Müller *Journal of Statistical Distributions and Applications* (2020) 7:5
https://doi.org/10.1186/s40488-020-00108-7

Journal of Statistical Distributions and Applications

RESEARCH Open Access

Item fit statistics for Rasch analysis: can we trust them?

Marianne Müller

Correspondence:
marianne.mueller@zhaw.ch
School of Engineering, Zurich University of Applied Sciences, Rosenstr. 3, 8400 Winterthur, Switzerland

Abstract

Aim: To compare fit statistics for the Rasch model based on estimates of unconditional or conditional response probabilities.

Background: Using person estimates to calculate fit statistics can lead to problems because the person estimates are biased. Conditional response probabilities given the total person score could be used instead.

Methods: Data sets are simulated which fit the Rasch model. Type I error rates are calculated and the distributions of the fit statistics are compared with the assumed normal or chi-square distribution. Parametric bootstrap is used to further study the distributions of the fit statistics.

Results: Type I error rates for unconditional chi-square statistics are larger than expected even for moderate sample sizes. The conditional chi-square statistics maintain the significance level. Unconditional outfit and infit statistics have asymmetric distributions with means slighly below 1. Conditional outfit and infit statistics have reduced Type I error rates.

Conclusions: Conditional residuals should be used. If only unconditional residuals are available parametric bootstrapping is recommended to calculate valid *p*-values. Bootstrapping is also necessary for conditional outfit statistics. For conditional infit statistics the adjusted rule-of-thumb critical values look useful.

Keywords: Rasch model, Chi-square test statistics, Outfit and infit statistics, Conditional probability

Introduction

Rasch models are increasingly used for the examination and development of measurement instruments in the health and psychological sciences (Belvedere and de Morton 2010; Bond and Fox 2015). They facilitate the detection of measurement problems like item bias or local dependence that may be overseen by traditional validation methods such as factor analysis and Cronbach's alpha coefficient. If the data from a questionnaire fit to the model expectations, a transformation of the ordinal score into an interval-level variable is available. To achieve all this, rigorous tests are essential, because the Rasch model makes some strong assumptions on the item response process.

To assess whether individual items fit the Rasch model fit statistics are widely used. Software like Winsteps (Linacre 2019), DIGRAM (Kreiner and Nielsen 2013) or the R packge

eRm (Mair et al. 2019) calculate infit and outfit mean squares. RUMM2030 (Andrich et al. 2010) uses item and total item-trait interaction chi-square statistics. To detect misfitting items the fit statistics are either compared to rule-of-thumb critical values or transformed to test statistics which can be compared with the values of the purported distribution. A few questions arise immediately: what are suitable critical values for sound decisions, are the distributional assumptions justified and what happens if the sample size increases? There are many different guidelines for acceptable ranges for mean squares and different recommendations as to which approach should be chosen when sample size is large.

The behaviour of chi-square statistics has not been widely tested. Hagell and Westergren (2016) have shown that Type I errors increase for $n \geq 500$. They studied situations with 25 items and sample sizes up to 2500, but conducted only one simulation for each situation. Other approaches to deal with sample size issues are drawing smaller random samples from a large sample or use an algebraic adjustment of the sample size before calculating p-values. These two procedures have been implemented in RUMM2030. Bergh (2015) compared the two approaches with each other. He found that for original sample sizes up to 21 000 and adjustments to sample sizes of 5000 both procedures work equally well. For adjustments to smaller sizes, the algebraic adjustment approach appeared less effective than random samples.

Simulation studies for outfit and infit statistics have shown several weaknesses. Means are not equal to the expected value of 1, distributions are asymmetric with extreme values more often occuring above 1, simple rule-of-thumb critical values for acceptable fit may be inappropriate (Smith 1991; Smith et al. 1998; Wang and Chen 2005). Wolfe (2013) examined the distributions of outfit and infit statistics under a limited number of conditions, and based on these results recommended bootstrapping to get adequate critical values.

There are two main problems when calculating and using fit statistics: the estimation of the residuals and the distribution of the fit statistics. All item fit statistics summarize standardized residuals which are based on estimates of response probabilities. Item and person parameter estimates are usually utilized here. Item estimates are consistent but person estimates are not. The latter are biased and the bias does not disappear with increased sample size. Due to the correspondence between person estimates and scores, estimates of conditional response probabilities given the total person score could be used to virtually eliminate the bias (Kreiner and Christensen 2011).

Because the exact distributions of the fit statistics are unknown for unconditional and conditional estimates, asymptotic distributions are used. It is unclear how reliable these approximations are. In this paper, we compare chi-square as well as outfit and infit statistics based on unconditional and conditional estimation procedures. Furthermore, bootstrap simulations are used to understand the distributions of these statistics.

Background

The Rasch model for dichotomous items (Rasch 1960) assumes that the response of a person to an item is stochastically independent of all other item responses for the same and other persons, and that the probability of a positive response to an item is equal to

$$P(X_{vi}=1)=\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)},\quad v=1,\ldots,n;\ i=1,\ldots,k. \tag{1}$$

where θ_v is a parameter characterizing the person v and β_i is an item parameter.

