JOURNAL OF APPLIED MEASUREMENT, *16*(2),

Chi-Squared Test of Fit and Sample Size— A Comparison between a Random Sample Approach and a Chi-Square Value Adjustment Method

Daniel Bergh
Karlstad University

Chi-square statistics are commonly used for tests of fit of measurement models. Chi-square is also sensitive to sample size, which is why several approaches to handle large samples in test of fit analysis have been developed. One strategy to handle the sample size problem may be to adjust the sample size in the analysis of fit. An alternative is to adopt a random sample approach. The purpose of this study was to analyze and to compare these two strategies using simulated data.

Given an original sample size of 21,000, for reductions of sample sizes down to the order of 5,000 the adjusted sample size function works as good as the random sample approach. In contrast, when applying adjustments to sample sizes of lower order the adjustment function is less effective at approximating the Chi-square value for an actual random sample of the relevant size. Hence, the fit is exaggerated and misfit under-estimated using the adjusted sample size function. Although there are big differences in Chi-square values between the two approaches at lower sample sizes, the inferences based on the *p*-values may be the same.

Requests for reprints should be sent to Daniel Bergh, Karlstad University, Centre for Research on Child and Adolescent Mental Health, SE-651 88 Karlstad, Sweden, e-mail: daniel.bergh@kau.se.

Introduction

Social and medical scientists sometimes rely on *p*-values rather than effect sizes or descriptions by other means (Lantz, 2013) when reporting and interpreting statistical analyses, which is also encouraged by journals arguing that they have limited space. However, it also seems to be common to use large samples, following the rationale that larger samples makes it easier to detect small effects (Veldhuizen, Pasker-De Jong, and Atsma, 2012). The combination of using large samples and relying solely on *p*-values for statistical interpretation is, however, not a good idea.

Significance tests are commonly sensitive to sample size; given that a sample is large enough even trivial differences will turn up as significant (Martin-Löf, 1973, 1974). Thus, there is an evident danger of drawing false conclusions based on the combination of large samples and mechanically relying on *p*-values as the only source of information, a phenomenon labeled "The large sample size fallacy" (Lantz, 2013). However, it is important to also recognize the opposite problem, i.e., that the analyst sometimes has too small samples, or too low power, in order to statistically identify substantial differences.

From a Rasch measurement perspective, the implications of "The large sample size fallacy" are somewhat different. Statistics, for instance Chi-square, are commonly used in order to analyze the concordance between the data and the expected Rasch model (Rasch, 1960). Thus, when using a large sample size, the parameters will be estimated with great precision, which further means that even very small differences between the expected Rasch model and the observed data will be readily exposed, and consequently, no items are likely to fit the model (Andrich, 1988). Put differently, when applying a large sample, the power to detect misfit is so great that even if observed and expected values are very close, all items will misfit (Andrich, Sheridan, and Luo, 2009). Therefore, using a large sample and mechanically relying on traditional fit statistics, will almost automatically discard any model tested.

In this study the concordance between observed data and the expected Rasch model is analysed by means of the Rasch model for ordered response categories, also called the polytomous Rasch model. The Rasch model for ordered response categories (Andrich, 1978; Wright and Masters, 1982) takes the general form:

$$\Pr\{x_{ni} = x\} = \frac{e^{-\tau_{1i}-\tau_{2i}\ldots-\tau_{xi}+x(\beta_n-\delta_i)}}{\sum_{x'=0}^{m_i} e^{-\tau_{1i}-\tau_{2i}\ldots-\tau_{xi}+x'(\beta_n-\delta_i)}}. \quad (1)$$

Thus, in the polytomous case a central concept is threshold. Given a situation with five response categories (0, 1, 2, 3, 4), a threshold specifies the point at which the probability for choosing one of two answers is equal, for instance an answer of 0 or 1. In the equation above the threshold parameter is denoted by τ and the item score by x in the numerator. Given that there is concordance between the expected Rasch model and the data, the item discriminations are the same, as is illustrated in Figure 1.

For illustrative purposes only, Table 1 shows an analysis based on simulations of perfect circumstances using different sample sizes, with 10 repetitions at each sample size (10 items in 10 class intervals ($df = 90$). The values in Table 1 are Chi-square values with their probabilities reflecting the statistical concordance between the expected polytomous Rasch model and the data.

