Journal Title
XX(X):1–??
©The Author(s) 2023
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
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]

Abstract

...

Keywords

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

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 infer 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

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 function 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.

Therefore, we explore a new possible goodness-of-fit test to test for model misspecification between the 2PL and 3PL in 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 were to be 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, this would mean that the test is robust against violations of model specification. If the interest is then in testing for differences between existing groups, it means that the test ignores model misspecification and tests solely for these group differences.

The second procedure of assessing model fit is through the use of fit indices. 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 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 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). One of the limitations for many fit indices was that they were unable to detect model misfit between the 2PL and 3PL. Fit indices from Structural Equation Modeling can also be used in IRT. Although research has been done to examine the performance of 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 additional insights in assessing model fit. However, there have been no studies that investigate the performance of the TLI and CFI to detect misfit between the 2PL and 3PL 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 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. Therefore, this coefficient can reflect the degree to which the observed score pattern frequencies are aligned with the expected score pattern frequencies under a given IRT model.

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 based on multi-group analysis, named the Multiple Group LR (MG LR) test, and assessing the performance of the TLI and CFI based on calculations not researched before in an IRT setting. Finally, we also research

the performance of a new fit index inspired by the identity coefficient by Zegers & ten Berge (1985), named the identity coefficient fit index (ICFI) in the context of IRT.

The present study

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

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?
3. Under random 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?
4. What is the performance of the TLI, CFI and ICFI in detecting model mismatch between the 2PL and 3PL in IRT?
5. When groups are based on true differences, what is the power of the MG LR test to detect these differences in IRT?

Results from the simulation studies are reported to answer the research questions. Furthermore, empirical data from ... are used to illustrate the value of the investigated model fit measures.

– IETS over discussie –

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, 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, the IRF for the 3PL (Birnbaum, 1968):

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

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 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 question 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:

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

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:

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

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:

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

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 $\hat{\boldsymbol{\nu}}$. 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 ratio 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, the observed value of Pearson's χ^2 test is based on the statistic below:

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

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 statistic also asymptotically follows a χ^2 distribution. The MG LR test that we develop and test in the present study is based on the likelihood ratio statistic:

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

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 existing groups, or they can be gained by randomly assigning observations to g groups. Wilks' theorem also applies to this likelihood statistic. Because all three aforementioned tests asymptotically follow a χ^2 distribution under the null hypothesis, they can be used to test for goodness-of-fit.

For fit indices, we research the performance of the TLI, CFI and ICFI. The TLI and CFI 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, 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. 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 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 result 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 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 the ICFI. Parallel to the Pearson's χ^2 test, the ICFI measures model fit by comparing the observed score pattern frequencies to the expected score pattern frequencies under a given 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 again the observed frequency for score pattern $\boldsymbol{x}$ and $\varepsilon_{\boldsymbol{x}}$ is the expected frequency for score pattern $\boldsymbol{x}$.

Simulation study I

In this simulation study, we consider IRT with unidimensional θ, dichotomous test items and the IRF for the 2- and 3PL. For the MG LR test, we investigate whether the test is robust 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 test and Pearson's χ^2 test. For the TLI, CFI and ICFI, we investigate their average values and variance under correct and incorrect model specification. The simulation study was conducted in R (R Core Team, 2022).

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 [-1.0, -0.5, 0.0, 0.5, 1.0]. As for the discrimination parameter 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 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 0.25 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 design results in a total of 30 distinct conditions. Furthermore, within each of these conditions, we then 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 model. 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 obtain the results of the three fit indices 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 have to be 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 under the given model, and (3) the MG LR test with a varying number of randomly assigned groups.

Performance metrics

We study performance of the different goodness-of-fit tests by estimating both empirical α (i.e., 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. 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 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, CFI and ICFI, we estimate the mean and standard error (SE) of each fit index under every condition. Then, we can inspect whether the average fit index 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 MG LR test with randomised groups and the three fit indices, 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 (2^{20}) in these conditions, 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. Here, the asymptotic property of the tests 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 MG LR 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 under the null hypothesis. However, this also means that for low sample

Table 1. Empirical α estimates for the 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 = MG LR test with $g = 2$; LR3 = MG LR test with $g = 3$; LR4 = MG LR test with $g = 4$; χ^2 = χ^2-difference test under the three-parameter logistic model with equal γ constraint; P-χ^2 = Pearson's χ^2 test.