Two different types of residuals are calculated during tests of fit of items to the Rasch model. Response residuals compare observed and expected for every combination of person and item. These residuals are given by

$$Z_{vi}=\frac{X_{vi}-E(X_{vi})}{\sqrt{Var(X_{vi})}}, \tag{2}$$

whith $E(X_{vi})=P(X_{vi}=1)=p_{vi}$ and $Var(X_{vi})=p_{vi}(1-p_{vi})$. Outfit and infit statistics are calculated as means of the squared residuals.

$$\text{Outfit}_i=\sum_v Z_{vi}^2/n,\qquad \text{Infit}_i=\frac{\sum_v Z_{vi}^2\cdot w_{vi}}{\sum_v w_{vi}}. \tag{3}$$

The weights w_{vi} used to calculate the infit statistics are equal to $Var(X_{vi})$. Two approaches are used to assess item fit: rule-of-thumb critical values for the mean squares or formal tests by dividing the mean squares by their standard errors and comparing the resulting test statistics with the normal distribution. The Wilson-Hilferty cube root transformation can be used to improve the approximation of a chi-square variable to the normal distribution (Wilson and Hilferty 1931).

Rule-of-thumb lower and upper limits for acceptable mean square fit values have been set by many researchers to 0.7 and 1.3. Linacre (2017) gave a detailed instruction on cut-off numbers suggesting values between 0.5 and 1.5 as acceptable. Adjustments have been proposed which take into account the sample size n. Smith et al. (1998) recommend critical values equal to $1\pm 6/\sqrt{n}$ for outfits and $1\pm 2/\sqrt{n}$ for infits. Unfortunately, simulation studies have shown that appropriate critical values also depend on the number of items and the difficulty of the item considered (Wang and Chen 2005).

A second type of residuals used are group residuals. They compare the total number of positive responses to an item in a group of persons to the expected number of responses in the same group. Item chi-square fit statistics are calculated as the sum of squared group residuals, where persons are grouped into class intervals g depending on their scores.

$$X_i^2=\sum_g\left[\frac{\sum_{v\in g}X_{vi}-\sum_{v\in g}E(X_{vi})}{\sqrt{\sum_{v\in g}Var(X_{vi})}}\right]^2 \tag{4}$$

The total "item-trait interaction" chi-square test statistic is the sum of the item chi-squares.

$$X^2=\sum_i X_i^2,\qquad df=k\cdot df_i \tag{5}$$

It is assumed that the test statistic defined in (4) follows a chi-square distribution with degrees of freedom df_i equal to the number of class intervals minus 1, and that the total "item-trait interaction" follows a chi-square distribution with $k\cdot df_i$ degrees of freedom. This test should show whether the data fit to the Rasch model for the classes along the scale.

The formulas for the fit statistics involve $E(X_{vi})$ and $Var(X_{vi})$ which are unknown and have to be estimated. The usual way to estimate $E(X_{vi})$ is to plug in the item and person parameter estimates. We call this the unconditional estimate leading to unconditional fit statistics.

$$\hat{E}(X_{vi}) = \hat{P}(X_{vi} = 1) = \frac{\exp\left(\hat{\theta}_v - \hat{\beta}_i\right)}{1 + \exp\left(\hat{\theta}_v - \hat{\beta}_i\right)} \tag{6}$$

Kreiner and Christensen (2011) showed that using person parameter estimates for the estimation of response probabilities lead to biased residuals and therefore biased outfit statistics. It is actually not needed to plug in person estimates, conditional estimates could be used instead. They are given by

$$\hat{E}\left(X_{vi}\right) = \hat{P}\left(X_{vi} = 1 | R_v = r\right) = \frac{exp\left(-\hat{\beta}_i\right)\gamma_{r-1}\left(\hat{\boldsymbol{\beta}}^{(i)}\right)}{\gamma_r\left(\hat{\boldsymbol{\beta}}\right)}, \tag{7}$$

where $\hat{\boldsymbol{\beta}}$ denotes the vector of item parameters, $\hat{\boldsymbol{\beta}}^{(i)}$ denotes the vector of item parameters without $\hat{\beta}_i$ and $\gamma_r(\hat{\boldsymbol{\beta}})$ is the elementary symmetrical function of order r of the item parameter estimates. The fit statistics based on (7) are called conditional fit statistics. Full details of the calculations can be found in Christensen and Kreiner (2013).

The widely used Rasch software Winsteps and the R package eRm calculate unconditional outfits and infits (3) based on (6). Other R packages such as mirt, LTM and irtoys use the same estimation approach. RUMM2030 estimates unconditional chi-square fit statistics (see (4) and (5)) also relying on (6). Only DIGRAM estimates conditional outfit and infit statistics utilizing (7). Most programs use chi-square or normal distributions for goodness of fit tests. Only LTM and DIGRAM allow simulations to get p-values. In this paper we want to answer the following questions: what are the consequences of using biased estimates and inadequate distributional assumptions and for which sample size do problems become serious?

Methods

Data sets were simulated to fit the Rasch model, with sample sizes between 150 and 10 000, and 10, 15 or 20 items. Person parameters came from a standard normal distribution. Item parameters were also chosen from a normal distribution or equidistantly fixed, ranging from −2 to +2 or from −2.5 to +2.5. Different conditions were studied because the range and the distribution of item parameters could have an effect on the results. Item parameters were estimated with conditional maximum likelihood, for the person parameters weighted maximum likelihood was used. Unconditional and conditional outfit, infit and chi-square fit statistics were calculated. The distributions of the p-values were then examined and proportions of p-values below 0.05, the type I error rates, were compared.