Thus, using a sample larger than 5,000 individuals would reveal significant results also in perfect circumstances.

A suggested strategy has therefore been to combine statistical analysis of fit with descriptive, graphical analysis (Andrich, 1988). In Figure 2 an example of that sort of analysis is provided by means of an item characteristic curve (ICC), for the worst fitting item of simulation 1 from Table 1, using 21,000 individuals. Even though this example constitutes the worst fitting item, the observations are located on the according to the Rasch model expected curve, i.e., the deviations between observed data and the expected Rasch

model are very small. Nevertheless, the statistical counterpart to the ICC shows misfit for all items applying the same sample size.

However, due to space limitations that is not always possible—the journal format sometimes requires a one-number solution describing model fit. Therefore, several different approaches to handle the sample size problem, and tangent issues, have been discussed elsewhere (see for instance: (Gustafsson, 1980; Tennant and Pallant, 2012; Wright and Masters, 1982; Wright and Linacre, 1994; Wright and Masters, 1990).

In the RUMM2030 program (Andrich, Sheridan, and Luo, 2013), a facility enabling sample size adjustment in the statistical analysis of fit has been implemented, implying that sample size, all other things being equal, is adjusted in the analysis. The Chi-square statistic for test of fit is conducted by comparing the total score of persons in approximately equal sized class intervals, with the sum of expected values. This is resulting in an approximate Chi-square statistic with *C-1* degrees of freedom and which is, in accordance with Andrich and Styles (2011 PG 81), denoted by:

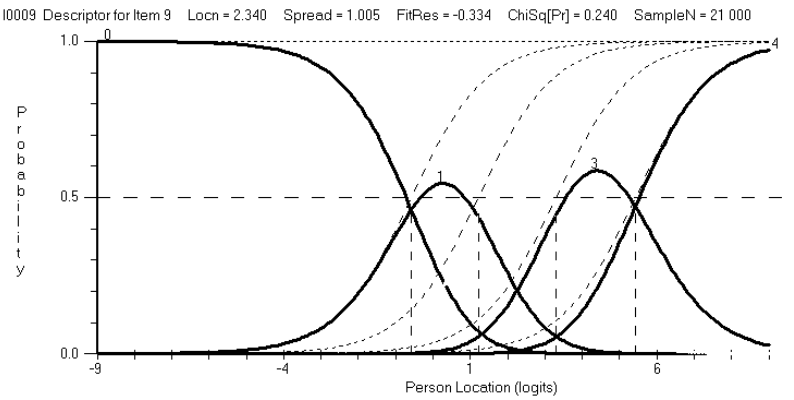

Figure 1. An example of a Category Probability Curve showing the latent dichotomous threshold characteristic curves with equal slopes.

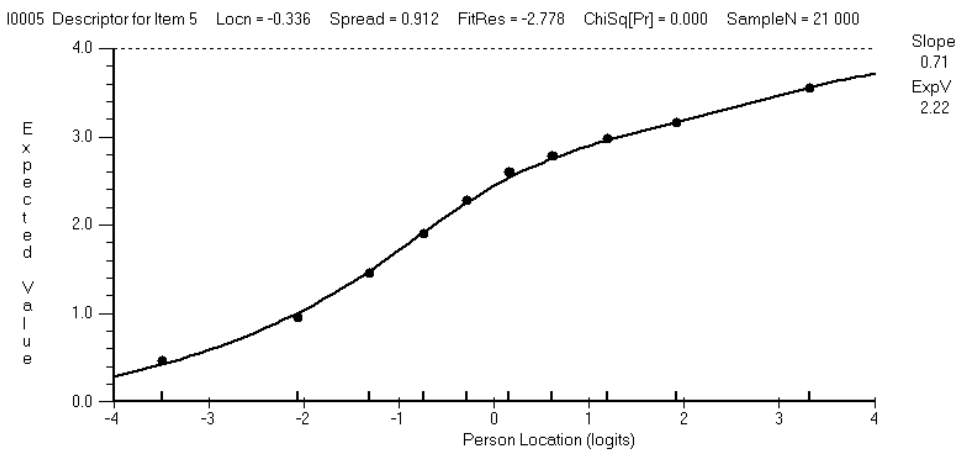

Figure 2. Item Characteristic Curve for the worst fitting item from simulation 1 Table 1, using 21,000 individuals.

