Psychological Test and Assessment Modeling, Volume 56, 2014 (2), 168-177

# Actual type-I- and type-II-risk of four different model tests of the Rasch model

*Karin Futschek*[1]

## Abstract

To prove whether the Rasch model holds for a certain achievement test, several model tests are available. Unfortunately the actual type-I- and type-II-risks of these tests are widely unknown. A simulation study was done to compare four model tests regarding their type-I- and type-II-risk: Andersen's Likelihood-Ratio test (Andersen, 1973), the *z*-Test of Fischer and Scheiblechner (1970) with estimation of Wald (1943), the Martin-Löf test (Martin-Löf, 1973) as well as the new approach of Kubinger, Rasch, and Yanagida (2009) who proposed a three-way nested analysis of variance. Different scenarios were simulated: No violation of the model, violation by one pair of DIF (differential item functioning) and violation of the model due to no one-dimensional but a multi-dimensional given ability. Depending on the scenario different model tests turned out to be advantageous. For all the simulated conditions it was shown, that the analysis of variance approach is an alternative to Andersen's Likelihood-Ratio test.

Keywords: type- I -and type-II-risk; Andersen's Likelihood-Ratio test; analysis of variance; *z*-Test of Fischer and Scheiblechner; Martin-Löf test

[1] *Correspondence concerning this article should be addressed to:* Karin Futschek, MSc, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Vienna Austria; email: karin.futschek@univie.ac.at

## Introduction

In the meantime, it is well-known, that only if the Rasch model holds for a psychological (achievement) test the sum of correct given answers is a sufficient estimator of the person's ability (Fischer, 1995). In order to test the Rasch model there are several model tests available, but their actual type-I- and type-II- risks are rarely examined. To help for choosing the best model test (and designing the needed sample size) when an achievement test shall be constructed, in this paper simulation studies were done to compare four model tests regarding their type-I-risk and their power: Andersen's Likelihood-Ratio test (Andersen, 1973), the *z*-Test of Fischer and Scheiblechner (1970) with estimation of Wald (1943) (first applied by Fischer und Ponocny-Seliger, 1998), the Martin-Löf test (Martin-Löf, 1973) as well as the approach of Kubinger, Rasch, and Yanagida (2009) who proposed to use a three-way nested analysis of variance. Except for the last one, all test statistics are only asymptotically $\chi^2$-distributed in case of the null-hypothesis, so that the type-I-risk might not hold; and there is no formula to calculate the type-II-risk, given the type-I-risk, the sample size and the effect size. The only way to determine the type-II-risk is via simulation. Therefore the model tests were compared for different kinds of model violations as well as for the case the null-hypothesis is true.

Alexandrowicz (2002) showed in a simulation study that the power of Andersen's Likelihood-Ratio test is more influenced by the sample size than it is by the kind of model violation when using the internal split criterion: "high vs. low score". Andersen's Likelihood-Ratio is constructed to test whether there applies some differential item functioning (DIF) regarding either an internal or an external criterion for splitting the sample of tested persons. It is a global model test. The *z*-Test of Fischer and Scheiblechner (1972) with the estimation of Wald proves for each item separately whether there is a DIF or not; while the Martin-Löf test (Martin-Löf, 1973) proves, whether two hypothesized subgroups of items measure the same ability (dimension). A simulation study of Verguts und De Boeck (2000) showed that the number of items affects the type-I-risk: Type-I-risk decreases with an increasing number of items (and small sample size of tested persons). For 24 items and 5000 persons the type-I-risk was 0%.

Kubinger, Rasch, and Yanagida (2009; see also Kubinger, Rasch, & Yanagida, 2011 as well as Rasch, Kubinger, & Yanagida, 2011) proposed a new method to test the Rasch model with regard to an external split criterion, with the purpose of calculating proper sample sizes for given type-I-risk, power and effect size. They suggested to use a three-way (nested) analysis of variance for mixed classification [i.e. (A $\succ$ **B**) x C]. A is a fixed factor and splits the data into two groups of tested persons. **B** represents the persons and is a random factor. It is nested in A because each person is assigned to only one group of A. (A $\succ$ **B**) is cross-classified with C, which represents the items. Given $H_0$: there is no interaction effect A x C, means specific objectivity holds and therefore the Rasch model. If the respectiv F-Test of the interaction term A x C is significant, the null hypothesis must be rejected because the data don't confim the Rasch model.