Parametric bootstrap was used to study the distributions of the fit statistics for various n (sample size) and k (number of items). First, a data set was generated from a Rasch model. Item and person parameters were estimated for this data set and fit statistics and p-values were calculated. Next, bootstrap samples were generated from a Rasch model with parameters equal to the estimates calculated in the first step. Fit statistics for the samples were used to get their empirical distribution. This is called the bootstrap distribution. The critical value from this distribution gives the bootstrap p-value, which was compared with the p-value based on the normal or the chi-square distribution.

All simulations and calculations were done with the R statistical package (R Core Team 2019) and the additional R packages eRm (Mair et al. 2019) and PP (Reif and Steinfeld 2019).

Results

Chi-square fit statistics

We start with simulations with normally distributed person parameters and fixed item parameters, equidistant in the interval [−2, +2]. Sample sizes vary between 150 and 2000, the number of items is 10, 15 or 20. The number of class intervals is three for $n \leq 250$, and seven for larger n. For each situation 1000 simulations are done. As the Rasch model is true for all the data sets, the proportions of p-values below 0.05 should be about 0.05. Table 1 shows the proportions of p-values below 0.05 for unconditional and conditional individual item and total tests. Let us look first at the unconditional tests. For n=200 and k=10, we have a proportion of 0.107 of p-values below 0.05. The proportion even reaches 1 for n $\geq$ 1500 and k=10. There are therefore too many significant results regarding the total test for $n \geq 200$, especially if there are not many items. For $n \geq 500$, the type I error rate is also increased for single item tests. The proportions vary between 0.047 and 0.119 for n=500 and k=10. Hence, the chi-square statistics appear to be okay for some items, but not for all. The conditional total and single item tests maintain the significance level.

Higher proportions can be found for unconditional tests for items located in the center. Figure 1a shows that the distribution of p-values for item 1 looks more or less uniform, whereas the same distribution for item 5 is skewed (Fig. 1b). In the case of conditional tests, the distributions look uniform as they should (Fig. 1c and d). As the type I error rates for the unconditional total tests are more affected than the error rates for single item test, the distributions of p-values for unconditional total tests are even more skewed. Simulations with fixed item parameters equidistant in the interval [−2.5, +2.5] or normally distributed item parameters lead to very similar results.

Table 1 Proportions of p-values < 0.05 for chi-square statistics

		Unconditional		Conditional	
n	k	items	total	items	total
150	10	0.035–0.087	0.074	0.043–0.053	0.058
	15	0.030–0.065	0.057	0.032–0.061	0.058
	20	0.032–0.060	0.051	0.036–0.056	0.062
200	10	0.031–0.092	0.107	0.043–0.050	0.050
	15	0.027–0.062	0.065	0.040–0.061	0.055
	20	0.030–0.063	0.053	0.036–0.068	0.059
250	10	0.031–0.099	0.153	0.036–0.051	0.047
	15	0.029–0.059	0.067	0.034–0.056	0.046
	20	0.031–0.061	0.064	0.039–0.062	0.062
500	10	0.047–0.119	0.228	0.043–0.064	0.059
	15	0.036–0.077	0.087	0.042–0.062	0.051
	20	0.036–0.059	0.055	0.036–0.062	0.050
1000	10	0.104–0.252	0.880	0.044–0.060	0.064
	15	0.047–0.110	0.305	0.036–0.059	0.057
	20	0.041–0.093	0.156	0.040–0.069	0.052
1500	10	0.190–0.416	1.000	0.036–0.057	0.053
	15	0.061–0.156	0.613	0.043–0.061	0.052
	20	0.048–0.112	0.304	0.037–0.067	0.048
2000	10	0.335–0.592	1.000	0.040-0.065	0.058
	15	0.083–0.226	0.907	0.036–0.060	0.058
	20	0.055–0.130	0.505	0.034–0.057	0.054

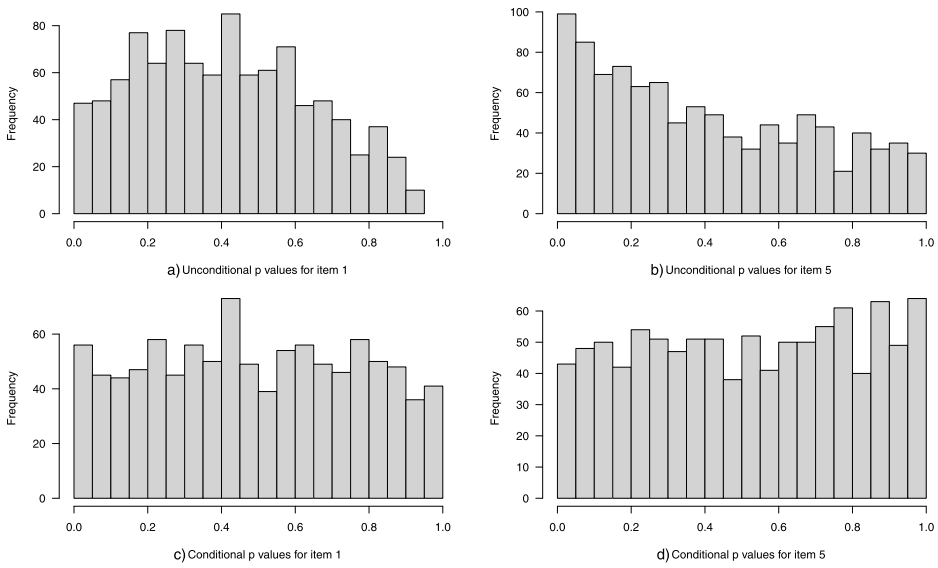

Fig. 1 Distribution of p-values for unconditional and conditional tests of an extreme item (item 1) and an item located in the center (item 5) with $n = 500$ and $k = 10$