Table 1

Total Chi-square values and Mean-square values from analyses based on simulations of perfect circumstances for different sample sizes with 10 repetitions for each sample size (10 items in 10 class intervals (df = 90)).

Sample	Sim 1	Sim 2	Sim 3	Sim 4	Sim 5	Sim 6	Sim 7	Sim 8	Sim 9	Sim 10	Mean- Square
21,000	251.87	243.61	272.18	358.05	318.58	332.89	270.75	312.79	287.99	241.66	3.21
Prob.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
15,000	195.05	219.76	188.50	227.88	233.43	178.97	186.23	178.13	260.56	209.08	2.31
Prob.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
10,000	175.25	203.24	203.55	191.70	198.04	213.38	172.05	142.11	157.14	187.0	2.05
Prob.	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000391	0.000016	0.000000	
5,000	111.14	152.84	127.37	150.22	117.5	158.04	127.25	164.8	98.45	134.37	1.49
Prob.	0.064755	0.000042	0.005879	0.000073	0.027411	0.000013	0.005999	0.00003	0.254393	0.001717	
3,000	109.93	111.04	120.75	99.97	103.25	93.28	123.89	132.27	99.06	82.26	1.19
Prob.	0.07532	0.065602	0.016955	0.221543	0.160497	0.385503	0.010385	0.002509	0.240928	0.707019	
1,000	89.85	104.51	103.71	83.57	107.46	110.36	99.92	105.86	106.17	76.38	1.10
Prob.	0.4847	0.140546	0.152954	0.670512	0.101138	0.071457	0.222591	0.121373	0.117245	0.846558	
500	96.17	71.23	91.02	114.05	80.29	78.45	71.43	81.07	76.92	63.23	0.92
Prob.	0.308752	0.927842	0.450057	0.044302	0.758630	0.802681	0.925412	0.7387717	0.835622	0.985520	

$$X^2_{C-1,i} \approx \sum_{c=1}^{C}\left(\left(\sum_{n\in c} X_{ni} - \sum_{n\in c} E[X_{ni}]\right)^2 / \sum_{n\in c} V[X_{ni}]\right). \quad (2)$$

Following Andrich and Styles (2011 PG 81), the Chi-square test of fit statistic can be adjusted to an equivalent effective sample of different size by applying the rationale:

$$\begin{aligned}
X^2_{C-1,i} &= \sum_{c=1}^{C}\left(\left(n_c\sum_{n\in c} X_{ni}/n_c - n_c\sum_{n\in c} E[X_{ni}]/n_c\right)^2 / n_c\sum_{n\in c} V[X_{ni}]/n_c\right),\\
&= \sum_{c=1}^{C}\left(n_c^2\left(\sum_{n\in c} X_{ni}/n_c - \sum_{n\in c} E[X_{ni}]/n_c\right)^2 / n_c\sum_{n\in c} V[X_{ni}]/n_c\right),\\
&= \sum_{c=1}^{C}\left(n_c^2\left(\sum_{n\in c} X_{ni}/n_c - \sum_{n\in c} E[X_{ni}]/n_c\right)^2 / \sum_{n\in c} V[X_{ni}]/n_c\right),\\
&= \sum_{c=1}^{C}\left(n_c\left(\bar{X}_{ci} - \bar{E}_{ci}\right)^2 / \bar{V}_{ci}\right),\\
&\approx \sum_{c=1}^{C}\left(\frac{N}{C}\left(\bar{X}_{ci} - \bar{E}_{ci}\right)^2 / \bar{V}_{ci}\right),\\
&= \frac{N}{C}\sum_{c=1}^{C}\left(\left(\bar{X}_{ci} - \bar{E}_{ci}\right)^2 / \bar{V}_{ci}\right).
\end{aligned} \quad (3)$$

The number of persons in class interval C is then denoted by $n_{c,}$

$$\bar{E}_{ci} = \sum_{n\in c} X_{ni}/n_c; \bar{V}_{ci} = \sum_{n\in c} V\left[X_{ni}\right]/n_{c,}$$

and by constructing class intervals,

$$n_c \approx \frac{N}{C}\ .$$

As the outcome of equation (3) is proportional to the original sample size (N), in order to adjust the analysis to a smaller equivalent effective sample size (n), the Chi-square value obtained using the original sample size should be multiplied by n/N (Andrich and Styles, 2011). Thus, by adjusting the analysis to a smaller equivalent effective sample size, the Chi-square test of fit is expected to be less sensitive to sample size (Andrich and Styles, 2011), but with the residuals reflecting the degree of precision available in the original sample.

