Psychological Test and Assessment Modeling, Volume 65, 2023 (2), 206-219

Visualizing Rasch item fit using conditional item characteristic curves in R

Ann-Sophie Buchardt[1], Karl Bang Christensen[1], and Sidsel Normann Jensen[1]

[1]Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

Abstract

New R computer routines that support rigorous statistical validation of psychological tests have recently appeared. We illustrate how Rasch item fit can be evaluated visually and propose an extension of existing implementations in R. We illustrate the utility using two short psychological tests.

Keywords: Rasch model, R, graphics, item fit

Author note

Correspondence concerning this article should be addressed to Karl Bang Christensen, Postboks 2099, Øster Farimagsgade 5 opg. B, 1014 København K, 15 Øster Farimagsgade 5.kach@sund.ku.dk

Introduction

Since the emergence of the Rasch (1960, 1980) model in 1960 many approaches for evaluating its empirical validity have been proposed. Numerous journal articles on testing the Rasch model have been published. A summary of the literature before 1995 is contained in two book chapters by Glas and Verhelst (1995a,b) and some updated model tests are described in more recent book chapters (Horton et al., 2013; Kreiner and Christensen, 2013b,a; Christensen and Kreiner, 2013). Examples of tutorials and reviews include Pallant and Tennant (2007); Tennant and Conaghan (2007); Hagquist et al. (2009); Aryadoust et al. (2021). Examples of recent literature include Komboz et al. (2018); Courtney et al. (2021); Debelak et al. (2022); Alexandrowicz (2022).

Rasch himself did not discuss the calculation and evaluation of item fit statistics, but stressed the use of graphical evaluations of model fit. The purpose of this article is to illustrate advantages of using graphical model checks for the Rasch model. The focus is on implementations in R (R Core Team, 2013).

Many statistical tests have been proposed for evaluating the fit of an empirical data set to the Rasch model (Glas and Verhelst, 1995a,b; Christensen and Kreiner, 2013). Some of these fit statistics have unknown asymptotic distributions (Müller, 2020b) and this of course complicates their use. However, even for item fit statistics with known asymptotic distributions or non-parametric item fit statistics evaluated using Markov chain Monte Carlo (MCMC) methods it is perhaps not optimal that evaluation of item fit relies on p-values (Wasserstein and Lazar, 2016; Betensky, 2019).

The R package `eRm` (Mair and Hatzinger, 2007) includes Rasch models, linear logistic test models, (linear) rating scale models and (linear) partial credit models in a conditional maximum likelihood (CML) implementation which relates directly to Rasch's original concept of specific objectivity. The introduction of open source tools for Rasch analysis is a great step forward, but many applied researchers and test developers prefer the existing stand-alone programs that they are used to working with. One reason for this is the graphical capabilities of these proprietary computer programs.

The R package `eRm` provides a single plot that can be used to inspect item fit. It is implemented for dichotomous items only. Our aim is to present a plot that can be used to inspect item fit for the dichotomous Rasch model, the rating scale model and the partial credit model. This is done using conditional item characteristic curves (ICC).

Conditional ICC

Visualization of Rasch item fit is often done using the ICC that shows the mean item score as a function of the underlying latent variable. The curve cannot be directly used to evaluate item fit because this latent variable is unobserved. A Conditional ICC (CICC) is a curve describing the expected item mean as a function of the total score. It is possible to make an empirical CICC based on the observed data as both the empirical expected item score and the total scores can be calculated from the data. This empirical curve can then be compared to the model-based CICC to visualize item fit.

Notation Let, for person v, $x_{vi} \in \{0, 1, \ldots, m_i\}$ denote the observed response to item i and $R_v = \sum_{i=1}^{k} X_{vi}$ denote the total score over all k items. Under the polytomous Rasch model

$$P(X_{vi} = x | \theta_v = \theta) = \frac{\exp(x\theta + \beta_{ix})}{\sum_{l=1}^{m_i} \exp(l\theta + \beta_{il})}, \tag{1}$$

the distribution of the total score is

$$P(R_v = r | \theta_v = \theta) = \frac{\exp(r\theta)\gamma_r(\boldsymbol{\beta})}{\prod_{i=1}^{k} \sum_{l=1}^{m_i} \exp(l\theta + \beta_{il})}, \tag{2}$$

