, $avgAUC$ measures the average performance of predicting item response by the model across all possible classification thresholds. Note that the AUC is usually used with cross-validation method in the ML literature. However, in this study, we used all data points to calculate the proposed AUC statistic like how other GOF indices were computed. The reason is that cross-validation procedures are meant for prediction evaluation measuring the potential errors for unseen data, while we emphasize model evaluation showing how a model fits the collected information.

Simulation Study

Data Generation

Following the study design of Xu et al. (2017), multiple types of unidimensional and multidimensional IRT (MIRT) models were adopted for data generation. The unidimensional IRT models were the 2PL and the 3PL models. To maintain simplicity without losing generalizability, only compensatory MIRT models were included in the simulations as this genre is more commonly seen (DeMars, 2016; Immekus et al., 2019). For compensatory MIRT models, a low ability in one dimension can be compensated by a high ability in other dimension(s) to reach a high expected probability of a correct answer.

Four types of model structures were adopted for data generation: (a) unidimensional structure, indicating that each of the items on the test measures one dimension; (b) multidimensional between-item structure (or equivalently simple structure), indicating that a test contains multiple unidimensional subscales; (c) partially cross-loading structure such that several items measuring a single dimension while others corresponding to multiple dimensions; (d) completely cross-loading structure (or equivalently within-item multidimensionality), implying that each of the items on the test measures more than one dimension.

Specifically, for the unidimensional structure, a unidimensional (1D) model with all items measuring the same dimension was adopted for data generation in the study.

Table 1. Description of the Naming Convention for Data Structures.

Data structures for data generation	
1D_2PL	One-dimensional 2PL model
1D_3PL	One-dimensional 3PL model
2DW_2PL	Two-dimensional within-item 2PL model
2DW_3PL	Two-dimensional within-item 3PL model
2DPS_2PL	Two-dimensional partially simple structure 2PL model
4DS_2PL	Four-dimensional simple structure 2PL model
4DS_3PL	Four-dimensional simple structure 3PL mode

Note. 1D = unidimensional; 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model; 2DW = two-dimensional within-item; 2DPS = two-dimensional partial simple structure model; 4DS = four-dimensional simple structure.

For the multidimensional between-item structure, a four-dimensional simple structure (4DS) was used. For the partially cross-loading structure, a two-dimensional partial simple structure model (2DPS) was used. For the completely cross-loading structure, a two-dimensional within-item (2DW) structure model was used in the simulation. Table 1 provides the labels of all the models used for data generation, resulting in 32 conditions in data generations: (a) eight latent factor structures were simulated: two unidimensional structures (1D_2PL and 1D_3PL models) and six multidimensional structures (2DW_2PL/3PL models, 2DPS_2PL/3PL models, and 4DS_2PL/3PL models); (b) two levels of sample size (i.e., 750 and 1,500); and (c) two levels of test length (i.e., 20 and 40) were used to generate data. Each condition was experimented with 500 replications, leading a total of 16 000 (500 $\times$ 32) generated data sets. For each data set, the Rasch, 2PL, and 3PL models with the same latent factor structures as the generated models were fitted. AIC, BIC, and *avgAUC* were then obtained from the three models fitted to each data set.

Models for Data Simulation

The general form of the 2PL/3PL IRT model used for data simulation was

$$P\left(X_{ij}=1 \mid \theta_i\right)=g_j+\left(1-g_j\right)\frac{\exp\left(\Sigma_k a_{jk}\theta_{ik}-b_j\right)}{1+\exp\left(\Sigma_k a_{jk}\theta_{ik}-b_j\right)}, \tag{3}$$

where χ^2 is an ability vector for person i; θ_{ik} denotes person i's latent trait for dimension k; a_{jk} is the item discrimination or item slope of item j for dimension k; b_j is the difficulty parameter for item j; and g_j is the guessing parameter for item j representing the lower bound of the item response function curve. The 1D model was obtained when $k=1$ and $g_j=0$ (1D_2PL) and when $k=1$ and g_j was freely estimated (1D_3PL). The 2DW models were obtained when $k=2$ and $g_j=0$ (2DW_2PL) and when $k=2$ and g_j were freely estimated (2DW_3PL). For 2DPS model was obtained

when $k = 2$, $g_j = 0$ (2PL) or freely estimated (3PL), and some a_{jk} are constrained to 0 so that around one-third item set measures the first dimension (six items for 20-item test and 13 items for 40-item test), another one-third items measure the second dimension, and the third set measures both dimensions (eight items for 20-item test and 14 items for 40-item test). The 4DS models were obtained when $k = 4$, and $g_j = 0$ (2PL) or freely estimated (3PL) with the constraint $a_{jk} = 0$ for a *j*-*k*-combo conditions such that each of the four different item sets measured its own single dimension. For example, for the 20-item test, there are five items loading on each dimension. For the 40-item test, there were 10 items loading on each dimension.

