Psychological Test and Assessment Modeling, Volume 53, 2011 (3), 283-294

# Performance of the bootstrap Rasch model test under violations of non-intersecting item response functions

*Moritz Heene*[1], *Clemens Draxler*[2], *Matthias Ziegler*[3] *&*
*Markus Bühner*[4]

## Abstract

The Rasch model is known in particular for its properties of parameter separability and specific objectivity. The extent to which this property is attained depends on the magnitude of the discrepancy between the data and the model. The use of reliable model fit tests which can detect model violations is therefore essential before a psychological test is used and inferences based on the requirements of the Rasch model are drawn. This paper provides a critical analysis of the performance of the parametric bootstrap model test (von Davier, 1997) in the presence of non-parallel item response functions as violations of a basic requirement of the dichotomous Rasch model. Based on results from simulated data it is shown that in general the bootstrap test leads too often to failures to reject non-fitting data.

Key words: Rasch model; Rasch model test; Rasch model violation; Bootstrap

[1] *Correspondence concerning this article should be addressed to:* Dr. Moritz Heene, Karl Franzens University, Department Psychology, Unit Psychological Diagnostics, Maiffredygasse 12b, A-8010 Graz, Austria; email: moritz.heene@uni-graz.at

[2] Ludwig Maximilian University, Munich, Germany

[3] Humboldt University, Berlin, Germany

[4] Karl-Franzens University, Graz, Austria

## Introduction

The Rasch model (Rasch, 1960) is frequently used in educational and psychological testing (e.g., Bond & Fox, 2007; De Boeck & Wilson, 2004). It is known for its characteristics of scaling items and persons on a joint unidimensional interval scale and the separability of item parameters and person parameters, within a conditional maximum likelihood approach: by conditioning on the sufficient statistics for the person parameters, the raw score *r*, item parameters can be estimated without the involvement of the person parameters. That is, because the latter disappear from the conditional likelihood equation (Rasch, 1960, 1977). This can briefly be indicated as follows. Consider *n* persons indexed by $v = 1,...,n$ responding to *k* binary items indexed by $i = 1,...,k$. The binary responses of the *n* persons to the *k* items are arranged in an $n \times k$ matrix. Let the response of person *v* to item *i* be modelled by the Bernoulli variable $X_{vi}$ which can take on the value $x_{vi} \in \{0,1\}$. The discrete probability distribution of $X_{vi}$ is determined by

$$P(X_{vi} = x_{vi}) = \frac{\exp\left[x_{vi}(\theta_v - \beta_i)\right]}{1 + \exp(\theta_v - \beta_i)}, \tag{1}$$

where $\theta_v$ is the persons parameter of person *v* and $\beta_i$ the item parameter of item *i*.

Let $r_v = \sum_{i=1}^{k} x_{vi}$, the score of person *v*, be defined. Then $r_v$ is a sufficient statistic for $\theta_v$. Given $r_v$, it follows that the conditional probability of the *k* responses of person *v*, the response pattern $\boldsymbol{x}_v$, is independent of $\theta_v$. It depends on the *k* item parameters $\beta_1,...,\beta_k$ only. That is,

$$P(\boldsymbol{x}_v | r_v) = \frac{\exp\left(\sum_{i=1}^{k} -x_{vi}\beta_i\right)}{\gamma_{r_v}}, \tag{2}$$

with $\gamma_{r_v}$ as the well-known elementary symmetric function of order $r_v$ of the item parameters $\exp(-\beta_1),...,\exp(-\beta_k)$. Multiplying (2) over all observed response patterns $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ (over all persons) gives the conditional likelihood function. Maximizing the conditional likelihood function yields a consistent estimator for the item parameters. This is known as the CML method. As is well-known from the theory of exponential families, parameter estimation essentially reduces to equating the observed values of the sufficient statistics to their expected values (under the model).