where $\boldsymbol{\beta}$ is the vector of all item (easiness) parameters and

$$\gamma_r(\boldsymbol{\beta}) = \sum_{(x_i)} \exp\left(\sum_{i=1}^{k} \beta_{ix_i}\right) \tag{3}$$

are so-called γ-functions (Andersen, 1995, formula 15.20). For dichotomous items these are known as elementary symmetric functions. In (3) the summation is over the set $\{(x_i) : \sum_{i=1}^{k} x_i = r\}$ of all response vectors with total score r and the γ-functions can be calculated using the recursive formula

$$\gamma_r(\boldsymbol{\beta}) = \gamma_r^{(i)}(\boldsymbol{\beta}) + \sum_{x=1}^{m_i} \exp(\beta_{ix})\gamma_{r-x}^{(i)}, \tag{4}$$

where $\gamma_{r-x}^{(i)}$ is the γ-function (3) evaluated without item i. This means that the conditional probabilities of item scores given total scores can be written

$$P(X_{vi} = x | R_v = r) = \frac{\exp(\beta_{ix})\gamma_{r-x}^{(i)}(\boldsymbol{\beta})}{\gamma_r(\boldsymbol{\beta})}. \tag{5}$$

Note that (5) is independent of the value of θ_v due to the sufficiency of the total score in the Rasch model. Based on (5) it is straight-forward to calculate the conditional expectation of item scores given total scores

$$E_{ri} = E(X_{rv}|R_v = r) = \sum_{x=0}^{m_i} x \frac{\exp(\beta_{ix})\gamma_{r-x}^{(i)}(\boldsymbol{\beta})}{\gamma_r(\boldsymbol{\beta})} \tag{6}$$

(Andersen, 1995, formula 15.22).

Grouping the total score For person v, we use S_v to denote her raw score. This leads to the definition of score groups as formalized below. Let $m_. = \sum_{i=1}^{k} m_i$ denote the maximum value of the total score and consider a disjoint union of the set of total scores

$$\{0,1,\ldots,m_.\} = I_1 \cup \ldots I_G,$$

where the sets $I_g = \{L_g,\ldots,U_g\}$, for $g = 1,\ldots,G$, defined using the convention $L_{g+1} = U_g + 1$. Use these to define the ordinal categorical variable

$$S_v = \begin{cases} 1 & \text{if } R_v \in I_1 \\ \vdots & \\ G & \text{if } R_v \in I_G. \end{cases}$$

Note that this definition contains as a special case a singleton I_g defined using $L_g = U_g$.

The Conditional item characteristic curve (CICC) The CICC describes the expected item scores as a function of the total score. It is a plot of the function $r \mapsto E_{ri}$. It can be used to inspect item fit visually when plotted together with empirical item means. These can be plotted for each value of the total score or for each value of the grouped total score.

Empirical item means and confidence intervals The empirical expected item score can be calculated from data. Let, for $r = 0,\ldots,m_.$, n_r denote the number of respondents with $R_v = r$ and let, for $g = 1,\ldots,G$, m_g denote the number of respondents with $S_v = g$. The average item scores at score group level g is

$$A_{gi} = \frac{1}{m_g} \sum_{v:S_v=g} x_{vi}, \tag{7}$$

and the variance is

$$V_{gi} = \frac{1}{n_g - 1} \sum_{v:S_v = g} (x_{vi} - A_{gi})^2. \tag{8}$$

Finally, confidence intervals computed at the 95% confidence level are

$$CI_{gi} = A_{gi} \pm 1.96 \frac{V_{gi}}{\sqrt{m_g}}. \tag{9}$$

Note that the formulas (7), (8), and (9), also, as a special case, describe averages and their confidence intervals for single score values.

Plotting the CICC The total score r is on the horizontal axis and the conditional expected item score E_{ri} on the vertical axis. The expected item score will always be 0 when the total score is 0 and m_i when the total score is $m_. = \sum_{i=1}^{k} M_i$. Between these two endpoints the curve will be monotone increasing. Let $\bar{r}_g = \frac{1}{m_g} \sum_{r \in I_g} n_r r$ denote weighted averages. Observed average item scores are plotted as $(\bar{r}_g, A_{gi})$ with error bars representing confidence intervals. This yields a visual illustration of item fit. Examples of CICCs can be seen in Figure 1 and Figure 2, which are based on two examples, which are presented in the following sections. When grouped total scores are used, the intervals $I_1, \ldots, I_G$ are illustrated on the plot as shaded areas.