Model Parameter Values for Data Simulation

For the unidimensional IRT model (1DS_2PL/3PL), a scalar of person latent trait parameters was drawn from a standard normal distribution. For the multidimensional IRT model, a vector of the person latent trait values was drawn from a multivariate normal distribution with the means center at zeros and the covariance matrix of Σ that has 1 s in the diagonals. The moderate factor correlations in Σ were set to 0.5 for the 2DW and 2DPS models and 0.7 and 0.5 for the 4DS model. The item difficulty parameters (b_j) were randomly generated from a truncated normal distribution with lower bound -2 and upper bound +2 using the R package *truncnorm* (Mersmann et al., 2018). The item slope or discrimination parameters (a_{jk}) for dimension k were generated independently from the truncated lognormal distribution with a mean of 0 and a standard deviation of .5 and truncated within [.5, 4] using the R package *EnvStats* (Millard, 2013). For 3PL models, guessing parameters (c_j) were randomly generated from uniform distribution with the range [0, .3]. All data sets were generated using the R package *mirt* (Chalmers, 2012). For each simulation model in each replication, item parameters were regenerated following the same procedures. The model estimation was also implemented by the *mirt* package (Chalmers, 2012), and the *caret* package (Kuhn, 2021) and the *pROC* package (Robin et al., 2011) in R (version 4.1.1; R Core Team, 2021) were used to calculate the confusion matrix and AUC, respectively.

For model estimation, the R (version 4.1.1; R Core Team, 2021) language with the *mirt* package was used with the standard expectation–maximization (EM) algorithm with a fixed quadrature for the 1DS, 2DS, and 2DPS models and quasi-Monte Carlo EM estimation for 4DS. The standard EM algorithm is generally effective with 1 to 3 factors, but methods such as the quasi-Monte Carlo EM algorithm and other Monte Carlo methods were used when the dimensions were three or more for efficacy. To minimize the chance of convergence problems in the 2PL and the 3PL model estimation, item parameter priors—lognormal for slope parameters: $a \sim lognormal(0, .5^2)$ and normal distribution for guessing parameters: $c \sim normal(0.18, 0.1)$—were used.

Result

Convergence Rates

Convergence rates for 88 of the 32 conditions with three analyzed models per condition (91.7%) were above 90%. The eight cases showing lower than 90% convergence rate were from the 4DS structure models. For all eight cases, the Rasch models were fitted with convergence rates from .18 to .56.

Comparing Relative Model–Data Fit Indices: Type I Error Rate and Power

To evaluate the performance of the relative model–data fit indices, the AIC, the BIC, and AUC values were calculated for all replications. For each replication, the best-fitting model was chosen according to each index (i.e., the selecting criteria were both the smallest AIC and BIC values and with the largest AUC value). The proportion of each analyzed model selected by each index across all replications are computed and compared. The fit indices which show higher proportions of models consistent with the generating model have better performance. Table 2 shows the proportions of replications for which each fit index selected the analyzed model as the best-fitting model. In general, AIC and BIC prefer the 2PL models as the best-fit model, while AUC performs better with more parameters and a more complex structure of the true generating model.

To illustrate, in row 1, under the conditions with a sample size of 750, number of items of 20, and true model being the unidimensional 2PL model (1D_2PL), AIC correctly selected the estimated 1D_2PL model as the best-fitting model in all of the replications, while BIC correctly selected the estimated 1D_2PL model as the best-fitting model in 97.4% of the replications, but incorrectly selected the unidimensional Rasch model as the best-fitting model in 2.6% of the replications. In addition, AUC incorrectly selected the unidimensional Rasch model as the best-fitting model in all the replications. The results show how often fit indices were able to determine the true generating model (i.e., power) and which estimation models were incorrectly selected as the best-fitting model when the correct model was not identified.

In general, the AIC showed sufficient power under conditions when the 2PL was the true model structure and when the data were generated from unidimensional (1D_2PL, see rows 1–4 in Table 2) or two-dimensional model structures (2DPS_2PL and 2DW_2PL, see rows 5–12 in Table 2). However, when the data were generated from high-dimensional structure (i.e., the 4DS_2PL model), the true model was less likely to be identified as the best-fitting model (over 80% proportions) when the number of items was relatively small ($J = 20$). Moreover, the AIC indices have no power selecting 3PL models as the AIC never selected 3PL models as the best-fitting model for all replications under all conditions of this study (see rows 17–32 in Table 2). The results showed that, regardless of whether the data were generated from a 2PL or a 3PL model, the AIC's choice across conditions exhibited a high consistency in preferring the unidimensional/multidimensional 2PL as the best-fitting model,

Table 2. Comparisons of Model Selection Methods.

