Journal Title
XX(X):1–8
©The Author(s) 2023
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Assessing Fit of IRT models using a Randomisation LR Test and Fit Indices

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

Abstract

Abstract text.

Keywords

IRT, Goodness-of-fit test, fit indices

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 – constructs 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 wrong model is used, 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 or questionnaire data.

[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

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 hypothese 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. These two tests, however, both suffer from issues. Pearson's test, for example, will not follow a χ^2-distribution when many score pattern frequencies are missing or low, which is often the case – especially when test length increases. The χ^2-difference test instead is difficult to use for both the two- (2PL) and three-parameter (3PL) model. This 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 χ^2-difference test 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).

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 some of the 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 some 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. The added value of the current study lies within the calculations of the fit indices. The calculations will be based on a complete-independence baseline model, which we argue is a more accurate baseline model. Furthermore, compared to Cai et al., we base the calculations on the χ^2 information statistic and not a limited information statistic. Both of these factors can influence the performance and therefore should be investigated further.

To summarise, in the field of IRT there are (a) not enough usable goodness-of-fit tests available to test the 2- and 3PL model and (b) scarce studies investigating the performance of the CFI and TLI, which we address.

The present study

In order to create a goodness-of-fit test to use 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 Likelihood Ratio (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 with a complete-independence baseline model 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

Theoretical background

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. In the present study, we consider IRT with unidimensional θ, dichotomous test items and two different IRFs. Below, you find the IRF for the three-parameter logistic model (3PL):

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

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 model 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 two-parameter logistic model (2PL) is a special case of the 3PL, where γ_i equals 0 for all items. The 2PL can then be specified even further to obtain the one-parameter logistic model (1PL), where it is assumed that α_i equals 1 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 patterns 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. In the current study, we assume θ to be 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. In the current study, 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. After a model has been fitted as described above, the next step is usually to test how well the model fits the data. There are multiple options to assess model fit. Commonly, both goodness-of-fit tests and fit indices are used.

HIER VERBETEREN:

The present study compares the performance of three goodness-of-fit tests: (1) a χ^2-difference test, (2) Pearson's χ^2-test and (3) our own developed test, a LR Randomisation test with the following formula:

$$\frac{max(\mathcal{L}_0)}{\prod_{j=1}^{g} max(\mathcal{L}_j)}, \tag{5}$$

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 each group, gained by randomly assigning the observations to g groups. According to Wilks' theorem (Wilks, 1938), this LR will then asymptotically follow a χ^2-distribution. This allows the test to be used for Null Hypothesis Significance testing. The χ^2-difference test is obtained by testing the 2PL under the 3PL. Finally, Pearson's χ^2-test is calculated through aggregating score patterns from the data and comparing the observed score pattern frequencies to the expected score pattern frequencies under a given model.

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

$$P(X_i = 1|\beta_i) = \frac{e^{-\beta_i}}{1 + e^{-\beta_i}}, \tag{6}$$

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 therefore the items will no longer correlate with one another (i.e., they 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. To estimate the indices, a saturated model is also required. In a saturated model there are as many parameters as data points, which leads to a perfect fit. We choose the following saturated model:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \pi_{\boldsymbol{x}}, \tag{7}$$

where there no longer is an IRF. Instead, perfect model fit is gained by allowing each score pattern to have their own parameter ($\pi_{\boldsymbol{x}}$), which is equal to the relative frequency of the score pattern.

– HIER DE MLE TOEVOEGEN VAN SAT MODEL –

Using these two models, the fit indices can be calculated through the following formulae:

$$\text{CFI} = 1 - \frac{\chi^2 - \mathit{df}}{\chi_0^2 - \mathit{df}_0}, \tag{8}$$

$$\text{TLI} = 1 - \frac{\chi^2/\mathit{df}}{\chi_0^2/\mathit{df}_0}, \tag{9}$$

where in both equations, the numerator is a χ^2-difference test between the tested model and the saturated model with *df* degrees of freedom, and the denominator is a χ^2-difference test between the baseline model and the tested model with df_0 degrees of freedom.

– hier iets toevoegen over hoe we gaan teseten –

Simulation study

Data generation

Data generation starts by first sampling the person parameters θ from a standard normal distribution. Then a model 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] for every repetition of five items. As for the discrimination parameter α_i, we choose repetitions of the values [0.7, 0.85, 1, 1.15, 1.3] per five items. Finally, for the pseudo-guessing parameter γ_i, we choose the values [.09, 0.12, 0.15, 0.18, 0.21] per five items. Then, probabilities are estimated for all items on a test, given θ and the chosen model and its item parameters. Finally, a matrix of simulated responses to a test is created by sampling from a binomial distribution for every item, given each person. We replicate each simulation condition (see below) 300 times.

Simulation design

In order to answer the research questions, we conduct a simulation study that varies 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 1*. This resulted in a total of 3 (test length) x 5 (sample size) x 2 (model type) = 30 conditions. In each replication of each condition, we fit the 2PL model and then obtain the results of the five goodness-of-fit tests (of which three are different versions of the LR Randomisation test) and the two fit indices.

Performance metrics

We study performance of the different goodness-of-fit tests by estimating both type I error and power. 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.

To measure the performance of the TLI and CFI, we estimate the mean and 95% confidence interval (C.I.) of each fit index under every condition. Then, we can inspect whether the average TLI and CFI values decrease in the conditions where the data is generated under the 3PL compared to when the data is generated under the 2PL.

Empirical example

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

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.

(Temporary) Results

Simulation study

Empirical example