Nonparametric tests for the Rasch model: explanation, development, and application of quasi-exact tests for small samples

Ingrid Koller[1*] & Reinhold Hatzinger[2]

[1]Department of Psychological Basic Research and Research Methods; Faculty of Psychology; University of Vienna

2 Institute for Statistics and Mathematics; Vienna University of Economics and Business

* Corresponding author: ingrid.koller@univie.ac.at

Abstract

Psychological test validation by means of item response models is commonly carried out with large samples. However, in practice, the use of large samples is not always feasible, e.g., it is not possible to get large samples because of not enough people showing the construct of interest, and collecting new data would incur high costs. Therefore, it would be preferable to a priori analyze newly-developed items using small samples. Ponocny (2001) introduced quasi-exact tests for the Rasch model based on Monte-Carlo simulations in order to sample random matrices with identical margins compared to the observed matrix. These tests allow the investigation of the model fit even in small samples. In this paper we describe some tests of Ponocny and two newly-developed tests based on the simulation algorithm of Verhelst (2008). Theoretical foundations, practical recommendations and an empirical example are given. The present study underlines the usefulness of these nonparametric tests and demonstrates how new tailored methods for small samples applied to an examination of item quality can easily be developed.

Keywords: Rasch model, nonparametric test, quasi-exact test, model test, small samples

1. Introduction

An important topic in psychology is the psychometric investigation of items contributing to a score. Different methods for the evaluation of items exist. From an item response model perspective the Rasch model (RM; Fischer & Molenaar, 1995; Rasch, 1960) offers the following mathematical properties which hold if the model fits the data: 1) *Local independence* means that each item's probability of being solved is independent of the probability from any other item given the latent trait. 2) *Undimensionality* or homogeneity of items means that in general all items of a test measure the same latent construct. 3) *Parallel and strictly increasing item characteristic curves* means that with increasing person ability the probability to solve an item is also increasing. 4) *Specific objectivity* implies that it is irrelevant which item is used to compare two persons with certain person abilities. It is also irrelevant which person is used to compare two items with specific item difficulties. An important aspect of specific objectivity is *measurement invariance* (e.g., Kim, Yoon, & Lee, 2012) which insists that different subgroup of persons must show the same conditional probabilities to solve items, given the latent trait. If external criteria, such as gender, are investigated this is also called the investigation of differential item functioning (e.g., Holland & Thayer, 1988).

Given that all previously mentioned properties are valid, the number of items solved (raw score) and the number of people with positive response to one item (sum score) contain all information about the item difficulty and the person ability, respectively. This property is called *sufficiency* and implies that it is irrelevant which items have been solved, the margins contain all information.

To examine the properties of the RM a variety of methods have been developed (e.g., Fischer & Molenaar, 1995; Suárez-Falcón & Glas, 2003). Usually first, the parameters of items and persons have to be estimated. Second, the model fit has to be examined, e.g., with the likelihood ratio test (LRT) suggested by Andersen (1973). But these significance tests require large samples to be asymptotically valid. If the sample is too small, various problems may occur in practice: First, the parameter estimates might be inaccurate (Fischer, 1981) which biases subsequent model tests. Second, model testing based on asymptotic distributions (e.g., the χ^2-distribution of the LRT's) may only be valid if sample size is reasonably large (e.g., Fischer & Molenaar, 1995; Ponocny, 2001).

Furthermore, practical problems often occur when analyzing a set of items. In practical applications, it is not always possible to get large samples because of complex study designs (e.g., experiments), not enough people showing the characteristics of interest (e.g., clinical studies), or financial limitations, which may hamper implementations of large scale assessments. Furthermore, a good practice in test evaluation is to investigate the quality of items more than once. As a result, many studies make use of small samples with the consequences that (a) the results of test evaluation may be incorrect, (b) it is not possible to evaluate the items with the RM, and (c) the results of the analysis cannot be cross validated.

A possible way to overcome these problems are quasi-exact tests which allow checking of the properties of the RM even if sample sizes are small. In this paper, we will describe the idea of exact tests and give a short introduction of the simulation method underlying the statistics. Second, quasi-exact tests are described, whereas the focus lay on some tests of Ponocny (2001) and two new tests developed. Theoretical foundations are outlined, a short practical guide and a practical example are given by a reanalysis of a data to the topic of dyscalculia. However, to show the comparableness with one parametric approach not a small sample is used, but a reasonably large sample was chosen which makes it is possible to apply both approaches.

