Journal Title
XX(X):1–9

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 measurement and analysing 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. Furthermore, IRT allows for more flexibility in analysing tests, where you can investigate all items separately instead of test-level approach of CTT. Models in IRT are models that measure latent traits (i.e., not directly observable features) through analysing answers to a test. An example of a commonly measured latent trait in multiple areas is intelligence. The goal in IRT is to find the model that correctly describes the scores on the test items. To

[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

achieve this, the model must fit. If a wrong model is used, it can lead to dire consequences such as faults in the validity of your measurement (...) and by extension your conclusion. Therefore, model-fit should be assessed.

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 (...). Another concern with the χ^2-difference test is power. When sample size increases, the power to reject any reasonable model that does not perfectly fit, also increases (...). 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. On their own, fit indices are not sufficient to infer whether a model fits well. However, taken together with goodness-of-fit tests, it allows you to have a good overview of model-fit. For example, when due to power, a goodness-of-fit test shows that the model should be rejected, fit indices can indicate whether the model is still reasonable. Multiple fit indices have been developed specifically for IRT, however most suffer from their own limitations (for an overviw, see (Nye et al., 2020)). Fit indices from Structural Equation Modeling can also be used in IRT. However, 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). Only one paper examines the CFI (Yang, 2020) and two papers investigate the TLI (Cai et al., 2021; Yang, 2020) in an IRT setting. 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 Randomisation test to perform well?
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?

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). Note that θ tends to be unidimensional, however it can be generalised to a multidimensional setting. An IRF is a mathematical equation that relates the probability to score a certain category on an item given θ. In the present study, we considered IRT with unidimensional θ, 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, θ, that you are trying to measure and the difficulty of the item, β_i. The two-parameter logistic model (2PL) is a generalisation of the 1PL:

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

where the probability of scoring a 1 now also depends on an item-dependent slope term α_i, which shows how well item i 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 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). 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 probabilities for each item:

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

where $\boldsymbol{X}$ is now a random vector indicating a scorepattern and $\boldsymbol{x_a}$ is the realisation of $\boldsymbol{X}$ 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. The joint marginal probability of all score patterns in a given sample will then become:

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

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. 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. Most 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 our own developed test (3) a Randomsation LR test with the following formula:

$$\frac{max(L_0)}{\prod_{j=1}^{g} max(L_j)}, \tag{6}$$

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 (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 underthe 2PL model, the 2PL model was tested under the 3PL model and 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:

$$P(X_i = 1|\beta_i) = \frac{e^{-\beta_i}}{1 + e^{-\beta_i}}, \tag{7}$$

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:

$$P(\boldsymbol{X} = \boldsymbol{x_a}) = \pi_{\boldsymbol{X}}. \tag{8}$$

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 ($\pi_{\boldsymbol{X}}$), which is equal to the frequency of the scorepattern. Using these two models, the fit indices can be calculated through the following formulae:

$$\text{CFI} = 1 - \frac{\chi^2 - df}{\chi_0^2 - df_0}, \tag{9}$$

$$\text{TLI} = 1 - \frac{\chi^2/df}{\chi_0^2/df_0}, \tag{10}$$

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, we chose the values -2, -1, 0, 1 and 2 for every repetition of five items. As for the discrimination parameter α, 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, 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 Randomisation LR test) and the 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 LR 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.