row	Data generating model	Test length	Simple size	AIC			BIC			AUC		
				*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL
1	1D_2PL	20	750	0	**1**	0	.026	**0.974**	0	1	**0**	0
2			1500	0	**1**	0	.002	**0.998**	0	1	**0**	0
3		40	750	0	**1**	0	0	**1**	0	1	**0**	0
4			1500	0	**1**	0	0	**1**	0	1	**0**	0
5	2DPS_2PL	20	750	0	**1**	0	0	**1**	0	0	**1**	0
6			1500	0	**1**	0	0	**1**	0	0	**1**	0
7		40	750	0	**1**	0	0	**1**	0	0	**1**	0
8			1500	0	**1**	0	0	**1**	0	0	**1**	0
9	2DW_2PL	20	750	0	**1**	0	.678	**0.322**	0	.998	**0.002**	0
10			1500	0	**1**	0	.116	**0.884**	0	1	**0**	0
11		40	750	0	**1**	0	.6	**0.4**	0	.282	**0**	.718
12			1500	0	**1**	0	.004	**0.996**	0	.048	**0**	.952
13	4DS_2PL	20	750	.89	**0.11**	0	.972	**0.028**	0	0	**1**	0
14			1500	.788	**0.212**	0	.868	**0.132**	0	0	**1**	0
15		40	750	.18	**0.82**	0	.566	**0.434**	0	0	**1**	0
16			1500	.114	**0.886**	0	.248	**0.752**	0	0	**1**	0
17	1D_3PL	20	750	0	1	**0**	.028	.972	**0**	1	0	**0**
18			1500	0	1	**0**	.002	.998	**0**	1	0	**0**
19		40	750	0	1	**0**	0	1	**0**	1	0	**0**
20			1500	0	1	**0**	0	1	**0**	1	0	**0**
21	2DPS_3PL	20	750	0	1	**0**	0	1	**0**	0	1	**0**
22			1500	0	1	**0**	0	1	**0**	0	1	**0**
23		40	750	0	1	**0**	0	1	**0**	0	1	**0**
24			1500	0	1	**0**	0	1	**0**	0	1	**0**

(continued)

Table 2. (continued)

row	Data generating model	Test length	Simple size	AIC			BIC			AUC		
				*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL
25	2DW_3PL	20	750	.002	.998	**0**	.656	.344	**0**	.998	.002	**0**
26			1500	0	1	**0**	.106	.894	**0**	.998	.002	**0**
27		40	750	0	1	**0**	.568	.432	**0**	.308	.002	**0.69**
28			1500	0	1	**0**	.006	.994	**0**	.052	0	**0.948**
29	4DS_3PL	20	750	.894	.106	**0**	.976	.024	**0**	0	1	**0**
30			1500	.804	.196	**0**	.886	.114	**0**	0	1	**0**
31		40	750	.178	.822	**0**	.632	.368	**0**	0	1	**0**
32			1500	.104	.896	**0**	.284	.716	**0**	0	1	**0**

Note. 1D_2PL and 1D_3PL are one-dimensional 2PL model and one-dimensional 3PL model, respectively; 2DW_2PL and 2DW_3PL indicate two-dimensional within-item 2PL model and two-dimensional within-item 3PL model, respectively; 2DPS_2PL and 2DPS_3PL denote two-dimensional partially simple structure 2PL model and two-dimensional partially simple structure 3PL model, respectively; 4DS_2PL and 4DS_3PL are four-dimensional simple structure 2PL model and four-dimensional simple structure 3PL model, respectively; AIC = Akaike's information criterion; BIC = Bayesian information criterion; AUC = area under curve; 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model. *Rasch, *2PL, and *3PL indicate calibrating models with the same model structure as the corresponding data generation models indicated in the rows. Values in cells where the calibrating model is the same as the data generating model are in bold.

except in the condition of true model being high-dimensional AIC preferred simple structure model (see rows 13–14 and 29–30 in Table 2).

Similar to the AIC, the BIC has similar issues when true models are 3PL models or have high-dimensional structures. First, the BIC never selected 3PL models as the best-fitting model for all replications under all conditions in this study, making the power of BIC all zero when the data were generated by unidimensional/multidimensional 3PL models. Second, the BIC preferred the Rasch model as the best-fitting model under the conditions of high-dimensional model structure than the AIC, which may attribute to the fact that the BIC has a higher penalty for model complexity. In addition, sample size appeared to have impact on the performance of the BIC. Regardless of whether the generating model was the 2PL or the 3PL model, the rate at which BIC selected 2PL models as the best-fit models increased as the sample size increased, except for the cases where the 2PL models were selected at a rate of 1. Thus, we found that BIC behave similarly with AIC in term of power but sample sizes have more influence on the selection of the BIC. These results were consistent with previous research done by Kang and Cohen (2007), in which data were generated from the unidimensional 1PL, 2PL, or 3PL model and analyzed using each of these three models. When data were generated by a 1PL or 2PL model, AIC and BIC could select the generating model as the best model in almost every replication. However, when data were generated by a 3PL model, the 3PL model was not selected by BIC for any of the replications and AIC also demonstrated a preference for selection of the simpler model. The study of Lin and Dayton (1997) also found that the BIC demonstrated a preference for selection of the simpler model. This may be related to the quality of the items, and that when using the 3PL model to generate data, a larger guessing parameter might enable the AIC and BIC more inclined to choose the correct 3PL model.

Compared with the information criterion (AIC, BIC), AUC showed some advantage under certain conditions such as high-dimensional structure with 2PL and some 3PL models. When the data were generated by the 2DW_3PL model and the number of items was 40, the true 2DW_3PL model more frequently obtained larger statistic values than other incorrect models, allowing the AUCs to have powers of 0.69 (for the sample size of 750) and 0.948 (for the sample size of 1,500) under this condition (see rows 27 and 28 in Table 2). The AUC also showed better performance than the AIC and the BIC in selecting the 4DS_2PL model as the best-fitting model when it was the true mechanism. However, when the data were generated by unidimensional 2PL/3PL models, AUC performed poorly by always selecting Rasch models as the best-fitting model.