1.1. Exact tests

Rasch was the earliest to propose the idea of an exact test, and this was followed by other researchers (e.g., Ponocny, 1996, 2001). It can be considered as a generalization of Fisher's exact test. The idea is based on the property of sufficient statistics. If the property is valid, all possible matrices with identical margins will have the same parameter estimates. This property allows us to examine the RM conformity also in small samples (e.g., Koller, Alexandrowicz, & Hatzinger, 2012; Ponocny, 2001). The procedure is simple and straightforward: There is an $r \times c$ observed matrix $\mathbf{A}_0$ (r = raw score of the persons, c = sum score of the items). First, all possible matrices with equal margins as in $\mathbf{A}_0$ are generated ($\mathbf{A}_1$, …, $\mathbf{A}_s$, …, $\mathbf{A}_S$). We denote all matching matrices (s = 1, …, S) with Ω_{rc}. In a next step a suitable test-statistic T for the observed matrix $\mathbf{A}_0$ (T_0) and all generated matrices $\mathbf{A}_s$ will be calculated (T_1, …, T_S). In the last step a model test is carried out by counting how often the T's show the same or a more extreme value compared to T_0. The relative frequency gives the resampling p-value under the null hypothesis of model conformity which will be compared with an *a-priori* defined nominal significance level α. The equation of model test is given as

$$p = \frac{1}{S}\sum_{s=1}^{S} t_s \qquad \text{where} \qquad t_s = \begin{cases} 1, & \text{if } T_s \geq T_0 \\ 0, & \text{elsewhere} \end{cases}. \tag{1}$$

Due to computational limitations full enumeration (i.e., to calculate all possible matrices with given margins) is not feasible in practice. Several authors address the problem by simulating possible matrices for a given observed matrix $\mathbf{A}_0$ (e.g., Chen & Small, 2005; Ponocny, 2001; Snijders, 1991; Verhelst, 2008). For example, Ponocny (2001) developed a Monte-Carlo simulation algorithm, which can only be applied to small matrices (i.e., a maximum of 100 subjects or items and 30 items or subjects). Verhelst (2008) introduced a

Markov Chain Monte-Carlo algorithm (MCMC), which allows the simulation of larger matrices and requires less calculation time compared to Ponocny's approach. The MCMC approach of Verhelst is implemented in the open-source software *R* (R-Core-Team, 2012) in the package *RaschSampler* (Verhelst, Hatzinger, & Mair, 2007) and is connected to the function *NPtest* in the package *eRm* (Mair, Hatzinger, & Maier, 2012). Because the exact tests are not based on all possible matrices but only on a reduced number of simulated matrices these tests are called *quasi-exact tests* instead of exact tests.

1.2. The simulation algorithm

In this section we will give a short description of the simulation algorithm of Verhelst (2008). For a detailed description see Verhelst et al. (2007) and Verhelst (2008).

In order to ensure valid test results, the following two conditions must be met: First, all binary matrices with equal margins (Ω_{rc}) have to occur with the *same probability*. Second, each matrix has to be *independent of the initial matrix* $\mathbf{A}_0$. If either one of the two conditions are not met, $\mathbf{A}_1$, …, $\mathbf{A}_{nsim}$ (*nsim* is the number of simulated matrices) might be too similar to $\mathbf{A}_0$ and therefore the test-statistics T_1, …, T_{nsim} will have a similar value which can affect the results. Next, we will describe how the two conditions can be achieved.

Procedure: In the first step two columns of $\mathbf{A}_0$ are randomly chosen (see Tab. 1). The possible raw scores are 0, 1, and 2. The rows of interest are those with the raw score 1 (pattern {10} and {01}). Thus, given the two columns there are five rows where an exchange of the patterns {10} and {01} are possible (highlighted in gray). In the next step entries in these rows will be randomly exchanged. This step leads to a new matrix $\mathbf{A}_1$ (right panel of Tab. 1) with the same margins as in $\mathbf{A}_0$.

