# Designing and Evaluating a Likelihood-Ratio Test for IRT models

*Methodology and Statistics for the Behavioural,*

*Biomedical and Social Sciences*

Nina van Gerwen (1860852)

Supervisor: Dave Hessen

Candidate Journal: *Applied Psychological Measurement*

13th of October, 2022

# 1 Introduction

In organisations, a considerable amount of effort tends to be dedicated to researching important organisational variables, such as job performance and personality traits. These types of variables are latent constructs. This entails that they are not directly observable. Thus, to measure latent constructs we have to infer them based on responses to a test or questionnaire. To achieve this, Item Response Theory (IRT) is often used. A large issue in IRT, however, is that you must use a model that correctly describes the answers to test items in order to gain correct conclusions. This means that you should investigate whether the model you use, fits the data you acquired. If a wrong model is used, it can lead to dire consequences such as faults in the validity of your measurement (**consq2**; **consq3**; Zhao & Hambleton, 2017) and by extension your conclusion. In an organisation, this could for example translate into issues such as discharging the wrong people. Hence, it is vital that you test the appropriateness of your model (i.e., model-fit).

Currently in IRT research, there are mainly two procedures that can be applied to every model to measure model-fit. Namely, a $\chi^2$-difference test and the Pearson's $\chi^2$-test.

These tests both suffer from issues. Pearson's test, for example, will not follow a $\chi^2$-distribution when many score pattern frequencies are missing or low. The $\chi^2$-difference test, instead, is difficult to use for the three-parameter (3PL) model. This is because the 3PL model can only be nested under the four-parameter logistic (4PL) model, which suffers from consistency and estimation issues (Barton & Lord, 1981; Loken & Rulison, 2010). Another issue with the $\chi^2$-difference test is that when sample size is large, the test will increase in power and it will reject models that are still reasonable (Curran et al., 1996). For this reason, fit indices were introduced to investigate whether a model is reasonable after being rejected by a $\chi^2$-difference test. Fit indices are standardised indicators of model appropriateness. For an overview of current fit indices in IRT and their limitations, see Nye et al. (2020). Besides these indices, 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 and Lewis, 1973) and the Comparative Fit Index (CFI; Bentler, 1990) in IRT. We found only one paper examining the CFI (Yang, 2020) and two papers investigating the TLI (Cai et al., 2021; Yang, 2020). However, we believe that the calculations for the CFI and TLI by Yang are not correct due to

the baseline model used, which would affect the results. To summarise, there are (a) too few goodness-of-fit tests available in IRT to test the 3PL model and (b) insufficient research on the CFI and TLI, which we will attempt to address.

Our proposed research would (a) develop and test the performance of a model-fit Likelihood Ratio (LR) test that is applicable to all IRT models and (b) test the performance of the CFI and TLI with a complete-independence baseline model. More specifically, we will answer the following three research questions:

1. Under which conditions will the $\chi^2$-test associated with our LR perform well?
2. How does the performance of our developed test compare to the performance of the $\chi^2$-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?

# 2 Analytic strategy

In order to answer the research questions, we will conduct a simulation study. When simulating data, we will vary the following four variables:

- Test length
- Sample size
- Model types
- Number of groups

For a complete overview of the conditions that will be used for the different variables, see *Table 1.*

Table 1

*Overview of Simulation Conditions for all Variables*

| Variable | 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 LR calculations |

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

Each particular condition will be replicated 500 times and in each replication the following LR will be calculated:

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

where $g$ is the group number, $L_0$ is the likelihood for the whole dataset and $L_j$ is the likelihood forthe dataset for each group in $g$, gained by randomly assigning the dataset to $g$ groups. According to Wilk's theorem (Wilks, 1938), this LR will then asymptotically follow a $\chi^2$-distribution with the following formula:

$$-2ln\frac{max(L_0)}{\prod_{j=1}^{g} max(L_j)} \rightarrow \chi^2(\Delta) \tag{2}$$

where $\Delta$ is the difference in the number of estimated parameters between the two likelihoods. This entails that the LR can be used for Null Hypothesis Significance testing. Then, due to the fact that we know the data-generating model, we can calculate the proportion of times that the test rejects the correct model (i.e.,$\beta$: type II error). Furthermore, we can also calculate the proportion of times that the test accepts a wrong model (i.e., empirical $\alpha$). Knowing these values, we can compare the multiple tests with one another, where a test with higher values for either proportions will be noted as performing worse. This will inform us of the scenarios where our test works properly and scenarios where the test lacks power. To measure

the performance of the TLI and CFI, we can calculate the proportion of times that the fit indices improve (i.e., approach the value 1) when the correct model is used compared to the other models. Finally, we will provide an empirical example by testing the LR test, CFI and TLI on an existing dataset and interpreting their results.

The proposed research would be conducted in R (R. C. Team, 2021) through the use of RStudio (R. Team, 2020). As for packages, we will use the MASS (Venables & Ripley, 2002), lavaan (Rosseel, 2012) and ltm (Rizopoulos, 2006) packages. Approval for the ethical consent for the simulated dataset has been requested. The approval for the empirical example, however, has to wait until a proper dataset has been found.

# References


Barton, M. A., & 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., & Lee, T. (2021). Incremental model fit assessment in the case of categorical data: Tucker–lewis index for item response theory modeling. *Prevention Science*, 1–12.

Curran, P., West, S., & Finch, J. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

Kang, T., & Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358. https://doi.org/10.1177/0146621606292213

Krammer, G. (2018). The andersen likelihood ratio test with a random split criterion lacks power. *Journal of modern applied statistical methods: JMASM*, *17*, eP2685. https://doi.org/10.22237/jmasm/1555594442

Loken, E., & 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. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*. https://doi.org/10.1080/00273171.2014.911075

Nye, C. D., Joo, S.-H., Zhang, B., & Stark, S. (2020). Advancing and evaluating irt model data fit indices in organizational research. *Organizational Research Methods*, *23*(3), 457–486. https://doi.org/10.1177/1094428119833158

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Rizopoulos, D. (2006). Ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Team, R. C. (2021). *R: A language and environmental for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Team, R. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth) [ISBN 0-387-95457-0]. Springer. https://www.stats.ox.ac.uk/pub/MASS4/

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. https://doi.org/10.1214/aoms/1177732360

Yang, X. (2020). *Comparing global model-data fit indices in item response theory applications* [PhD dissertation]. Florida State University, College of Education.

Zhao, Y., & 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*. https://doi.org/10.3389/fpsyg.2017.00484