Journal Title
XX(X):1–8

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

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

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

**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?

[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

2. How does the performance of the Randomisation test compare to the performance of a $\chi^2$-difference and Pearson's $\chi^2$-test?
3. What is the performance of the TLI and CFI with a complete-independence baseline model in IRT?

## 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 $\theta$, (2) independence of observations and (3) the response to an item can be modeled by an item response function (IRF). Note that $\theta$ tends to be unidimensional, however it can be generalised to a multidimensional setting. An IRF is a mathematical equation that calculates the probability to score a certain category on an item given $\theta$. In the present study, we considered IRT with unidimensional $\theta$, dichotomous test items and the following three IRF:

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

which is known as the one-parameter logistic model (1PL), 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, $\theta$, that you are trying to measure and the difficulty of the item, $\beta_i$.

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

this is a generalisation of the 1PL, known as the two-parameter logistic model (2PL), where the probability of scoring a 1 now also depends on an item-dependent intercept term $\alpha_i$, which shows how well an item discriminates between individuals who score a 0 and individuals who score a 1. This IRF can be generalised even further to the three-parameter logistic model (3PL):

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

where the probability of scoring a 1 on item $i$ is further dependent on an item-specific lower asymptote $\gamma_i$, 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).

Then, due to the assumption of conditional independence, we can model the probability of a complete score pattern to $k$ items simply by factoring the probability for each item:

$$P(\boldsymbol{X_a} = \boldsymbol{x_a}|\theta_a, \boldsymbol{\nu}) = \prod_{i=1}^{k} P(X_i = 1|\theta_a, \boldsymbol{\nu}) \tag{4}$$

where $\boldsymbol{X}$ is now a random variable indicating a scorepattern (i.e., a vector of 0's and 1's) and $\boldsymbol{x_a}$ is the realisation of $\boldsymbol{X_a}$ for person $a$. 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. Then, the joint marginal probability of all score patterns in a given sample will become:

$$\int \prod_{i=1}^{k} \{P(\boldsymbol{X_a} = \boldsymbol{x_a}|\theta_a, \boldsymbol{\nu})\}\, d\phi(\theta)\, d\theta \tag{5}$$

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. In the current study, we assume $\theta$ 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.

– perhaps talk about conditional independence now and maximum likelihood estimation –

then talk about model fit testing and fit indices

For the calculation of fit indices, two more IRF are considered...

Complete independence:

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

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:

$$P(X_i = 1) = \frac{n_{X_i=1}}{N} \tag{7}$$

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

CFI:

$$CFI = 1 - \frac{\chi^2 - df}{\chi_0^2 - df_0} \tag{8}$$

TLI:

$$TLI = 1 - \frac{\chi^2/df}{\chi_0^2/df_0} \tag{9}$$

where the numerator is a $\chi^2$-difference test between the tested model and the saturated model with $df$ degrees of freedom and the denominator is a $\chi^2$-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.

- $\chi^2$-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 $\chi^2$-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:

$$\frac{max(L_0)}{\prod_{j=1}^{g} max(L_j)} \tag{10}$$

    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 $g$ grouops. According to Wilk's theorem (**?**),

this LR will then asymptotically follow a $\chi^2$-distribution. This allows the Randomisation test to be used for Null Hypothesis Significance testing.

## Data generation

Data generation was done by first sampling person parameters (i.e., $\theta$) 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 $\theta$ 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

## References

Barton, M. A. and Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, 107(2):238.

Cai, L., Chung, S. W., and Lee, T. (2021). Incremental model fit assessment in the case of categorical data: Tucker–lewis index for item response theory modeling. *Prevention Science*, pages 1–12.

Crişan, D. R., Tendeiro, J. N., and Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional irt models. *Applied Psychological Measurement*, 41(6):439–455. PMID: 28804181.

Curran, P., West, S., and Finch, J. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1):16–29.

Darrell Bock, R. and Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, 35(2):179–197.

Jiao, H. and Lau, A. C. (2003). The effects of model misfit in computerized classification test. In *Annual meeting of the National Council of Educational Measurement, Chicago, IL*.

Kang, T. and Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, 31(4):331–358.

Krammer, G. (2018). The andersen likelihood ratio test with a random split criterion lacks power. *Journal of modern applied statistical methods: JMASM*, 17:eP2685.

Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63(3):509–525.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement Interdisciplinary Research and Perspectives*, 11:71–101.

Maydeu-Olivares, A. and Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 49.

Nye, C. D., Joo, S.-H., Zhang, B., and Stark, S. (2020). Advancing and evaluating irt model data fit indices in organizational research. *Organizational Research Methods*, 23(3):457–486.

Orlando, M. and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1):50–64.

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Team, R. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Team, R. C. (2021). *R: A Language and Environmental for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60 – 62.

Yang, X. (2020). *Comparing Global Model-Data Fit Indices in Item Response Theory Applications*. PhD dissertation, Florida State University, College of Education.

Zhao, Y. and Hambleton, R. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, 8.