Designing and Evaluating a Likelihood-Ratio Test for IRT models

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Nina van Gerwen (1860852)
Dave Hessen
Applied Psychological Measurement

13th of October, 2022

- Real life problem: For every type of research, it is important to know whether the model you are using actually correctly illustrates the response process. In other words, it is important to know whether the model fits the data. If the wrong model is used, this can lead to dire consequences such as false conclusions. Furthermore, it also entails that you are make inefficient use of your data. So in order to test whether the model fits the data, there exist goodness-of-fit tests in almost every field of statistics. Most indicators of model fit, such as the AIC/BIC/DIC are relative: they can be used to compare two models on their fit. In Structural Equation Modeling, there also exist fit indices (e.g., SRMR, RMSEA, CFI/TLI), which when all taken together with their rules of thumb can also indicate model fit. However, for IRT research, there exists no goodness-of-fit LR test that is generally applicable to all IRT models (besides the χ^2 test of a model vs. alternative model). Instead, some models have specific LR tests that tend to suffer from different issues (e.g., Andersen's LR test for all Rasch models, which has been shown to lack power (Krammer, 2018)). Therefore, we have come up with a LR test, also based on the proof behind the χ^2 test for a model vs. alternative model, that would be applicable to all IRT models in order to improve model-fit research.
  - Furthermore, goodness-of-fit tests all sufer from specific issues, such as a sensitivity to larger sample sizes. Therefore, there exist fit indices, which can help determine your model fit. Compared to SEM research, IRT models, however, have a lack of fit indices. For a relatively recent overview of fit indices used in IRT, see Nye et al. (2020).
- Research questions: What are the statistical properties (robustness, power, empirical α) of the designed LR test that is applicable to all IRT models?
  - Extra possible research question: something fit index related
- Analytic strategy: To research the statistical properties of our test, a simulation study will be conducted. First, data will be simulated according to certain IRT models. Then, knowing the true model, we can test both the real and other IRT models to the data and see whether our goodness-of-fit test has the abillity to determine when we used the right or wrong IRT model. Empirical α can be determined by calculating how many times the test rejects the model that was used to simulate the data. Power can be determined by calculating what percentage of wrong models are correctly rejected by the test. Robustness ??
  - When simulating data, we will for sure vary: the amount of items in the test (e.g., 5 - 10 - 20) and sample size (e.g., 20 - 50 - 100 - 200 - 500). Other factors we can vary in order to increase generalizability are: the parameters used for the IRT model (e.g., discrimination parameter of 0.4 and 0.7 and difficulty parameter with intervals of 0.5 versus intervals of 1.0), different types of IRT models (e.g., Graded Response Model - Rasch Model, Dichotomous / Polytomous IRT),

and the amount of group randomization used in the LR test (e.g., randomized into 2 - 3 - 4 groups).

- Ethical Consent: due to the fact that the data will be simulated, there should be no issue with either license of the data or ethical consent.

The Likelihood-Ratio test formula:

$$-2ln(\frac{L_{total}}{L_{half1} \cdot L_{half2}}) \rightarrow \chi^2(k) \tag{1}$$

Part I
Introduction

In organisations, a lot of effort tends to be dedicated to researching important organising variables, such as job performance and personality traits. These types of variables are latent constructs, in other words they are not directly visible. In order to measure these constructs, we infer the latent constructs based on responses to items on a test or questionnaire. To achieve this, Item Response Theory (IRT) is often used. A large issue in the field of IRT is that you must use a model that correctly describes the answering process of the test in order to gain correct conclusions. This means that you should test whether the model you used 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 (Zhao & Hambleton, 2017) and invalid conclusions (citation). In an organisation, this could translate into issues such as discharging the wrong people. Hence, it is important that you test the appropriateness of your model (i.e., model-fit).

Currently in IRT research, there are not a lot of procedures to measures model-fit widely available compared to other fields of statistics (e.g., SEM). To illustrate, there exists only one generally applicable model-fit test in IRT. Namely, a comparative χ^2 LR test between two models. The issue with this test, however, is that (a) it requires the null model to be nested under the alternative model and (b) the test of model-fit is relative, it only informs researchers whether the null model is more appropriate than the alternative model. Furthermore, other available tests tend to be specialised, such that the tests work only on certain types of IRT models (e.g., Andersen's (modified) Likelihood-Ratio (LR) test, which only works for Rasch models and has been shown to lack power (Krammer, 2018)). As for global indicators of model-fit, there exist fit indices. – explanation of fit indices – However, there are very few fit indices available for IRT research compared to Structural Equation Modelling. For an overview of current fit indices that are available, see Nye et al. (2020). The paper also shows us that the available fit indices have large limitations, such as being unable to detect minor forms of misfit and faults due to multidimensionality. To summarise, there is a large lack of different measures of model fit in IRT research that we will attempt to address.

Our proposed research would develop and test the performance of a model-fit LR test that is applicable to all IRT models. In more detail, we will answer the following two research questions.

1. What is the robustness (i.e., empericial α and power) of the χ^2 test associated with our LR?

2. How does the performance of our developed test compare to the performance of other available IRT tests?

Part II

Analytic strategy

In order to answer the research question, we will conduct a simulation study. First, data will be simulated according to certain IRT models. Then we will fit multiple types of IRT models to the data and calculate the following LR:

$$-2ln(\frac{L_0}{L_1 \cdot L_2}) \rightarrow \chi^2(2k) \tag{2}$$

where the nominator is the likelihood for the whole dataset, and the denominator is the likelihood for the two halved datasets, which are gained by randomly assigning the whole dataset to two groups. Due to the fact that the test will asymptomatically follow a χ^2 distribution as proven by Wilks' theorem (Wilks, 1938), it 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.,β: type II error). Furthermore, we can also calculate the proportion of times that the test accepts a wrong model (i.e., empirical α). 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.
When simulating data, we will for sure vary: the amount of items in the test (e.g., 5 - 10 - 20) and sample size (e.g., 20 - 50 - 100 - 200 - 500). Other factors we can vary in order to increase generalisability are: the parameters used for the IRT model (e.g., discrimination parameter of 0.4 and 0.7 and difficulty parameter with intervals of 0.5 versus intervals of 1.0), different types of IRT models (e.g., Graded Response Model - Rasch Model, Dichotomous / Polytomous IRT), and the amount of group randomisation used in the LR test (e.g., randomised into 2 - 3 - 4 groups). The more factors we cross-examine, the more information we will receive about the scenarios in which the LR test works and does not work.
The proposed research would be conducted in R (R. C. Team, 2021) through the use of RStudio (R. Team, 2020). For packages, we will use the MASS (Venables & Ripley, 2002), lavaan (Rosseel, 2012) and ltm packages.

References

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

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

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/

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

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