The RUMM2030 adjustment facility has been available for long time, but still there seems to be a lack of studies describing the empirical operational characteristics of the function, i.e., the empirical consequences of adjusting a sample to a smaller effective sample size in the statistical analysis of fit. Alternatively, a random sample approach could be adopted in order to handle the sample size problem. The purpose of this study was to analyze and to compare these two strategies as test of fit approximations, using simulated data.

Methods

Frame of reference

The analyses conducted in accordance with the purpose of this paper are based on simulated data with different degrees of fit using RUMMss simulation package (Marais and Andrich, 2012). For illustrative purposes and as a frame of reference, simulated data of perfect concordance to the polytomous Rasch model was simulated. Using 10 items, and with a person mean of 0.0, a standard deviation of 2.0, item locations were simulated to range between −3 and 3 logits. The first threshold was set to −3 logits and the last threshold to 3. Item discriminations were set to 1.00 for all items. As extreme person locations would violate the model characteristics, extreme scores were excluded in the simulation procedure.

The simulation procedure was conducted for different sample sizes (21,000, 15,000, 10,000, 5,000, 3,000, 1,000 and 500) and was repeated 10 times for each sample size. Total Chi-square values were calculated for each simulation. As an overall indicator of the level of fit, mean-square values were calculated by the summated total Chi-square values divided by the total *df*, implying an expected value of 1 (this analysis is reported in Table 1). Thus, this fit statistic is similar to that of Outfit, often referred to in Winsteps contexts, but sometimes also called the reduced Chi-squared or the mean square weighted deviation.

Two scenarios

In addition, two scenarios were simulated. First, a realistic situation of fitting data relatively well targeted and with item discriminations of 1.00 across items, but allowing for extreme scores was simulated. In the second scenario, a realistic situation of misfitting data was simulated. Item discriminations were then simulated to vary substantially across items, ranging from 0.5 for

poor discriminations to 1.5 for high discriminations. Thus the model was violated by varying the item discriminations. In the two scenarios item difficulties were set to range between −3.00 and 3.00 logits. Also in both of the scenarios, the estimated model includes 7 polytomous items (with 5 response categories) and with persons grouped into 10 equal sized class intervals based on the person locations of the whole sample ($df = 63$). The number of items used here is analogous with prior and ongoing work on real data not yet published, but is also considered to be common item compositions in social and medical sciences. The response format implies 28 item thresholds in total, and threshold values were simulated to (−1.00, −0.300, 0.300 and 1.00) be equal for each item.

In order to facilitate the study of trends of Chi-square values when moving from one sample size to a smaller, starting with an original sample size of 21,000, the sample was adjusted to 19,000, 15,000, 12,000, 10,000, 7,000, 5,000, 3000, 2,000, 1,000, 750, 500, 300, and 100 for each of the two simulated scenarios, using the RUMM2030 sample size adjustment function. By adjusting the analysis to several different effective sample sizes, the observation of inconsistencies in Chi-square value trends is possible.

However, in order to qualify the analysis, the Chi-square values obtained using the adjustment function need to be compared to external values, i.e., values that are obtained not using the adjustment function, but values that would be expected provided the specific sample size and level of fit. Thus, following statistical principles, averaged Chi-square values based on sets of random samples are considered to be a good Chi-square value approximation. Therefore, 10 random samples with replacement were drawn for each sample size and averaged total Chi-square values calculated. These were then compared with the total Chi-square values obtained using the sample size adjustment function in the RUMM2030 software, for each of the different sample sizes, and for each of the two scenarios. In addition, mean-square values were calculated by dividing the summated total Chi-square values with the total number of *df*, implying an expected value of 1. This calculation was performed for random samples as well as adjusted samples. It should, however, be noticed that when generating random samples similar to the original sample in terms of size, the two will be highly dependent on each other, i.e., largely containing the same individuals.

The concordance between the Rasch model and the observed data was analysed by means of the polytomous Rasch model, also called the Rasch model for ordered response categories. All analyses were conducted using the RUMM2030 software (Andrich et al., 2013).