Motivating example 1: The abbreviated mental test score

The abbreviated mental test score (AMTS; Hodkinson, 1972) is a test for rapidly assessing elderly patients for the possibility of dementia. It was first used in 1972, and is now sometimes also used to assess for mental confusion and other cognitive impairments. It consists of ten items and is scored by awarding one point for each correct answer. Scores of six or less suggests mental impairment in patients. The data set is included in the `iarm` package:

```
library(iarm)
```

Initial item analysis is done using conditional item parameter estimation as implemented in the `eRm` package:

```
it.AMTS <- amts[,4:13]
it.AMTS <- na.omit(it.AMTS)
mod.AMTS <- RM(it.AMTS, sum0 = FALSE)
```

Item fit statistics are computed based on comparison of observed and expected item-restscore correlation (Christensen and Kreiner, 2013, section 5.4). When item fit is tested for ten items there is an elevated risk of type I errors due to multiple testing. For this reason the Benjamini-Hochberg correction is used to control the false discovery rate (FDR) associated with the multiple tests. The R code for calculating the item fit statistics can be seen in appendix A. The item fit statistics use a recent R package, `iarm`, computing item fit statistics with known asymptotic distributions (Müller, 2020a).

Item	Wording	Obs.	Exp.	P*
age	What is your age?	0.8376	0.7511	0.1295
time	What is the time to the nearest hour?	0.7087	0.7366	0.7470
address	Give the patient an address, and ask him or her to repeat it at the end of the test.	0.6157	0.6874	0.5212
name	What is the name of the office or doctor you are seeing today?	0.7973	0.7511	0.5455
year	What is the current year?	0.8189	0.7350	0.1295
dob	What is your date of birth?	0.7884	0.7812	0.9115
month	Name the actual month.	0.8842	0.7303	0.0004
firstww	In what year did World War 2 end?	0.6170	0.7409	0.2235
monarch	Name the current President/Prime Minister.	0.7736	0.7342	0.5455
countbac	Count backwards from 20 down to 1	0.5869	0.7303	0.1295

Table 1: Item wording and evaluation of item fit for the abbreviated mental test score (AMTS) items. *: corrected *p*-values using the Benjamini-Hochberg correction.

The results are shown in Table 1 and indicate, based on the small adjusted *p*-value, problematic fit of the item “month” (item 7). Visualization of the item fit can be done using the plotting function `CICCplot` that is part of the `RASCHplot` R-package. This package can be downloaded to your computer using the following R code:

```
install.packages("devtools")
devtools::install_github("ERRTG/RASCHplot")
library(RASCHplot)
```

and the plot can be generated using the `eRm` object `mod.AMTS` with the following code

```
CICCplot(model = mod.AMTS,
         which.item = 7,
         lower.groups = c(0,3,6,8,9))
```

The plot generated by the code above is shown in Figure 1.

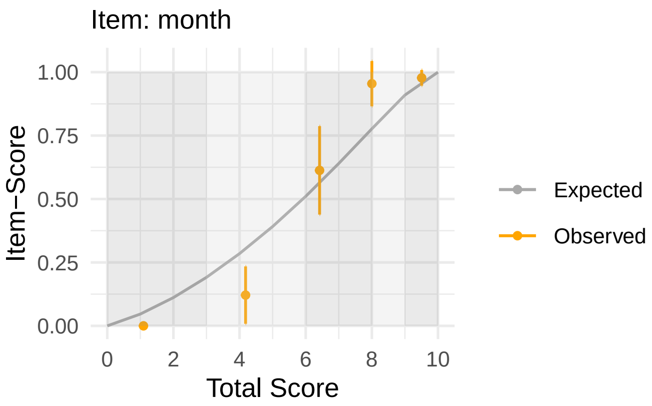

Figure 1: The conditional ICC plot for item 7 ('month') of the Abbreviated Mental Test Score (AMTS).