In order to achieve the two previous mentioned conditions a *burn-in period* and a *step size* have to be set up. A burn-in period means that a predefined number of simulated matrices is not used for the calculation of test-statistics. All simulated matrices of the burn-in period will be left out and the following matrix will be chose to be the first simulated matrix for the test-statistic. This procedure guarantees that the simulated matrices are (pseudo-) independent of the initial matrix. The second important parameter of the simulation is the *step size*. As we have seen in Table 1, a new matrix is generated by simply exchanging two columns. Therefore, the new matrix $\mathbf{A}_2$ is very similar to $\mathbf{A}_1$. To overcome this problem not every simulated matrix is used. If a step size of 32 is defined, 31 simulated matrices are discarded and the 32nd matrix is used as an *effective matrix*. Finally, it is necessary to determine how many effective matrices should be chosen for the calculation of the test-statistics. If we define a burn-in of 100, a step size of 32, and 100 effective matrices, then in total $100 \times 32 + 100 \times 32 = 6400$ matrices are simulated. Thus, for the calculation of the quasi-exact test only the 100 effective matrices will be used. The impact of different parameter settings (e.g., burn-in) on the MCMC-algorithm has been studied by Verhelst (2008).

Table 1. Example for the generation of a new matrix by permutation of the elements of two columns.

	A$_0$					**A**$_1$	
Person	***i***	***j***	***r***		***i***	***j***	***r***
1	0	0	0		0	0	0
.	1	0	1		1	0	1
.	1	0	1		1	0	1
.	0	1	1		0	1	1
v	0	0	0	→	0	0	0
.	1	1	2		1	1	2
.	0	1	1		0	1	1
.	1	0	1		1	0	1
.	1	1	2		1	1	2
n	0	0	0		0	0	0

Note. *i* … item i, *j* … item j, *r* is the raw score; Only rows with sub-total 1 can be permuted – they are highlighted in grey.

2. Quasi-Exact Tests

Different authors have developed quasi-exact tests for the Rasch model (e.g., Chen & Small, 2005; Christensen & Kreiner, 2010; Ponocny, 1996; Ponocny, 2001; Verhelst et al., 2007). The focus of this paper is to describe some of Ponocny's tests as well as to develop and describe two new tests. All tests are implemented in the *R* package *eRm* and this will be used for the application with a real data set.

2.1. Tests for local independence and homogeneity between items

T_{11}: A global test for inappropriate inter-item correlations.

The test-statistic T_{11} (Ponocny, 2001) investigates whether there exists a violation of local independence and/or homogeneity between items. It tests whether the inter-item correlations are too high or too low. The statistic can be written as

$$T_{11}(\mathbf{A}) = \sum_{ij} |r_{ij} - \tilde{r}_{ij}| \,, \tag{2}$$

where r_{ij} is the inter-item correlation between the items *i* and *j*, and where the summation is over all item pairs $(r_{12}, r_{13}, r_{23}, \dots, r_{ij})$. Next, the average $\tilde{r}_{ij}$ for the simulated matrices are formed by summing the individual r_{ij} from the simulated matrices and dividing by the number of simulated matrices $(\sum_{s=1}^{nsim} r_{ij}/nsim)$. This average quantity forms the expected $\tilde{r}_{ij}$. Then the absolute deviation between the observed r_{ij} and the expected $\tilde{r}_{ij}$ will be calculated and summed up for all pairs of items in a test $(|r_{12} - \tilde{r}_{12}| + |r_{13} - \tilde{r}_{13}|, \dots, |r_{ij} - \tilde{r}_{ij}|)$ which results in T_0 for the observed matrix. The second stage is to obtain the test-statistics $T_1, \dots, T_{nsim}$ for each of the simulated matrices. We get these test-statistics by replacing the observed correlation r_{ij} by the correlations of the simulated matrices (**A**1, …, $\mathbf{A}_{nsim}$) while $\tilde{r}_{ij}$ is the same value for all test-statistics. The model test is given in Equation (3) and is defined as the relative frequency of test-statistics T_s of the simulated matrices which have the same or a larger value compared to the test-statistic resulting from the observed matrix $\mathbf{A}_0$.

$$p = \frac{1}{nsim} \sum_{s=1}^{nsim} t_s \qquad \text{where} \qquad t_s = \begin{cases} 1, & \text{if } T_s(\mathbf{A}_s) \geq T_0(\mathbf{A}_0) \\ 0, & \text{elsewhere} \end{cases} \tag{3}$$

It has to be noted that the test detects correlations that are, both, too small or too large. Therefore, a significant result means that one or both properties could be violated.