One problem of this approach is that the assumption of normal distribution is violated, because the data are dichotomous. Besides this, there is only a single observation in each cell. Hence it is necessary to test via simulation studies whether both these facts affect type-I- and type-II-risk. The authors showed that the actual type-I-risk is close to the nominal risk, given there is no main effect of A. Is there however a main effect of A, then the type-I-risk is artificial too high for the interaction effect A x C. In a second paper (Kubinger, Rasch, & Yanagida, 2011) the authors showed restrictively that the type-I-risk of the interaction effect AxC, given there is no main effect A, grows with increasing sample size and longer test length, as well as with a greater range of the item parameters. The type-I-risk just holds if the item parameters lie between -3 and 3 and are rather unimodally distributed, the person ability parameters are normally distributed with a standard deviation not bigger than 1.5, the number of items is not bigger than 100, and the sample size is not bigger than 300. Finally, the power of the F-test is greater if the model violation regards to items with average difficulty and if there are more items violating the model – and, of course, the larger the sample is.

## Method

The simulation was performed using the package "Extended Rasch Modeling" (eRm; Mair, Hatzinger, & Maier, 2011) of the statistic software R (R Development Core Team, 2012). To speed up the computing time the package "Simple Network of Workstations" (snow; Tierney, Rossini, Li, & Sevcikova, 2012) was used.

From all possible scenarios five representative ones were chosen, concerning their practical relevance, to determine the type-I-risk and the power of the four model tests: Andersen's Likelihood-Ratio test, the three-way analysis of variance approach by Kubinger, Rasch, and Yanagida, the *z*-test of Fischer and Scheiblechner and the Martin-Löf test. While for all variations of the five scenarios the number of items was constantly 20 and the significance level was always 5 %, the number of persons, the effect size and the kind of model violation varied. The numbers of persons were 100, 200, and 300. The scenarios used for the simulations were a) no model violation, b) model violation through one pair of differential item functioning (DIF), which varied between 0.75 and 3, which is one half of the standard deviation and two standard deviations of the person ability parameters, and c) model violation due to a multi-dimensional ability with a latent correlation of 0.5, which represents a realistic association between two latent dimensions measured by a possible achievement test. The following table (Table 1) shows all simulations.

Each simulation was repeated 10,000 times. Just the simulations of multidimensionality were repeated 1,000 times only.

The parameters of the simulation were determined according to practical relevant conditions. Therefore 20 items – a typical length of achievement tests – were chosen. The item parameters were set between -3 and +3. The person ability parameters were normally distributed with mean zero and standard deviation 1.5. As a consequence 2 % of all simulees would solve no item and 2 % solve all items. There are slightly more items with medium difficulty in order to better discriminate between people of average ability.

**Table 1:**
Simulation scenarios and applied model tests

| **simulation** | **number of persons** | | |
|---|---|---|---|
| | 100 | 200 | 300 |
| No violation<br>(type-I-risk) | Andersen's Likelihood-Ratio test<br>Three-way analysis of variance<br>*z*-test of Fischer and Scheiblechner<br>Martin-Löf test | | |
| Effect size ½ *sd* of 1 DIF-pair | Andersen's Likelihood-Ratio test<br>analysis of variance design<br>*z*-test of Fischer and Scheiblechner | | |
| Effect size 1 *sd* of 1 DIF-pair | | | |
| Effect size 2 *sd* of 1 DIF-pair | | | |
| Multidimensionality<br>latent correlation 0.5 | Andersen's Likelihood-Ratio test<br>analysis of variance design<br>Martin-Löf test | | |

To calculate the actual type-I-risk 10,000 Rasch model conform datasets were simulated. The item parameters were: -3, -2.5, -2, -1.5, -1.2, -0.9, -0.75, -0.5, -0.3, -0.1, 0.1, 0.3, 0.5, 0.75, 0.9, 1.2, 1.5, 2, 2.5, 3. The datasets were analyzed by Andersen's Likelihood-Ratio test using the internal split criterion, the Martin-Löf test using a random split criterion, the *z*-test of Fischer und Scheiblechner and the three-way analysis of variance with a random group factor. If the result of any simulation is significant, a type-I-error is committed.