The distributions of the simulated unconditional test statistics coincide with the proposed chi-square distributions as long as $n < 500$. For larger n the discrepancy between empirical and assumed distribution becomes larger and larger. Because the empirical distribution is shifted to the right, the type I errors increase. For $n \geq 1000$ this affects items independently of their location. Figure 2 shows the unconditional (a-c) and conditional (d-f) distributions for item 5 with $n = 500$ and item 1 with $n = 1000$ and $n = 2000$, $k = 10$. Chi-square density curve and histogram agree for the conditional test statistic in all situations.

Table 2 contains unconditional and conditional chi-square fit statistics, and p-values based on the chi-square distribution as well as based on the bootstrapping procedure with $n = 1000$, $k = 10$ and fixed item parameters in the interval [-2, +2]. The two unconditional item fit statistics for item 5 and item 6 show misfit and the total test also rejects the Rasch model if the chi-square distribution is used. The bootstrap p-values are larger and do not indicate any misfit. As for the conditional tests, the chi-square distribution and bootstrap p-values are quite similar.

Outfit and infit statistics

Simulations are done again with normally distributed person parameters and fixed item parameters, equidistant in the interval [−2, +2]. Sample sizes vary between 150 and 10 000, the number of items is 10, 15 or 20. For each situation 1000 simulations are done. The Wilson-Hilferty transformation has been used for the unconditional fit statistics, but not for the conditional values because there was no apparent improvement of the approximation to the normal distribution. Table 3 shows mean values of outfit and infit statistics, and the proportions of p-values below 0.05. The unconditional outfit and infit statistics are biased, their means are smaller than the expected mean of 1. The size of the bias depends on the number of items. Type I error rates are increased for unconditional statistics if $n \geq 500$, especially if there are not many items. Mean values are okay for the conditional

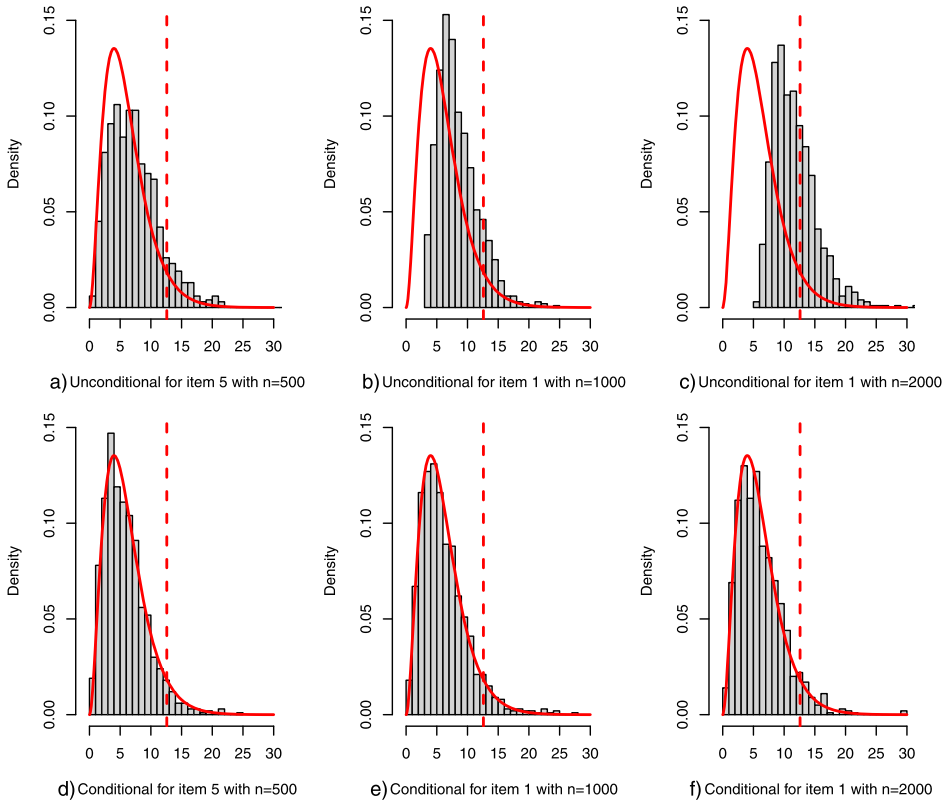

Fig. 2 Distribution of the unconditional and conditional test statistic. The red curve is the proposed chisquare distribution, the red vertical line is the 95% percentile. Values above this critical value will lead to rejecting the null hypothesis

tests, but the error rates appear to be too small. This is particularly true for items with large or small difficulties (see Fig. 3).