***T*₁: A test at the item level for too many equal response patterns {00} and {11}.**

The statistic T_1 (Ponocny, 2001) is a test for pairs of items. It allows us to investigate if the extreme patterns {00} and {11} occur more often than it would be expected by the Rasch model. If the frequency is too high, the correlation between two items will be too high which suggests a violation of the local dependency assumption. The basic test-statistic can be written as

$$T_1(\mathbf{A}) = \sum_{v=1}^{n} \delta_{ijv} \quad \text{where} \qquad \delta_{ijv} = \begin{cases} 1, & \text{if } x_{vi} = x_{vj} \\ 0, & \text{if } x_{vi} \neq x_{vj} \end{cases}, \tag{4}$$

where x_{vi} is the response of person *v* on item *i* and x_{vj} is the response of person *v* on item *j*. To get T_1 the equally patterns have to be summed up. The model test (Eq. 3) is defined as the relative frequency of T_s which show the same or a higher number of equal patterns as in T_0.

If more than two items are investigated, then the test is only useful for detecting violations in the same direction. Otherwise, if one pair of items shows too high a number of equal patterns and another pair of items shows too low a number of equal patterns, low sum of one pair will compensate for the high sum of the other and the test will hide such model violations.

***T*₁*ₘ*: A test at the item level for too few equal response patterns {00} and {11}.**

With a modification of the model test in Equation (4) it is possible to investigate a violation of the assumption of homogeneity between items (too low correlation between items). The model test for the statistic T_{1m} (*m* means multidimensionality) is given in Equation (5).

$$p = \frac{1}{nsim}\sum_{s=1}^{nsim} t_s \qquad \text{where} \qquad t_s = \begin{cases} 1, & \text{if } T_s(\mathbf{A}_s) \leq T_0(\mathbf{A}_0) \\ 0, & \text{elsewhere} \end{cases} \tag{5}$$

Here, all T_s will be summed up, which have the same or a smaller number of equal patterns as in T_0. If there are too few equal patterns the correlation between items is too small and the assumption of homogeneity is violated.

$T_{1\ell}$: A test at the item level for too many equal response patterns {11}.

Sometimes it is assumed that people might have learned from a previously solved item. In this case there should be more patterns {11} than expected, but not {00}. If T_1 is applied, the power to detect this kind of violation is small, because in T_1 it is assumed, that also the frequency of the pattern {00} is increased and therefore also this pattern will be summed up. But learning has no effect on the frequency of pattern {00}. So the statistic T_1 has to be modified so that only the patterns {11} are summed up. The equation for $T_{1\ell}$ (ℓ for learning) is given in Equation (6). The model test is the same as mentioned for T_1 (Eq. 3).

$$T_{1\ell}(\mathbf{A}) = \sum_{v=1}^{n} \delta_{ijv} \quad \text{where} \qquad \delta_{ijv} = \begin{cases} 1, & \text{if } x_{vi} = x_{vj} = 1 \\ 0, & \text{elswhere} \end{cases} \tag{6}$$

T_2: A test at the item level for a too large variance in the raw score of a subscale.

The test T_2 (Ponocny, 2001) uses another approach to test if there is a violation of the local independence assumption. This statistics tests whether the variance of a subscale is too large. The test-statistic can be written as

$$T_2(\mathbf{A}) = Var(r_v^{(I)}) \qquad \text{where} \quad (r_v^{(I)}) = \sum_{i \in I} x_{vi} \ , \tag{7}$$

where $Var(r_v^{(I)})$ is the variance of a subscale *I* with a minimum of at least two items. Following the variance addition theorem, the variance of a scale consisting of two subscales *I* and *J* is defined as $Var(r_v^{(I)}) + Var(r_v^{(J)}) + 2 \times Cov(r_v^{(I,J)})$. Therefore, the variance of a scale is increasing with increasing covariance of *I* and *J*. The covariance is the unstandardized association of two variables. This means that with increasing association of two raw scores of *I* and *J* also the variance of the scale is increasing. Furthermore, with increasing association also the dependence between items is increasing.

These considerations lead to the following test procedure. First, the variance of the raw score of a specified subscale is calculated and the result is the test-statistic for the observed matrix. This step has to be repeated for all simulated matrices. The model test (Eq. 3) is defined as the relative frequency of T_s which show the same or a larger variance as in T_0. The annotations are the same as for T_1.

T

T_{2m}: A test at the item level for a too low variance in the raw score of a subscale.

As for T_1, with a modification of the model test of T_2 it is possible to investigate the assumption of homogeneity between items. The model test for the modified statistic T_{2m} is given in Equation (5) and defines the *p*-value as the relative frequency of T_s which show the same or a lower variance as in T_0.

2.2. Measurement invariance

T_{10}: A global test for the investigation of measurement invariance.

