Journal Title
XX(X):1–16

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Assessing Fit of Item Response Theory Models using a Likelihood Ratio Randomisation Test and Fit Indices

Nina van Gerwen [1] and Dave Hessen[2]

Abstract

...

Keywords

item response theory, goodness-of-fit tests, fit indices, power, empirical alpha

Introduction

Item Response Theory (IRT) is an often used tool for designing and analysing tests or questionnaires in psychological, educational and organisational practices. The models in IRT measure latent traits – characteristics that are not directly observable (e.g., attitudes, intelligence, etc.) – through analysing answers to a test or questionnaire. The goal in IRT is to find the most parsimonious model that best describes the scores on the test items. To achieve this, the model must show a good fit. If a model is used that does not fit the

[1]Utrecht University, NL
[2]Utrecht University, NL

Corresponding author:
Nina van Gerwen, Utrecht University, Faculty of Social Sciences, Department of Methodology and Statistics, Padualaan 14, Utrecht, 3584 CH, NL.
Email: n.l.vangerwen@uu.nl

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

data, it can lead to dire consequences such as faults in the validity of your measurement (Crişan et al., 2017; Jiao & Lau, 2003; Zhao & Hambleton, 2017) and by extension your conclusion. Therefore, model fit should be assessed after fitting an IRT model to test/questionnaire data.

There are mainly two procedures that can be applied to quantify model fit. First, model fit can be assessed through goodness-of-fit tests. Goodness-of-fit tests are statistical hypothesis tests that describe how well the observed data follows the expected data under a given model.

Commonly used goodness-of-fit tests in IRT are a χ^2-difference test and Pearson's χ^2-test. The χ^2-difference and Pearson's χ^2-test, however, both suffer from issues. Pearson's test, for example, will not follow a χ^2-distribution when many score patterns are missing or have a low frequency, which is often the case – especially as test length increases. The χ^2-difference test instead is difficult to use for two often used IRT models: the two- (2PL) and three-parameter logistic (3PL) model. These models both specify the probability of scoring a dichotomous multiple choice item correctly as a logistic distribution dependent on item characteristics and a latent variable. The reason the χ^2-difference test is difficult to use for these models, is because the 3PL and its generalisations tend to suffer from consistency and estimation issues (Barton & Lord, 1981; Loken & Rulison, 2010). Another concern with the tests is power, which can be viewed as both a blessing and a curse. This is because many goodness-of-fit tests suffer from a lack of power when sample size is low. However, when sample size increases, the power to reject any reasonable model that does not perfectly fit, also increases (Curran et al., 1996). These findings suggest that there is a lack of usable goodness-of-fit tests to test certain IRT models.

Consequently, fit indices were introduced as a second procedure to assess model fit. Fit indices are mathematical descriptives that indicate how well a model fits to the data. Note, however, that fit indices alone are not sufficient to infer whether a model fits well as they are not an inferential statistic and only describe the overall fit. Taken together with goodness-of-fit tests, fit indices do allow researchers to have a more comprehensive overview of model fit. For example, when due to a lot of power, a goodness-of-fit test shows that the model should be rejected, fit indices can indicate whether the model is still reasonable.

Previous research in IRT has seen the development of multiple fit indices. (e.g., Y-Q_1; Yen, 1981, RMSEA$_n$; Maydeu-Olivares & Joe, 2014). For an overview of the use and limitations of several IRT fit indices, see Nye et al. (2020). Fit indices from

Structural Equation Modeling can also be used in IRT. Although research has been done to examine the performance of a few fit indices from Structural Equation Modeling in IRT (e.g., RMSEA & SRMR; Maydeu-Olivares & Joe, 2014), hardly any research has been done towards the use of the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) and the Comparative Fit Index (CFI; Bentler, 1990) in an IRT context. Only one paper examines the CFI (Yang, 2020) and two papers investigate the TLI (Cai et al., 2021; Yang, 2020) in an IRT setting. Both studies concluded that the two fit indices can add addititional insights in assessing model fit. However, there have been no studies that investigate the performance of the TLI and CFI when calculated through the χ^2 statistic and with a baseline model that allows for independence between test items. The calculations of the fit indices influence their performance and this should therefore be investigated further.

Furthermore, we research another possible fit index not investigated before in the field of IRT. This fit index is based on the identity coefficient by Zegers & ten Berge (1985), which indicates the degree to which two variables of absolute scales are identical.