It can be shown that the Rasch model implies monotonically increasing and non-intersecting item response functions (IRFs) (Fischer, 1995; Karabatsos, 2001). The extent to which this property is attained depends on the magnitude of the discrepancy between the data and the model. Many of the statistical tests for the Rasch model are based on the asymptotic distribution of the conditional maximum likelihood estimators from which various $\chi^2$ distributed test statistics can be derived. Glas and Verhelst (1995) give a review of these tests. In recent years exact tests have also been proposed (Chen & Small, 2005, Verhelst, 2008; Ponocny, 2001). Exact tests are based on the uniform distribution of the $n \times k$ matrix with given margins. By conditioning on all margins of the matrix the joint conditional probability of all entries (all responses of the n persons to the k items) is independent of both the person and item parameters. It is equal to the reciprocal of the number of matrices consistent with the given, fixed margins. Unfortunately there do not exist satisfactory combinatorial or analytical methods to determine this number, so that Monte Carlo methods have been applied to approximate the exact distribution of any statistic by drawing random samples from the uniform distribution.

Other statistics frequently applied but not based on the CML approach are response residual-based statistics of model fit (Glas & Verhelst, 1995; Rost & von Davier, 1994; Smith, Schumacker, & Bush, 1998; Suarez-Falcon & Glas, 2003; Wang & Chen, 2005; Wright, 1994; Wright & Stone, 1979). In particular, the widely-used mean square infit and outfit statistics (Wright & Stone, 1979) have been criticized, because their sensitivity to detect model misfit is affected by sample size and item variance. Therefore, setting a common cutoff-value that applies to every item is inappropriate (Karabatsos, 2000; Smith, 1994, 1996; Smith et al., 1998; Wang & Chen, 2005). Moreover, as Karabatsos (2001) has shown, true residuals between the observed and expected item responses may be underestimated: if the data contain noise in the form of systematic measurement disturbances, the estimated parallel IRFs also contain noise, which will be absorbed to an unknown extent due to the use of already minimized residuals in the context of the maximum-likelihood algorithm.

Two test statistics which are of particular interest in this paper assess the divergence between the frequencies of the observed response patterns and their expected frequencies under a certain item response model. The first one is the Pearson $\chi^2$-statistic which aggregates the differences between the observed and expected frequencies defined by a fitted Rasch model over all possible response patterns, defined as follows

$$\chi^2 = \sum_{i=1}^{m^k} \frac{(X_i - E_i)^2}{E_i} \tag{3}$$

whereby $X_i$: observed response pattern,

$E_i$: expected response pattern.

The second one is the $\chi^2$-statistic by Read and Cressie (1988), *CR*(2/3), defined as

$$CR\left(\frac{2}{3}\right) = 1.8 \sum_{i=1}^{m^k} X_i \left[\left(\frac{X_i}{E_i}\right)^{\frac{2}{3}} - 1\right] \tag{4}$$

However, the particular problem with both statistics is that the number of possible response patterns of *k* items with *m* response categories increases exponentially with increasing number of items. As a consequence the expected frequencies become smaller in general. In most cases of empirical research by far not all possible response patterns are observed, and the expected frequencies are therefore very low. In these cases, the test statistic is no longer $\chi^2$-distributed with degrees of freedom equal to the number of possible response patterns minus 1 and consequential statistical inferences are invalid (Koehler & Larntz, 1980).

## The bootstrap test for the Rasch model

In order to avoid the problem that the asymptotic distribution of $\chi^2$-statistics does not hold, von Davier (1997) proposed a parametric bootstrap method to generate the unknown sampling distribution of a test statistic $T$ as follows. In the first step, person and item parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ are estimated and the test statistic $T_{obs}$ for the given data set is calculated. In the second step, a new data-set is generated according to the model equation and $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$. Then, model parameters are estimated for the newly simulated data set and the test statistic, $T$, is calculated. This step is repeated B times to generate the sampling distribution for $T$. Usually, the number $B$ of bootstrap samples has to be large, that is $B \geq 1000$ (Efron & Tibshirani, 1996). The $p$-value of the observed test statistic $T_{obs}$ is then estimated using $\#\left\{T \geq T_{obs}\right\} / B$.

