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*

FETC Approval: 22-1818

13th of October, 2022

1 Introduction

In Item Response Theory (IRT), it is of vital importance to use a model that correctly describes the answers to test items. 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. 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 χ^2-difference test and Pearson's χ^2-test. These tests 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 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 concern with the χ^2-difference test is power. When sample size is large, the test has too much power and it will reject models that are still reasonable (Curran et al., 1996). Therefore, fit indices were introduced to investigate whether a model is reasonable. Fit indices are standardised indicators of model-fit. For an overview of fit indices in IRT and their limitations, 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 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 incorrect due to a wrong baseline model, affecting 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 address.

Our proposed research would (a) develop and test the performance of a goodness-of-fit Likelihood Ratio (LR) test 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 three research questions:

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

2 Analytic strategy

In order to answer the research questions, we shall conduct a simulation study, varying four variables: test length, sample size, model types and number of groups. For an overview of the conditions we will use for the different variables, see *Table 1.*

Table 1

Overview of Simulation Conditions for Each 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 condition will be replicated 500 times and the following LR will be calculated:

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

where L_0 is the likelihood for the whole dataset and L_j is the likelihood for each group, gained by randomly assigning the observations to g groups. According to Wilk's theorem (Wilks, 1938), this LR will then asymptotically follow a χ^2-distribution. This allows the LR to be used for Null Hypothesis Significance testing. Performance of the test can then be operationalised by both type I error and power. Power can be calculated by fitting and testing a different model to the data than the model used to generate the data. Type I error can be calculated when fitting and testing the model that was used to generate the data. With these values, we can study what sample size and test length is necessary for our χ^2-test to have enough power. Furthermore, we can compare multiple tests with one another, where a test with lower power will be noted as performing worse. To measure the performance of the TLI and CFI, we can calculate the proportion of times that the fit indices improve when the correct model is used compared to other models. Finally, we will provide an empirical example by testing our χ^2-test, CFI and TLI on the LSAT dataset from Darrell Bock and Lieberman (1970).

The proposed research would be conducted in R (R. C. Team, 2021) through the use of RStudio (R. Team, 2020) with the *MASS* (Venables & Ripley, 2002), *lavaan* (Rosseel, 2012) and *ltm* (Rizopoulos, 2006) packages. Ethical approval for the example has been requested.

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.

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional irt models [PMID: 28804181]. *Applied Psychological Measurement*, *41*(6), 439–455. https://doi.org/10.1177/0146621617695522

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

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

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

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