Assessing Fit of Item Response Theory Models using a Multi-Group Likelihood Ratio Test and Fit Indices

Journal Title
XX(X):2–22

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

Nina van Gerwen [1]

Prepared using sagej.cls

Abstract

In Item Response Theory (IRT), it is important to investigate the appropriateness of the models you fit to the data to ensure valid measurements and conclusions. The appropriateness of a model is usually assessed through goodness-of-fit tests and fit indices. However, for dichotomous IRT models, common goodness-of-fit tests suffer from various issues (Barton & Lord, 1981; Loken & Rulison, 2010). Previous research has also shown that many fit indices in IRT are also unable to determine model misspecification when fitting the two-parameter logistic model (2PL) (Nye et al., 2020). We build upon previous research by investigating a goodness-of-fit test based on multiple group analysis, the MG LR test, and three scarcely studied fit indices: (1) Tucker Lewis index (TLI; Tucker & Lewis, 1973), (2) Comparative Fit index (CFI; Bentler, 1990) and (3) Identity Coefficient Fit index (ICFI; Zegers & ten Berge, 1985). We examine the performance of these fit measures through a simulation study. The results suggests that the ICFI, Pearson's χ^2 and the MG LR test are unable to determine model misspecification when fitting the 2PL. In contrast, the TLI and CFI showed good performance in determining model misspecification in dichotomous IRT. A follow-up simulation study showed that the MG LR test is effective in detecting group differences, while insensitive to incorrect model specification. We believe this to be a unique feature of the test and speculate its value as an instrument in detecting group differences. We also recommend that the TLI and CFI become more standard in use in order to improve model appraisal in IRT.

Keywords

item response theory, goodness-of-fit tests, fit indices, power, group differences

Introduction

Item Response Theory (IRT) is an often used instrument for designing and analysing tests/questionnaires in psychological, educational and organisational practices. The models in IRT infer latent traits – characteristics that are not directly observable (e.g.,

[1]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

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

There are mainly two procedures that can be applied to quantify the appropriateness of a model. 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 follow the expected data given a model.

Common goodness-of-fit tests in IRT are a χ^2-difference test and Pearson's χ^2 test. These tests suffer from issues however. Pearson's χ^2 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 dichotomous IRT models: the two- (2PL) and three-parameter logistic (3PL) model. These models both specify the probability of scoring a dichotomous item correctly as a logistic function of a latent trait and two or three item characteristics. The issue with these models and their generalisations is that they suffer from consistency and estimation issues (Barton & Lord, 1981; Loken & Rulison, 2010). As a result, the χ^2-difference test is hard to use with the models. 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 dichotomous IRT models.

Therefore, we explore a new possible goodness-of-fit test to assess model misspecification in dichotomous IRT. The test we are interested in originates from multi-group analyses, where it can be used to detect group differences in existing groups. However, what if there were no existing groups, and groups are made by random assignment instead. In this scenario, if a wrong model is specified, would the test have power to reject this model in favour of the true model? We expect that the test will have power to detect this model misspecification. If the test does not turn out to have any power, it would mean that the test is insensitive to incorrect model specification. If the interest is then in testing for differences between existing groups, it means that the test ignores model (mis)specification and tests solely for group differences.

The second procedure of assessing model fit is through fit indices. Fit indices are mathematical descriptives that indicate how close the fit of a model is to a perfect fit. Note, however, that fit indices alone are not sufficient to infer whether a model fits well as they are not inferential statistics 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 within 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 fit indices in IRT, see Nye et al. (2020). One limitation of numerous fit indices was that they exhibited poor performance when the 2PL was fitted to the data. In addition to these indices, fit indices from Structural Equation Modeling (SEM) can also be used in IRT. Although the performance of multiple fit indices from SEM have been researched in IRT (e.g., RMSEA & SRMR; Maydeu-Olivares & Joe, 2014), hardly any past research investigated the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) and Comparative Fit Index (CFI; Bentler, 1990). We found only one paper that examined the CFI (Yang, 2020) and two papers that investigated the TLI (Cai et al., 2021; Yang, 2020) in IRT. Furthermore, these papers came to different conclusions. Where Yang (2020) concluded that the CFI and TLI showed unsatisfactory performance in distinguishing data following the 2PL and 3PL, Cai et al. (2021) concluded that the TLI showed promise as a valuable index in evaluating IRT models. A possible reason for these mixed results could be due to differences in baseline model and calculations. Previous research has exhibited the importance of baseline models for incremental fit indices such as the TLI and CFI (Widaman & Thompson, 2003; Van Laar & Braeken; 2021). Since the calculations and baseline model influence the performance of the fit indices, we were inspired to research the TLI and CFI further in dichotomous IRT. We examine the TLI and CFI to a greater degree by investigating their performance while calculated through the χ^2 statistic and with a baseline model that assumes independence between test items.