To summarise, in the field of IRT there are not enough usable goodness-of-fit tests available to test the 2- and 3PL model and scarce studies investigating the performance of the TLI and CFI. We address these two issues by developing and evaluating a new Likelihood Ratio (LR) goodness-of-fit test, named the LR Randomisation test, and assessing the performance of the TLI and CFI based on calculations not researched before in an IRT setting.

The present study

In order to evaluate the LR Randomisation test for the 2- and 3PL in IRT and to better understand the possible uses of the CFI and TLI in an IRT setting, the present study answers the following three research questions through a simulation study:

1. What sample size is necessary at different test lengths for the LR Randomisation test to perform well?
2. How does the performance of the LR Randomisation test compare to the performance of a χ^2-difference and Pearson's χ^2-test?
3. What is the performance of the TLI and CFI in IRT?

Results from the simulation study are reported to answer the research questions. Furthermore, empirical data from the Law School Admission Test (LSAT) are used to illustrate the value of the LR Randomisation test, the TLI and the CFI.

Methods

Before we share the methodology of the simulation studies, let us first examine a brief summary on the statistical theory associated with our study. In IRT, the goal is to find a model that best describes scores on test items. To achieve this, IRT presupposes three assumptions: (1) conditional independence of items given the latent trait, denoted by θ, (2) independence of observations and (3) the response to an item can be modeled by an item response function (IRF). An IRF is a mathematical equation that relates the probability to score a certain category on an item to θ. Below, you find the IRF for the 3PL (Birnbaum, 1968):

$$P(X_i = 1|\theta, \alpha_i, \beta_i, \gamma_i) = \gamma_i + (1 - \gamma_i) \cdot \frac{e^{\alpha_i\theta - \beta_i}}{1 + e^{\alpha_i\theta - \beta_i}}, \tag{1}$$

where X_i is a random variable indicating the response to item i. The probability of scoring a 1 on item i in the 3PL depends on (a) the latent variable θ, (b) the location parameter of the item β_i, which denotes how difficult the item is, (c) the scaling parameter of the item α_i, which shows how well item i discriminates between individuals who score a 0 and individuals who score a 1, and (d) an item-specific lower asymptote γ_i, which indicates whether there is a baseline probability of scoring a 1 (e.g., a multiple choice test with 4 options has a .25 baseline probability of scoring a 1). The 2PL is a special case of the 3PL, where γ_i equals 0 for all items.

Combining an IRF with the assumption of conditional independence allows us to model the probability of a complete score pattern to k items simply by factoring the probabilities for each item:

$$P(\boldsymbol{X} = \boldsymbol{x}|\theta, \boldsymbol{\nu}) = \prod_{i=1}^{k} \{P(X_i = 1|\theta, \boldsymbol{\nu})\}^{x_i} \cdot \{1 - P(X_i = 1|\theta, \boldsymbol{\nu})\}^{1-x_i}, \tag{2}$$

where $\boldsymbol{X}$ is now a random vector indicating a score pattern and $\boldsymbol{x}$ is the realisation of $\boldsymbol{X}$. Note that $\boldsymbol{\nu}$ is a vector containing item parameters for all k items. We can take this even further by taking into account the assumption that persons are randomly sampled from a population. The joint marginal probability of a score pattern of a randomly sampled individual then becomes:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \int \prod_{i=1}^{k} \{P(\boldsymbol{X} = \boldsymbol{x}|\theta, \boldsymbol{\nu})\} \, \phi(\theta) \, d\theta, \tag{3}$$

where $\phi(\theta)$ is the univariate density of the latent variable θ. In order to solve this equation, the density of θ has to be specified. For example, $\phi(\theta)$ can be specified as a standard normal distribution. With the joint marginal probability and the assumption of independence of observations, we can construct a likelihood function and estimate $\boldsymbol{\nu}$ through marginal maximum likelihood estimation:

$$\mathcal{L}(\boldsymbol{\nu}) = \prod_{\boldsymbol{x}} (P(\boldsymbol{X} = \boldsymbol{x}))^{n_{\boldsymbol{x}}}, \tag{4}$$

where $n_{\boldsymbol{x}}$ is the frequency of score pattern $\boldsymbol{x}$. Maximisation of $\mathcal{L}(\boldsymbol{\nu})$ would then lead to the estimation of the item parameters $\hat{\boldsymbol{\nu}}$. This is generally how in IRT a model is fitted to data.