Results

Figure 3 and Figure 4 depict the distribution of persons relative to item thresholds in the two simulated scenarios of relatively well-targeted fitting and misfitting data, respectively. In the case of fit (Figure 3) the results imply a mean of 0 and standard deviation of 2 for persons and items respectively. The scenario of misfit, as is shown in Figure 4, reveals a person mean of 0.1 and a standard deviation of 1.64, and similar figures for items (0 and 1.76 respectively).

In Table 2, the comparison between the random sample approach and the RUMM2030 adjustment function is displayed for fitting data. A general pattern observable implies decreasing Chi-square values when moving from one sample size to a smaller, which is true for adjusted samples as well as random samples of different sizes. At sample sizes of 19,000 and 15,000 the random sample/adjustment function ratio is smaller than 1, but at all other sample sizes the ratio is bigger than 1, i.e., the total Chi-square value obtained using the random sample approach is bigger than that obtained using the RUMM2030 adjustment facility. At sample sizes between 19,000 and 5,000, the differences between Chi-square values obtained using the random sample approach and the adjustment facility approach are small (the ratio is close to 1). However, at sample sizes smaller than 5,000, the ratio is generally and gradually increasing as turning from one sample size to a smaller. Based on the *p*-values, when using the adjustment function data start to fit the model at a sample size of

about 2,000 individuals, while the corresponding sample size for random samples is 1,000, which is also reflected in the mean-square values. From Table 2 it can also be seen that the mean-squares are varying substantially for adjusted Chi-squares when adjusting to sample sizes between 2,000 and 100 individuals, while they are largely constant for random samples. However, using a sample size of 1,000, and based on *p*-values, the analyst would reach the same conclusion of whether the data fits the model or not, regardless of which of the two methods used. As shown in Table 3, this is also confirmed at the item level, given samples of 1,000 but not 3,000. In Table 3, it should also be noticed that the relationship between items for adjusted samples, i.e., the best or worst fitting items are the same at different adjustments, but not for random samples due to the nature of the simulated data (well fitted = small differences) in combination with random variation.

Applying misfitting data, the overall pattern is the same, but less pronounced. For instance, at sample sizes between 19,000 and 1,000 the differences between the random sample approach

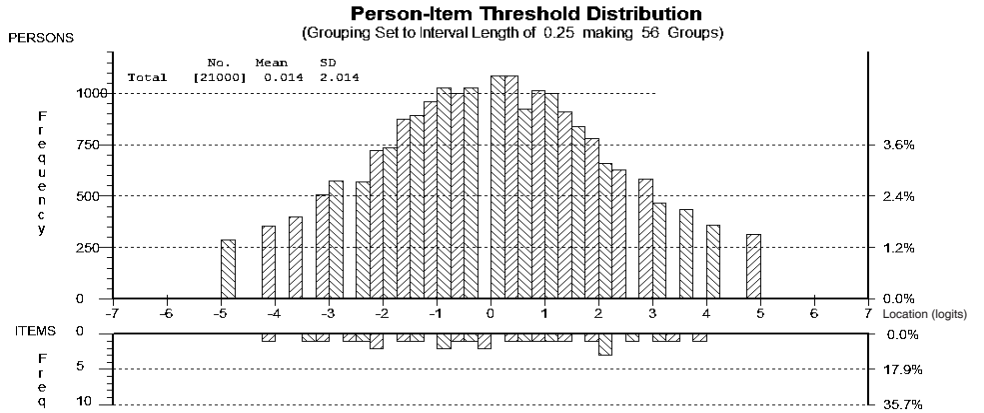

Figure 3. Person-Item Threshold Distribution, a realistic situation with fitting data simulated.