To simulate the power datasets were generated which contradict the assumption of the Rasch model insofar as some DIF applies.

The two (random) groups of a dataset differed according to three conditions: There are three different effect sizes of violation of the model. In the first condition the difference of item difficulty of item 9 and 12 was 0.75, which is one half of the standard deviation of the person ability parameters. In the second condition item 7 and 14 were affected with a DIF of 1.5 which equals a standard deviation of the person ability parameters. In the third condition item 4 and 17 were affected with a DIF of three which means two standard deviations. Table 2 shows the item difficulties for all conditions for both groups. Differences are printed in bold.

For both groups in all three conditions a set of Rasch model conform datasets was simulated. For the scenarios of 100 simulees the group was split into two subsets of 50 each: For one set of 50 simulees data were simulated with the first group of item parameters and for

**Table 2:**
Model violation due to DIF

| | DIF | | | | | |
|---|---|---|---|---|---|---|
| | 0.75 | | 1.5 | | 3 | |
| | Group | | | | | |
| Item | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | -3 | -3 | -3 | -3 | -3 | -3 |
| 2 | -2.5 | -2.5 | -2.5 | -2.5 | -2.5 | -2.5 |
| 3 | -2 | -2 | -2 | -2 | -2 | -2 |
| 4 | -1.5 | -1.5 | -1.5 | -1.5 | **-1.5** | **1.5** |
| 5 | -1.2 | -1.2 | -1.2 | -1.2 | -1.2 | -1.2 |
| 6 | -0.9 | -0.9 | -0.9 | -0.9 | -0.9 | -0.9 |
| 7 | -0.75 | -0.75 | **-0.75** | **0.75** | -0.75 | -0.75 |
| 8 | -0.5 | -0.5 | -0.5 | -0.5 | -0.5 | -0.5 |
| 9 | **-0.375** | **0.375** | -0.3 | -0.3 | -0.3 | -0.3 |
| 10 | -0.1 | -0.1 | -0.1 | -0.1 | -0.1 | -0.1 |
| 11 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| 12 | **0.375** | **-0.375** | 0.3 | 0.3 | 0.3 | 0.3 |
| 13 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| 14 | 0.75 | 0.75 | **0.75** | **-0.75** | 0.75 | 0.75 |
| 15 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| 16 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| 17 | 1.5 | 1.5 | 1.5 | 1.5 | **1.5** | **-1.5** |
| 18 | 2 | 2 | 2 | 2 | 2 | 2 |
| 19 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 |
| 20 | 3 | 3 | 3 | 3 | 3 | 3 |

the other 50 simulees data with the second group of item parameters. These datasets were merged so that the final dataset contradicts the Rasch model. 10,000 datasets were generated for each of the three conditions and each number of persons. These datasets were analyzed by Andersen's Likelihood-Ratio test using the group variable as split criterion (in contrast to the simulation of the type-I-risk, where the internal split criterion was used), the *z*-test of Fischer and Scheiblechner and the three-way analysis of variance using the same group-variable as a factor, which was used as the split criterion of Andersen's Likelihood-Ratio test. A significant result means that the violation of the Rasch model is discovered.

The determined DIFs reflect that a pair of DIF with a difference of two standard deviations means that one item difficulty is one standard deviation above-average in one group and accordingly below-average in the other group; and for a second pair of items the relationship between groups is quite contrary. A proper model test should be able to detect at least model violations of such a magnitude.

The Martin-Löf test does not aim for detecting model violation by some DIF. However, it aims for detection of two groups of items, each of them fitting the Rasch model meaning each of them measures uni-dimensionally, but they do not if pooled; thereby, both uni-dimensional abilities may correlate more or less. Hence, in order to evaluate the power of the Martin-Löf test multi-dimensional data were simulated. As indicated above, two latent dimensions with a correlation coefficient of 0.5 were used. Therefore two different item groups with the following difficulty parameters each were determined: 3, -2, -1, -0.4, -0.1, 0.1, 0.4, 1, 2 and 3. The distributions of the person ability parameters were the same as always, but now there were two of them correlating as described, therefore data were drawn out of a bivariate normal distribution with correlation 0.5. The resulting data were analyzed by the Martin-Löf test using the two item groups as a split criterion, Andersen's Likelihood-Ratio test using the internal split criterion and the three-way analysis of variance using a random group factor.