From the figure it is seen that the misfit seems to be due to over discrimination. The R code above illustrates the basic use of `CICCplot`, where only one argument, `model`, needs to be specified to construct a figure. This `model` argument should be an object of class `Rm` or `eRm` returned from one of the estimation functions, `RM`, `RSM`, or `PCM`, in the `eRm` package. Such an object contains all the relevant information of the fitted model for further use. Additional optional arguments include `which.item` which is either an integer or a vector stating the item(s), for which a CICC plot should be constructed or the character string `all` for constructing CICC plots for all items in the data. The name of the variable containing the item responses can also be used. Item one (the first column in the input matrix of items) is the default. The vertical axis ranges from zero to $m_. = \sum_{i=1}^{k} m_i = 10$; `lower.groups` is a vector used for grouping the set of possible total scores into intervals, for which empirical expected item-scores are calculated and added to the plot. The vector contains the lower points of the intervals, into which the set of possible total scores should be divided. If zero is not included in the vector, it will be added automatically. In the code shown above `lower.groups` is defined as the vector `c(0,3,6,8,9)` and this means that the score groups are $\{0,1,2\}$, $\{3,4,5\}$, $\{6,7\}$, $\{8\}$, and $\{9,10\}$. These are visualized as shaded areas in Figure 1. The default is `lower.groups = "all"` which results in empirical expected item-scores for every possible total score in the figure. No method for automatically choosing intervals is implemented. When multiple items are provided for `which.item`, the intervals provided for lower.groups are used for all items. If different intervals are preferred, the CICCplot function should be run separately for each item; error.bar is an indicator describing whether or not error bars on the empirical CICC should be added to the figure. The default value is TRUE; grid.items is a logical argument. If grid.items = TRUE, the items selected by which.item will be arranged in grids with at most four plots per grid. The default value is FALSE. For a description of the additional settings which are possible to adjust in the figures i.e. plot-title and axis-labels, please see the help page for the function by typing ?CICCplot in R. Finally, we note that the CICCplot function creates a ggplot graphic. This means that additional layers can be added when CICCplot is followed by '+' and functions from the ggplot2 package. This is, however, not possible when grid.items = TRUE.

Motivating example 2: The hospital anxiety and depression scale

The hospital anxiety and depression scale (HADS; Zigmond and Snaith, 1983) aims to measure symptoms of anxiety and depression. It was designed as a brief instrument used to assess symptoms of anxiety and depression and contains 14 items often scored as two seven-item subscales: “depression” (even numbered items) and “anxiety” (odd numbered items). In what follows we use data reported by Pallant and Tennant (2007). The R code below reads the data and fits the polytomous Rasch model

```
HADS <- "https://raw.githubusercontent.com/ERRTG/ERRTG.github
    .io/master/HADS.csv"
it.HADS <- read.csv(HADS)
it.HADS.A <- it.HADS[,c(1,3,5,7,9,11,13)]
mod.HADS <- PCM(it.HADS.A, sum0 = FALSE)
```

Again, the item fit statistics based on comparison of observed and expected item-restscore correlation are reported (along with the false discovery rate). Details about computation of these are reported in Appendix A.

The results are reported in Table 2 and show evidence of misfit for the items AHADS7 and AHADS13 based on the small adjusted *p*-values for the two items.

Visualization of item fit for these two items can be done using this R code:

```
CICCplot(mod.HADS,
         which.item = c(4,7),
         lower.groups = c(0,3,6,9,13,18),
         grid.items = TRUE)
```

Item	Wording	Obs.	Exp.	P*
AHADS1	I feel tense or "wound up"	0.6422	0.5999	0.3357
AHADS3	I get a sort of frightened feeling as if something awful is about to happen	0.6191	0.6072	0.7440
AHADS5	Worrying thoughts go through my mind	0.6792	0.5983	0.0406
AHADS7	I can sit at ease and feel relaxed	0.4054	0.5747	0.0085
AHADS9	I get a sort of frightened feeling like "butterflies" in the stomach	0.6617	0.5787	0.0406
AHADS11	I feel restless as I have to be on the move	0.4987	0.5953	0.0406
AHADS13	I get sudden feelings of panic	0.6998	0.5901	0.0085

Table 2: Item wording and evaluation of item fit for the hospital anxiety and depression scale (HADS) items. *: corrected *p*-values using the Benjamini-Hochberg correction

Here a model object is passed from the `PCM()`-function in the `eRm`-package and the `ggarrange()` function from the `ggpubr` package is used to arrange the two ggplots. The code generates the plot shown in Figure 2.

