

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

Journal Title

XX(X):2–19

©The Author(s) 2023

Reprints and permission:

[sagepub.co.uk/journalsPermissions.nav](http://sagepub.co.uk/journalsPermissions.nav)

DOI: 10.1177/ToBeAssigned

[www.sagepub.com/](http://www.sagepub.com/)

SAGE

Nina van Gerwen <sup>1</sup>

---

## Abstract

In Item Response Theory (IRT), it is important to investigate goodness-of-fit of the models you fit to the data in order to have valid measurements and conclusions. Goodness-of-fit is usually measured through goodness-of-fit tests and fit indices. However, for two commonly used IRT models, commonly used goodness-of-fit tests suffer from various issues (citation). Previous research has also shown that many fit indices in IRT are also unable to determine model misspecification when working with these models (citation). Therefore, we further the research in this area by researching a new goodness-of-fit test based on multiple group analysis, the MG LR test, and three fit indices that have not been in this setting before: (1) the Tucker Lewis index (citation), (2) Comparative Fit index (citation) and (3) the Identity Coefficient Fit index (citation). We examine the performance of these different measures of model fit through a simulation study. Results from the first simulation study suggests that the ICFI, Pearson's  $\chi^2$  and the MG LR test are unable to determine model misspecification when the 2PL is fit to the data. In contrast, the TLI and CFI showed good performance in determining model misspecification in this context. A follow-up simulation study showed that the MG LR test is very effective in detecting group differences, while insensitive to violations of model misspecification. We believe this to be a unique feature of the test and speculate its value as a tool in detecting group differences, whilst noting its limitations. We recommend that the TLI and CFI become more commonly used as in Structural Equation Modeling in order to improve model appraisal in IRT. Recommendations for future research in measuring model fit are made.

## Keywords

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

---

<sup>1</sup>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

## 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 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 follow the expected data under a given model.

Commonly used goodness-of-fit tests in IRT are a  $\chi^2$ -difference test and Pearson's  $\chi^2$  test. However, these tests both suffer from issues. Pearson's test, for example, will not follow a  $\chi^2$  distribution when many score patterns are missing or have a low frequency, which is often the case – especially as test length increases. The  $\chi^2$ -difference test 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 item correctly as a logistic function of item characteristics and a latent variable. The reason the  $\chi^2$ -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, 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 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 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 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 limitations of many fit indices was that they exhibited poor performance when the 2PL was fitted to the data. 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 in estimating model fit when the 2PL is fitted to data and when calculated through the  $\chi^2$  statistic and with a baseline model that assumes independence between test items. These factors influence the performance of the fit indices and should therefore be investigated further.

Finally, 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 vectors are identical. Therefore, this coefficient can reflect the degree to which the observed score pattern frequencies are aligned with the expected score pattern frequencies given a certain model.

To summarise, in the field of IRT there are not enough measures of model fit available

to use when the 2PL is fitted to the data 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. Specifically, we answer and discuss the following four 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  $\chi^2$  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 group assignment, how does the performance of the MG LR test compare to the performance of a  $\chi^2$ -difference and Pearson's  $\chi^2$  test in detecting model misspecification?
4. What is the performance of the TLI, CFI and ICFI in assessing model fit when the 2PL is fitted?

## 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  $\theta$ , (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  $\theta$ . 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  $\theta$ , (b) the location parameter of the item  $\beta_i$ , which denotes how difficult the item is, (c) the scaling parameter of the item  $\alpha_i$ , which shows how well item  $i$  discriminates between individuals with lower and higher  $\theta$ , and (d) an item-specific lower asymptote  $\gamma_i$ , 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  $\gamma_i$  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:

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

where  $\mathbf{X}$  is now a random vector indicating a score pattern and  $\mathbf{x}$  is the realisation of  $\mathbf{X}$ . 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:

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

where  $\phi(\theta)$  is the univariate density of the latent variable  $\theta$ . In order to solve this equation, the density of  $\theta$  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:

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

where  $n_{\mathbf{x}}$  is the frequency of score pattern  $\mathbf{x}$ . 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  $\chi^2$ -difference test and the Pearson's  $\chi^2$  test. The  $\chi^2$ -difference test uses the following LR statistic:

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

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 Wilks' theorem (Wilks, 1938), under the null hypothesis this likelihood statistic will asymptotically follow a  $\chi^2$  distribution as sample size increases. Alternatively, the observed value of Pearson's  $\chi^2$  statistic is:

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

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

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

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 likelihood statistic. Because all three aforementioned tests asymptotically follow a  $\chi^2$  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:

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

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  $\beta_i$  is:

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

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:

$$P(\mathbf{X} = \mathbf{x}) = \pi_{\mathbf{x}}, \quad (9)$$

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

$$\hat{\pi}_{\mathbf{x}} = \frac{n_{\mathbf{x}}}{n},$$

where  $n_{\mathbf{x}}$  is the number of observations with score pattern  $\mathbf{x}$  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:

$$\text{CFI} = 1 - \frac{\max\{(\chi_T^2 - df_T), 0\}}{\max\{(\chi_T^2 - df_T), (\chi_0^2 - df_0), 0\}}, \quad (10)$$

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

In both equations,  $\chi_T^2$  is the result of a  $\chi^2$ -difference test between the tested model and the saturated model with  $df_T$  degrees of freedom, and  $\chi_0^2$  is the outcome of a  $\chi^2$ -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  $\chi^2$  test, the ICFI measures model fit by comparing the observed score pattern frequencies to the expected score pattern frequencies under a given model:

$$\text{ICFI} = \frac{2 \cdot \sum_{\mathbf{x}} n_{\mathbf{x}} \cdot \varepsilon_{\mathbf{x}}}{\sum_{\mathbf{x}} n_{\mathbf{x}}^2 + \sum_{\mathbf{x}} \varepsilon_{\mathbf{x}}^2}, \quad (12)$$

where  $n_{\mathbf{x}}$  is again the observed frequency for score pattern  $\mathbf{x}$  and  $\varepsilon_{\mathbf{x}}$  is the expected frequency for score pattern  $\mathbf{x}$ .



## Simulation study I

In this simulation study, we consider IRT with unidimensionality, 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  $\chi^2$ -difference test and Pearson's  $\chi^2$  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  $\theta$  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  $\beta_i$ , we choose the values [-1.0, -0.5, 0.0, 0.5, 1.0]. As for the discrimination parameter  $\alpha_i$ , 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  $\gamma_i$ . Then, probabilities are estimated for all items on a test, given  $\theta$ , 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. 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  $\chi^2$ -difference test, obtained by testing the

2PL under the 3PL with the constraint that all  $\gamma_i$  have to be equal through the *mirt* package (Chalmers, 2012), (2) Pearson's  $\chi^2$  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  $\alpha$  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  $\alpha$  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 values, we compare the different type of tests with one another under every condition, where a test with lower power or higher empirical  $\alpha$  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 assess 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  $\alpha$  and power estimates for Pearson's  $\chi^2$  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 at the  $\alpha = .05$  level for the different  $\chi^2$ -based goodness-of-fit tests. Here, the asymptotic property of the tests can be observed clearly. As sample size increases, empirical  $\alpha$  values approach the nominal level of .05. At sample sizes larger than 1000, the rejection rates of the MG LR and  $\chi^2$ -difference test seem to not deviate far from the nominal level for different levels of test lengths. Therefore, we can state that these two test statistics indeed asymptotically follow their supposed  $\chi^2$  distribution under the null hypothesis. However, this also means that for

**Table 1.** Empirical  $\alpha$  estimates for the different goodness-of-fit tests

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

low sample sizes, the test statistic of the MG LR test does not yet completely follow a  $\chi^2$  distribution, resulting in higher rejection rates. For Pearson's  $\chi^2$  test, the table also shows that as the number of items increases, the empirical  $\alpha$  increases dramatically from 0.06 to 0.22. This is in accordance with the issues we discussed in the introduction, where Pearson's  $\chi^2$  test does not follow a  $\chi^2$  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  $\chi^2$ -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  $\chi^2$  test reflect that the test seems to not be able to detect any model misspecification with power estimates only slightly larger than the empirical  $\alpha$  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  $\alpha$  level of .05. This means that when not under the null hypothesis, the LR

statistic of our test (eq. 7) still asymptotically follows a  $\chi^2$  distribution when groups are based on random assignment. These results entail that the test cannot be used as a goodness-of-fit test to detect model misspecification under randomised groups. This implies that the test is insensitive to violations of model specification in testing for group differences. Therefore, we explore how effective the test is at detecting group differences in a follow-up simulation study under correct and incorrect model specification.

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

Conditions		Goodness-of-fit test				
<i>I</i>	<i>N</i>	LR2	LR3	LR4	$\chi^2$	P- $\chi^2$
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  $g = 4$ ;  $\chi^2 = \chi^2$ -difference test under the three-parameter logistic model with equal  $\gamma$  constraint; P- $\chi^2$  = Pearson’s  $\chi^2$  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. Both tables show that as sample size increases, the fit index values 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. Nonetheless, even though the various values 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 when the 2PL is fit to the data. 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	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; 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 good 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 freedom of the hypothesised model are small and correlations among observed variables are low (Wang & Wang, 2019; Widaman & Thompson, 2003). In the current study, we rounded the average to 0 if it was negative.