After a model has been fitted as described above, the next step is to test how well the model fits the data. There are multiple options to assess model fit. Usually, both goodness-of-fit tests and fit indices are used. Commonly used goodness-of-fit tests are the χ^2-difference test and the Pearson's χ^2-test. The χ^2-difference test uses the following likelihood statistic:

$$\frac{\mathcal{L}_0}{\mathcal{L}_a}, \tag{5}$$

where $\mathcal{L}_0$ is the likelihood of the model you fit (i.e., the null model) and $\mathcal{L}_a$ is the likelihood of an alternative model under which the null model has to be nested. According to Willks' theorem (Wilks, 1938), under the null hypothesis this likelihood statistic will asymptotically follow a χ^2 distribution as sample size increases. Alternatively, Pearson's χ^2-test is based on the statistic below:

$$\sum_{\boldsymbol{x}} \frac{(O_{\boldsymbol{x}} - E_{\boldsymbol{x}})^2}{E_{\boldsymbol{x}}}, \tag{6}$$

where $O_{\boldsymbol{x}}$ is the observed frequency for score pattern $\boldsymbol{x}$ and $E_{\boldsymbol{x}}$ is the expected frequency for score pattern $\boldsymbol{x}$ given a model. This statistic also asymptotically follows a χ^2 distribution. The goodness-of-fit test that we develop and test in the present study is based on the hereunder likelihood statistic:

$$\frac{max(\mathcal{L}_0)}{\prod_{j=1}^{g} max(\mathcal{L}_j)}, \tag{7}$$

where L_0 is the likelihood of the chosen model for the whole dataset and L_j is the likelihood of the chosen model for group j, which is gained by randomly assigning the

observations to g groups. Wilks' theorem also applies to this likelihood statistic. Because all three aforementioned tests asymptotically follow a χ^2 distribution, they can be used for Null Hypothesis Significance Testing.

For fit indices, we research the performance of the TLI and CFI. These indices are estimated through the use of a baseline and saturated model. Our baseline model is a complete-independence model, which has the following IRF:

$$P(X_i = 1|\beta_i) = \frac{e^{-\beta_i}}{1 + e^{-\beta_i}}, \tag{8}$$

where the probability of scoring a 1 on item i is dependent only on the difficulty of the item and no longer on a latent variable. We argue that this is an appropriate baseline model, because the IRF entails that the joint probability distribution is simply the product of the marginal probability distributions and the items are independent. In this baseline model, the probability of scoring a 1 on item i is simply the proportion of people who score a 1 on item i. From this, the maximum likelihood estimate of β_i, denoted by $\hat{\beta}_i$, can then mathematically be derived to:

$$\hat{\beta}_i = ln(\frac{n - n_i}{n_i}),$$

where n_i is the number of observations who scored a 1 on item i and n is the total number of observations. In the saturated model there are as many parameters as data points, which leads to a perfect fit. We choose the following saturated model:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \pi_{\boldsymbol{x}}, \tag{9}$$

where there no longer is an IRF. Instead, perfect model fit is gained by allowing each score pattern to have their own parameter ($\pi_{\boldsymbol{x}}$). The maximum likelihood estimate of $\pi_{\boldsymbol{x}}$, denoted by $\hat{\pi_{\boldsymbol{x}}}$, is then the relative frequency of the score pattern:

$$\hat{\pi_{\boldsymbol{x}}} = \frac{n_{\boldsymbol{x}}}{n},$$

where $n_{\boldsymbol{x}}$ is the number of observations with score pattern $\boldsymbol{x}$ and n is the total number of observations. Using both the baseline and saturated model, the fit indices can be calculated through the following formulae:

$$\text{CFI} = 1 - \frac{\max\{(\chi^2_T - df_T), 0\}}{\max\{(\chi^2_T - df_T), (\chi^2_0 - df_0), 0\}}, \tag{10}$$

$$\text{TLI} = \frac{\chi_0^2/df_0 - \chi_T^2/df_T}{\chi_0^2/df_0 - 1}. \tag{11}$$

In both equations, χ_T^2 is the result of a χ^2-difference test between the tested model and the saturated model with df_T degrees of freedom, and χ_0^2 is the outcome of a χ^2-difference test between the baseline model and the saturated model with df_0 degrees of freedom.