Additionally, we research another possible fit index not investigated before in IRT. This fit index is based on the identity coefficient by Zegers & ten Berge (1985), which indicates the degree to which two vectors are identical. Therefore, this coefficient can reflect the degree to which the observed score pattern frequencies follow the expected score pattern frequencies given a model.

To summarise, in dichotomous IRT there are not enough measures of model fit available 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 based on multi-group analysis, named the Multiple Group LR (MG LR) test. We also develop and evaluate a new fit index inspired by the identity coefficient by Zegers & ten Berge (1985), named the identity coefficient fit index (ICFI), and assess the performance of the TLI and CFI based on new calculations. Specifically, we answer and discuss the following four research questions through multiple simulation studies and an empirical example:

1. What sample size is necessary at different test lengths for the MG LR test to approximately follow a χ^2 distribution?
2. What is the power of the MG LR test to detect model misspecification when groups are based on random assignment when fitting the 2PL?
3. Under random group assignment, how does the performance of the MG LR test compare to the performance of a χ^2-difference and Pearson's χ^2 test in detecting model misspecification when working with the 2PL?
4. What is the performance of the TLI, CFI and ICFI in assessing model fit when the 2PL is fitted to the data?

Methods

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, (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 the latent trait. The IRF for the 3PL (Birnbaum, 1968) is:

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

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 trait 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 signifies how well item i discriminates between individuals with lower and higher θ, 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 question with 4 options has a 0.25 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 score pattern to k items 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 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. By assuming that persons are randomly sampled from a population, the joint marginal probability of a score pattern of an individual becomes:

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

where $\phi(\theta)$ is the univariate density of the latent trait θ. To solve this equation, the density of θ has to be specified. For example, $\phi(\theta)$ can be specified as a standard Gaussian distribution. With the joint marginal probability and 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 maximum likelihood estimates of the item parameters $\hat{\boldsymbol{\nu}}$. Generally this is how a model is fitted to test data in IRT.

After a model has been fitted, 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. Common goodness-of-fit tests are a χ^2-difference test and Pearson's χ^2 test. The χ^2-difference test uses the following LR statistic:

$$-2 \ln \left\{ \frac{\max(\mathcal{L}_0)}{\max(\mathcal{L}_a)} \right\}, \tag{5}$$

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 is nested. According to Willks' theorem (Wilks, 1938), under the null hypothesis this LR statistic will asymptotically follow a χ^2 distribution as sample size increases. Alternatively, the observed value of Pearson's χ^2 statistic is:

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

where $n_{\boldsymbol{x}}$ is the observed frequency for score pattern $\boldsymbol{x}$ and $\varepsilon_{\boldsymbol{x}}$ is the expected frequency for score pattern $\boldsymbol{x}$ given a model. This χ^2 statistic also asymptotically follows a χ^2 distribution as sample size approaches infinity. The MG LR test that we develop and evaluate in the present study is based on the LR statistic:

$$-2 \ln \left\{ \frac{\max(\mathcal{L}_0)}{\prod_{j=1}^{g} \max(\mathcal{L}_j)} \right\}, \tag{7}$$

where $\mathcal{L}_0$ is the likelihood of the chosen model for the whole dataset and $\mathcal{L}_j$ is the likelihood of the chosen model for group j. Group j can be an existing group, or the groups can be gained by randomly assigning observations to g groups. Wilks' theorem also applies to this LR statistic. Because all three aforementioned tests asymptotically follow a χ^2 distribution under the null hypothesis, they can be used to assess goodness-of-fit.