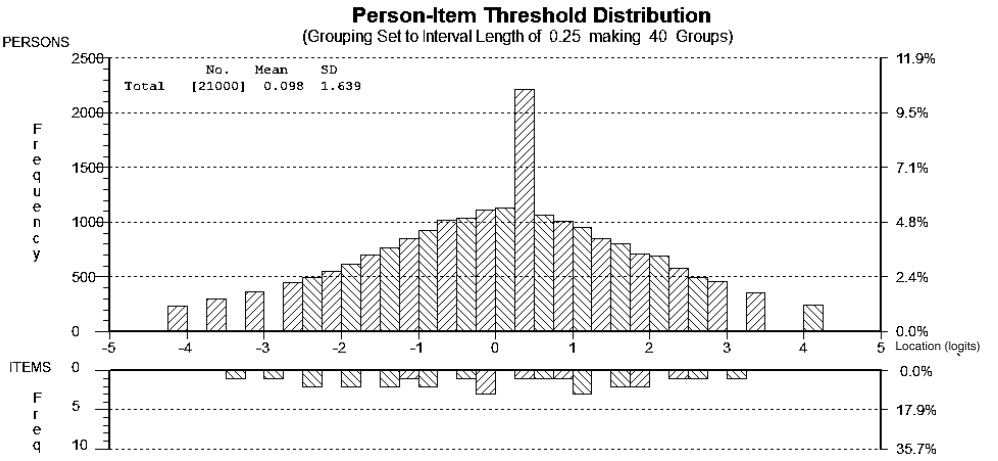

Figure 4. Person-Item Threshold Distribution, a realistic situation with misfitting data simulated.

Table 2

Comparisons between total Chi-square values based on different sample sizes using the RUMM2030 adjustment function, and average total Chi-square values based on 10 random samples for each sample size (significant differences are bolded), the corresponding Mean-square values also provided.*

Fitting data (7 polytomous items, 10 class intervals, *df* = 63).

Sample size	Chi-square original sample	Prob.	Adjusted Chi-square	Prob. Adjusted Chi-square	Mean-square adjusted Chi-square	Averaged Chi-square Random samples	Prob. random samples	Mean-Square random samples	Ratio averaged Chi-square/ adjusted Chi-square
21,000	661.8	0.000000							
19,000			616.4	0.000000	9.78	601.7	0.000000	9.55	0.98
15,000			486.6	0.000000	7.72	482.1	0.000000	7.65	0.99
12,000			389.3	0.000000	6.17	394.4	0.000000	6.26	1.01
10,000			324.4	0.000000	5.15	337.0	0.000000	5.35	1.04
7,000			227.1	0.000000	3.60	248.7	0.000000	3.95	1.10
5,000			162.2	0.000000	2.57	197.9	0.000000	3.14	1.22
3,000			**97.3**	0.003592	1.54	**138.1**	0.000002	2.21	1.42
2,000			**64.9**	0.410968	1.03	**103.9**	0.000905	1.65	1.60
1,000			**32.4**	0.999502	0.51	**69.8**	0.259725	1.11	2.15
750			**24.3**	0.999997	0.39	**68.8**	0.287520	1.09	2.83
500			**16.2**	1.000000	0.26	**59.5**	0.601819	0.94	3.67
300			**9.7**	1.000000	0.15	**54.9**	0.756467	0.87	5.66
100			**3.2**	1.000000	0.05	**52.2**	0.832237	0.83	16.31

*Independent samples *t*-tests between Chi-Square values obtained using the random sample approach and the RUMM2030 adjustment facility, for each sample size. *P*-values are considered as significant when $p < 0.001$

Table 3

Comparisons between individual item Chi-Square values based on adjusted samples and averages based on 10 random samples using 3,000, 2,000 and 1,000 individuals. Comparisons leading to different conclusions are bolded. $p < 0.05$ are considered as significant.

Fitting data

	Adjusted sample to 3,000	Averaged Chi-square random samples $n = 3,000$	Adjusted sample to $n = 2,000$	Averaged Chi-square random samples $n = 2,000$	Adjusted sample to $n = 1,000$	Averaged Chi-square random samples $n = 1,000$
Item	Chi-square (probability)	Chi-square (Probability)	Chi-square (probability)	Chi-square (probability)	Chi-square (probability)	Chi-square (probability)
1	9.795 (0.367311)	16.7714 (0.052419)	6.53 (0.685903)	13.6122 (0.136804)	3.265 (0.952851)	8.2023 (0.513895)
2	**14.66 (0.100699)**	**20.1706 (0.016888)**	9.774 (0.369127)	15.8291 (0.070537)	4.887 (0.844064)	10.092 (0.343089)
3	18.176 (0.033183)	25.7939 (0.002208)	12.118 (0.206766)	14.619 (0.101947)	6.059 (0.734024)	9.8855 (0.359834)
4	**16.831 (0.051439)**	**20.6051 (0.014524)**	11.22 (0.260908)	16.3276 (0.060348)	5.61 (0.77821)	9.875 (0.360699)
5	**15.194 (0.085751)**	**19.8239 (0.019031)**	**10.129 (0.340131)**	**17.1208 (0.046858)**	5.065 (0.828643)	9.7523 (0.370908)
6	**13.932 (0.124778)**	**21.3927 (0.011016)**	9.288 (0.411146)	13.0817 (0.158944)	4.644 (0.864182)	12.349 (0.194342)
7	8.731 (0.462492)	14.4169 (0.108249)	5.82 (0.757741)	13.3012 (0.149444)	2.91 (0.967747)	9.6307 (0.381201)