Besides the TLI and CFI, we also research the performance of a new fit index based on the identity coefficient (ICFI). In a similar fashion to the Pearson's χ^2 test, the NFI measures model-fit by comparing the observed score pattern frequencies to the expected score pattern frequencies under a given model. The NFI can be calculated through the following formula:

$$\text{NFI} = \frac{2 \cdot \sum_{\boldsymbol{x}} O_{\boldsymbol{x}} \cdot \sum_{\boldsymbol{x}} E_{\boldsymbol{x}}}{\sum_{\boldsymbol{x}} \{O_{\boldsymbol{x}}\}^2 + \sum_{\boldsymbol{x}} \{E_{\boldsymbol{x}}\}^2} \tag{12}$$

where ... . The NFI is bound between 0 and 1, as proven in Appendix A.

Simulation study I

In the present study, we consider IRT with unidimensional θ, dichotomous test items and the IRF for the 2- and 3PL. In order to evaluate the performance of the LR Randomisation test and the TLI and CFI in this setting, we conduct a simulation study. Specifically, for the LR Randomisation test we investigate whether whether the test is robust and at which rate the test is able to reject a misspecified model under different conditions. Furthermore, we research how well the LR Randomisation test performs compared to a χ^2-difference test and Pearson's χ^2-test. For the TLI and CFI, we investigate their values under correct and incorrect model specification. The aim of this is to examine whether the fit indices are a useful addition for assessing model fit in IRT modelling.

Data generation

Data generation starts by first sampling the person parameters θ from a standard normal distribution. Then, either the 2PL or 3PL is chosen as basis for the data generation (see below). We choose to keep item parameters static over all simulations. For the difficulty parameter β_i, we choose the values [-1, -.5, 0, .5, 1]. As for the discrimination parameter α_i, we choose from the values [0.7, 0.85, 1, 1.15, 1.3]. To create a more realistic scenario, these two parameters were then matched with one another in order to make sure that for

every item difficulty, there are low and high discriminating items. Finally, when the data generating model is the 3PL, we choose the value [.25] for each item for the pseudo-guessing parameter γ_i. Then, probabilities are estimated for all items on a test, given θ, the chosen model and item parameters. Finally, a matrix of simulated responses to a dichotomous test is created by sampling from a binomial distribution using the estimated probabilities.

Simulation design

In the simulation study, we vary four factors. We examine three conditions of test length ($I = 5, 10, 20$), five conditions of sample size ($N = 200, 300, 500, 1000, 1500$), and two conditions for the model on which data generation is based (2PL, 3PL). This simulation design results in a total 30 conditions. Furthermore, within each of these conditions, we then vary the LR Randomisation test based on the number of groups ($G = 2, 3, 4$). In each replication of each condition, we fit the 2PL model. The 2PL is fitted using functions from the *ltm* package in R (Rizopoulos, 2006), which approximates marginal maximum likelihood estimation through the Gauss-Hermite quadrature rule. Then we obtain the results of the three fit indices and the three different types of goodness-of fit tests: (1) a χ^2-difference test, which is obtained by testing the 2PL under the 3PL with the constraint that all γ_i have to be equal through the mirt package (citation), (2) Pearson's χ^2-test, which is estimated through aggregating score patterns from the data and comparing the observed score pattern frequencies to the expected score pattern frequencies under the given model, and (3) the LR Randomisation test with a varying number of groups. We replicate each simulation condition 300 times.

Performance metrics

We study performance of the different goodness-of-fit tests by estimating both type I error and power. Both values are estimated by obtaining the detection rate, which is the number of times the null hypothesis is rejected divided by the total number of replications per condition. Power is estimated when generating data under the 3PL and fitting the 2PL model. Type I error is estimated when generating data under the 2PL and fitting the 2PL model. With these values, we compare the different type of tests with one another under every condition, where a test with lower power or higher type I error is noted as performing worse. For all tests, we chose a level of significance of $\alpha = .05$.

To measure the performance of the TLI and CFI, we estimate the mean and standard

error (SE) of each fit index under every condition. Then, we can inspect whether the average TLI and CFI values decrease in the conditions where data is generated under the 3PL, compared to when data is generated under the 2PL.

Results

To asses the performance of the LR Randomisation test and the three fit indices, we conducted the above described simulation study. Non-convergence of the models occured $< .001$ percent of the time. Furthermore, when test length was 20, the empirical α and power estimates for Pearson's χ^2-test were not calculated. This is because of the large amount of possible score patterns (2^20) in this condition, which required too much computational power for the scope of the current study.