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

Conditions		TLI	CFI	ICFI
<i>I</i>	<i>N</i>	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.

The ICFI shows less promise compared to the TLI and CFI in its ability to determine model fit. The values seem to not differ in incorrect model specification compared to correct model specification. In fact, the average values are usually slightly higher under incorrect model specification. Therefore, the fit index was unable to determine whether there was model misspecification in the current setting. This finding corresponds with the results given by the Pearson’s  $\chi^2$  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.

Simulation Study II

Because of the result we found in Simulation Study I, where under model misspecification the MG LR test still asymptotically follows a  $\chi^2$  distribution, we were inspired to run this 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 the correct model is fitted to the data compared to an incorrect model. In this simulation study, we again only consider IRT with unidimensionality, dichotomous test items and the IRF for the 2- and 3PL. 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  $\theta$ , (2) differences in the item parameters  $\alpha_i$  and  $\beta_i$  and (3) differences in the item parameters  $\alpha_i$ ,  $\beta_i$  and  $\gamma_i$  (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  $\theta$  conditions, the item parameters are the same for all  $G$  groups. However, the mean of the normal distribution for  $\theta$  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 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,  $\alpha_i$ , all initial values (identical to Simulation Study I) are lowered by 0.10 for each consecutive group. Whereas for the difficulty parameter per item,  $\beta_i$ , the values are lowered by 0.25 per consecutive group. So when  $G = 2$ , group one has  $\alpha_1 = 0.70$  and  $\beta_1 = -1$ , so group two has  $\alpha_1 = 0.60$  and  $\beta_1 = -1.25$ . For the conditions where  $\gamma_i$  differs, we chose to lower the  $\gamma_i$  values by .05 per consecutive group. Finally, we no longer estimate both the  $\chi^2$ -difference and Pearson's  $\chi^2$  test, as these are used to detect model misspecification, whereas we are now interested in using the MG LR test to detect 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 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 differences. Non-convergence of the models occurred < .001 percent of the time.

**Table 5.** Power estimates for the MG LR test under group differences in  $\theta$

Conditions		2PL			3PL		
		<i>G</i>			<i>G</i>		
<i>I</i>	<i>N</i>	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.

For an average differences of 0.25 in  $\theta$  between each consecutive 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. The power estimates are slightly lower in conditions when the data is generated under the 3PL. This means that incorrect model specification 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.** Power estimates for the MG LR test under group differences in item parameters

Conditions		2PL			3PL $\alpha_i$ & $\beta_i$			3PL $\alpha_i$ & $\beta_i$ & $\gamma_i$		
		$G$			$G$			$G$		
		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.

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. 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 the  $\gamma_i$  parameters change per group, we find that the power values are much higher compared to their corresponding conditions when only  $\alpha_i$  and  $\beta_i$  parameters 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 latent distribution or item parameters, especially with sample sizes larger than 500 and more

than two groups and that incorrect model specification leads to only a slight loss in power.

## Discussion

The current study investigated existing and new measures of model fit in the context of IRT to test for model misspecification and group differences. 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  $\chi^2$  distribution. The MG LR test, however, 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  $\chi^2$  distribution. Similarly, we observed that model fit measures that assess goodness-of-fit through estimating the difference between expected and observed score pattern frequencies (i.e., the ICFI and Pearson's  $\chi^2$  test) had no power to detect model misspecification when fitting the 2PL with a unidimensional latent variable.

Contrary to these results, we did find that the TLI and CFI were very sensitive to detecting model misspecification in this context. This is an important result considering the results given by Nye et al. (2020), who found very few fit indices in IRT that were sensitive to this particular misfit. 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. The phenomenon might be due to our shared choice of baseline model. Previous research by Widaman & Thompson (2003) and Van Laar & Braeken (2021) has already discussed the importance of baseline models for these fit indices. Future research could look into the relationship between baseline model, test length and TLI/CFI values to a greater extent. We must also remember that compared to Cai et al., we used a different statistic for the calculation of the fit indices which could have affected the results. 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 Structural Equation Modeling literature. However, our results indicate that the behaviour of the CFI follows the same trend as the TLI in IRT.

Further exploring possible uses 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 Psychology - 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 test is that the test is not able 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 also 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/or multidimensional IRT and continuous settings such as in Structural Equation Modelling. 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 summarise, preliminary evidence brought about our belief that the TLI and CFI show promise as valuable aids for model evaluation in IRT research that uses the 2PL. On the contrary, the ICFI, Pearson's  $\chi^2$  test and the MG LR test do not seem to be able to help in this research setting. 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 here 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 through the following [link](#).