For fit indices, we research the performance of the TLI, CFI and ICFI. The TLI and CFI are estimated by comparing the tested model to a baseline and saturated model. Our baseline model is a complete-independence model with 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 β_i and no longer on a latent trait. We argue that this is an appropriate baseline model, because the IRF entails that the joint probability distribution is 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 the proportion of people who score a 1 on item i. From this, it follows that the maximum likelihood estimate of β_i is:

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

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

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

where there no longer is an IRF. Instead, perfect fit is gained by allowing each score pattern to have its own parameter $\pi_{\boldsymbol{x}}$. The maximum likelihood estimate of $\pi_{\boldsymbol{x}}$ 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. With a baseline and saturated model, the TLI and CFI can be calculated through the following formulae:

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

$$\text{TLI} = \frac{\chi^2_0/df_0 - \chi^2_T/df_T}{\chi^2_0/df_0 - 1}. \tag{11}$$

In both equations, χ^2_T is the observed value of the LR statistic of a χ^2-difference test between the tested model and the saturated model with df_T degrees of freedom, and χ^2_0 is the observed value of the LR statistic of a χ^2-difference test between the baseline model and the saturated model with df_0 degrees of freedom. We also research the performance of the ICFI. Parallel to the Pearson's χ^2 test, the ICFI gauges model appropriateness by comparing the observed score pattern frequencies to the expected score pattern frequencies given a model:

$$\text{ICFI} = \frac{2 \cdot \sum_{\boldsymbol{x}} n_{\boldsymbol{x}} \cdot \varepsilon_{\boldsymbol{x}}}{\sum_{\boldsymbol{x}} {n_{\boldsymbol{x}}}^2 + \sum_{\boldsymbol{x}} {\varepsilon_{\boldsymbol{x}}}^2}, \tag{12}$$

where $n_{\boldsymbol{x}}$ is the observed frequency for score pattern $\boldsymbol{x}$ and $\varepsilon_{\boldsymbol{x}}$ is the expected frequency for score pattern $\boldsymbol{x}$ given a model.

Simulation study I

In this simulation study, we consider dichotomous IRT with a unidimensional latent trait. For the MG LR test, we investigate its asymptotic properties under the null hypothesis

and at which rate the test is able to reject a misspecified model under different conditions when groups are made through random assignment. We then compare this rate against the rejection rates of a χ^2-difference and Pearson's χ^2 test. For the TLI, CFI and ICFI, we investigate their mean and variance under correct and incorrect model specification. The simulation study was conducted in R (R Core Team, 2022).

Data generation

Data are generated by first sampling θ from a standard Gaussian distribution. Then, either the 2PL or 3PL is chosen as basis for the data generation (see below). Item parameters are kept static over all conditions. For β_i, repetitions of the values $[-1.0, -0.5, 0.0, 0.5, 1.0]$ are chosen. As for α_i, we choose from the values $[0.7, 0.85, 1.0, 1.15, 1.3]$. To create a more realistic scenario, these two parameters are then matched with one another in order to make sure that for every item difficulty, there are low and high discriminating items. When the data generating model is the 3PL, we set γ_i to 0.25 for each item. Then, probabilities are estimated for all items on a test, given θ, the chosen model and its item parameters. By sampling from a binomial distribution with the estimated probabilities, a matrix of simulated responses to a dichotomous test is created.