and the adjustment facility approach are small, i.e., the ratio is close to 1, as is shown in Table 4. According to the *p*-values, using the adjustment function would reveal non-significant results with adjustments to 750 individuals, while the corresponding sample size for random samples is 300.

The two scenarios taken together

Figure 5 shows the two scenarios simultaneously displayed in one single graph. From Figure 5, it is evident that the two scenarios seem to work in a similar manner given sample sizes of between 19,000 and 5,000 individuals. However, applying fitting data, the random sample/adjusted sample ratio increases gradually as the sample size is adjusted to each lower level. For instance, the ratio increase seems to be particularly pronounced when moving between adjustments to 500 and 300 individuals. Given the situation with misfit, the ratio increase is at its highest levels at the same points, but is less pronounced.

Discussion

The purpose of this study was to analyze and to compare the RUMM2030 adjust sample size function and a random sample approach as test of fit approximations, using simulated data with different degrees of fit.

The overall pattern reveals that the Chi-square values obtained using random samples are bigger compared to the adjusted ones. Given sample sizes ranging between 19,000 and 5,000, the random sample approach and the adjust sample approach seem to work in similar manner, i.e., the random sample/adjustment ratio is close to 1, regardless of the nature of data used. Put differently, given that the original (N = 21,000) sample size is adjusted to a larger proportion than approximately 0.24, the adjustment function and the random sample approach seem to work similarly, which is true for fitting as well as misfitting data. Therefore, the adjust sample size function is considered to work well as a relevant test of fit approximation in these circumstances.

However, for adjustments to smaller samples, the ratio increases substantially with decreasing sample size, in particular when using fitting data. Thus, in these circumstances the random sample/adjustment ratio is substantially different from 1 when adjusting to a sample size equivalent to 3,000 individuals or smaller, indicating that the Chi-square values are not comparable. For adjustments to each smaller sample size than approximately 14 percent of the original (equivalent to about 3,000 individuals) the ratio increases substantially. Thus, in these circumstances, the RUMM2030 adjustment function is not effective at approximating the Chi-square value for an actual random sample of relevant size.

The differences between the results of using the two methods may highlighted by focusing on their respective Chi-square values and cor-

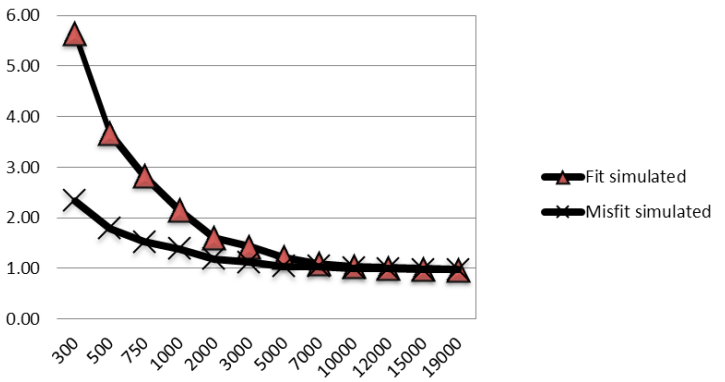

Figure 5. The Random sample/adjustment function ratio at different sample sizes. The two scenarios displayed simultaneously.

Table 4

Comparisons between total Chi-square values based on different sample sizes using the RUMM2030 adjustment function and average total Chi-square values based on 10 random samples for each sample size (significant differences are bolded), the corresponding Mean-square values also provided.*

Misfitting data (7 polytomous items, 10 class intervals, *df* = 63).