Note that the selection of the Rasch model as the best-fitting model using the maximum average AUC as a criterion for the entire test does not mean that every item in the test obtained the maximum AUC value when analyzed using the Rasch model. To explore the relationship between item characteristics and AUC values, we further analyzed the item-level AUC under the condition that the sample size was 750 and the test length was 20, with data generated by the 1D_2PL model.

Table 3. The Description of Item Level AUC and Item Parameters in the 500 Repetitions Under the Condition that Sample Size of 750, Test Length of 20, and Data Generated by the 2PL Model.

Calibrating model	Frequency	AUC	Discrimination	Difficulty
Rasch	4803	.743	.861	–0.313
2PL	2937	.861	1.734	–0.374
3PL	2260	.827	1.213	1.109

Note. AUC = area under curve; 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model.

In the 500 replications, we recorded the calibrating model chosen for each item according to whose AUC value was the largest. Table 3 shows the frequency of the Rasch, 2PL and 3PL models being selected as the best-fitting model among 10,000 items (20 items $\times$ 500 repetitions), as well as the average AUC, discrimination and difficulty parameters of the items for different calibrating model. The results of Table 3 showed that when the difficulty parameter was closer to zero, AUC preferred the Rasch model as the best item fitting model, and it was chosen the most often. When the discrimination parameter was relatively large, AUC preferred the 2PL model as the best item fitting model, which indicates that the prediction accuracy of item responses estimated by the model were also higher; thus, the average AUC value of the items selected for the 2PL model was the largest. In fact, for these 10,000 items under this condition, the AUC values of the items are significantly correlated with their discrimination parameters ($r = .369$, $p < .01$). When the data were generated by the 2PL model, the Rasch model was selected as the best-fitting model at the test level due to it obtaining the largest average AUC values, probably because the test had the most items with moderate difficulty. That said, AUC perform best on model selection when the test has most moderately difficult items.

Comparing Global Fit Indices: Average Values and Rejection Rates

To examine whether AUC could be used for model evaluation, we also examined several frequently used GOF indices—M2, CFI, and TLI for comparison. We selected GOF indices from two perspectives: (a) AUC has been recognized as a special GOF index in real settings; for example, Pham et al. (2021) outlined ‘‘. . . has the highest goodness-of-fit (AUC = 0.970) . . .’’ and Finlayson et al. (2018) stated ‘‘. . . goodness of fit, with an AUC of 0.73. . .’’; (b) although indices such as RMSEA, CFI, and TLI have been used in confirmatory factor analysis more, IRT literature uses them to represent GOF as well (Chernyshenko et al., 2001; Huggins-Manley & Han, 2017). The comparison is not primarily targeting classification accuracy. Instead, the accuracy (e.g., AUC) and mean squared error distance (e.g., standardized

root mean square residual, SRMR; see Asparouhov & Muthén, 2018) are both ways of evaluating the model fit.

The mean values and rejection rates of the RMSEA, CFI, TLI, and average AUC among the 500 replicates under each condition are shown in Table 4, with the cells where the calibrating models were the data generating models are in bold. In the columns corresponding to RMSEA in Table 4, the rejection rates refer to the frequency of using M2 statistics to reject the null hypothesis among the 500 repetitions by setting the nominal significance level at 0.05. For TLI and CFI, values less than 0.9 were considered unacceptable. Only the mean values of average AUC values are presented in Table 4 to determine the cutoff line that requires further discussion.

M2 statistics seem impractical when dealing with high-dimensional data regarding empirical Type I error rates. The rejection rates of M2 under the 1D_2PL and the 1D_3PL model cases were under the .05 significance level; the finding was similar to what Xu et al. (2017) reported in their study. However, when the generative model was the 2PL model and has a multidimensional structure, or the generative model was the 3PL model with four dimensions, the empirical Type I error rates of M2 were much higher than 0.05 and were nearly 1 with the increase in sample size and number of items.

The power of M2 was the rate of correctly rejecting the null hypothesis, while the estimate model did not match the generating model. The power reached 1 when calibrated with Rasch models regardless of the generated data structures. However, M2 did not show enough power to detect misfits when calibrating data generated by the unidimensional 3PL model with unidimensional 3PL, and it was below .1.

TLI and CFI performed similarly in that they both had large empirical Type I error rates when the generated model was the 4DS_2PL model. They were insensitive to model misspecification except when the generated model structure was four-dimensional, which is partly related to the well-fitted threshold set at 0.9.

As shown in Table 4, for the calibrating models identical to the data generating models, the mean values of CFI and TLI are all greater than 0.95 and the mean values of RMSEA are less than 0.05, representing an excellent fit. For the same data, the more complex the calibrating model is for the same data, the larger the corresponding CFI and TLI values and the smaller the RMSEA values are, even though the TLI already penalizes more complex models. The results make RMSEA, CFI, and TLI inappropriate for selecting more concise generative models.

Unlike other absolute GOF indices, for the same data, the degree of fit measured by the average AUC (i.e., the average model predicts accuracy) may not improve when a more complex estimate model is used. The average AUC also did not improve as the sample size increased. As shown in Table 4, when the structure of the generated data is 2DW, the AUC values are higher than other generating structures, all greater than 0.8. This suggests that thresholds for absolute fit criteria of average AUC may need to be set for each combination of conditions, which requires further discussion.