The actual type-I-risk of Andersen's Likelihood-Ratio test and the Martin-Löf test was computed as the proportion of significant results. For the power calculation, of course, the same quotient of the number of significant results and the number of analyzed data sets was used. The same is true as concerns the *z*-test of Fischer and Scheiblechner, but as it tests each item 200,000 significance tests (10,000 for each of the 20 items) would result for every simulation scenario. For this reason, for the calculation of the power only those two tests were taken into account in the following, the difficulties of which were transposed between the groups: that is, there are 20,000 significant tests.

To evaluate the approach of the three-way analysis of variance it was counted how often the interaction term A x C was significant, whereas all simulations were excluded when the main effect A was significant (see above). Beside this type-I-risk and the power of this approach were calculated analogously.

## Results

Results are presented in Table 3. The number of valid repetitions, the actual type-I-risk and the simulated power for each model test are shown. In case of Andersen's Likelihood-Ratio test and the Martin-Löf test the number of repetitions deviates from 10,000 (in the cases of no model violation or model violation by DIF) and from 1,000 (violation because of multidimensionality) if for a certain simulated data set the Rasch model parameters were not able to be estimated. This could happen if an item was never or always solved in one group. Equivalently the number of repetitions of the *z*-test of Fischer and Scheiblechner deviates from 200,000 (no model violation) or 20,000 (model violation), if it was not always possible to estimate the parameters of the Rasch model. In case of the three-way analysis of variance the difference of 10,000 (no violation or violation by DIF) and of 1,000 (violation because of multidimensionality) and the value of valid repetitions is the number of significant main effects A.

For instance, in the case of no model violation and 100 simulees 9,998 data sets could be analysed and Andersen's Likelihood-Ratio test applied, respectively. 360 of these were significant. Hence the actual type-I-risk is 3.6 %.

The *z*-test of Fischer and Scheiblechner holds the type-I-risk which actually lays between 4 % and 4.6 %. The actual type-I-risk of Andersen's Likelihood-Ratio test and the three-way analysis of variance lay between 3.6 % and 5.6 % thus they do not cover the obligatory 20 %-robustness (that is 0.04 – 0.06) due to Rasch & Guiard (2004).

The Martin-Löf test was never significant under the condition of no violation of this simulation study. A closer look at the p-values showed, that they are far from the critical value of significance, even if the p-values decrease with increasing number of simulees (cf. Table 4).

**Table 3:**
Valid repetitions, number of significant results, the actual type-I-risk and the power, respectively for all model tests and sample sizes in all scenarios.

| | | Simulees | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 100 | | | 200 | | | 300 | | |
| | | valid repetitions | sign. results | type-I-risk / power | valid repetitions | sign. results | type-I-risk / power | valid repetitions | sign. results | type-I-risk / power |
| Type -I- risk | LR-Test | 9998 | 360 | 0.036 | 10000 | 394 | 0.0394 | 10000 | 457 | 0.0457 |
| | 3VA | 9999 | 454 | 0.0454 | 10000 | 563 | 0.0563 | 10000 | 363 | 0.0363 |
| | *z*-test | 185141 | 7409 | 0.04 | 194549 | 8567 | 0.044 | 197543 | 9148 | 0.0463 |
| | ML- test | 9999 | 0 | 0 | 10000 | 0 | 0 | 10000 | 0 | 0 |
| 0.75 DIF | LR-test | 9980 | 2238 | 0.2242 | 9995 | 4176 | 0.4178 | 10000 | 5990 | 0.599 |
| | 3VA | 9996 | 2523 | 0.2524 | 10000 | 4828 | 0.4828 | 9998 | 6264 | 0.6265 |
| | z-test | 20000 | 942 | 0.0471 | 20000 | 1190 | 0.0595 | 20000 | 1218 | 0.0609 |
| 1.5 DIF | LR-test | 9984 | 8290 | 0.8303 | 9997 | 9886 | 0.9889 | 10000 | 9998 | 0.9998 |
| | 3VA | 9998 | 8768 | 0.877 | 10000 | 9941 | 0.9941 | 9999 | 9998 | 0.9999 |
| | *z*-test | 19996 | 1252 | 0.0626 | 20000 | 2029 | 0.1015 | 20000 | 2748 | 0.1374 |
| 3 DIF | LR-test | 9987 | 9987 | 1 | 9998 | 9998 | 1 | 10000 | 10000 | 1 |
| | 3VA | 9996 | 9996 | 1 | 10000 | 10000 | 1 | 9998 | 9998 | 1 |
| | *z*-test | 19996 | 5657 | 0.2829 | 20000 | 11023 | 0.5512 | 20000 | 15304 | 0.7652 |
| multi | ML-test | 994 | 32 | 0.0322 | 1000 | 315 | 0.315 | 1000 | 699 | 0.699 |
| | LR-test | 1000 | 28 | 0.028 | 1000 | 39 | 0.039 | 1000 | 45 | 0.045 |
| | 3VA | 995 | 69 | 0.0721 | 995 | 48 | 0.0505 | 947 | 47 | 0.0496 |