Simulation design

In total, we vary four factors. The four factors and their conditions we examine are: 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 design results in a total of 30 conditions. Furthermore, within each of these conditions, we vary the MG LR test based on the number of groups $(G = 2, 3, 4)$. We replicate each simulation condition 300 times. In each replication of each condition, we fit the 2PL. The 2PL is fitted using functions from the *ltm* package (Rizopoulos, 2006), which approximates marginal maximum likelihood estimation through the Gauss-Hermite quadrature rule. Then we assess the performance of the TLI, CFI and ICFI and the three different types of goodness-of fit tests: (1) a χ^2-difference test, obtained by testing the 2PL under the 3PL with the constraint that all γ_i are equal through the *mirt* package (Chalmers, 2012), (2) Pearson's χ^2 test, estimated through aggregating score patterns from the data and comparing the observed score pattern frequencies to the expected score pattern frequencies given the fitted 2PL, and (3) the MG LR test with a varying number of randomly assigned groups.

Performance metrics

We study performance for the different goodness-of-fit tests by estimating both empirical α and power. Empirical α and power 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. Empirical α is estimated when generating data under the 2PL and fitting the 2PL. Power is estimated when generating data under the 3PL and fitting the 2PL. With these, we compare the different goodness-of-fit tests with one another, where a test with lower power or higher empirical α can be seen as performing worse. For all tests, we chose a level of significance of $\alpha = 0.05$.

To measure the performance of the TLI, CFI and ICFI, we estimate their mean and standard error (SE) in each condition. Then, we can inspect whether the means of the fit indices decrease in the conditions where data are generated under the 3PL, compared to when data are generated under the 2PL.

Results

To assess the performance of the MG LR test with randomised groups and the TLI, CFI and ICFI, we conducted the above described simulation study. Non-convergence of the models occurred $< .001$ percent of the time. When test length was 20, the empirical α and power estimates for Pearson's χ^2 test were not estimated. This is because of the large amount of possible score patterns in these conditions (2^{20}), which required too much computational power for the scope of the current study. Furthermore, negative values sometimes occurred for the TLI at higher test lengths under model misspecification. Previous literature has shown that this can occur when the degrees of freedom of the hypothesised model are small and correlations among observed variables are low (Wang & Wang, 2019; Widaman & Thompson, 2003). After estimation, if the mean was negative, we set it to 0.

Table 1 presents the empirical α estimates for the different χ^2-based goodness-of-fit tests. Here, the asymptotic property of the tests can be observed clearly. As sample size increases, empirical α approximations approach the nominal level of 0.05. At sample sizes larger than 1000, the rejection rates of the MG LR and χ^2-difference test do not deviate far from the nominal level for different levels of test lengths. However, this means that for low sample sizes, the LR statistic of the MG LR test does not yet completely follow a χ^2 distribution, resulting in higher rejection rates. For Pearson's χ^2 test, the table shows that as the number of items increases, the empirical α increases greatly from

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

Conditions		Goodness-of-fit test				
I	N	LR2	LR3	LR4	$\Delta\chi^2$	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 = MG LR test with $g = 2$; LR3 = MG LR test with $g = 3$; LR4 = MG LR test with $g = 4$; $\Delta\chi^2$ = χ^2-difference test under the three-parameter logistic model with equal γ_i constraint; P-χ^2 = Pearson's χ^2 test.

0.06 to 0.22. This is in accordance with the issues we discussed in the introduction, where Pearson's χ^2 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 when $I = 20$ & $N = 1500$. The results for Pearson's χ^2 test reflect that the test seems to not be able to detect any model misspecification with power estimates only slightly larger than the empirical α estimates. For the MG LR test, the table shows a peculiar result. Namely, the power estimates of the test still converge to the nominal level of 0.05. This means that when not under the null hypothesis, the LR statistic of the MG LR test still asymptotically follows a χ^2 distribution when groups are based on random assignment. This result indicates that the MG LR test cannot be used as a goodness-of-fit test to detect model misspecification under randomised groups. Moreover, the result also

implies that the MG LR test is insensitive to violations of correct model specification in testing for group differences. Therefore, we explore how effective the MG LR test is at detecting group differences and the effect of incorrect model specification on detecting group differences in a follow-up simulation study.

