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

Journal Title

XX(X):1–7

©The Author(s) 2023

Reprints and permission:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/ToBeAssigned

www.sagepub.com/

SAGE

Nina van Gerwen¹ and Dave Hessen²

Abstract

Abstract text.

Keywords

IRT, Goodness-of-fit test, fit indices

Introduction

Item Response Theory (IRT) is an often used tool for measuring and analysing tests or questionnaires in psychological, educational and organisational practices due to the advantages it holds compared to Classical Test Theory (CTT). For example, compared to CTT, IRT has been shown to be better at individual change detection when using larger tests (Ruslan et al., 2016). Furthermore, IRT allows for more flexibility in analysing tests, where you can investigate all items separately instead of the test-level approach in CTT. The models used in IRT measure latent traits - traits that are not directly observable (e.g., attitudes, intelligence, etc.) - through analysing answers to a test/questionnaire. The goal in IRT is to find the model that best describes the scores on the test items. To achieve this,

¹Utrecht University, NL

²Utrecht 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.i.vangerwen@uu.nl

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 the data gained from a questionnaire.

There are mainly two procedures that can be applied to measure model-fit. First, model-fit can be assessed through goodness-of-fit tests. Commonly used goodness-of-fit tests are a χ^2 -difference test and Pearson's χ^2 -test. These 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. The χ^2 -difference test instead is difficult to use for the three-parameter (3PL) model. This is because generalisations of the 3PL model 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).

Therefore, 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. Finally, theory and logic should also be involved when determining which model should be used as previously recommended by Bryant (2021). 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, such as the SRMR. Although research has been done to examine the performance of some the Structural Equation Modeling fit indices 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. However, Cai et al. researched the performance of the TLI with a different

information statistic, whereas we are more interested in the performance of the 'standard' TLI. Furthermore, we believed that the calculations for the CFI and TLI by Yang were incorrect due to a wrong baseline model, affecting their results.

To summarise, there were (a) too few goodness-of-fit tests available in IRT to test the 3PL model and (b) scarce studies investigating the CFI and TLI, which we addressed.

The present study

In order to create a goodness-of-fit test to use for the 3PL in IRT and to better understand the possible uses of the CFI and TLI in an IRT setting, the present study answered 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 were reported to answer the research questions. Furthermore, empirical data from XX study were used to illustrate the value of the LR Randomisation test, the TLI and the CFI.

Methods

Statistical background

Before we share the methodology of the current study, let us first examine a brief summary on the statistical theory associated with our study. In IRT, the goal is to find the 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 considered IRT with unidimensional θ , dichotomous test items and three different IRF. Below, you find the IRF for the three-parameter logistic model (3PL):

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

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 model depends on (a) the latent variable, θ , (b) the intercept of the item β_i , which denotes how difficult the item is, (c) the slope 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 specification 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. The 1PL model is also commonly known as the Rasch model.

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 of independence of observations and the assumption that persons are randomly sampled from a population. The joint marginal probability of all score patterns in a given sample will then become:

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

where $\phi(\theta)$ is the univariate density of the latent variable θ . In order to solve this equation, the density of θ has to be specified. In the current study, we assume θ to always be a standard normal distribution. With the joint marginal probability, 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 estimation of the item parameters $\hat{\boldsymbol{\nu}}$. This is generally how an IRT model is fitted to data. In the current study, models were fitted using functions from the *ltm* package in R (Rizopoulos, 2006), which approximates marginal maximum likelihood estimation through the Gauss-Hermite quadrature rule.

Assessing Model Fit

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. The present study compared 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:

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

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 was calculated in the following ways: the 1PL model was tested under the 2PL model, the 2PL model was tested under the 3PL model. For the 3PL model, the test was not used due to the limitations mentioned before. Finally, the 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 the model. As for fit indices, we researched the performance of the TLI and CFI in an IRT context. These models make use of a baseline model, our baseline model is a complete-independence model with the following IRF:

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

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 a correct 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). Furthermore, the models also make use of a

saturated model, where there are as many parameters as data points, leading to a perfect fit:

$$P(\mathbf{X} = \mathbf{x}) = \pi_{\mathbf{x}}. \quad (7)$$

In the saturated model, there is no longer an IRF. Instead, perfect model fit is gained by allowing each score pattern to have their own parameter ($\pi_{\mathbf{x}}$), which is equal to the relative frequency of the score pattern. Using these two models, the fit indices can be calculated through the following formulae:

$$\text{CFI} = 1 - \frac{\chi^2 - df}{\chi_0^2 - df_0}, \quad (8)$$

$$\text{TLI} = 1 - \frac{\chi^2/df}{\chi_0^2/df_0}, \quad (9)$$

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 tested model and the complete independence model with df_0 degrees of freedom.

Data generation

Data generation was done by first sampling person parameters, θ , from a standard normal distribution. Then, a model was chosen as basis for the data generation. We chose to keep item parameters static over all simulations. For the difficulty parameter β_i , we chose the values -2, -1, 0, 1 and 2 for every repetition of five items. As for the discrimination parameter α_i , we chose repetitions of the values 0.7, 0.85, 1, 1.15 and 1.3 per five items. Finally, for the pseudo-guessing parameter γ_i , we chose XXX. Then, probabilities were calculated for all items on a test, given θ and the chosen model's item parameters. Finally, a matrix of simulated responses to a test was created by sampling from a binomial distribution for every item, given each person. We replicated each simulation condition (see below) 500 times.

Simulation design

In order to answer the research questions, we conducted a simulation study that varied 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 3 (model type) = 45 conditions.

In each replication of each condition, we calculated five goodness-of-fit tests (3 of which are different versions of the LR Randomisation test) and the two fit indices. Performance of the different goodness-of-fit tests was then studied by estimating both type I error and power. Power was estimated when fitting and testing a different model to the data than the model used to generate the data. Type I error was estimated when fitting and testing the model that was used to generate the data. With these values, we compared the different type of tests with one another, where a test with lower power or higher type I error was noted as performing worse. To measure the performance of the TLI and CFI, we calculated the proportion of times that the fit indices improved when the correct model was used compared to another model.

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	20 - 50 - 100 - 200 - 500	The total number of observations that will be available for each item
Model type	1PL - 2PL - 3PL	The models that we will use as the basis for both data generation and model-fitting
Number of groups	2 - 3 - 4	The number of groups that the total dataset gets divided into for the Randomisation LR test calculations

Note. 1PL = one-parameter logistic model; 2PL = two-parameter logistic model; 3PL = three-parameter logistic model.

Empirical example

To examine the possible use 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 ($N = \dots$). The dataset has XX items on a dichotomous scale that together measure XXX. We fitted the 3PL model with XX.