
Designing and Evaluating a Goodness-of-Fit Test for IRT models

Journal Title

XX(X):1–5

©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

This part will contain information about: IRT, fit indices, goodness of fit tests, issues, etc.

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 simulation studies:

1. What sample size is necessary at different test lengths for the Randomisation test to perform well?

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.l.vangerwen@uu.nl

2. How does the performance of the 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?

Statistical background

Item response functions The below defined item response functions were used in the current study:

1PL:

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

where X_i is a random variable indicating the response to item i . The probability of scoring a 1 on item i in the 1PL model depends on the latent variable, θ , that you are trying to measure and the difficulty of the item, β_i .

2PL:

$$P(X_i = 1|\theta, \alpha_i, \beta_i) = \frac{e^{\alpha_i \theta + \beta_i}}{1 + e^{\alpha_i \theta + \beta_i}} \quad (2)$$

where the probability of scoring a 1 now also depends on an item-dependent intercept term α_i , which shows how well an item discriminates between individuals who score a 0 and individuals who score a 1,

3PL:

$$P(X_i = 1|\theta, \alpha_i, \beta_i, \gamma_i) = \gamma_i + (1 - \gamma_i) \cdot \frac{e^{\alpha_i \theta - \beta_i}}{1 + e^{\alpha_i \theta - \beta_i}} \quad (3)$$

where the probability of scoring a 1 on item i is also dependent on 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).

Complete independence:

$$P(X_i = 1|\beta_i) = \frac{e^{\beta_i}}{1 + e^{\beta_i}} \quad (4)$$

In the complete independence model, 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. This 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).

Saturated model:

$$P(X_i = 1) = \frac{n_{X_i=1}}{N} \quad (5)$$

Fit indices The current study investigated the CFI and TLI, which can be calculated through the following formulae:

CFI:

$$CFI = 1 - \frac{\chi^2 - df}{\chi_0^2 - df_0} \quad (6)$$

TLI:

$$TLI = 1 - \frac{\chi^2/df}{\chi_0^2/df_0} \quad (7)$$

where 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.

Goodness-of-fit tests We compared the performance of the following three goodness-of-fit tests.

- χ^2 -difference test
 - The 1PL model was tested under the 2PL model. The 2PL model was tested under the 3PL model. For the 3PL model, this test was not used.
- Pearson's χ^2 -test
 - Calculated through observing the differences in the observed and expected frequency of score patterns: $\sum_{i=1}^n \frac{(Obs_i - Exp_i)^2}{Exp_i}$
- Randomisation test
 - This is the test that we have developed and tested in the current study. The formula of the test statistic is:

$$\frac{\max(L_0)}{\prod_{j=1}^g \max(L_j)} \quad (8)$$

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 Wilk's theorem (?), this LR will then asymptotically follow a χ^2 -distribution. This allows the Randomisation test to be used for Null Hypothesis Significance testing.

Methods

Data generation

Data generation was done by first sampling person parameters (i.e., 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 an overview of the chosen item parameters per model, see *Table 2*. Then, probabilities were calculated for all items on a test, given θ and the chosen model's item parameters. Finally, a matrix of 0's and 1's was created by sampling from a binomial distribution for every item, given each person. We replicated each simulation condition (see below) 500 times. Each replication, the item parameters remained the same and only new person parameters were sampled.

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) x 3 (number of groups) = 135 conditions.

In each replication of each condition, we calculated the three goodness-of-fit tests and two fit indices. Performance of the three 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 test calculations

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

Results

Empirical example

Discussion