The statistic T_{10} (Ponocny, 2001) is a global test for the investigation whether there exists a violation of the assumption of measurement invariance. The statistic can be written as

$$T_{10}(\mathbf{A}) = \sum_{ij} |n_{ij}^{ref} \times n_{ji}^{foc} - n_{ji}^{ref} \times n_{ij}^{foc}|. \tag{9}$$

Here, the sample has to be divided into a *reference* group and a *focal* group (e.g., males and females). First, the number of people in the reference group (ref) who solved item *i* but not item *j* (n_{ij}^{ref}) is multiplied with the number of people in the focal group (foc) who solved item *j* but not item *i* (n_{ji}^{foc}). The result is subtracted from the product of the two remaining possibilities ($n_{ji}^{ref} \times n_{ij}^{foc}$). This step is done for all possible item combinations and summed over all pairs of items. The model test (Eq. 3) is defined as the relative frequency of T_s which have the same or a higher value as in T_0. Note, it is not possible to investigate more than two groups of people.

This test-statistic can be seen as a nonparametric competitor to the parametric LRT described by Andersen (1973). The LRT investigates whether there is a significant difference between the item difficulty parameters of different groups ($\beta_i^{ref} - \beta_i^{foc}$). Due to the approximate relationship between n_{ij}/n_{ji} and $\exp(\beta_i)$ / $\exp(\beta_j)$ (e.g., Fischer & Molenaar, 1995) T_{10} also investigates differences between parameters.

T_4: A test at the item level for the detection of too many or too few positive answers within a group.

The statistic T_4 (Ponocny, 2001) investigates if one or more items within a predefined subgroup (e.g., females) are too easy or too difficult compared to another subgroup (e.g., males). The test-statistic can be written as

$$T_4(\mathbf{A}) = \sum_{v \in G_g} x_{vi}, \tag{10}$$

where the answers x_{vi} of an item of interest in the predefined subgroup (g = 1, …, G) are summed up. Assuming that the item is too easy (i.e., higher number of correct responses as expected) in the investigated group, the model test is given by Equation (3). Assuming that the item is too difficult (i.e., smaller number of correct responses as expected) in the investigated group the model test is given by Equation 5.

If more than two subgroups are defined, a significant result means that the item is significantly easier (more difficult) than the RM assumed. In this case all subgroups of interest should be investigated separately.

If more than one item is of interest, the number of positive answers for each item has to be summed up to a global sum. Here it is important to note that all investigated items have to show the violation in the same direction. Otherwise, if one item is too easy (higher sum) and the other item is too difficult (lower sum), then the two effects may cancel, with the test giving a false conclusion of no violation of measurement invariance.

Furthermore, it is not possible to sum up over all items. The margins are fixed and also the sum over all items is fixed. If all items are summed up all simulated matrices achieved the same result and the test-statistic is always one.

2.3. A new test for multidimensional functioning subscales

T_{md}: A global test for inappropriate correlations of subscales.

Based on the Martin-Loef test (as cited in Fischer & Molenaar, 1995) we further develop the quasi-exact test T_{md} (*md* means MultiDimensionality) for the detection of multidimensional functioning subscales. Conceptually, the new approach can be compared to the idea of Verguts and Boeck (2001) who developed a test based on the Mantel-Haenzel statistic of Rosenbaum (1984) for the investigation of monotonicity and conditional independence. The basic idea is that the data are divided into two subscales *I* and *J*. If the RM holds, the two raw scores $r_v^{(I)}$ and $r_v^{(J)}$ should be positive associated. If the correlation is too low, it is an indicator that there exist items which show low discrimination (monotonicity), and/or multidimensionality between items. The test-statistic can be written as

$$T_{md}(\mathbf{A}) = \ Cor(r_v^{(I)}, r_v^{(J)}) \qquad \text{where} \ \ (r_v^{(I)}) = \textstyle\sum_{i \in I} x_{vi} \ . \tag{8}$$

The model test is given in Equation (5) and is defined as the relative frequency of T_s which have the same or a smaller correlation value as in T_0. Note, it is not possible to investigate more than two subscales.