Von Davier (1997) investigated the performance of this bootstrap test in an extensive Monte Carlo simulation study, systematically varying response format, sample size, number of items, and number of responses generating latent classes. He concluded that both the Pearson chi-square and Cressie-Read statistic performed well in terms of detection rates of the true mixture distribution model based on the conventional alpha-level of five percent. Meanwhile, the bootstrap test is used not only to determine the true number of latent classes in the context of mixture distribution models, but also to compare uni- and multidimensional models (Carstensen & Rost, 2003; Rost & Carstensen, 2002). Or in general as a global Rasch model test, irrespective of any alternative model comparisons, to test the null hypothesis that the observed data matrix has been generated under the Rasch model (Rizopoulos, 2008), that is, that the Rasch model can reproduce the observed pattern frequencies.

Quite recently, Tollenaar and Mooijart (2003) investigated Type I error rates and the power of different bootstrapped fit statistics with regard to first-order Markov models for categorical data. They concluded that bootstrapping the usual fit statistics leads to Type I error rates being larger than the stated alpha level and that the parametric bootstrap has low power in situations with small sample sizes.

Despite its application as a global model test, it should nevertheless be noted that the original formulation of the bootstrap test refers to the hypothesis of population homogeneity (Langeheine, Pannekoek & van de Pol, 1996; von Davier, 1997). The null hypothesis of both chi-square test statistics does, however, not directly refer to a hypothesis of population homogeneity but is, in fact, quite general. The Pearson $\chi^2$- as well as the CR(2/3)-statistic tests the null hypothesis that the frequency distribution of the observed response patterns is consistent with the theoretical distribution of the expected response patterns under an item response model, for example, the Rasch model or a mixed Rasch model assuming more than one class. In particular, the null hypothesis of a one-class solution states that the observed data have been generated under the Rasch model with parameter values as the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$. What is then obvious is that the theoretical distribution of the expected response patterns is not independent from all other model assumptions, for example, non-intersecting item response functions. Consequently, the bootstrap model test cannot be regarded as a test of population homogeneity solely and it is thus reasonable to investigate its power with respect to the violation of non-intersecting item response functions. However, no study to date has investigated the performance of the bootstrap test with regard to model violations of non-intersecting IRFs. The purpose of this paper is therefore to investigate the performance of the bootstrap test to detect this kind of model violation.

## Method

### Data simulations

WinGen 2.5.4.414 (Han, 2007) was used to simulate 2700 data matrices according to a fully-crossed 3 × 3 × 3 design. The design included three levels of test length (10, 20 and 30 items; see von Davier, 1997 for a similar test length condition), three levels of sample size (150, 500 and 2500) and three levels of intersecting IRFs, low, medium, and high degree. Thus, the two-parametric logistic model (Birnbaum, 1968) was used to simulate data with crossing IRFs. Following Suarez-Falcon and Glas (2003) for each discrimination parameter a log normal distribution with common mean, but different standard deviations was specified: LN(0, 0.12), LN(0, 0.25), and LN(0, 0.50), respectively. The person and item parameters have been drawn from a standard normal distribution. Note that although tests of ten to thirty items would be seen as short in the context of educational and psychological measurement, the experimental design followed that of von Davier (1997), who used similar scale lengths. The normality of the person parameters as well as parameter recovery was verified for a few randomly chosen data sets from different simulation conditions by using the Shapiro-Wilk test as implemented in R and the R package “Latent trait models” (Rizopoulos, 2008), respectively.

For the simulations with $N = 500$ and $N = 2500$ item parameters of the corresponding conditions of the data matrices with $N = 150$ were used to imitate replication studies of the same items in independent samples. Finally, for each of the resulting 27 conditions, 100 replications were generated[5].

The bootstrap test was performed with the maximum possible number of 999 bootstrap replications in Winmira 2001 (von Davier, 2001) for each of the 2700 data files.

[5] The simulation code for generating data sets with defined model violations and Rasch-fitting data used in this study can be obtained from the first author.