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

Conditions		Goodness-of-fit test				
I	N	LR2	LR3	LR4	$\Delta\chi^2$	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 = MG LR test with $g = 2$; LR3 = MG LR test with $g = 3$; LR4 = MG LR test with $g = 4$; $\Delta\chi^2$ = χ^2-difference test under the three-parameter logistic model with equal γ_i constraint; P-χ^2 = Pearson's χ^2 test.

For the fit indices, *Table 3* and *Table 4* exhibit the means and standard errors for the TLI, CFI and ICFI under correct and incorrect model specification respectively. The tables show that as sample size increases, the fit index values move closer to 1 (i.e., perfect 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. Nonetheless, the means for the TLI and CFI are higher under correct model specification compared to incorrect model specification for every single condition (e.g., for $N = 400$ & $I = 10$, the mean of the TLI is 0.11 under correct model specification, whereas it is 0.00 under incorrect model specification). This means that for any test length the TLI and CFI are able to detect model misspecification when working with the 2PL. However, the tables

also show that for cases where there is a small sample size and few items, there are high standard errors for the TLI and CFI. This indicates that when test length and sample size are low, point estimates for the TLI and CFI are not accurate enough to safely decide which model fits best. For this, methods such as bootstrapping should be employed.

Table 3. TLI, CFI and ICFI values under correct model specification

Conditions		TLI	CFI	ICFI
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; ICFI = identity coefficient fit index; M = mean; SE = standard error.

The ICFI shows less promise compared to the TLI and CFI in its ability to determine model fit in dichotomous IRT. The tables show that the means are either equal or slightly higher under incorrect model specification. The interpretation of this is that when data is generated under the 3PL, the ICFI shows either no difference or a small improvement in the fit of the 2PL compared to when data is generated under the 2PL, which is a counterintuitive result. Therefore, we argue that the ICFI is unable to determine whether there was model misspecification. This finding corresponds with the results from the Pearson's χ^2 test. Together, they strongly indicate that model fit measures based on observed and expected score-pattern frequencies are not able to detect model misspecification when working with the 2PL.

Table 4. TLI, CFI and ICFI values under incorrect model specification

Conditions		TLI	CFI	ICFI
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; ICFI = identity coefficient fit index; M = mean; SE = standard error.

Simulation Study II

Due to the result we found in Simulation Study I, where under model misspecification the MG LR test asymptotically follows a χ^2 distribution, we were inspired to run a follow-up simulation study. Here, we investigate how effective the MG LR test is at detecting different types of group differences under correct and incorrect model specification for different test lengths and sample sizes. Besides general effectiveness, we are interested in ascertaining whether there is a difference in effectiveness of the MG LR test to detect group differences when an incorrect model is fitted to the data compared to the correct model. We only consider IRT with unidimensionality and dichotomous test items. The simulation study was conducted in R.

Similar to Simulation Study I, we vary the four factors test length ($I = 5, 10, 20$), sample size ($N = 200, 300, 500, 1000, 1500$), model type (2PL, 3PL) and number of groups ($G = 2, 3, 4$). We also investigate a fifth factor: type of group differences, which has three conditions: (1) differences in θ, (2) differences in the item parameters α_i and β_i and (3) differences in the item parameters α_i, β_i and γ_i (only applicable to the

3PL condition). This results in a total of 225 distinct rejection rates. Each condition is replicated 300 times.

Data generation for Simulation Study II works in the same way as in Simulation Study I with one key difference. Now, data generation is no longer according to the same parameters described in Simulation Study I for the whole dataset. Instead, the parameters are dependent on two factors: the number of groups and the type of group differences. In the differences in θ conditions, the item parameters are the same for all G groups. However, the mean of the Gaussian distribution for θ depends on group membership. For each consecutive every group past the first group, there is a mean difference of -0.25 with the previous group. For example, when $G = 3$, group one has a mean of 0, group two a mean of -0.25, and group three a mean of -0.50. In the differences in item parameter conditions, the person parameters are all sampled from the same standard Gaussian distribution. However, the item parameters now depend on group membership. For α_i, all initial values (identical to Simulation Study I) are lowered by 0.10 for each consecutive group. Whereas for β_i, the values are lowered by 0.25 per consecutive group. To illustrate, when $G = 2$, group one has $\alpha_1 = 0.70$ and $\beta_1 = -1$, and group two has $\alpha_1 = 0.60$ and $\beta_1 = -1.25$. For the conditions where γ_i differs, we lower the γ_i values by .05 per consecutive group. We also no longer estimate the χ^2-difference and Pearson's χ^2 test, as these are used to detect model misspecification, whereas we are now interested in detecting group differences. Performance is measured by the detection rate, which is considered a power estimates against group differences.