Table 1 presents the empirical rejection rates test at the $\alpha = .05$ level for the different χ^2-based goodness-of-fit tests. The asymptomatic can be observed clearly. As sample size increases, empirical α values approach the nominal level of .05. At large sample sizes, the rejection rates of the LR Randomisation and χ^2 difference test seem to not deviate far from the nominal level for different levels of test lengths. Therefore, we can state that these two tests statistics indeed asymptotically follow their supposed χ^2 distribution. However, this also means that for low sample sizes, the test statistic does not yet completely follow a χ^2 distribution, resulting in higher rejection rates. For Pearson's χ^2 test, the table also shows that as the number of items increases, the empirical α increases dramatically from 0.06 to 0.22. This is in accordance with the issues we discussed in the introduction, where the test does not follow a χ^2 distribution when many score patterns are low or missing. These results can be seen as confirmation for the theory and design behind our study. Important to remember, however, is that empirical data is almost never as clean.

Next, we present the power estimates of the different goodness-of-fit tests in *Table 2*. For the χ^2-difference test, the table presents clearly that as sample size and/or test length increase, the power to reject the tested model also increases in favour of the true model up to a power of 1.00 with a test length of 20 and sample size of 1500. For the LR Randomisation test, the table shows that the power estimates still converge to the nominal alpha level of .05. This means that when not under the null hypothesis, the LR statistic of our test still asymptotically follows a χ^2-distribution.

Even though usually these tests only follow their distribution under the null hypothesis.

Although an interesting finding, these results entail that the test can not be used for assessing model-fit in its current state. Therefore, we were inspired to run a follow-up simulation study that we discuss below.

NOTE: why is power of pearson's χ^2 test also so low/barely increasing?

Table 3 and *Table 4* exhibit the mean and standard error for the TLI, CFI and our NFI under correct and incorrect model specification respectively. Both tables show that as sample size increases, the values on the fit indices move closer to 1 (i.e., a good model fit). However, if test length increases, fit index values move closer to 0 instead. This can be interpreted in the following way: as test length increases, the fit of the baseline model becomes relatively better compared to the tested model. This result has not been found before in IRT literature. However, it might be due to our choice of baseline model. This result therefore further strengthens the argument made by (Van Laar & Braeken, 2021) who emphasized that the choice of baseline model is important when working with fit indices.

Even though fit indices value tend to 0 as test length increases, the values for the TLI and CFI are nevertheless on average higher under correct model specification compared to incorrect model specification (e.g., for $N = 400$ & $I = 10$, the mean of the TLI is 0.11 under correct model specification, whereas it is .00 under incorrect model specification). This means that these fit indices can be used for model-fit decisions. However, test length would have to be taken into account if a 'rule-of-thumb' is desired.

The tables also show that for cases where there is a small sample size and few test items, there are high standard errors. This indicates that when assessing model-fit in low sample size settings, a point estimate might not be enough. Other methods such as bootstrapping might be required in order to be able to decide which model fits best. For high sample sizes and/or larger test lengths, this should not pose an issue.

For the TLI, negative values sometimes occurred under model misspecification, this can occur when the degrees of the hypothesized model are small and correlations among observed variables are low (Wang & Wang, 2019; Widaman & Thompson, 2003). In the current study, we rounded these slightly negative values to 0.

The new fit index, based on the cocurrence of score patterns and their expected occurence, shows less promise. The values seem to not differ in incorrect model specification compared to correct model specification. Therefore, this fit index was not able to decide which of the two models fit the data better in the current study.

Table 1. Empirical Alpha estimates for different goodness-of-fit tests

Conditions		Goodness-of-fit test				
I	*N*	LR2	LR3	LR4	χ^2	P-χ^2
5	200	0.09	0.06	0.10	0.05	0.06
	300	0.07	0.08	0.07	0.04	0.04
	500	0.07	0.07	0.06	0.03	0.05
	1000	0.04	0.06	0.07	0.03	0.06
	1500	0.07	0.07	0.06	0.02	0.04
10	200	0.08	0.11	0.11	0.03	0.22
	300	0.04	0.05	0.09	0.04	0.20
	500	0.04	0.05	0.06	0.06	0.17
	1000	0.02	0.06	0.06	0.03	0.12
	1500	0.03	0.05	0.04	0.02	0.11
20	200	0.07	0.08	0.12	0.06	–
	300	0.05	0.08	0.10	0.04	–
	500	0.07	0.08	0.06	0.03	–
	1000	0.05	0.05	0.05	0.02	–
	1500	0.04	0.05	0.05	0.03	–

Note. Fitted model = two-parameter logistic model; Data generating model = two-parameter logistic model; *I* = test length; *N* = sample size; LR2 = LR Randomisation test with $g = 2$; LR3 = LR Randomisation test with $g = 3$; LR4 = LR Randomisation test with $g = 4$; χ^2 = χ^2-difference test under the three-parameter logistic model with no constraints; P-χ^2 = Pearson's χ^2 test.