## Results

To compute the general detection rates of model violations and failure-to-reject rates the recommendations of von Davier (1997) were applied to accept the Rasch model if neither the Pearson $\chi^2$ statistic nor the Cressie-Read $\chi^2$ statistic rejects the model on the conventional 5 per cent alpha level. Table 1 shows the overall detection rates of model misfit separated by sample size, test length, and degree of model violation.

**Table 1:**
Bootstrap Rasch model rejection rates for different sample sizes and different degrees of model violations

| Item Number | Sample Size | Degree of model violation | | |
|---|---|---|---|---|
| | | **weak** | **moderate** | **strong** |
| 10 | 150 | .07 | .02 | .12 |
| | 500 | .09 | .04 | .48 |
| | 2500 | .17 | .13 | 1 |
| 20 | 150 | .05 | .05 | .05 |
| | 500 | .07 | .05 | .03 |
| | 2500 | .06 | .07 | .06 |
| 30 | 150 | .03 | .07 | 0 |
| | 500 | .06 | .04 | .02 |
| | 2500 | .08 | .04 | 0 |

Note. Rasch model rejection if *p*-value for Pearson-$\chi^2$ and Cressie-Read$\chi^2$ < .05.

The table shows that only under the condition of test length with 10 items did detection rates increase as the sample size increased, and therefore statistical power of the bootstrap model test increased. Moreover, only under the condition of 10 items a relationship between degree of model violation and rejection rates in all sample size conditions can be observed. Nevertheless, except for the condition of sample size $N = 2500$, item number $k = 10$, and strong model violation, the rejection rates were alarmingly low. Recall that the number of possible response patterns under the more realistic condition with 20 and 30 dichotomous items is $2^{20}$ and $2^{30}$, respectively, resulting in extremely sparse data. Consequently, even with a sample size of 2500, the sampling distribution of the Pearson $\chi^2$- and the CR(2/3)-statistic is not well approximated, yielding very low rejection rates.

Of course, the presented analyses which were carried out solely with non-fitting data, teaches us nothing about the performance of the bootstrapped fit statistics with respect to model fit comparisons between data matrices generated by the Rasch model and those generated by the 2-PL model. The results presented above with their surprisingly low detection rates must arouse suspicion that the bootstrapped test statistics might also have low power to distinguish between 2-PL and Rasch-fitting data matrices. However, the power rates under the various experimental conditions are still unknown and should be investigated. Consequently, additional data fitting the Rasch model were simulated under a fully crossed 3 × 3 design. In accordance with the previous simulation conditions, test length was varied (10, 20 and 30 items) as well as sample size (150, 500 and 2500). Person and item parameters were both drawn from a standard normal distribution. For the simulations with $N = 500$ and $N = 2500$, item parameters analogous to those of the data sets with $N = 150$ were used to imitate replications of the same items in independent samples. Finally, 100 replications per condition were used resulting in 900 data sets. Again, the bootstrap test was performed with 999 replications per data set and, following von Davier (1997), the model was accepted if neither the Pearson $\chi^2$ nor the Cressie-Read-statistic rejected the Rasch model at the five percent level.

With the 2700 2-PL-generated data sets and the 900 Rasch-fitting data sets, Receiver Operator Characteristic (ROC) analyses were used to evaluate the bootstrap test in its ability to detect 2-PL data. Hence, for different sample size and test length conditions, the data-generating model (coded 1: Rasch model and 2: 2-PL model) served as the state variable and model fit as the test variable (coded 0: data rejected as Rasch-fitting; coded 1: data accepted as Rasch-fitting).

Under each experimental condition, the ROC analysis of the bootstrap test estimated $H(c)$, the probability that a 2-PL data set will be rejected $(c = 0)$. Hence, $H(c)$ denotes the sensitivity – the probability that the bootstrap test correctly identifies a 2-PL model, that is, the "hit rate" using $c = 0$ as the critical value. Thus, $1 - H(c)$ represents the "miss-rate", that is, the probability that the bootstrap test incorrectly identifies a 2-PL-fitting data set as Rasch-fitting. The ROC analysis also estimated $F(c)$, the probability that a Rasch-fitting data set has a value equal to $c = 0$ which refers to the probability that a Rasch-fitting data set will be incorrectly identified as non-fitting, that is, the "false alarm" rate. Consequently, $1 - F(c)$ is the specificity, that is, the probability that the bootstrap test correctly classifies a Rasch-generated data set as Rasch-fitting.