Table 4. Comparisons of Absolute GOF Indices.

row	Generated model	Test length	Simple Size	RMSEA			CFI			TLI			Average AUC		
				*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL
1	1D_2PL	20	750	.044/1.000	**.004/.042**	.001/.032	.948/.004	**.999/.000**	.999/.000	.948/.004	**1.001/.000**	1.006/.000	0.784	0.78	0.772
2			1500	.044/1.000	**.004/.036**	.000/.000	.947/.006	**.999/.000**	1.000/.000	.947/.006	**1.000/.000**	1.004/.000	0.782	0.779	0.773
3		40	750	.035/1.000	**.002/.022**	.000/.002	.966/.000	**1.000/.000**	1.000/.000	.966/.000	**1.001/.000**	1.008/.000	0.77	0.768	0.763
4			1500	.035/1.000	**.002/.026**	.000/.000	.966/.000	**1.000/.000**	1.000/.000	.966/.000	**1.000/.000**	1.004/.000	0.771	0.769	0.766
5	2DPS_2PL	20	750	.071/1.000	**.025/.880**	.002/.036	.900/.468	**.988/.000**	.999/.000	.899/.486	**.986/.000**	1.003/.000	0.796	0.824	0.802
6			1500	.072/1.000	**.026/.992**	.001/.014	.901/.444	**.989/.000**	1.000/.000	.900/.462	**.987/.000**	1.002/.000	0.797	0.826	0.808
7		40	750	.063/1.000	**.023/.998**	.000/.010	.916/.074	**.989/.000**	1.000/.000	.916/.076	**.989/.000**	1.004/.000	0.779	0.807	0.79
8			1500	.063/1.000	**.023/1.000**	.000/.000	.916/.026	**.990/.000**	1.000/.000	.916/.026	**.989/.000**	1.003/.000	0.779	0.808	0.795
9	2DW_2PL	20	750	.042/1.000	**.013/.372**	.000/.006	.986/.000	**.998/.000**	1.000/.000	.986/.000	**.998/.000**	1.003/.000	0.868	0.868	0.864
10			1500	.042/1.000	**.016/.740**	.004/.114	.986/.000	**.998/.000**	1.000/.000	.985/.000	**.998/.000**	1.000/.000	0.868	0.867	0.864
11		40	750	.033/1.000	**.012/.612**	.000/.000	.991/.000	**.999/.000**	1.000/.000	.991/.000	**.999/.000**	1.004/.000	0.86	0.86	0.862
12			1500	.034/1.000	**.016/.968**	.000/.000	.991/.000	**.998/.000**	1.000/.000	.991/.000	**.998/.000**	1.002/.000	0.861	0.86	0.864
13	4DS_2PL	20	750	.103/1.000	**.051/1.000**	.027/.926	.381/1.000	**.863/.770**	.966/.008	.367/1.000	**.846/.850**	.956/.034	0.745	0.812	0.738
14			1500	.104/1.000	**.050/1.000**	.017/.898	.372/1.000	**.869/.748**	.985/.000	.358/1.000	**.854/.834**	.981/.000	0.745	0.818	0.76
15		40	750	.076/1.000	**.041/1.000**	.017/.838	.677/1.000	**.915/.194**	.984/.000	.675/1.000	**.910/.276**	.982/.000	0.729	0.79	0.746
16			1500	.078/1.000	**.038/1.000**	.011/.852	.659/1.000	**.924/.076**	.994/.000	.657/1.000	**.920/.120**	.993/.000	0.729	0.792	0.758
17	1D_3PL	20	750	.044/1.000	.005/.044	**.002/.050**	.946/.006	.998/.000	**.999/.000**	.946/.006	1.000/.000	**1.005/.000**	0.782	0.779	0.77
18			1500	.043/1.000	.003/.038	**.000/.000**	.949/.002	.999/.000	**1.000/.000**	.948/.004	1.000/.000	**1.004/.000**	0.783	0.779	0.774
19		40	750	.036/1.000	.002/.018	**.000/.000**	.966/.000	1.000/.000	**1.000/.000**	.965/.000	1.001/.000	**1.008/.000**	0.77	0.768	0.764
20			1500	.036/1.000	.002/.038	**.000/.000**	.966/.000	1.000/.000	**1.000/.000**	.965/.000	1.000/.000	**1.004/.000**	0.77	0.768	0.766
21	2DPS_3PL	20	750	.072/1.000	.026/.916	**.003/.054**	.900/.462	.988/.000	**.999/.000**	.899/.488	.986/.000	**1.003/.000**	0.797	0.824	0.803
22			1500	.072/1.000	.026/.994	**.001/.012**	.900/.502	.988/.000	**1.000/.000**	.898/.534	.986/.000	**1.002/.000**	0.796	0.825	0.808
23		40	750	.063/1.000	.023/.996	**.000/.008**	.916/.070	.989/.000	**1.000/.000**	.916/.074	.988/.000	**1.004/.000**	0.779	0.808	0.791
24			1500	.063/1.000	.023/1.000	**.000/.000**	.917/.038	.990/.000	**1.000/.000**	.916/.038	.989/.000	**1.002/.000**	0.779	0.808	0.794

(continued)