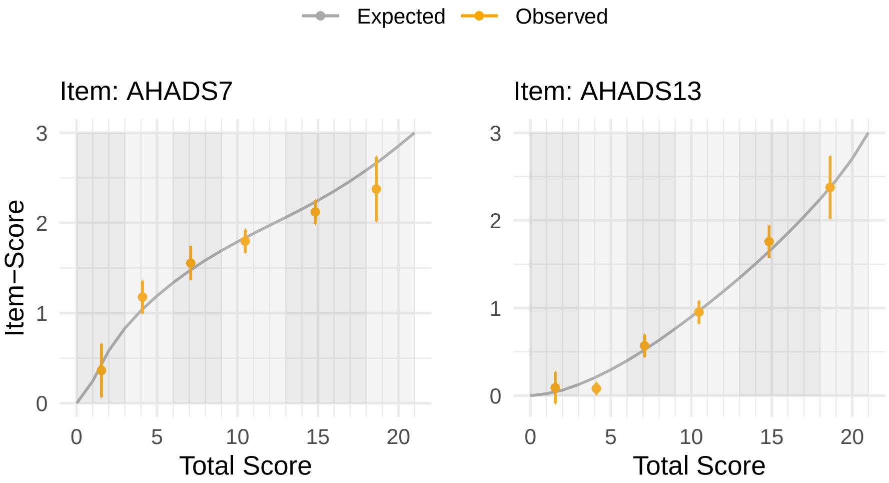

Figure 2: The conditional ICC plot for items 7 and 13 of the Hospital Anxiety and Depression Scale (HADS).

Again, lower.groups define the score groups that are shown as shaded areas in both plots. The figure shows no clear trend of misfit for the item AHADS13, except for a single score group where the empirical item mean differs significantly from model expectations. For the item AHADS7 five of the six score groups follow a pattern of *under-discrimination*, where observed items scores are too high for low score values and too low for high score values. This matches the statistically significant difference between item-restscore correlations reported in Table 2.

Discussion

The eRm package is the most widely used implementation of Rasch models in R and it has been extended in several ways. Recently Alexandrowicz (2022) proposed an extension of the plotGOF() function and, in a similar vein, here we have extended the graphical capabilities of eRm as regards item fit evaluation. The eRm package provides the function plotICC that plots the item characteristic curves. Using the argument empICC an empirical ICC can be plotted. This plot is implemented for dichotomous Rasch models and plots the emprical item mean in each score group at the single location of the latent variable associated with the score value. The argument empCI can be used to add confidence intervals to the plot. The R function CICCplot described here extends this plot in two ways: (i) by providing a plot for polytomous items, and (ii) by making it possible to imposed a grouping on the total score in the plot. This is helpful in choosing the plot that best communicates the level of fit or misfit.

For dichotomous items the plot generated using CICCplot when no grouping is imposed is very similar, but not identical, to the plot included in eRm. These plots will show the same trend, but one shows the expected item score as a function of the latent variable and the other as as function of the total score. Hence, the specific shape of the curve can differ, but the relation between the empirical and model based curves will be the same. Another difference between the two plots is that the plot proposed here does not present empirical confidence intervals for score groups where all observed responses are zero.

Calculations of the expected conditional item scores in CICCplot are based on the estimated parameters calculated by the eRm package. The eRm package is able to handle incomplete data when estimating dichotomous and polytomous Rasch models using RM(), RSM(), or PCM(), and therefore it is also possible to draw Conditional ICCs for models fitted to incomplete data. The empirical conditional item scores and corresponding confidence intervals are computed based on complete cases. The generalization to an R function that can handle incomplete responses is an important topic for future research.

The conditional ICC is implemented using the recursive formula (4) for the γ polynomials and therefore, the output is somewhat computationally expensive. In order to speed up the process we have used the function operator `memoise()` from the `memoise` package. It *memoises* a function, meaning that the function will remember previous inputs and return cached results. The γ-function always returns the same output if the same arguments are passed. In other words, it is a *pure* function, that is, it does not depend on any data change during a call, it only depends on its input arguments. Therefore, memoising the γ-function should not cause a problem. It should be noted that memoisation is a tradeoff of memory versus speed: a memoised function run faster, but because it stores all of the previous inputs and outputs, it uses more memory.

The use of `ggplot` requires specific technical knowledge, which is a potential weakness of this implementation.