Both $F(c)$ and $H(c)$ define a two-dimensional graph, with $F(c)$ on the x-axis and $H(c)$ on the *y*-axis. The coordinate $\left\{\hat{F}(c), \hat{H}(c)\right\}$ contains the estimates of the false alarm and the hit rate, conditional on the value of c. Therefore, the ROC curve is represented by a line connecting the set of coordinates over all possible values of *c*. Thus, a model test with high sensitivity and specificity would yield a ROC which would curve close to the upper-left corner of the graph, whereas a completely random guess would result in a diagonal line from the left bottom to the top right corner, the line of no-discrimination. Therefore, the sensitivity and specificity are jointly represented by the area a under the ROC curve with $a \in \{0,1\}$, whereby a perfect classification would be represented by an area of one, a completely random guess by an area of 0.5.

Table 2 shows the resulting areas under the curve obtained for different experimental conditions.

**Table 2:**
Receiver-operator analysis table (area under the curve) for various item- and sample-size-conditions

| Test Length | Sample Size | | |
|---|---|---|---|
| | 150 | 500 | 2500 |
| 10 | .480 | .582* | .672** |
| 20 | .470 | .470 | .502 |
| 30 | .497 | .480 | .455 |

In general, and in accordance with the expectation following from the already low rejection rates with 2-PL data only, the bootstrap tests performed poorly and did not distinguish sufficiently between 2-PL- and Rasch-generated data. Under the conditions of short test length (10 items) and large sample sizes (500, 2500) only the null hypothesis (i.e., that the true area under the curve equals .50) had to be rejected. However, the related areas under the curve were far from being either statistically or practically satisfying. For all remaining conditions, the probability that the bootstrap test would distinguish between 2-PL model and Rasch model data was not significantly different from random guessing.

## Conclusion and discussion

This simulation study indicates that there is much room for improvement within the bootstrap model test. As for 2-PL data only, failure-to-reject rates were extremely high in almost all of the cases leading to the conclusion that – except for short test lengths, strong model violations, and large sample size – the bootstrap test performs even worse than pure chance. This implies that for the specified conditions, the bootstrap test leads to over-optimistic decisions about the fit of the data to the dichotomous Rasch model.

Consequently, the ROC analyses revealed that the bootstrap test performed very poorly in distinguishing between 2-PL and Rasch-fitting data matrices. Although the bootstrap test distinguished between those models statistically significant under the condition of ten items and sample sizes of $N$ =500 and $N$ = 2500, the related areas under the curve indicated a very low discrimination power to distinguish between both models.

It appears that the bootstrap test suffers from a severe statistical power problem, thereby failing too frequently to reject a non-fitting model. A possible explanation for the bad performance given in this paper is an inappropriate choice of the statistic. Theoretically a statistic referring to all possible response patterns must have power against model violations as simulated in this study. Practically the great number of possible response patterns obtained already for a small number of items has the effect that the expected frequencies of the response patterns are generally very low and thus barely distinguishable between the tested model and the alternative, even for the strongest model deviations considered here. Consequently the power of the test must be low. Only for very small item numbers (for instance 10 as in this study) and large sample sizes the performance gets better so that power rates reach practically acceptable values. Thus, the bootstrap test in its present form seems to have limited use in detecting violations of the central requirement of non-intersecting IRFs and it is therefore difficult to call it a test of the Rasch model. One might argue that the bootstrap test was originally suggested to compare models with increasing number of latent classes and not to test the central assumption of the Rasch model of non-intersecting IRFs. But note that the hypothesis of both $\chi^2$-statistics about the reproducibility of the observed response patterns under a certain model is a general hypothesis about the differences between observed and expected response pattern frequencies and is therefore by no means restricted to conditions where only the true number of latent classes is crucial. As mentioned in the introduction, this is exactly the reason why the bootstrap test is used in the context of comparing the data-model-fit of uni- and multidimensional Rasch models (Carstensen & Rost, 2003; Rost & Carstensen, 2002) or the data-model-fit without any reference to alternative models (Rizopoulos, 2008). Given the results of this study one can therefore raise doubts whether the bootstrap model test is – in its present form – really a test of the Rasch model. When used as a global model test we strongly recommend at least complementing the bootstrap test procedure by using the likelihood ratio test by Andersen (1973) as well as using powerful tests of item fit as suggested by Ponocny (2001), both implemented in the R package “Extended Rasch modeling” (Mair & Hatzinger, 2007).