Table 4. (continued)

row	Generated model	Test length	Simple Size	RMSEA			CFI			TLI			Average AUC		
				*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL	*Rasch	*2PL	*3PL
25	2DW_3PL	20	750	.042/1.000	.013/.396	**.000/.004**	.985/.000	.998/.000	**1.000/.000**	.985/.000	.998/.000	**1.003/.000**	0.867	0.867	0.863
26			1500	.041/1.000	.015/.706	**.003/.090**	.986/.000	.998/.000	**1.000/.000**	.986/.000	.998/.000	**1.000/.000**	0.867	0.866	0.864
27		40	750	.034/1.000	.013/.634	**.000/.000**	.991/.000	.999/.000	**1.000/.000**	.991/.000	.998/.000	**1.004/.000**	0.861	0.86	0.862
28			1500	.034/1.000	.016/.966	**.000/.000**	.991/.000	.998/.000	**1.000/.000**	.991/.000	.998/.000	**1.002/.000**	0.86	0.859	0.863
29	4DS_3PL	20	750	.103/1.000	.051/1.000	**.026/.916**	.372/1.000	.859/.780	**.964/.012**	.358/1.000	.842/.866	**.955/.034**	0.744	0.811	0.736
30			1500	.104/1.000	.050/1.000	**.017/.870**	.361/1.000	.866/.750	**.986/.000**	.346/1.000	.850/.858	**.982/.000**	0.744	0.818	0.76
31		40	750	.076/1.000	.041/1.000	**.017/.848**	.675/.996	.914/.204	**.984/.000**	.673/.998	.909/.292	**.982/.000**	0.729	0.789	0.745
32			1500	.077/1.000	.038/1.000	**.010/.826**	.669/1.000	.925/.054	**.994/.000**	.667/1.000	.921/.084	**.994/.000**	0.729	0.792	0.758

Note. The values before the slash represent the average model fit; the values after the slash represent the proportion of models having unacceptable model fit among all repetitions. Under RMSEA columns, rejection rates are in terms of the M2 statistic. 1D_2PL and 1D_3PL are one-dimensional 2PL model and one-dimensional 3PL model, respectively; 2DW_2PL and 2DW_3PL indicate two-dimensional within-item 2PL model and two-dimensional within-item 3PL model, respectively; 2DPS_2PL and 2DPS_3PL denote two-dimensional partially simple structure 2PL model and two-dimensional partially simple structure 3PL model, respectively; 4DS_2PL and 4DS_3PL are four-dimensional simple structure 2PL model and four-dimensional simple structure 3PL model, respectively; GOF = goodness of fit; RMSEA = root mean square error approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; AUC = area under curve; 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model. *Rasch, *2PL, and *3PL indicate calibrating models with the same model structure as the corresponding data generation models indicated in the columns. Values in cells where the calibrating model is the same as the data generating model are in bold.

Discussion

AUC is the standard choice in assessing classification accuracies, mainly produced by predictive models. It circumvents the necessity of subjective threshold decisions. This study is an initial attempt to systemically investigate the behaviors of AUC in the context of psychometric models, mainly IRT ones as the simulations demonstrated. Based on the present findings, we do not recommend using AUC solely for evaluating psychometric models, neither relative model comparisons nor absolute fit assessments. As the simulation results entailed, none of the statistics delivers a panacea to various model-fitting scenarios; again, this finding verifies the recommended practice of using multiple GOF statistics when judging modeling qualities.

In terms of relative comparisons, AIC, BIC, and AUC perform inconsistently across simulated conditions; it is reasonable because the mechanisms are determined by the indices' components and calculation rules. AIC and BIC are based on likelihood and inflated by the numbers of parameters, while AUC is modeling-free and parameter numbers are off consideration. Theoretically, since the mechanisms are substantively different, using AUC to select models, like AIC and BIC do, may not be reasonable, yet such adoptions have been found in the literature. In other words, AIC and BIC are informing researchers how good a model fits for a specific misclassification cost, and AUC indicates how good the model would work, on average, across all misclassification costs. Surprisingly, AUC did outperform its counterparts in many situations. Jointly using AIC, BIC, and AUC can be a solution, which has already been reported in published studies (Xiao et al., 2019).

Presumably, AUC could be regarded as an absolute fit index as it essentially utilizes information from residuals akin to the famous SRMR. Following the same fashion, AUC was expected to have a proposed cutoff (i.e., rule of thumb) to recognize a model that fits the data well. In terms of numeric ranges, AUC was not as ''delicate'' as other indices (e.g., CFI and TLI) of which perfect conditions (i.e., the chosen model was the true model) produce values from 0.95 to 1. The highest AUC, seen from the simulation, rarely reaches 0.90, which clearly conflicts with what researchers usually think is a delicate threshold. Furthermore, finding a cutoff to discover a true model seems not plausible. To be concrete, selecting any values between 0.8 and 0.9 does not grant a correct decision in Table 4.