Results

To research whether the MG LR test has power to discern whether there are group differences and the effect incorrect model specification has on the power of the test, we ran the above described simulation study where data is generated according to different types of group differences. Non-convergence of the models occurred $< .001$ percent of the time.

For an average difference of 0.25 in θ between each consecutive group, *Table 5* presents the power of the MG LR test. The table shows that the power to reject the null hypothesis approaches 1.00 as sample size increases. The table also shows a main effect of the number of groups, where the power to reject the null hypothesis increases as the number of groups increases. The power estimates are slightly lower in conditions when the data is generated under the 3PL. This means that incorrect model specification

Table 5. Power estimates for the MG LR test under group differences in θ

Conditions		Data Generating Model: 2PL, G			Data Generating Model: 3PL, G		
I	N	2	3	4	2	3	4
5	200	0.13	0.28	0.37	0.15	0.24	0.31
	300	0.15	0.32	0.46	0.14	0.23	0.41
	500	0.20	0.42	0.71	0.17	0.36	0.59
	1000	0.40	0.79	0.98	0.33	0.64	0.90
	1500	0.65	0.96	1.00	0.47	0.85	0.98
10	200	0.13	0.26	0.40	0.18	0.23	0.44
	300	0.17	0.24	0.44	0.15	0.30	0.50
	500	0.18	0.43	0.72	0.19	0.34	0.63
	1000	0.42	0.80	0.98	0.38	0.66	0.92
	1500	0.62	0.97	1.00	0.49	0.89	1.00
20	200	0.13	0.23	0.35	0.10	0.25	0.39
	300	0.14	0.23	0.41	0.13	0.26	0.37
	500	0.17	0.40	0.65	0.15	0.36	0.57
	1000	0.38	0.76	0.97	0.26	0.68	0.89
	1500	0.57	0.94	1.00	0.48	0.88	1.00

Note. Fitted model = two-parameter logistic model; 2PL = two-parameter logistic model; 3PL = three-parameter logistic model; I = test length; N = sample size; G = number of groups under which the data was generated.

leads to a small loss in power. Finally, the table shows no main effect of test length on the power of the MG LR test to detect group differences.

Table 6 presents the power of the MG LR test when differences are based on the item parameters. Similar results can be found here, where power increases as either sample size or the number of groups increases. There is also no main effect of test length on power. Incorrect model specification again leads to a slight decrease in power estimates compared to correct model specification. However, looking at the conditions where α_i, β_i and γ_i change per group, we find that the power estimates are much higher compared to the corresponding conditions where solely α_i and β_i change per group. This entails that the power of the MG LR test is dependent on the amount of group differences in item parameters, even if these group differences cannot properly be reflected in the fitted model. The overarching results are that the MG LR test seems to be very perceptive to group differences in either the distribution of the latent trait or item parameters, especially with sample sizes larger than 500 and more than two groups, and

Table 6. Power estimates for the MG LR test under group differences in item parameters