Table 4 contains 2.5 and 97.5% percentiles for outfit and infit statistics. Ranges get more narrow for unconditional and conditional estimates, but unconditional fit statistics are not symmetric around 1. The critical values 0.5–1.5 proposed by Linacre are only valid for very small sample sizes and few items ($n \leq 150$, $k \leq 10$). The usual rule-of-thumb of

Table 2 Comparison of chi-square and bootstrap *p*-values with $n = 1000$

	Unconditional			Conditional		
	X_i^2	p-Chi	p-Boot	X_i^2	p-Chi	p-Boot
Item 1	4.642	0.591	0.891	1.627	0.951	0.953
Item 2	4.793	0.571	0.863	2.302	0.890	0.896
Item 3	5.471	0.485	0.816	2.416	0.878	0.873
Item 4	5.754	0.451	0.789	4.861	0.562	0.528
Item 5	13.752	0.033	0.181	4.422	0.620	0.612
Item 6	14.734	0.022	0.154	6.483	0.371	0.348
Item 7	7.290	0.295	0.674	11.559	0.073	0.073
Item 8	10.436	0.107	0.347	4.641	0.591	0.578
Item 9	9.568	0.144	0.346	7.659	0.264	0.256
Item 10	6.391	0.381	0.633	3.100	0.796	0.785
Total	82.830	0.027	0.770	49.070	0.842	0.815

Table 3 Mean values and proportions of p-values < 0.05 for unconditional and conditional outfits and infits

		Unconditional				Conditional			
n	k	Out	p < 0.05	In	p < 0.05	Out	p < 0.05	In	p < 0.05
150	10	0.91	0.014–0.062	0.94	0.018–0.077	1.00	0.033–0.052	1.00	0.000–0.052
	15	0.95	0.010–0.040	0.96	0.006–0.057	1.00	0.021–0.046	1.00	0.000–0.054
	20	0.96	0.011–0.049	0.97	0.001–0.057	1.00	0.021–0.053	1.00	0.000–0.052
200	10	0.91	0.020–0.070	0.94	0.022–0.082	1.00	0.028–0.048	1.00	0.000–0.051
	15	0.94	0.013–0.043	0.96	0.007–0.064	1.00	0.021–0.047	1.00	0.000–0.050
	20	0.96	0.009-0.053	0.97	0.000-0.057	1.00	0.020–0.055	1.00	0.000–0.060
250	10	0.91	0.023–0.060	0.94	0.047–0.079	1.00	0.023-0.048	1.00	0.000–0.044
	15	0.95	0.016–0.048	0.96	0.006–0.058	1.00	0.022–0.044	1.00	0.000–0.038
	20	0.96	0.013–0.055	0.97	0.002–0.066	1.00	0.017–0.053	1.00	0.000–0.055
500	10	0.91	0.058–0.109	0.94	0.110–0.148	1.00	0.026–0.041	1.00	0.000–0.044
	15	0.95	0.022–0.063	0.96	0.022–0.063	1.00	0.024–0.046	1.00	0.000–0.046
	20	0.96	0.019–0.062	0.97	0.008–0.067	1.00	0.019–0.052	1.00	0.000–0.056
1000	10	0.91	0.149–0.190	0.94	0.236–0.403	1.00	0.018–0.043	1.00	0.000–0.046
	15	0.95	0.058–0.080	0.96	0.086–0.115	1.00	0.021–0.054	1.00	0.000–0.047
	20	0.96	0.029–0.067	0.97	0.034–0.081	1.00	0.020–0.055	1.00	0.000–0.054
1500	10	0.91	0.223–0.293	0.94	0.319–0.672	1.00	0.015–0.053	1.00	0.000–0.060
	15	0.95	0.093–0.123	0.96	0.128–0.237	1.00	0.013–0.050	1.00	0.000–0.046
	20	0.96	0.053–0.091	0.97	0.069–0.110	1.00	0.021–0.060	1.00	0.000–0.067
2000	10	0.91	0.319–0.402	0.94	0.442–0.839	1.00	0.020–0.043	1.00	0.000–0.048
	15	0.95	0.123–0.160	0.96	0.166–0.352	1.00	0.021–0.045	1.00	0.000–0.049
	20	0.96	0.065–0.104	0.97	0.090–0.139	1.00	0.021–0.058	1.00	0.000–0.060
3000	10	0.91	0.452–0.563	0.94	0.585–0.975	1.00	0.020–0.052	1.00	0.000–0.050
	15	0.95	0.152–0.254	0.96	0.208–0.621	1.00	0.012–0.048	1.00	0.000–0.049
	20	0.96	0.090–0.145	0.97	0.110–0.278	1.00	0.018–0.049	1.00	0.000–0.048
10000	10	0.91	0.930–0.980	0.94	0.981–1.000	1.00	0.016–0.045	1.00	0.000–0.045
	15	0.95	0.490–0.749	0.96	0.629–1.000	1.00	0.017–0.056	1.00	0.000–0.061
	20	0.96	0.256–0.501	0.97	0.337–0.976	1.00	0.018–0.068	1.00	0.000–0.058

0.7–1.3 is valid for n around 200, whereas the adjusted crical values proposed by Smith et al. (1998) fit quite well for conditional infits over the range of sample sizes considered.

Next, histograms of the standardized outfit and infit values are compared with the standard normal distribution. The deviation from the expected mean for unconditional outfit values is obvious in Fig. 4 (upper row). The Wilson-Hilferty transformation makes the approximation even worse for larger n. The conditional statistics are unbiased but there are some large outliers and too many values in the center, the calculated standard errors seem to be too large. Standard errors are also too large for unconditional outfit statistics. This can be seen if the outfits are centered around zero. For items with small ore large difficulties the situation is more extreme.

Parametric bootstrap is expected to produce smaller p-values for conditional statisics as the test based on the normal approximation. This is verified in Table 5 for the conditional outfit statistics with $n = 2000$ and $k = 10$.