The causes of AUC's incapacities in the psychometric model evaluations can be scattered and indeed demand more investigations in the future, but some obviously suspicious can be reasoned in this discussion. AUC is critically a type of discrimination indicating the possibility that a ''presence'' will receive a higher predicted value than an ''absence'' (Hosmer & Lemeshow, 2000, p.162). Indirectly using the actual probability values, AUC is insensitive to transformations of the predicted probabilities that preserve their ranks (Ferri et al., 2005), making it plausible that a poor model possesses a good discrimination power, and vice versa. In the case where probabilities for presences are only moderately higher than those for absences, a well-fitted model can have low AUC (Lemeshow & Hosmer, 1982): if in a test where some items do show the patterns as mentioned earlier, average AUC is likely to fail

to discriminate. AUC weights false positive and false negative errors equally, while in many applications of distribution modeling, these two errors may not be equally important. In situations where misclassification costs are inconsistent, summarizing over all possible threshold values is questionable. The way IRT likelihood function weights the two errors is unknown, and therefore, summarizing corresponding performance over all possible thresholds may not embrace the features of psychometric models. However, as the sample size increases, it becomes too sensitive to reject the fitted model, which may have a tolerable or negligible degree of the misfit. Previous work suggested that AIC and BIC showed very poor performances in finding the correct 3PL model (Kang & Cohen, 2007; Lin & Dayton, 1997). We believe that this may be related to the quality of the items and that when using the 3PL model to generate data, a larger guessing parameter might enable the AIC and BIC to be more inclined to choose the correct 3PL model.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Z.J. was supported by National Natural Science Foundation of China for Young Scholars under Grant 72104006 and Peking University Health Science Center under Grant BMU2021YJ010.

ORCID iD

Dexin Shi https://orcid.org/0000-0002-4120-6756

Notes

1. Equation (1) is the summation of the areas of a series of trapezoids. Since the positive and negative instances in this case have different scores (predicted probabilities), the shape enclosed by the two successive points and the X-axis in the ROC plane is actually a special case of trapezoids—rectangles. More complicated situations considering positive and negative instances with same scores can be seen in Tom Fawcett (2006).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus*. http://www.statmodel.com/download/SRMR2.pdf

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1–15. https://doi.org/10.1348/000711002159617

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), Article 238.

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). *Model-based collaborative filtering analysis of student response data: Machine-learning item response theory*. International Educational Data Mining Society.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk*, *24*(1), 20–46.

Bozdogan, H. (1987). Model selection and Akaike s information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*, 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2 tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194. https://doi.org/10.1348/000711005X66419

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data* (CRESST Report 839). National Center for Research on Evaluation, Standards, and Student Testing. https://eric.ed.gov/?id=ED555726

Carlson, S. E. (2018). *Identifying students at risk of dropping out: Indicators and thresholds using ROC analysis* (Doctor of Education (EdD), 114). George Fox University. https://digitalcommons.georgefox.edu/edd/114

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, *41*, 372–387. https://doi.org/10.1177/0146621617692079

Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., Ma, H., & Hu, G. (2019, November 3–7). Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2397–2400). Association for Computing Machinery.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562.

DeMars, C. E. (2016). Partially compensatory multidimensional item response theory models: Two alternate model forms. *Educational and Psychological Measurement*, *76*(2), 231–257. https://doi.org/10.1177/0013164415589595

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Lawrence Erlbaum Associates. (Original work published 2000)

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Ferri, C., Flach, P., Hernández-Orallo, J., & Senad, A. (2005). Modifying ROC curves to incorporate predicted probabilities. In *Proceedings of the Second Workshop on ROC Analysis in Machine Learning* (pp. 33–40). http://dmip.webs.upv.es/ROCML2005/papers/ferriCRC.pdf

Finlayson, K. J., Parker, C. N., Miller, C., Gibb, M., Kapp, S., Ogrin, R., & . . .Edwards, H. E. (2018). Predicting the likelihood of venous leg ulcer recurrence: The diagnostic accuracy of a newly developed risk assessment tool. *International Wound Journal*, *15*(5), 686–694.

Gonzalez, O. (2021). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods*, *26*(2), 236–254. https://doi.org/10.1037/met0000317

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression (2nd ed.)*. New York, NY: John Wiley and Sons.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Huggins-Manley, A. C., & Han, H. (2017). Assessing the sensitivity of weighted least squares model fit indexes to local dependence in item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 331–340.

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, *4*, Article 45. https://www.frontiersin.org/article/10.3389/feduc.2019.00045

Jiang, Z., Fitzgerald, S. R., & Walker, K. W. (2019a). Modeling time-to-trigger in library demand-driven acquisitions via survival analysis. *Library & Information Science Research*, *41*(3), Article 100968.

Jiang, Z., Walker, K., & Shi, D. (2019b). Applying AdaBoost to improve diagnostic accuracy. *Methodology*, *15*, 77–87.

Johnson, E., & Semmelroth, C. (2010). The predictive validity of the early warning system tool. *NASSP Bulletin, 94*(2), 120–134. https://doi.org/10.1177/0192636510380789

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X2 Item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358. https://doi.org/10.1177/0146621606292213

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*(7), 499–518. https://doi.org/10.1177/0146621608327800

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). *How deep is knowledge tracing?*https://arxiv.org/abs/1604.02416

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*(370), 336–344. https://doi.org/10.1080/01621459.1980.10477473

Kuhn, M. (2021). *caret: Classification and regression training* (R package version 6.0-90). https://CRAN.R-project.org/package=caret

Le, C. V., Pardos, Z. A., Meyer, S. D., & Thorp, R. (2018). Communication at scale in a MOOC using predictive engagement analytics. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren & B. du Boulay (Eds.), *Artificial intelligence in education* (pp. 239–252). Springer. https://doi.org/10.1007/978-3-319-93843-1_18

Lee, Y. (2019). Estimating student ability and problem difficulty using item response theory (IRT) and TrueSkill. *Information Discovery and Delivery*, *47*(2), 67–75. https://doi.org/10.1108/IDD-08-2018-0030

Lemeshow, S., & Hosmer, D. W. Jr. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, *115*(1), 92–106. https://doi.org/10.1093/oxfordjournals.aje.a113284

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, *22*, 249–264.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*, 71–101.