		Data Generating Model								
		2PL			3PL			3PL (γ_i)		
Conditions		G			G			G		
I	N	2	3	4	2	3	4	2	3	4
5	200	0.12	0.31	0.36	0.15	0.24	0.30	0.19	0.46	0.57
	300	0.20	0.34	0.43	0.14	0.23	0.39	0.28	0.51	0.80
	500	0.22	0.43	0.73	0.18	0.38	0.56	0.41	0.80	0.96
	1000	0.46	0.82	0.98	0.33	0.68	0.86	0.76	1.00	1.00
	1500	0.72	0.98	1.00	0.51	0.85	0.98	0.93	1.00	1.00
10	200	0.14	0.30	0.48	0.19	0.31	0.50	0.31	0.49	0.75
	300	0.21	0.30	0.51	0.18	0.35	0.62	0.36	0.64	0.88
	500	0.22	0.45	0.80	0.21	0.39	0.65	0.49	0.85	0.99
	1000	0.53	0.88	0.98	0.46	0.76	0.93	0.87	1.00	1.00
	1500	0.76	0.98	1.00	0.63	0.90	1.00	0.99	1.00	1.00
20	200	0.13	0.23	0.35	0.10	0.25	0.39	0.21	0.54	0.75
	300	0.14	0.23	0.41	0.13	0.26	0.37	0.32	0.63	0.89
	500	0.17	0.40	0.65	0.15	0.36	0.57	0.46	0.87	1.00
	1000	0.38	0.76	0.97	0.26	0.68	0.89	0.87	1.00	1.00
	1500	0.57	0.94	1.00	0.48	0.88	1.00	0.97	1.00	1.00

Note. Fitted model = two-parameter logistic model; 2PL = two-parameter logistic model.; 3PL = three-parameter logistic model where α_i and β_i differ per group; 3PL (γ_i): three-parameter logistic model where α_i, β_i and γ_i differ per group; I = test length; N = sample size; G = number of groups under which the data was generated.

that incorrect model specification leads to only a slight loss in power.

Empirical example

To demonstrate how a χ^2-difference test, the TLI and CFI and theory can be used in tandem to perform model appraisal in dichotomous IRT, we estimated and interpreted their values on Section 6 of the Law School Admission Test (LSAT) dataset given by Bock & Lieberman (1970). The dataset contains 1000 observations on 5 homogenous Figure Classification items on the unidimensional subject of Debate. The items were multiple choice, where a 1 denoted a correct answer and a 0 denoted a wrong answer. We fitted four models in total: the 2PL, the 3PL with no constraints, the 3PL with the constraint that all γ_i are equal and the 3PL where we set all γ_i to 0.25. Afterwards, we evaluated the appropriateness of each model by calculating the TLI and CFI and by

performing two χ^2-difference tests with the 2PL as the null model and the 3PL with no constraints and the 3PL with the constraint that all γ_i are equal as the alternative models.

For the 2PL, the TLI equaled 0.57 and the CFI equaled 0.81. For the 3PL with no constraints, the TLI lowered to 0.09 and the CFI lowered 0.70. A χ^2-difference test between these two models showed a non-significant result $(p = 1.00)$. This means that there is no evidence that the 3PL fits significantly better than the 2PL. For the 3PL with equal γ_i, the TLI was 0.53 and the CFI 0.79. The χ^2-difference test between these two models also showed a non-significant result $(p = 1.00)$. Therefore, we still have no evidence against the 2PL. Finally, for the 3PL where we set all γ_i to 0.25, the TLI was 0.56 and the CFI was 0.80.

From these results, we conclude the following: allowing each item to have a free γ_i does not lead to according to the χ^2-difference test. The TLI and CFI show that it leads to a worse fit. However, setting the constraint that the pseudo-guessing parameters have to be equal or setting them to a set value of 0.25 does not lead to a much worse fit compared to the 2PL according to the TLI and CFI. Yet the χ^2-difference test shows that losing one degree of freedom and constraining all γ_i to be equal does not lead to a significant improvement of model fit compared to the 2PL. However, we have to also take into account the fact that the items were multiple choice and in theory do have a baseline probability of being scored correctly. Therefore, we conclude that the most reasonable models are the 3PL with equal γ_i or the 3PL with γ_i set to 0.25.

Discussion

The current study investigated existing and new measures of model fit to test for model misspecification and group differences in dichotomous IRT. Through the first simulation study, we observed that the proposed MG LR test requires sample sizes larger than 1000 in order to properly follow a χ^2 distribution. The MG LR test was not able to detect any model misspecification under random group assignment. Rather, under incorrect model specification the MG LR test still asymptotically followed a χ^2 distribution. Similarly, we observed that fit measures that asses goodness-of-fit through estimating the difference between expected and observed score pattern frequencies (i.e., the ICFI and Pearson's χ^2 test) had no power to detect model misspecification when fitting the 2PL.