2.4. A new developed test for too low discrimination of items

***T_{pbis}*: A test at the item level for inappropriate patterns.**

The last statistic is a monotone transformation of the point biserial correlation. In the point biserial correlation the sample is divided by an item of interest in a group of subjects who respond positively on the item and a group who respond negatively on the item. The discrimination of the item will be investigated by comparing the mean raw scores of the remaining items between the two subgroups. The statistic is

$$r_{bis} = \frac{\bar{r}_1 - \bar{r}_0}{s_r} \sqrt{\frac{n_0 n_1}{n(n-1)}}, \tag{11}$$

where $\bar{r}_1$ is the mean raw score of the remaining items of the people who responded positively on the item and $\bar{r}_0$ is the mean raw score of the remaining items of the people who responded negatively on the item. The term s_r is the standard deviation of the remaining items for the total sample and n_0, n_1, and n are the number of people in group 0, group 1, and of the whole sample, respectively. The terms s_r and n are constants and can be removed. This results in $T_{pbis} = (\bar{r}_1 - \bar{r}_0)\sqrt{n_0 n_1}$. The square root is a monotone function $(n_0, n_1 > 0)$ and can also be removed, leading to

$$T_{pbis}(\mathbf{A}) = \frac{\sum r_1}{n_1} - \frac{\sum r_0}{n_0} = \frac{n_0 \sum r_1 - n_1 \sum r_0}{n_0 n_1} n_0 n_1 = n_0 \sum r_1 - n_1 \sum r_0. \tag{12}$$

Each summed raw score will thus be multiplied with the number of people in the other group (e.g., $n_0 \sum r_1$) and the difference between the two terms gives the test-statistic T_{pbis}. An item which shows a high discrimination should have a large value. Therefore, the model test is defined by Equation (5) and tests how many of the T_s have the same or a smaller value as in T_0.

3. Application to a real data set and comparison with parametric tests

To illustrate these tests with a practical example, we used a subset of the data to the topic of dyscalculia of Koller and Alexandrowicz (2010) who analyzed the neuropsychological test battery for number processing and calculation in children (ZAREKI-R; Von Aster, Weinhold Zulauf, & Horn, 2006). The authors checked the assumptions of the RM for the 16 subscales with a sample of 341 second- to fourth-grade children, applying parametric model tests.

At the global level Koller and Alexandrowicz (2010) used the LRT described by Andersen (1973), applying five different splitting criteria (raw score median, gender, class, time, and a random split). At the item level, they used a graphical model check, the Wald test (Wald, 1943), and χ^2-based item fit statistics (Wright & Stone, 1979).

In this study we reanalyzed the subtest "perceptive quantity estimation" (*k* = 5 items) with a subset of *n* = 221 second- to third-grade children (second grade: *n* = 122, *males* = 51; third grade: *n* = 99, *males* = 44) applying quasi-exact tests. Note, because the goal was to see if the quasi-exact tests agree with the parametric LRT we have chosen a reasonably sized sample with which it was possible to apply both methods.

In this subtest, the children have to estimate the number of objects which are presented separately for two seconds (first two items with black dots) or five seconds (the next two items with 3D objects) each. For the fifth item children have to compare the quantities of the two previously shown 3D objects (Question: "Were there more …?").

The analysis of showed that the RM does not fit the data. For the global test, the splitting criteria raw score median and gender showed significant violations of model fit. Furthermore, the results showed that item one was easier for girls, item four was easier for the higher scoring group, and item five was easier for the lower-scoring group. Also, the item fit statistic showed a significant result for item five, i.e., the item was more often solved by the lower scoring group than by the higher scoring group.

Koller and Alexandrowicz (2010) discussed some potential explanations of these model violations: Thus, there are differences particularly with item one and item two (black points). Faster children appeared to count some of the points (less than 15 points) and estimated the rest, whereas slower children could only use an estimation strategy. This led to two very different strategies, i.e., the items can be solved based on two different abilities. Item five has a different structure compared to the rest of the scale. The challenge of this item is to compare previously seen quantities and not to estimate quantities. Additionally, the question irritated children. The authors observed that the lower scoring group often gave incorrect responses to items three and four, but a correct response to item five and vice versa.

Furthermore, we detected local dependence between groups of items. The items one and two involve less than 15 objects and both include simple black points, and the items three and four includes both 3D objects with a similar alignment of objects. Additionally, we detected a violation of unidimensionality between the groups "black points" (1, 2) and "3D-objects" (3, 4).

3.1. Procedure

For the quasi-exact tests the MCMC-method was used with a burn-in period of 1,000, a step size of 64, and 1,000 effective matrices for each test. For each test a new starting value for the simulation of matrices was used to simulate different matrices for each statistic. First, the assumption of homogeneity and local independence between items were tested with T_{11} at the global level and with T_1 and T_{1m} at the item level. To test whether two item subsets were multidimensional T_{md} was used. We have chosen two splitting criteria. The median sum of the item scores (which splits the data set into a group of easier items and into a group of more difficult items) was used as internal criterion. The external criterion was based on the theoretical assumptions explained above. However, the first group of items were items one, two, and five (black dots and question) and the second group were the two 3D items. To test measurement invariance the global statistic T_{10} and T_4 for each item were used with three splitting criteria: raw score median, gender, and class. To test for the presence of item discrimination compatible with the RM we used T_{pbis}.