Finally, some limitations of this study must be noted and suggestions for additional future research are given. In particular, this study focused on various degrees of slope variation for different test length and sample sizes, but did not evaluate effects of multidimensionality or violations of the local stochastic independence assumption. It is known that such model violations can cause crossing IRFs (Hoskens & De-Boeck, 1997; Masters, 1988; Tuerlinckx & De Boeck, 2001a, 2001b; Yen, 1993) but are too often ignored in the application of the two-parametric logistic model (Andrich, 2004; Lumsden, 1978). Therefore, future research should evaluate the performance of the bootstrap test under these conditions because it cannot be concluded from the present results that the bootstrap test is insensitive to intersecting IRFs due to such measurement disturbances.

## References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-40.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1; Supplement:I-7).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-545). Reading, Massachusetts: Addison-Wesley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah: Erlbaum.

Carstensen, C., & Rost, J. (2003). *MULTIRA – A program system for multidimensional Rasch models* (Version 1.66). Kiel: Institute for Science Education.

Chen, Y. G., & Small, D. (2005). Exact tests for the Rasch model via sequential importance sampling. *Psychometrika, 70*(1), 11-30.

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models*. New York: Springer.

Efron, B., & Tibshirani, R. (1996). Computer-intensive statistical methods. In *Wiley series in probability and statistics; advances in biometry* (pp. 131-147): John Wiley and Sons, Inc.; John Wiley and Sons Ltd.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 15-38 ). Berlin: Springer.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models – Foundations, recent developments, and applications* (pp. 69-95). New York: Springer.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Hoskens, M., & De-Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2*(3), 261-277.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*(2), 152-176.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*(4), 389-423.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298.

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association, 75,* 336-344.

Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping Goodness of Fit Measures in Categorical Data Analysis. *Sociological Methods & Research, 24,* 492-516.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31*(1), 19-26.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15-29.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 29*(9), 1-20.

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika, 66*(3), 437-459.

Ponocny, I., & Ponocny-Seliger, E. (1999). T-Rasch: Non-parametric Rasch analysis: Assessment Systems Corporation.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. New York: Springer.

Rizopoulos, D. (2008). ltm: Latent trait models under IRT (Version 0.8-6). Vienna: R Foundation for Statistical Computing.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*(1), 42-56.

Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement, 18,* 171-182.

Smith, R. M. (1994). Detecting item bias in the Rasch rating scale mode. *Educational and Psychological Measurement, 54,* 886-896.

Smith, R. M. (1996). A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement, 56,* 403-418.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2,* 66-78.

Suarez-Falcon, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 56*(1), 127-143.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology, 56*(2), 271-288.

Tuerlinckx, F., & De Boeck, P. (2001a). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*(2), 181-195.

Tuerlinckx, F., & De Boeck, P. (2001b). Non-modeled item interactions lead to distorted discrimination parameters: A case study. *Methods of Psychological Research Online, 6*(2), 159-174.

Verhelst, N. D. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika, 73,* 705-728.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research, 2*(2), 29-48. Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/

von Davier, M. (2001). Winmira 2001 (Version 1.45). Kiel: IPN – institute for science education.

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational & Psychological Measurement, 65*(3), 376-404.

Wright, B. D. (1994). Reasonable mean-square fit values (Publication no. 8:3 p.370). Retrieved 26.1.2009, from Rasch Measurement Transactions: http://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Yen, W.-M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.