Journal Title
XX(X):??–??
©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] and Dave Hessen[2]

Abstract

...

Keywords

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

Introduction

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

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

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

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

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

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

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

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

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

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

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

The present study

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

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

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

Methods

Before we share the methodology of the simulation study, let us first examine a brief summary on the statistical theory associated with our study. In IRT, the goal is to find

a model that best describes scores on test items. To achieve this, IRT presupposes three assumptions: (1) conditional independence of items given the latent trait, denoted by θ, (2) independence of observations and (3) the response to an item can be modeled by an item response function (IRF). An IRF is a mathematical equation that relates the probability to score a certain category on an item to θ. Below, you find the IRF for the 3PL (Birnbaum, 1968):

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

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

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

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

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

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

where $\phi(\theta)$ is the univariate density of the latent variable θ. In order to solve this equation, the density of θ has to be specified. For example, $\phi(\theta)$ can be specified as a standard normal distribution. With the joint marginal probability and the assumption

of independence of observations, we can construct a likelihood function and estimate $\boldsymbol{\nu}$ through marginal maximum likelihood estimation:

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

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

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

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

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

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

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

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

where L_0 is the likelihood of the chosen model for the whole dataset and L_j is the likelihood of the chosen model for group j, which is gained by randomly assigning the observations to g groups. Wilks' theorem also applies to this likelihood statistic. Because all three aforementioned tests asymptotically follow a χ^2 distribution, they can be used for Null Hypothesis Significance Testing.

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

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

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

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

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

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

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

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

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

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

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

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.

Simulation study I

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

Data generation

Data generation starts by first sampling the person parameters θ from a standard normal distribution. Then, either the 2PL or 3PL is chosen as basis for the data generation (see below). We choose to keep item parameters static over all simulations. For the difficulty parameter β_i, we choose the values [-2, -1, 0, 1, 2]. As for the discrimination parameter α_i, we choose repetitions of the values [0.7, 0.85, 1, 1.15, 1.3]. To create a more realistic scenario, these two parameters were then matched with one another in order to make sure that for every item difficulty, there are low and high discriminating items. Finally, when the data generating model is the 3PL, we choose the values [.09, 0.12, 0.15, 0.18, 0.21] per five items 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: test length, sample size, model types and number of groups. For an overview of the conditions we used for the factors, see *Table* **??**. The simulation design results in a total of 3 (test length) x 5 (sample size) x 2 (model type)

Table 1. Overview of Simulation Conditions for Each Factor

Factor	Conditions	Description
Test length	5 - 10 - 20	The total number of items that the test will consist of
Sample size	100 - 200 - 500 1000 - 1500	The total number of observations that are available for each item
Model type	2PL - 3PL	The models that we will use as the basis for data generation
Number of groups	2 - 3 - 4	The number of groups that the data gets divided into for the LR Randomisation test calculations

Note. 2PL = two-parameter logistic model; 3PL = three-parameter logistic model.

= 30 conditions. In each replication of each condition, we fit the 2PL model. Models are fitted using functions from the *ltm* package in R (Rizopoulos, 2006), which approximates marginal maximum likelihood estimation through the Gauss-Hermite quadrature rule. Then we obtain the results of the TLI, CFI and the three different types of goodness-of fit tests: (1) a χ^2-difference test, which is obtained by testing the 2PL under the 3PL with no constraints, (2) Pearson's χ^2-test, which is estimated through aggregating score patterns from the data and comparing the observed score pattern frequencies to the expected score pattern frequencies under the given model, and (3) the LR Randomisation test with a varying number of groups. We replicate each simulation condition 300 times.

Performance metrics

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

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

Results

Below you will find four tables that give an indication of how the results will be shared. Please note that these results are synthetic and **not** the final results.