Contrary to these results and the results given by Yang (2020), we found that the TLI and CFI were very sensitive to detecting model misspecification in dichotomous IRT. This finding coincides with the results given by Cai et al. (2021), and we believe that it

might be due to our shared choice in baseline model. Previous research by Widaman & Thompson (2003) and Van Laar & Braeken (2021) has already discussed the importance of baseline models for incremental fit indices. However, against the expectations of Cai et al. (2021), who speculated that the TLI would not depend heavily on model size (i.e., test length), we discovered that the TLI and CFI estimates lower substantially as test length increases. Whether this is the effect of using a different statistic for the calculation of the TLI or the shared choice of baseline remains unknown. Future research could look into the relationship between baseline model, test length and performance of the TLI and CFI to a greater extent to uncover this. Fortunately, the effect of test length does not affect the fit indices' sensitivity to model misfit. It only poses issues for any rule-of-thumbs associated with the usage of the TLI and CFI. Cai et al. further speculated that the CFI would not have a similar interpretation in IRT as it does in SEM literature, yet our results indicate that the behaviour of the CFI follows the same trend as the TLI in IRT. All things considered, the ability of the TLI and CFI to detect model misspecification in dichotomous IRT remains a relevant finding, especially when considering the results given by Nye et al. (2020) who found few fit indices that were sensitive to these types of misspecification. We strongly recommend that future studies in IRT incorporate the TLI and CFI in evaluating model appropriateness.

Further exploring the value of the MG LR test, a follow-up simulation study showed that the MG LR test is effective at detecting group differences. The test is able to detect small group differences in the population mean - according to Cohen's (1988) standards for effect size in behavioural sciences - above the 0.80 power threshold once sample size rose above 1000 with more than two groups. Taken together, it means that the MG LR test could be a useful addition to the multi-group analysis literature. This is because the MG LR test tests only for group differences and mostly ignores model specification, where incorrect model specification leads only to a small loss in power. The largest limitation of the MG LR test is that it is unable to determine where these group differences lie. For this, a researcher has to investigate the latent construct and the items of the questionnaire through Measurement Invariance techniques.

Naturally, we have to bear in mind the context of the present study. We studied these different measures of model fit and group differences solely in IRT with dichotomous items and unidimensional latent traits. For the MG LR test, this signifies that we do not know whether the test is also sensitive for group differences beyond this setting. Future work could further investigate the asymptotic properties and power of the test for group differences in polytomous and multidimensional IRT and continuous settings such as in

SEM. Furthermore, as briefly mentioned before, we worked with simulated data where we made sure that no assumptions were violated. In real-life settings, this may almost never happen and this should be taken into account with usage of the measures reported here. To illustrate this, we showcased an example of how a χ^2-difference test, TLI and CFI can be used in tandem for model evaluation in dichotomous IRT.

To summarise, preliminary evidence brought about our belief that the TLI and CFI show promise as valuable aids for model evaluation in dichotomous IRT. On the contrary, the ICFI, Pearson's χ^2 test and the MG LR test are not able to help in assessing model fit to be able to help in this research setting. However, when researching differences between groups in unidimensional dichotomous IRT, the MG LR test can be used to detect whether these differences exist while disregarding model specification. Model evaluation in IRT remains an issue that requires more light to be shed upon in future research. Hopefully, the conducted research is able to improve model evaluation ever so slightly in the social and behavioural sciences.

Supplemental material

All annotated code and results used in the current study can be found online on the Open Science Framework by clicking here.

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.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6):1–29.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic press.

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.

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*.

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.

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

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.

Van Laar, S. and Braeken, J. (2021). Understanding the comparative fit index: It's all about the base! *Practical Assessment, Research & Evaluation*, 26(1).

Widaman, K. F. and Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological methods*, 8(1):16.

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.