Simulation Study II

To further investigate the possible uses of our LR test, we decided to investigate the uses of the LR statistics with existing groups. Again we investigate the same conditions of test length ($I = 5, 10, 20$), sample size ($N = 200, 300, 500, 1000, 1500$) and the number of groups ($G = 2, 3, 4$).

However, compared to Simulation Study I, data generation is now from a multigroup setting. More specifically, we vary data generation in two ways compared to the first simulation study.

Compared to Simulation Study I, the only difference is that now we generate data from a multigroup setting. This means that we again vary the number of groups ($G = 2, 3, 4$) and this time also vary the item parameters per group. XX. The only differences between SS 1 are in data generation.

Table 2. Power estimates for different goodness-of-fit tests

Conditions		Goodness-of-fit test				
I	*N*	LR2	LR3	LR4	χ^2	P-χ^2
5	200	0.10	0.09	0.12	0.06	0.06
	300	0.09	0.08	0.14	0.08	0.06
	500	0.07	0.09	0.12	0.11	0.04
	1000	0.06	0.05	0.08	0.12	0.07
	1500	0.05	0.07	0.05	0.19	0.07
10	200	0.09	0.14	0.20	0.25	0.29
	300	0.11	0.11	0.13	0.26	0.21
	500	0.05	0.09	0.08	0.32	0.22
	1000	0.08	0.07	0.07	0.58	0.16
	1500	0.06	0.06	0.06	0.75	0.22
20	200	0.10	0.11	0.16	0.52	–
	300	0.08	0.10	0.11	0.66	–
	500	0.05	0.08	0.08	0.89	–
	1000	0.03	0.05	0.06	0.97	–
	1500	0.03	0.05	0.07	1.00	–

Note. Fitted model = two-parameter logistic model; Data generating model = three-parameter logistic model; *I* = test length; *N* = sample size; LR2 = LR Randomisation test with $g = 2$; LR3 = LR Randomisation test with $g = 3$; LR4 = LR Randomisation test with $g = 4$; χ^2 = χ^2-difference test under the three-parameter logistic model with no constraints; P-χ^2 = Pearson's χ^2 test.

Results

Empirical example

To examine the possible uses of our LR Randomisation test, the TLI and CFI, we estimated and interpreted their values on a real-life dataset, gained from XX, which contains information about XXX ($N = ...$). The dataset has XX items on a dichotomous scale that together measure XXX. We fitted the 3PL model with XX.

Discussion

The current study investigated earlier existing and new possible measures of model-fit in the context of IRT.

Limitations

- didn't test 3PL - didn't test multidimensional IRT - didn't test polytome IRT

Table 3. TLI and CFI values under correct model specification

Conditions		TLI	CFI	NFI
I	*N*	M (SE)	M (SE)	M (SE)
5	200	0.63 (0.19)	0.75 (0.12)	0.98 (0.01)
	300	0.74 (0.14)	0.82 (0.09)	0.98 (0.01)
	500	0.84 (0.09)	0.90 (0.06)	0.99 (0.00)
	1000	0.92 (0.05)	0.94 (0.03)	0.99 (0.00)
	1500	0.95 (0.03)	0.96 (0.02)	1.00 (0.00)
10	200	0.06 (0.06)	0.23 (0.05)	0.69 (0.04)
	300	0.11 (0.05)	0.27 (0.04)	0.77 (0.03)
	500	0.19 (0.04)	0.33 (0.04)	0.85 (0.02)
	1000	0.32 (0.03)	0.45 (0.03)	0.91 (0.01)
	1500	0.42 (0.03)	0.53 (0.03)	0.94 (0.01)
20	200	0.04 (0.02)	0.14 (0.02)	0.03 (0.01)
	300	0.05 (0.02)	0.14 (0.02)	0.04 (0.02)
	500	0.05 (0.01)	0.14 (0.01)	0.06 (0.02)
	1000	0.07 (0.01)	0.16 (0.01)	0.11 (0.02)
	1500	0.08 (0.01)	0.17 (0.01)	0.17 (0.03)

Note. Fitted model = two-parameter logistic model; Data generating model = two-parameter logistic model; *I* = test length; *N* = sample size; TLI = tucker-lewis index; CFI = comparative fit index; M = mean; SE = standard error.

future research can do these, also the test can be used in any field of statistics. not just IRT.