Note. LR-Test = Andersen's Likelihood-Ratio test; 3VA = three-way analysis of variance; z-test = z-test of Fischer und Scheiblechner; ML-test = Martin-Löf test; multi = condition of multidimensionality.

**Table 4:**
The p-values of the Martin-Löf test: minimum, maximum, and mean for the scenario of no model violation (10,000 repetitions)

| | min | max | $\overline{p}$ |
|---|---|---|---|
| 100 persons | 0.7603 | 1 | 0.9992 |
| 200 persons | 0.5697 | 1 | 0.9972 |
| 300 persons | 0.4071 | 1 | 0.9949 |

The power of Andersen's Likelihood-Ratio test and the three-way analysis of variance was – in the case of DIF – always very similar, although the power of the latter was slightly higher. For a DIF pair of 0.75 it varies between 0.22 and 0.63, for a DIF pair of 1.5 between 0.83 and almost 1 depending on the sample size. In case of a DIF pair of 3 the power was 1 for all sample sizes for both tests. Even for small samples of 100 simulees the power of Andersen's Likelihood-Ratio test and the three-way analysis of variance was beyond 0.8 in case of a DIF pair of at least 1.5. The power of the *z*-test of Fischer and Scheiblechner was extremely low in case of a pair of small model deviations. In case of a pair of 0.75 DIF and a pair of 1.5 DIF it varies between approximately 0.05 and 0.14, increasing when the number of simulees and the DIF are higher. In the simulations with a pair of 3 DIF the power fluctuates between 0.28 and 0.77, which is higher but still very low compared to the other tests under the same conditions.

The simulation of multi-dimensional data showed that the power of Andersen's Likelihood-Ratio test and of the three-way analysis of variance is only as much as the actual type-I-risk. The power of the Martin-Löf test strikingly increases with larger sample size. In case of 100 simulees it is very low and amounts about only 0.03, increases to 0.32 in case of 200 persons and reaches almost 0.7 if the sample size is 300.

## Discussion

As expected, the four model tests have different strengths and weaknesses. Depending on the model violation they were more or less suitable to detect it. As usual, a bigger effect size (greater DIF) and larger sample size lead to a higher test power.

For the simulated scenarios the *z*-test of Fischer and Scheiblechner hold the type-I-risk. In some conditions the actual type-I-risk of Andersen's Likelihood-Ratio test and the three-way analysis of variance was below the minimum level of 0.04 due to the 20 %-robustness of Rasch & Guiard (2004). The type-I-risk of the Martin-Löf test was 0 %. Obviously, the Martin-Löf test needs greater sample sizes for getting significant according to a type-I-error. This result conforms to those of Verguts und De Boeck (2000). They simulated data and analysed 24 items and 5,000 persons on the basis of a type-I-risk of 0% .

If the data refer to a multi-dimensional (i.e. two-dimensional) ability the power of this test depends very strongly on the sample size: For 300 simulees and a latent correlation of both ability dimensions of .5 the power of the Martin-Löf test reaches .7.

Andersen's Likelihood-Ratio test and the three-way analysis of variance have on the other side no power to detect multi-dimensionality. Concerning Andersen's Likelihood-Ratio test this is not surprising (see Stelzl, 1979, Wollenberg, 1979, Formann, 1981 and Alexandrowicz, 2002).

In this study it was shown that Andersen's Likelihood-Ratio test and the three-way analysis of variance have a power greater than 80 % if the model violation is 1.5 for a single pair of DIF, even for the smallest sample of 100 simulees, whereby the power of the three-way analysis of variance is slightly higher.