Discussion

Residual-based fit statistics are widely used to assess Rasch model fit. At the same time, there are concerns about the quality of these indicators. Some people argue completely

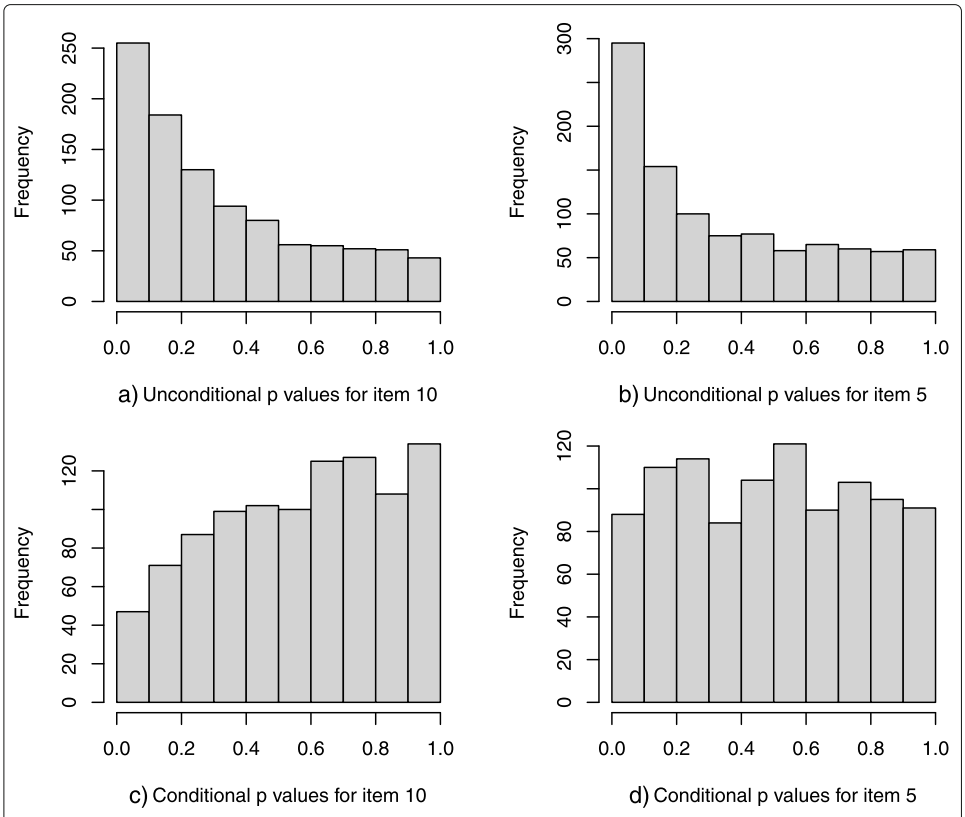

Fig. 3 Distribution of *p*-values for unconditional and conditional outfit statistics of difficult item (item 10) and an item located in the center (item 5) with $n = 1000$ and $k = 10$

against the use of any residual fit statistic (Karabatsos 2000), others have developped alternatives such as likelihood-based fit statistics or graphical approaches to assess item fit (Orlando and Thissen 2000; Yu 2020). There is no doubt that other approaches can give valuable information about possible problems of items. Nevertheless, this paper is focused on residual-based fit statistics because they are so popular and we would like to help to improve their usage.

Two different estimates of residuals are considered, one based on biased and not consistent person parameters, the other based on scores. For a sample size of 200 or more, the unconditional total item-trait interaction chi-square test which uses the person parameter estimates shows increased Type I error rates. Unconditional single item chi-square statistics become unreliable for $n \geq 500$. This is in accordance with the results of Hagell and Westergren (2016), but is now supported by many more simulations. The usually assumed chi-square approximations are inadequate and the parametric bootstrap confirms these results. In the case of unconditional tests, chi-square *p*-values are much smaller than bootstrap *p*-values. So this means that even for moderate sample sizes, the Rasch model is rejected too often and/or too many items falsely show misfit.

The conditional chi-square tests which are based on scores remain valid. The proportion of *p*-values below 0.05 is close to 0.05 and the chi-square distribution and the bootstrap distribution quite agree for conditional tests. As long as there are no conditional fit statistics implemented in RUMM2030, users have to be careful to not misinterpret seemingly significant results unless the sample size is small.

Table 4 2.5 — 97.5% Percentiles for unconditional and conditional outfits and infits