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 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 model misspecification between the 2PL and 3PL with power estimates only slightly larger than the empirical α estimates. For the MG LR test, the table shows a very interesting result. Namely, the power estimates of the test still converge to the nominal α level of .05. This means that when not under the null

hypothesis, the likelihood ratio statistic of our test (eq. 7) still asymptotically follows a χ^2 distribution when groups are based on random assignment. These results entail that the test cannot be used to detect model misspecification when working with randomised groups. However, this means the test is robust against violations of model specification. Therefore, we explore how good the test is at detecting group differences in Simulation Study II.

Table 2. Power estimates for the 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 = MG LR test with $g = 2$; LR3 = MG LR test with $g = 3$; LR4 = MG LR test with $g = 4$; χ^2 = χ^2-difference test under the three-parameter logistic model with equal γ constraint; P-χ^2 = Pearson's χ^2 test.

For the fit indices, *Table 3* and *Table 4* exhibit the mean and standard error for the TLI, CFI and ICFI 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. This phenomenon might be due to our choice of baseline model and further strengthens the argument made by (Van Laar & Braeken, 2021) who emphasised that the choice of baseline model is important when

working with fit indices. However, even though fit indices value tend to 0 as test length increases, the values for the TLI and CFI are on average 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 .00 under incorrect model specification). This means that these fit indices can be used for model fit decisions between the 2- and 3PL in IRT. However, test length would have to be taken into account if a 'rule-of-thumb' is desired.

Table 3. TLI, CFI and ICFI 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.

The tables also show that for cases where there is a small sample size and few test items, there are high standard errors for the TLI and CFI. This indicates that when assessing model fit through use of the TLI and CFI in low sample size settings, point estimates might not be enough. Other methods such as bootstrapping might be required in order to be able to properly 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 at higher test lengths under model misspecification. Previous literature has shown that this can occur when the degrees of the hypothesized model are small and

Table 4. TLI, CFI and ICFI 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.

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 ICFI shows less promise compared to the TLI and CFI. The values seem to not differ in incorrect model specification compared to correct model specification. In fact, the average values are usually higher under incorrect model specification. Therefore, the fit index was not able to decide which of the two models fit the data better in the current study. This finding is in accordance with the results given by 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 fit between the 2PL and 3PL in IRT.

Simulation Study II

In this simulation study, we again only consider IRT with unidimensional θ, dichotomous test items and the IRF for the 2- and 3PL. For the MG LR test, we investigate at which rate the test is able to reject a model under different conditions when groups are based on true differences. The simulation study was conducted in R.

Similar to Simulation Study I, we again vary the four factors test length ($I = 5, 10, 20$), sample size ($N = 200, 300, 500, 1000, 1500$), the model type (2PL, 3PL) and the number of groups ($G = 2, 3, 4$). We also investigate a fifth factor: 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 group differences factor, and the number of groups factor G. In the differences in θ conditions, the item parameters are the same for all G groups. However, the mean of the normal distribution depends on group membership. Between every 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 opposite happens. Here, the person parameters are all sampled from the same standard normal distribution. However, the item parameters now depend on group membership. For the discrimination parameter per item, α_i, all initial values (identical to Simulation Study I) are lowered by 0.10 for each consecutive group. Whereas for the difficulty parameter per item, β_i, the values are lowered by 0.25 per consecutive group. So when $G = 2$, group one has $\alpha_1 = 0.7$ and $\beta_1 = -1$, so group two has $\alpha_1 = 0.65$ and $\beta_1 = -1.25$. For the conditions where γ_i differs, we chose to lower the γ_i values by .05 per consecutive group. Finally, we no longer estimate both the χ^2-difference and Pearson's χ^2 test, as these are used to detect model misspecification, whereas we are now in the field of detecting group differences. Performance is again measured by the detection rate. However, now they are all considered power estimates against group differences.

Results

To research whether the MG LR test has power to discern whether there are group differences, we ran the above described simulation study. Non-convergence of the models occurred $< .001$ percent of the time.

For an average differences of 0.25 in θ between each group, *Table 5* presents the power of the MG LR test under the different conditions. The table shows exactly what one would expect of a power study, where 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. There do not seem to be large differences in the power of the MG LR test depending on the data generating model. The power estimates seem to be slightly lower when the data was generated under the 3PL compared to the 2Pl. This would entail that as model complexity increases, the MG LR test loses some of its power. When considering test length, it seems as if there is no main effect of test length on the power of the MG LR test.