Conclusion

TLI and CFI have possible use in IRT and should be used more commonly, such as in a similar fashion as in SEM. χ^2 based goodness-of-fit tests can be used here and there.

Model appraisal in IRT remains an issue that requires more light to be shed upon in future research.

Table 4. TLI and CFI values under incorrect model specification

Conditions		TLI	CFI	NFI
I	*N*	M (SE)	M (SE)	M (SE)
5	200	0.33 (0.31)	0.55 (0.19)	0.98 (0.01)
	300	0.49 (0.24)	0.66 (0.16)	0.99 (0.00)
	500	0.66 (0.18)	0.77 (0.12)	0.99 (0.00)
	1000	0.81 (0.11)	0.87 (0.07)	1.00 (0.00)
	1500	0.86 (0.08)	0.91 (0.05)	1.00 (0.00)
10	200	0.00 (0.04)	0.11 (0.03)	0.76 (0.05)
	300	0.00 (0.04)	0.13 (0.03)	0.83 (0.03)
	500	0.00 (0.03)	0.17 (0.03)	0.89 (0.02)
	1000	0.07 (0.03)	0.24 (0.02)	0.94 (0.01)
	1500	0.15 (0.03)	0.31 (0.02)	0.96 (0.01)
20	200	0.00 (0.01)	0.06 (0.01)	0.04 (0.02)
	300	0.00 (0.01)	0.06 (0.01)	0.06 (0.03)
	500	0.00 (0.01)	0.07 (0.01)	0.10 (0.03)
	1000	0.00 (0.01)	0.07 (0.01)	0.17 (0.03)
	1500	0.00 (0.01)	0.08 (0.01)	0.24 (0.03)

Note. Fitted model = two-parameter logistic model; Data generating model = three-parameter logistic model; *I* = test length; *N* = sample size; TLI = tucker-lewis index; CFI = comparative fit index; M = mean; SE = standard error.

References

Barton, M. A. and Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8.

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

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Cai, L., Chung, S. W., and Lee, T. (2021). Incremental model fit assessment in the case of categorical data: Tucker–lewis index for item response theory modeling. *Prevention Science*, pages 1–12.

Crişan, D. R., Tendeiro, J. N., and Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional irt models. *Applied Psychological Measurement*, 41(6):439–455. PMID: 28804181.

Curran, P., West, S., and Finch, J. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*,

1(1):16–29.

Darrell Bock, R. and Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, 35(2):179–197.

Hambleton, R. K. and Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice*, 12(3):38–47.

Jiao, H. and Lau, A. C. (2003). The effects of model misfit in computerized classification test. In *Annual meeting of the National Council of Educational Measurement, Chicago, IL*.

Kang, T. and Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, 31(4):331–358.

Krammer, G. (2018). The andersen likelihood ratio test with a random split criterion lacks power. *Journal of modern applied statistical methods: JMASM*, 17:eP2685.

Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63(3):509–525.

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

Maydeu-Olivares, A. and Joe, H. (2014a). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 49.

Maydeu-Olivares, A. and Joe, H. (2014b). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 49(4):305–328.

Nye, C. D., Joo, S.-H., Zhang, B., and Stark, S. (2020). Advancing and evaluating irt model data fit indices in organizational research. *Organizational Research Methods*, 23(3):457–486.

Orlando, M. and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1):50–64.

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Stone, B. M. (2021). The ethical use of fit indices in structural equation modeling: Recommendations for psychologists. *Frontiers in Psychology*, 12.

Team, R. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Team, R. C. (2021). *R: A Language and Environmental for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60 – 62.

Yang, X. (2020). *Comparing Global Model-Data Fit Indices in Item Response Theory Applications*. PhD dissertation, Florida State University, College of Education.

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

Zegers, F. E. and ten Berge, J. M. (1985). A family of association coefficients for metric scales. *Psychometrika*, 50:17–24.

Zhao, Y. and Hambleton, R. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, 8.