		Outfit			Infit		
n	k	uncond	cond	$1 \pm 6/\sqrt{n}$	uncond	cond	$1 \pm 2/\sqrt{n}$
150	10	0.60–1.33	0.67–1.55	0.51–1.49	0.78–1.10	0.84–1.17	0.84–1.16
	15	0.65–1.33	0.70–1.46		0.81–1.12	0.85–1.16	
	20	0.69–1.34	0.72–1.43		0.83–1.12	0.86–1.15	
200	10	0.64–1.27	0.70–1.46	0.56–1.42	0.80–1.07	0.86–1.15	0.86–1.14
	15	0.69–1.27	0.73–1.39		0.83–1.09	0.87–1.14	
	20	0.71–1.28	0.75–1.36		0.85–1.10	0.88–1.13	
250	10	0.65–1.21	0.73–1.40	0.62–1.38	0.81–1.06	0.88–1.13	0.87–1.13
	15	0.71–1.24	0.75–1.34		0.84–1.08	0.88–1.12	
	20	0.74–1.24	0.77–1.32		0.86–1.09	0.89–1.12	
500	10	0.71–1.11	0.80–1.26	0.73–1.27	0.84–1.03	0.91–1.09	0.91–1.09
	15	0.77–1.15	0.82–1.25		0.87–1.05	0.92–1.09	
	20	0.79–1.16	0.83–1.22		0.89–1.05	0.92–1.08	
1000	10	0.76–1.05	0.85–1.18	0.81–1.19	0.86–1.01	0.94–1.06	0.94–1.06
	15	0.81–1.08	0.87–1.16		0.89–1.02	0.94–1.06	
	20	0.83–1.10	0.88–1.16		0.91–1.03	0.94–1.06	
1500	10	0.77–1.02	0.87–1.14	0.84–1.15	0.87–1.00	0.95–1.05	0.95–1.05
	15	0.82–1.05	0.89–1.13		0.90–1.01	0.95–1.05	
	20	0.85–1.07	0.89–1.13		0.92–1.02	0.95–1.05	
2000	10	0.75–1.04	0.89–1.12	0.86–1.13	0.87–0.99	0.96–1.05	0.96–1.04
	15	0.84–1.04	0.89–1.13		0.91–1.01	0.96–1.04	
	20	0.86–1.05	0.91–1.11		0.92–1.02	0.96–1.04	
3000	10	0.81–0.99	0.91–1.10	0.89–1.11	0.88–0.99	0.96–1.04	0.96–1.04
	15	0.85–1.02	0.92–1.09		0.91–1.00	0.97–1.03	
	20	0.87–1.03	0.92–1.09		0.93–1.02	0.97–1.03	
10 000	10	0.83–0.97	0.95–1.05	0.94–1.06	0.89–0.97	0.98–1.02	0.98–1.02
	15	0.88–0.99	0.95–1.05		0.92–0.99	0.98–1.02	
	20	0.90–1.00	0.96–1.05		0.94–1.00	0.98–1.02	

The unconditional outfit and infit statistics have means slightly smaller than the expected value of 1, at least if the number of items is small. Type I error rates are increased for unconditional statistics if $n \geq 500$. Therefore, too many items are regarded as misfitting or the Rasch model as a whole is falsely rejected. Other authors have also noticed these problems a long time ago (Smith 1991; Wang and Chen 2005).

Mean values are okay for the conditional tests, but the error rates appear to be too small. The calculated standard errors are too large and therefore the standardized values become too small. The reason is that the squared residuals used in Eq. (3) are not independent as assumed. Correlations tend to be negative especially for items with large or small difficulties. The resulting true variance is therefore smaller than the estimated variance. The adjusted rule-of-thumb ($1 \pm 2/\sqrt{n}$) appears to be reasonable if applied to conditional infit statistics, whereas the adjusted rule-of-thumb for outfit statistics ($1 \pm 6/\sqrt{n}$) does not seem to be valid. As Winsteps and the R package eRm only calculate unconditional outfit and infit statistics, their results can become unreliable for sample sizes above 250. The R package iarm (Müller 2020) can be used to estimate conditional fit statistics which have correct mean values and to apply bootstrapping for the p-values. DIGRAM estimates conditional fit statistics as well. The user should also apply the recently implemented bootstrapping procedure to get p-values not relying on invalid distributional assumptions.

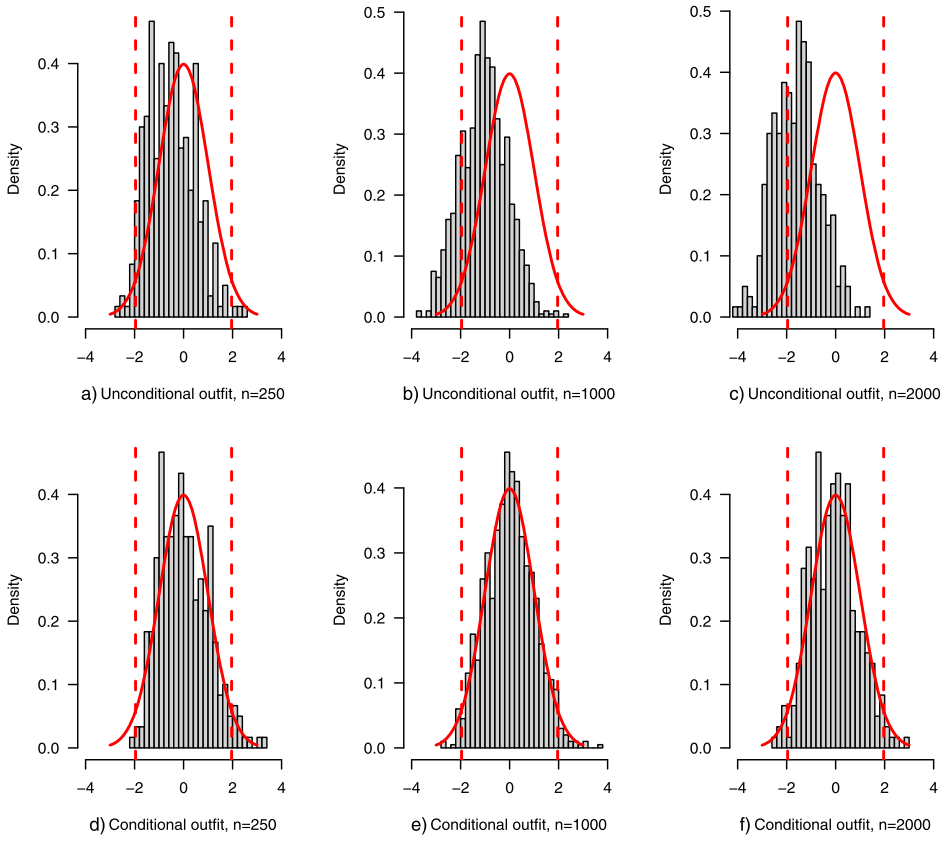

Fig. 4 Distribution of the unconditional and conditional standardized outfit statistic for item 5. The red curve is the standard normal distribution, the red vertical lines are the 2.5% and 97.5% percentiles. Values outside these critical values will lead to rejecting the null hypothesis