The power of the *z*-test of Fischer and Scheiblechner tends to be very low for small sample sizes. In case of 300 simulees and a DIF pair of 3, it is still below 80 %.

Simulations showed that for the given scenarios the three-way analysis of variance could always be used instead of Andersen's Likelihood-Ratio test, if the dataset is split with regard to an external criterion. However, further research is needed in order to evaluate both approaches' power when the given results for the three-way analysis of variance are compared to Andersen's Likelihood-Ratio test if the internal split criterion is used for this test.

To identify multi-dimensionality the Martin-Löf test should be applied. However for small samples its power is very low.

The *z*-test of Fischer and Scheiblechner, hardly used in practice (cf. Kubinger, 2005), has only unsatisfactory power with respect to a single item, especially in the case of small sample sizes and small effects.

In this paper the simulation was performed for restrictive scenarios. That is, our results can be generalized to other scenarios only with big reservation; better, to research other scenarios in detail.

## Acknowledgement

Many thanks to Jan Steinfeld.

## References

Alexandrowicz, R. (2002). *Die Teststärke des Likelihood-Quotienten-Tests nach Andersen bei der Überprüfung der Modellgültigkeit des dichotomen logistischen Modells nach Rasch.* [Power of Andersen's Likelihood-Ratio test testing the dichotomous logistic Rasch model.] Vienna: unveröffentlichte Dissertation University of Vienna.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Fischer, G. H. (1995). Derivations of the Rasch Model. In G. H. Fischer, & I. W. Molenaar (Hrsg.), *Rasch Models Foundations, Recent Developments, and Applications* (pp. 69-95). New York: Springer.

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling. Handbook of the usage of LpcM-WiN 1.0.* Groningen: PROGRAMMA.

Fischer, G. H., & Scheiblechner H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch [Algorithms and programs for the probabilistic Rasch model]. *Psychologische Beiträge, 12,* 23-51.

Kubinger, K. D. (2005). Psychological Test Calibration using the Rasch Model – Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5,* 377-394.

Kubinger, K. D., Rasch, D., & Yanagida, T. (2009). On designing data-sampling for Rasch model calibrating an achievement test. *Psychology Science Quarterly, 51 ,* 270-384.

Kubinger, K. D., Rasch, D., & Yanagida, T. (2011). A new approach for testing the Rasch model. *Educational Research and Evaluation, 17,* 321-333.

Mair, P., Hatzinger, R., & Maier, M. (2011). *eRm: Extended Rasch Modeling. R package version 0.14-0.* http://CRAN.R-project.org/package=eRm.

Martin-Löf, P. (1973). Statistika Modeller: Anteckningar från seminarier Lasåret 1969-1970, utarbetade av Rolf Sunberg [Statistical Models: Notes from seminars of the academic year 1969-1970, elaborated by Rolf Sunberg]. Obetydigt ändrat nytryck, oktober 1973. *Institutet för säkringsmatematik och matematisk statistik vid Stockholms universitet.*

R Development Core Team. (2012). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing http://www.R-project.org/.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science, 46,* 175-208.

Rasch, D., Kubinger, K. D. & Yanagida, T. (2011). *Statistics in Psychology – Using R and SPSS.* Chichester: Wiley.

Stelzl, I. (1979). Ist der Modelltest des Rasch-Modells geeignet, Homogenitätshypothesen zu prüfen? Ein Bericht über Simulationsstudien mit inhomogenen Daten. [Is the model test of the Rasch model suitable to check hypotheses of one-dimensionality? Report about a simulation study with multi-dimensional data.] *Zeitschrift für experimentelle und angewandte Psychologie, 26,* 652-672.

Tierney, L., Rossini, A. J., Li, N., & Sevcikova, H. (2012). *snow: Simple Network of Workstations. R package version 0.3-10.* http://CRAN.R-project.org/package=snow.

Verguts, T., & De Boeck, P. (2000). A note on the Martin-Löf test for unidimensionality. *Methods of Psychological Research Online, 5,* 77-82.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426-482.

Wollenberg, A. L. van den (1979). *The Rasch model and time-limit tests.* Proefschrift Univ. Nijmegen, Nijmegen.