Detecting item misfit in Rasch models

Magnus Johansson, PhD

2025-03-05

Psychometrics have long relied on rule-of-thumb critical values for goodness of fit metrics. With powerful personal computers it is both feasible and desirable to use simulation methods to determine appropriate cutoff values. This paper evaluates the use of an R package for Rasch psychometrics that has implemented functions to simplify the process of determining simulation-based cutoff values. Through six simulation studies, comparisons are made between information-weighted conditional item fit ("infit") and item-restscore correlations using Goodman and Kruskal's γ. Results indicate the limitations of small samples (n < 500) in correctly detecting item misfit, especially when a larger proportion of items are misfit and/or when misfit items are off-target. Infit with simulation-based cutoffs outperforms item-restscore with sample sizes below 500. Both methods result in problematic rates of false positives with large samples (n >= 1000). Large datasets should be analyzed using nonparametric bootstrap of subsamples with item-restscore to reduce the risk of type-1 errors. Finally, the importance of an iterative analysis process is emphasized, since a situation where several items show underfit will cause other items to show overfit. Underfit items should be removed one at a time, and a re-analysis conducted for each step to avoid erroneously eliminating items.

Introduction

This paper presents a series of simulations conducted to evaluate methods to detect item misfit due to multidimensionality in Rasch models. First, conditional item infit and outfit (Müller, 2020) will be under scrutiny. Second, item infit will be compared to the item-restscore method (Christensen & Kreiner, 2013; Kreiner, 2011). Third, a bootstrap method for item-restscore will be presented and tested. This paper is intended for a target group of those who make practical use of Rasch analysis and wish to better understand the expected performance of methods available. As such, we refer readers interested in mathematical and statistical descriptions of the methods to referenced papers detailing this aspect. Only two simple performance metrics will be presented in the results: correct detection rate and false positive rate, both in percentages.

The evaluation of item fit under the Rasch model has, in the majority of published psychometric papers, been conducted using various more or less arbitrary rule-of-thumb critical values. Regarding mean squared (MSQ) item residuals, which should ideally be centered around 1.0, there are two sources often cited. One is the book by Bond and Fox (2015), which has garnered around 12 000 citations according to Google Scholar. It contains a table with rule-of-thumb recommendations for various settings, ranging from 0.8–1.2 to 0.5–1.7. Another frequently seen source, which is not an actual peer-reviewed publication and thus lacks citation counts, is the webpage at https://rasch.org/rmt/rmt162f.htm, where Mike Linacre states 0.5-1.5 to be "productive for measurement". Neither of these sources seem to rely on simulation studies to support their recommendations. While it is reasonable to accept a non-perfect fit to the Rasch model and also describe what one defines as acceptable levels of misfit, such recommendations would seem less arbitrary if related to simulations showing the range of item fit values found when simulating data that fit the Rasch model.

Müller (2020) used simulation to show how the range of critical values for conditional item infit varies with sample size. The expected average conditional item infit range was described by Müller as fairly well captured by Smith's rule-of-thumb formula $1 \pm 2/\sqrt{n}$ (R. M. Smith et al., 1998), where n denotes the sample size. However, the average range does not apply for all items within a dataset, since item location relative to sample mean location also affects the expected model fit for individual items. This means that some items within a set of items varying in location are likely to have item fit values outside Smith's average value range while still fitting the Rasch model. Although primarily affecte by sample size, each item has its variations in the range of expected item fit.

While evaluation of item fit is an essential part of evaluating unidimensionality, it is recommended to use multiple methods. Standardized residuals are frequently analyzed, commonly with principal component analysis (PCA) and an analysis of residual correlations amongst item pairs, often referred to as Yen's Q3. Chou and Wang (2010) showed that the critical value for PCA of residuals to support unidimensionality suggested by Smith (2002), using the largest eigenvalue < 1.5, is not generally applicable since it is affected by both test length and sample size. Christensen and colleagues (2017) used simulation methods to illustrate the expected range of residual correlations under different conditions. Both of these papers provide important information about the dubiousness of using rule-of-thumb critical values when the empirical distribution of a statistic is not known, but they leave practitioners without tools to determine appropriate cutoffs to apply in practical analysis work.

It is here proposed that by using parametric bootstrapping one can establish item fit critical cutoff values that are relevant for a specific sample and item set. The procedure uses the estimated properties of the available data and simulates multiple new response datasets that fit the Rasch model to determine the range of plausible item fit values for each item. The R package `easyRasch` (Johansson, 2024a) includes a function to determine item infit and outfit cutoff values using this method and will be tested in the simulation studies in this paper.