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

Conditions		2PL			3PL		
		G			*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; I = test length; N = sample size; G = number of groups under which the data was generated.

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 again power increases as either sample size or the number of groups increases. Furthermore, there does not seem to be a main effect of test length on power. Finally, when looking at the conditions where also the γ_i parameters change per group, we find that the power values are much higher compared to their corresponding conditions when only α_i and β_i parameters change per group. This entails that the power of the MG LR test is dependent on the amount of group differences also. The overarching result is that the MG LR test seems to be very perceptive to group differences in either latent distribution or item parameters, especially with sample sizes larger than 500 and more than two groups.

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

Conditions		2PL			3PL α_i & β_i			3PL α_i & β_i & γ_i		
		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; Data generating model = two-parameter logistic model; *I* = test length; *N* = sample size; *G* = number of groups under which the data was generated.

Empirical example

To examine the possible uses of our MG LR 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 existing and new possible measures of model fit in the context of IRT to test for model misspecification and group differences. Through the first simulation study, we observed out that the proposed goodness-of-fit test - the MG LR test - requires sample sizes larger than 1000 in order to properly follow a χ^2 distribution. Furthermore, the MG LR test was not able to detect any model misspecification between the 2PL and 3PL when groups are based on random assignment. Instead, under model misspecification the MG LR test still asymptotically followed a χ^2 distribution. Conversely, we found that the TLI and CFI were very sensitive to detect misfit between the 2PL and 3PL in one-dimensional IRT. This is an important result when taking into account the results given by Nye et al. (2020), who found very few fit indices that were sensitive to this.

However, against the expectations of Cai (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 became lower as test length increases. Although this could be partly explained by the differences in calculations and choice of null model, it remains an important finding. It would entail that model size should be taken into account when using these fit indices to choose between multiple models. It also poses issues for a one-size-fits-all cut off rule for these indices. Finally, the simulation showed that model fit measures that asses model fit through estimating the difference between expected and observed scorepatterns (i.e., the ICFI and Pearson's χ^2 test) were not able to detect misspecification between the 2PL and 3PL.

Further exploring possible uses of the MG LR test, a follow up simulation study showed that the MG LR test is very effective at detecting group differences. The test is able to detect small group differences in the population mean - according to Cohen's standards for effect size (1988) - 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 is

a useful addition to the multi-group analysis literature, where the test can test for group differences whilst ignoring model specification. – hier moet wat citaties bij –

Although we have come to positive findings the present study has its limitations also.

Although we found positive results for possible uses of the TLI and CFI, it is important to remember that the present study researched these measures in a specific setting. Namely, we considered only unidimensional dichotomous IRT. Although the TLI has also been researched before in a multidimensional setting

Finally, as 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 is important to keep in mind when using the measures researched here.

For the MG LR test, a large limitation is the fact that although the test is sensitive to group differences, the test is not able to tell the user where these differences lie. Furthermore, this test was only researched in XXX even though the test is applicable to more fields. Future research could investigate whether the test still ignores model misspecification in Structural Equation Modelling.

To summarize, preliminary evidence brought about our belief that the TLI and CFI show promise as valuable aids for model evaluation in IRT research between the 2PL and 3PL. On the contrary, Pearson's χ^2 test and the MG LR test do not seem to be able to help between deciding between these two models. However, for this purpose a χ^2-difference test remains a faithful measure of goodness-of-fit. When researching differences between groups in IRT, the MG LR test can be used to detect whether these differences exist while disregarding model specification. However, the test is not able to show where these differences then lie. For this, researchers should use Differential Item Functioning methods such as the Mantel-Haenszel approach. Model evaluation in IRT remains an issue that requires more light to be shed upon in future research and hopefully, the conducted research here is able to improve model evaluation ever so slightly in the social and behavioural sciences.

Limitations

- didn't test 3PL - didn't test multidimensional IRT - didn't test polytome IRT - model differences: does not show where these differences are

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

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. Methods that assess model fit through comparing observed and expected score patterns do not seem to be viable for assessing fit between 2PL and 3PL in IRT.

Finally, the MG LR test has a lot of potential for assessing group differences. This is because the test is robust against model misspecifications, and instead only has power against group differences. However, the test is unable to assess where the group differences lie.

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