In the context of model testing we are particularly interested in not rejecting the null hypothesis (i.e., the RM fits the data). Thus a nominal significance level of $\alpha = 10\%$ was chosen. Therefore the type-II error plays an important role because it decreases if the type-I error increases. Furthermore, we investigated the fit of the RM with a series of tests (multiple testing). Thus, α was corrected using the Bonferroni method using the recommendations by Koller et al. (2012) who have formulated a possible way for dichotomous items (see Tab. 2).

Table 2. Corrected alpha (α*) for quasi-exact tests (Koller et al., 2012, p. 169)

Assumptions of model fit	*Tests*	corrected alpha
Measurement invariance	T_{10}	α/q
	T_4 (explorative)	α/q
	T_4	$\alpha/(q$*number of tests)
LD and/or heterogeneity	T_{11}	no correction
	T_1, T_{1m}, $T_{1\ell}$	α/number of groups of items
	T_2, T_{2m}	α/number of tests
Multidimensional subscales	T_{md}	α/q
Discrimination	T_{pbis}	α/k

Note: α … alpha; q … number of splitting criteria; k … number of items in the test; LD …local dependence; number of tests … e.g., a test consists of six items whereas three items were assumed to be easier for a specified group and only this hypothesis will be investigated. Therefore, the number of conducted tests is one; number of groups of items … with six items it is possible to investigate 15 groups of items.

3.2. Results: Model investigation with quasi-exact tests

The results showed that that the assumptions of local independence and/or homogeneity are violated (T_{11}: p <.001). Table 3 gives the results of T_1 and T_{1m}. We conclude that items one and two as well as items three and four violated the assumption of local independence (α^* <.01), which confirmed the theoretical considerations (see above). Furthermore, items three and five and items four and five showed a tendency for a violation of homogeneity between items. Additionally, the undimensionality of subscales was tested twice (α^* <.05). Both results showed that the subscales consisted of more than one dimension (T_{md} sum score: p = .032; T_{md} hypothesis: p < .001).

Table 3. Results (p-values) of the test-statistics T_1 and T_{1m}: local dependence and heterogeneity

Items	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
T_1	<.001	.944	.677	.797	.944	.944	.873	<.001	.996	.974
T_{1m}	>.999	.122	.505	.349	.052	.119	.233	>.999	.011	.044

Note. T_1 … tests the assumption of local independence between items; T_{1m} … tests the assumption of homogeneity between items; corrected alpha: α^* = .10/10 = .01

The assumption of measurement invariance was violated for the splitting criteria of raw score median and gender (see Tab. 4). On the level of items it can be seen, that item four was more difficult for the lower scoring group and item five was easier for the lower scoring group. In the splitting criterion gender the result showed that item one was easier for females than for males.

Table 4. Results (p-values) of the test-statistics T_{10} and T_4 for the investigation of measurement invariance.

Tests	T10	T4				
Splitting criteria/items	all	1	2	3	4	5
Score MD	<.001	.058*	.097	.944	.007*	<.001
Gender	.014	.007	.109	.859	.044*	.344
Grade	.250	.024	.801	.029	.651	.559

Notes: Score median (Score MD): 71 people solved less than four items and 150 people solved more than four items; gender: females = 126, males = 95; grade: 2nd grade = 122, 3rd grade = 99; T_4 … all tested items were analyzed whether or not there was an excess of ones. But, if the result was a quantile value near 1, then a second test which examined whether there was an excess of zeros was carried out. If this result gave more information about the item it was reported with an ^{x} by the value; corrected alpha: α^* = .10/3 = .033.

Finally, the discrimination of items was tested with T_{pbis} (α* <.02) and the results showed that only the fifth item showed a significant violation (p < .001; p-values of the remaining items range from .303 - .998).

4. Summary and Discussion

The present study discussed quasi-exact tests for the RM. So far, these tests are not well known and rarely used in practical applications. However, these tests allow the testing of the RM assumptions in small samples which is an important advantage compared to other well established model tests for the RM. For example, it is possible to investigate the item quality in complex experimental studies, in clinical studies, and in an earlier phase of construction of tests where samples are usually small.