Similar developments, moving from rule-of-thumb towards adaptive critical values, have recently taken place in the related field of confirmatory factor analysis. McNeish and Wolf

(2024) have created an R package called `dynamic` that uses simulation to determine appropriate critical values for commonly used model fit metrics for models using ordinal or interval data.

It is important to note that the conditional item fit described by Müller (2020) and implemented in the `iarm` R package (Mueller & Santiago, 2022) should not be confused with the unconditional item fit implemented in software such as Winsteps and RUMM2030, as well as all R packages except `iarm`. Unconditional item fit can result in unreliable item fit in sample sizes as small as 200 with an increasing probability of problems as sample size increases. Readers are strongly recommended to read Müller's paper to fully understand the issues with unconditional item fit. Additionally, the experienced Rasch analyst will perhaps wonder why the Wilson-Hilferty transformed Z statistic (often abbreviated ZSTD), which is based on unconditional MSQ is not included in this analysis. This is also explained in Müller's paper, where she describes both the notorious problems with sample size and shows that conditional item fit makes ZSTD superfluous. The `easyRasch` package, which is used in this paper, uses the `iarm` implementation of conditional item fit.

Currently, there are no published studies on the performance the item-restscore method, as described by Kreiner and Christensen (Christensen & Kreiner, 2013; Kreiner, 2011), in detecting misfitting items. Comparing it with an improved version of the long used item infit/outfit methods seemed like a good setting to evaluate item-restscore. The conditional likelihood ratio test (Andersen, 1973) is included in Study 6, since it is a global test of fit that many are likely to be familiar with. As such, it also serves as a point of reference.

There are six simulation studies included in this paper:

1. Conditional item infit and outfit
2. Item-restscore
3. Comparing infit and item-restscore
4. Bootstrapped item-restscore
5. Varying the number of items
6. Conditional likelihood ratio test

Methods

This is a general description of the methods used. Each study included in this paper has its own brief introduction and method section. A reproducible manuscript with R code and data is available on GitHub: https://github.com/pgmj/rasch_itemfit. First, a note on terminology. Non-parametric bootstrapping is synonymous with sampling with replacement. With this method, the original response data is directly sampled from multiple times. Parametric bootstrapping is synonymous with simulation, since new data is generated based on person and item parameter estimates from the original response data.

The simulation of response data used three steps: First, a vector of theta values (person scores on the latent variable's logit scale) was generated using `rnorm(mean = 0, sd = 1.5)`. Second, a set of item locations ranging from -2 to 2 logits were generated for dichotomous items, using `runif(n = 20, min = -2, max = 2)`. The same set of item locations were used for all studies except Study 5, which adds 20 more item locations. Third, the theta values were used to simulate item responses for participants, using `sim.xdim()` from the `eRm` package (Mair & Hatzinger, 2007), which allows the simulation of multidimensional response data. The sigma matrix used by `sim.xdim()` was specified to use 0.15 on the off-diagonal, where values are between 0 and 1 with lower values indicating stronger multidimensionality. Multiple datasets with 10 000 respondents each were generated using the same item and person parameters, varying the targeting of the misfitting item(s) and number of the misfitting item(s). More details are described in the separate studies. The parametric bootstrapping procedure was implemented using random samples from the simulated datasets. Sample size variations tested are also described in each study.

The general procedure for the parametric bootstrap was as follows:

1. Estimation of item locations based on simulated item response data, using Conditional Maximum Likelihood (CML, Mair & Hatzinger, 2007).
2. Estimation of sample theta values using weighted maximum likelihood (Warm, 1989).
3. Simulation of new response data that fit the Rasch model, using the estimated item locations and theta values.
4. Estimation of the dichotomous Rasch model for the new response data using CML.
5. Based on step 4, calculation of conditional item infit and outfit (Mueller & Santiago, 2022; Müller, 2020) and/or item-restscore metrics (Kreiner, 2011; Mueller & Santiago, 2022).

Steps three and four were iterated over, using sampling with replacement from the estimated theta values as a basis for simulating the response data in step three. Summary statistics were created with a focus on the percentage of correct detection of misfit and false positives. A complete list of software used for the analyses is listed under Section **??**.

Study 1: Item infit and outfit

Assessing item fit to the Rasch model using infit and outfit was first proposed by Wright and Panchapakesan (1969). For a historical perspective, see Smith and colleagues (1998). Item mean square standardized residuals are either unweighted, which is referred to as "outfit", or information weighted, also known as "infit" (Ostini & Nering, 2006, pp. 86–87). Outfit is sensitive to outliers, while infit is much less affected by outliers. Both infit and outfit are based on individual response residuals. Conditional item infit and outfit are expected to be near 1, with higher values indicating an item to be underfitting the Rasch model (often due