If only infit statistics are relevant, the adjusted rule-of-thumb given by Smith et al. (1998) could be used instead.

Conclusions

It is time to update the Rasch software. Large chisquare fit statistics are not just a matter of large sample sizes, problems start with n as small as 200. It is therefore crucial to use conditional estimates. The chisquare distribution can then be used as an approximation for the distribution of the test statistic. As for outfit and infit statistics, standard error calculations are not reliable. Parametric bootstrap should be used to get correct *p*-values.

Table 5 Normal approximation and bootstrap *p*-values for conditional outfits, $n = 2000$

	p-Normal	p-Boot
Item 1	0.247	0.147
Item 2	0.860	0.803
Item 3	0.265	0.241
Item 4	0.724	0.695
Item 5	0.449	0.460
Item 6	0.478	0.442
Item 7	0.196	0.141
Item 8	0.410	0.337
Item 9	0.322	0.228
Item 10	0.129	0.102

Acknowledgements
Not applicable.

Authors' contributions
All contributions were made bei MM. The author read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The author declares that she has no competing interests.

Received: 6 May 2020 Accepted: 9 August 2020
Published online: 28 August 2020

References

Andrich, D., Sheridan, B., Luo, G.: RUMM 2030. Rasch unidimensional measurement models software. RUMM Laboratory, Perth, WA, Australia (2010)
Belvedere, S., de Morton, N.: Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. J. Clin. Epidemiol. **63**, 1287–1297 (2010)
Bergh, D.: Chi-squared test of fit and sample size- a comparison between random sample approach and a chi-square value adjustment method. J. Appl. Meas. **16**, 204–217 (2015)
Bond, T. G., Fox, C. M.: Applying the Rasch model: Fundamental measurement in the human sciences. 3rd edition. Routledge, New York (2015)
Christensen, K., Kreiner, S.: Item fit statistics. In: Christensen, K., Kreiner, S., Mesbah, M. (eds.) Rasch Models in Health. Wiley, (2013)
Hagell, P., Westergren, A.: Sample size and statistical conclusions from tests of fit to the Rasch model according to the Rasch unidimensional measurement model (RUMM) program in health outcome measurement. J. Appl. Meas. **17**(4), 416–431 (2016)
Karabatsos, G.: A critique of Rasch residual fit statistics. J. Appl. Meas. **1**, 152–176 (2000)
Kreiner, S., Christensen, K. B.: Exact evaluation of bias in Rasch model residuals. In: Baswell, A. R. (ed.) Advances in Mathematics Research, pp. 19–40. Nova Science Publishers, Inc., (2011)
Kreiner, S., Nielsen, T.: Item analysis in DIGRAM: guided tours. Research Report 13/06, Copenhagen: Department of Biostatistics, University of Copenhagen (2013)
Linacre, M.: Teaching Rasch measurement. Trans. Rasch Meas. **31**, 1630–1631 (2017)
Linacre, J. M.: Winsteps®Rasch measurement computer program. Beaverton, Oregon: Winsteps.com (2019)
Mair, P., Hatzingen, R., Maier, M.: eRm: Extended Rasch Modeling. R package version 1.0-0 (2019)
Müller, M.: iarm: Item Analysis in Rasch Models. R package version 0.4.1 (2020)
Orlando, M., Thissen, D.: Likelihood-based item-fit indices for dichotomous item response theory models. Appl. Psychol. Meas. **24**(1), 50–64 (2000)
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019)
Rasch, G.: Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago (1960)
Reif, M., Steinfeld, J.: PP: Estimation of person parameters for the 1,2,3,4-PL model and the GPCM. R package version 0.6.2, 15 (2019)
Smith, R. M.: The distributional properties of Rasch item fit statistics. Educ. Psychol. Meas. **51**, 541–565 (1991)
Smith, R., Schumacker, R., Bush, M.: Using item mean squares to evaluate fit to the Rasch model. J. Outcome Meas. **2**, 66–78 (1998)
Wang, W., Chen, C.: Item parameter recovery, standard error estimates, and fit statistics of the winsteps program for the family of Rasch models. Educ. Psychol. Meas. **65**, 376–404 (2005)
Wilson, E., Hilferty, M.: The distribution of chi-square. Proc. Natl. Acad. Sci. U. S. A. **17**, 684–688 (1931)
Wolfe, E.: A bootstrap approach to evaluation person and item fit to the Rasch model. J. Appl. Meas. **14**(1), 1–9 (2013)
Yu, C.: Objective measurement: How Rasch modeling can simplify and enhance your assessment. In: Khine, M. (ed.) Rasch measurement: Applications in quantitative educational research, pp. 47–73. Springer, Singapore, (2020)

Publisher's Note