However, only a few studies (e.g., Koller, 2010; Ponocny, 2001; Suárez-Falcón & Glas, 2003) investigated the test performance (type I error robustness and power behavior) of some tests (T_1, $T_{1\ell}$, T_{10}, T_4) . Based on the results of these studies it can be recommended that for tests using the whole sample (e.g., T_1, $T_{1\ell}$) the minimum number of 30 respondents should be chosen. For tests using split criteria (e.g., T_{10}, T_4) the minimum number of respondents should be increased by 30 times the number of subgroups. For two subgroups the minimum number would then be 60 (30 times 2). Furthermore, the results of these studies showed that even in small samples large model violations can be detected and that the power is increasing if the sample size and the number of items which show violations are increased.

The application on a reasonably large data set to the topic of dyscalculia showed that the analysis with quasi-exact tests led to the same results as the described parametric approach by Koller and Alexandrowicz (2010). For example, the reanalysis showed that item five should form another dimension (compare quantities) and should not be included in a test for estimating quantities.

However, quasi-exact tests are applicable not only for the investigation of RM conformity. For example, (Ponocny (2002)) used some of the test-statistics for the investigation of appropriateness of item response models for measuring change (e.g., undimensionality of change). Furthermore, it is possible to develop new test-statistics for other statistical investigations. For example, in this study we described a monotonically transformation of the well-known point biserial correlation which is originally not a test for the investigation of RM conformity.

The study gives an overview of several existing tests and describes new tests, gives detailed explanations of test-statistics, and describes some practical guidelines. Future studies should explore test performance (Type I error robustness and power behavior) using intensive Monte-Carlo simulation experiments.

Acknowledgments: We would like to thank Regina Dittrich, Brain Francis, and Wolfgang Wiedermann for their critical comments and ideas for improvements.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140.

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

Christensen, K. B., & Kreiner, S. (2010). Monte Carlo tests of the Rasch model based on scalability coefficients. *British Journal of Mathematical and Statistical Psychology, 63*(1), 101-111.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46*(1), 59-77.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & I. H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing Measurement Invariance Using MIMIC Likelihood Ratio Test With a Critical Value Adjustment. *Educational and Psychological Measurement, 72*(3), 469-492.

Koller, I. (2010). *Item response models in practise: Testing the assumptions in small samples and comparing different models for repeated measurements.* (unpublished doctoral thesis), University of Klagenfurt, Austria.

Koller, I., & Alexandrowicz, R. (2010). Eine psychometrische Analyse der ZAREKI-R mittels Rasch Modellen [A psychometric analysis of the ZAREKI-R using Rasch-models]. *Diagnostica, 56*(2), 57-67. doi: 10.1026/0012-1924/a000003

Koller, I., Alexandrowicz, R., & Hatzinger, R. (2012). *Das Rasch Modell in der Praxis: Eine Einführung mit eRm [The Rasch model in practical applications: An introduction with eRm]*. Vienna: facultas.wuv, UTB.

Mair, P., Hatzinger, R., & Maier, M. J. (2012). eRm: Extended Rasch Modeling. R package version 0.15.1. http://CRAN.R-project.org/package=eRm

Ponocny, I. (1996). Kombinatorische Modelltests für das Rasch-Modell [Combinatorial goodness-of-fit tests for the Rasch model]. *Unpublished doctoral dissertation, University of Vienna, Austria*.

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

Ponocny, I. (2002). On the applicability of some IRT models for repeated measurement designs: Conditions, consequences, and goodness-of-fit tests. *Methods of Psychological Research Online, 7*(1), p22-40.

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

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. *Kopenhagen: Danish Institute for Educational Research*.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*(3), 425-435.

Snijders, T. A. (1991). Enumeration and simulation methods for 0–1 matrices with given marginals. *Psychometrika, 56*(3), 397-417.

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

Verguts, T., & Boeck, P. (2001). Some Mantel-Haenszel tests of Rasch model assumptions. *British Journal of Mathematical and Statistical Psychology, 54*(1), 21-37.

Verhelst, N. D. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika, 73*(4), 705-728.

Verhelst, N. D., Hatzinger, R., & Mair, P. (2007). The Rasch sampler. *Journal of Statistical Software, 20*(4), 1-14.

Von Aster, M., Weinhold Zulauf, M., & Horn, R. (2006). *Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern (ZAREKI-R) [Neuropsychological Test Battery for Number Processing and Calculation in Children]*. Frankfurt am Main: Harcourt Test Services.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54*(3), 426-482.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago, IL: Mesa Press.