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

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*(1), 49–57. https://doi.org/10.1177/014662168500900105

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). *truncnorm: Truncated normal distribution* (Version 1.0-8). https://CRAN.R-project.org/package=truncnorm

Millard, S. P. (2013). *EnvStats: An R package for environmental statistics*. Springer. https://doi.org/10.1007/978-1-4614-8456-1

Nicholls, G., Wolfe, H., Besterfield-Sacre, M., & Shuman, L. (2010). Predicting stem degree outcomes based on eighth grade data and standard test scores. *Journal of Engineering Education*, *99*(3), 209–223. https://doi.org/10.1002/j.2168-9830.2010.tb01057.x

Niemeijer, K., Feskens, R., Krempl, G., Koops, J., & Brinkhuis, M. J. S. (2020, March 23–27). Constructing and predicting school advice for academic achievement: A comparison of item response theory and machine learning techniques. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 462–471). Association for Computing Machinery. https://doi.org/10.1145/3375462.3375486

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Park, J. Y., Cornillie, F., van der Maas, H. L., & Van Den Noortgate, W. (2019). A multidimensional IRT approach for dynamically monitoring ability growth in computerized practice environments. *Frontiers in Psychology*, *10*, Article 620. https://doi.org/10.3389/fpsyg.2019.00620

Pham, B. T., Luu, C., Van Phong, T., Trinh, P. T., Shirzadi, A., Renoud, S., & . . .Clague, J. J. (2021). Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling? *Journal of Hydrology*, *592*, Article 125615.

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). *Deep knowledge tracing*. https://arxiv.org/abs/1506.05908

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, *137*, 91–103. https://doi.org/10.1016/j.compedu.2019.04.009

Rasch, G. (1960). *Studies in mathematical psychology: I—Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

R Core Team. (2021). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), Article 77. https://doi.org/10.1186/1471-2105-12-77

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Silva, W., Spalenza, M., Bourguet, J. R., & de Oliveira, E. (2020). Towards a tailored hybrid recommendation-based system for computerized adaptive testing through clustering and IRT. In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020)* (*Vol. 1*, pp. 260–268). https://doi.org/10.5220/0009419902600268

Steiger, J. H., & Lind, J. C. (1980) May. *Statistically based tests for the number of common factors* [Paper presentation]. Annual Meeting of the Psychometric Society, Iowa City, IA, United States.

Stuit, D., O'Cummings, M., Norbury, H., Heppen, J., Dhillon, S., Lindsay, J., & Zhu, B. (2016). *Identifying early warning indicators in three Ohio school Districts* (REL 2016-118). Regional Educational Laboratory Midwest. https://ies.ed.gov/ncee/edlabs

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., & . . .Hu, G. (2018, April 26). Exercise-enhanced sequential modeling for student performance prediction. In. *Proceedings of the AAAI Conference on Artificial Intelligence* (*Vol. 32*, No. 1). https://ojs.aaai.org/index.php/AAAI/article/view/11864

Svicher, A., Romanazzo, S., De Cesaris, F., Benemei, S., Geppetti, P., & Cosci, F. (2019). Mental Pain Questionnaire: An item response theory analysis. *Journal of Affective Disorders*, *249*, 226–233.

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 148–177). SAGE. http://dx.doi.org/10.4135/9780857020994.n7

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Walker, K. W., & Jiang, Z. (2019). Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach. *The Journal of Academic Librarianship*, *45*(3), 203–212.

Windle, M., & Windle, R. C. (2017). The measurement of adolescent alcohol problems via item response theory and their 15-year prospective associations with alcohol and other psychiatric disorders. *Alcoholism: Clinical and Experimental Research*, *41*(2), 399–406.

Wright, B., & Masters, G. (1982). *Rating Scale analysis*. MESA Press.

Wu, R., Xu, G., Chen, E., Liu, Q., & Ng, W. (2017, April 3–7). Knowledge or gaming? Cognitive modelling based on multiple-attempt response. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 321–329). https://doi.org/10.1145/3041021.3054156

Wu, Z., Ioannidis, N. M., & Zou, J. (2020). Predicting target genes of non-coding regulatory variants with IRT. *Bioinformatics*, *36*(16), 4440–4448.

Xiao, Z., Shi, Z., Hu, L., Gao, Y., Zhao, J., Liu, Y., Xu, Q., & Huang, D. (2019). A new nomogram from the seer database for predicting the prognosis of gallbladder cancer patients after surgery. *Annals of Translational Medicine*, *7*(23), Article 738. https://doi.org/10.21037/atm.2019.11.112

Xu, J., Paek, I., & Xia, Y. (2017). Investigating the behaviors of M2 and RMSEA2 in fitting a unidimensional model to multidimensional data. *Applied Psychological Measurement*, *41*(8), 632–644.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262. https://doi.org/10.1177/014662168100500212