Table 2. Empirical Alpha estimates for different goodness-of-fit tests

Conditions		Goodness-of-fit test				
I	*N*	LR2	LR3	LR4	χ^2	P-χ^2
5	100	0.08	0.07	0.06	0.06	0.04
	200	0.07	0.04	0.05	0.05	0.07
	500	0.04	0.06	0.07	0.05	0.06
	1000	0.05	0.04	0.05	0.06	0.06
	1500	0.06	0.05	0.03	0.04	0.05
10	100	0.09	0.04	0.05	0.06	0.08
	200	0.08	0.07	0.06	0.06	0.04
	500	0.06	0.05	0.03	0.04	0.05
	1000	0.07	0.04	0.05	0.05	0.07
	1500	0.06	0.06	0.04	0.03	0.06
20	100	0.09	0.06	0.04	0.07	0.08
	200	0.07	0.04	0.05	0.05	0.07
	500	0.06	0.05	0.03	0.04	0.06
	1000	0.05	0.04	0.06	0.03	0.07
	1500	0.04	0.06	0.05	0.04	0.05

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

Table 3. Power estimates for different goodness-of-fit tests

Conditions		Goodness-of-fit test				
I	*N*	LR2	LR3	LR4	χ^2	P-χ^2
5	100	0.20	0.29	0.32	0.30	0.28
	200	0.35	0.37	0.42	0.45	0.50
	500	0.56	0.45	0.61	0.60	0.69
	1000	0.79	0.75	0.81	0.76	0.82
	1500	0.85	0.89	0.93	0.90	0.87
10	100	0.23	0.31.	0.29	0.28	0.30
	200	0.39	0.43	0.47	0.45	0.46
	500	0.61	0.62	0.59	0.64	0.65
	1000	0.80	0.84	0.87	0.86	0.72
	1500	0.85	0.88	0.92	0.89	0.80
20	100	0.17	0.19	0.22	0.30	0.10
	200	0.36	0.38	0.43	0.48	0.31
	500	0.57	0.61	0.59	0.70	0.47
	1000	0.81	0.83	0.91	0.82	0.61
	1500	0.88	0.87	0.94	0.91	0.75

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

Table 4. TLI and CFI values under correct model specification

Conditions		TLI	CFI
I	*N*	M (SE)	M (SE)
5	100	0.95 (0.04)	0.94 (0.03)
	200	0.96 (0.03)	0.95 (0.03)
	500	0.95 (0.02)	0.93 (0.02)
	1000	0.97 (0.02)	0.96 (0.02)
	1500	0.96 (0.02)	0.98 (0.02)
10	100	0.92 (0.03)	0.97 (0.01)
	200	0.98 (0.03)	0.96 (0.02)
	500	0.93 (0.04)	0.94 (0.03)
	1000	0.95 (0.04)	0.93 (0.02)
	1500	0.96 (0.02)	0.98 (0.03)
20	100	0.98 (0.03)	0.97 (0.02)
	200	0.95 (0.03)	0.93 (0.02
	500	0.96 (0.02)	0.98 (0.01)
	1000	0.97 (0.03)	0.94 (0.04)
	1500	0.96 (0.02)	0.98 (0.02)

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.

Table 5. TLI and CFI values under incorrect model specification

Conditions		TLI	CFI
I	*N*	M (SE)	M (SE)
5	100	0.91 (0.03)	0.92 (0.02)
	200	0.92 (0.04)	0.91 (0.01)
	500	0.90 (0.03)	0.88 (0.02)
	1000	0.89 (0.01)	0.90 (0.03)
	1500	0.91 (0.02)	0.92 (0.01)
10	100	0.90 (0.01)	0.91 (0.04)
	200	0.89 (0.03)	0.92 (0.03)
	500	0.90 (0.02)	0.91 (0.02)
	1000	0.91 (0.03)	0.90 (0.04)
	1500	0.87 (0.04)	0.89 (0.01)
20	100	0.89 (0.02)	0.91 (0.02)
	200	0.91 (0.02)	0.90 (0.02)
	500	0.92 (0.03)	0.91 (0.03)
	1000	0.90 (0.03)	0.92 (0.02)
	1500	0.89 (0.04)	0.90 (0.01)

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.