We stress that the visualization is intended to go hand in hand with robust statistical indices of item fit and note that the choice of intervals when grouping the total score is important. We recommend that groups be chosen to have a size where the error bars are not too wide or too narrow to convey information about item fit. The visualizations are useful as a visual supplement to robust statistical tests of items in order to understand the nature and magnitude of the misfit. Thus, for the single misfitting AMTS item identified by the item fit statistics in Table 1 the misfit was seen to be substantial in the plot in Figure 1. Similarly, for one of the misfitting items (`AHADS7`) in Table 2 a tendency towards under-discrimination was seen. However, for the other misfitting HADS item (`AHADS13`) the source of misfit was a single score group. This illustrates how visualization helps understand evidence from Rasch analysis beyond interpretation of significance.

Appendix A

R code for calculating the item fit statistics If the necessary R packages are not downloaded to your computer use this R code:

```
install.packages(c("eRm", "iarm", "ggpubr"))
```

The evaluation of item fit in the AMTS can then be done using this R code:

```
library(eRm)
library(iarm)
```

```
it.AMTS <- amts[,4:13]
mod.AMTS <- RM(it.AMTS, sum0 = FALSE)
item_restscore(mod.AMTS)
```

For the HADS data evaluation of item fit can then be done using this R code:

```
item_restscore(mod.HADS)
```

References

Alexandrowicz, R. W. (2022). GMX: Extended Graphical Model Checks. *Psychological Test and Assessment Modeling*, 64:215–225.

Andersen, E. B. (1995). *Polytomous Rasch Models and their Estimation*, pages 271–291. Springer New York, New York, NY.

Aryadoust, V., Ng, L. Y., and Sayama, H. (2021). A comprehensive review of rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 38(1):6–40.

Betensky, R. A. (2019). The p-Value Requires Context, Not a Threshold. *The American Statistician*, 73(sup1):115–117.

Christensen, K. B. and Kreiner, S. (2013). Item fit statistics. In *Rasch Models in Health*, chapter 5, pages 83–104. John Wiley & Sons, Inc.

Courtney, M. G. R., Chang, K. C. T., Mei, B., Meissel, K., Rowe, L. I., and Issayeva, L. B. (2021). autopsych: An r shiny tool for the reproducible rasch analysis, differential item functioning, equating, and examination of group effects. *PLOS ONE*, 16:e0257682.

Debelak, R., Stobl, C., and Zeigenfuse, M. D. (2022). *An Introduction to the Rasch Model with Examples in R*. Chapman and Hall/CRC.

Glas, C. A. W. and Verhelst, N. D. (1995a). Testing the Rasch model. In Fischer, G. H. and Molenaar, I. W., editors, *Rasch Models: Foundations, Recent Developments, and Applications*, pages 69–95. Springer New York, New York, NY.

Glas, C. A. W. and Verhelst, N. D. (1995b). Tests of fit for polytomous Rasch models. In Fischer, G. H. and Molenaar, I. W., editors, *Rasch Models: Foundations, Recent Developments, and Applications*, pages 325–352. Springer New York, New York, NY.

Hagquist, C., Bruce, M., and Gustavsson, J. P. (2009). Using the rasch model in nursing research: an introduction and illustrative example. *International journal of nursing studies*, 46:380–93.

Hodkinson, H. M. (1972). Evaluation of a Mental Test Score for Assessment of Mental Impairmet in the Elderly. *Age and Ageing*, 1(4):233–238.

Horton, M., Marais, I., and Christensen, K. B. (2013). Dimensionality. In *Rasch Models in Health*, pages 137–158. John Wiley Sons, Inc.

Komboz, B., Strobl, C., and Zeileis, A. (2018). Tree-based global model tests for polytomous rasch models. *Educational and Psychological Measurement*, 78.

Kreiner, S. and Christensen, K. B. (2013a). Overall tests of the rasch model. In *Rasch Models in Health*, pages 105–110. John Wiley Sons, Inc.

Kreiner, S. and Christensen, K. B. (2013b). Two tests of local independence. In *Rasch Models in Health*, pages 131–136. John Wiley Sons, Inc.

Mair, P. and Hatzinger, R. (2007). Extended Rasch modeling: The eRm Package for the application of IRT models in R. *Journal of Statistical Software*, 20(9).

Müller, M. (2020a). *iarm: Item Analysis in Rasch Models*. R package version 0.4.2.

Müller, M. (2020b). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications*, 7(1).

Pallant, J. F. and Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, 46:1–18.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press.

Tennant, A. and Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheumatism*, 57:1358–1362.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133.

Zigmond, A. S. and Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, 67:361–370.