Sample size	Chi-square original sample	Prob.	Adjusted Chi-square	Prob. Adjusted Chi-square	Mean-square adjusted Chi-square	Averaged Chi-square Random samples	Prob. random samples	Mean-Square random samples	Ratio averaged Chi-square/ adjusted Chi-square
21,000	2,222.26	0.000000							
19,000			2,057.0	0.000000	32.65	2,010.3	0.000000	31.91	0.98
15,000			1,623.9	0.000000	25.77	1,583.3	0.000000	25.13	0.97
12,000			1,299.2	0.000000	20.62	1,290.6	0.000000	20.49	0.99
10,000			1,082.7	0.000000	17.19	1,082.5	0.000000	17.18	1.00
7,000			757.9	0.000000	12.03	787.9	0.000000	12.51	1.04
5,000			541.3	0.000000	8.59	560.3	0.000000	8.89	1.04
3,000			324.8	0.000000	5.16	364.4	0.000000	5.78	1.12
2,000			216.5	0.000000	3.43	256.9	0.000000	4.08	1.19
1,000			108.3	0.000342	1.71	149.7	0.000000	2.38	1.38
750			**81.2**	0.061139	1.29	**123.8**	0.000008	1.96	1.52
500			**54.1**	0.779418	0.86	**96.8**	0.003984	1.54	1.79
300			**32.5**	0.999492	0.51	**75.9**	0.127668	1.21	2.34
100			**10.8**	1.000000	0.17	**56.6**	0.702201	0.90	5.24

*Independent samples *t*-tests between Chi-Square values obtained using the random sample approach and the RUMM2030 adjustment facility, for each sample size. P-values are considered as significant when $p < 0.001$

responding *p*-values. For instance, using random samples and a sample size range between 1000 and 100, reveal Chi-square values close to the number of degrees of freedom (i.e., mean-squares close to 1). However, the corresponding Chi-square values for adjusted samples of equal size are much lower (*p*-values approximating 1.00) than the degrees of freedom, implying mean-square values much lower than 1. For instance, provided the original sample of 21,000 and adjusting to a size of 100 implies a mean-square value 20 times lower than the expected value of 1, the value of a corresponding adjustment to a sample of 300 is about 7 times. Based on these findings it can be concluded that the adjustment of sample sizes mechanically will exaggerate the level of fit, and underestimate misfit, potentially leading to spurious acceptance of data as fitting the model when it actually misfits.

Given the characteristics of the adjustment function, the results should not be surprising; when applying a big sample (e.g., 21,000 individuals) the parameters are estimated with great precision, and consequently with very small standard errors. Thus, adjusting the sample to a much smaller effective size, the same characteristics are applied, but in addition with much less power to detect misfit. Consequently, there seems to be an evident risk of reinforcing fit, in comparison to random samples of equal size. Therefore, the assumptions behind the adjustment function may be considered unrealistic in that data that fits so well with a small sample size would be hard to find in any real situation.

Despite the diverging results between the two methods, it is important to recognize that the inference based on *p*-values may be the same provided big reductions in sample size (to 1,000 or lower in this case). Thus, in the case of fitting data, the conclusion would be that the data fit the polytomous Rasch model, regardless of whether analyst is using a random sample approach or adjusts the sample.

Even though it may be argued that big sample size reductions using the adjustment function will exaggerate the actual level of fit compared to a random sample approach, using the adjustment will keep the item characteristics also with big reductions, also with small differences between items. This means that it is possible to identify which item shows the best or worst fit. Thus, the overall level of fit may be exaggerated, but the relationship between the items will remain as in the original data size. Due to random variation, using a small number of random samples may imply that the relationship between items will be changed, although providing a more realistic level of overall fit.

Using big datasets, the solution is not to avoid Chi-square for statistical test of fit. The use of Chi-square is widespread within the social sciences, and therefore it is easily communicated to a broad audience. Following statistical principles, also in the statistical analysis of fit, one solution would be to adopt both a random sample approach and to employ a sample size adjustment. Ideally, using large samples graphical analysis of the concordance between the expected Rasch model and the observed data may combined with statistical analyses of item fit in order to identify the relationship between items. As statistical significance tests are not meaningful using very large data, the sample may be adjusted to a level facilitating the analysis of item fit, but where a random sample approach may be used in order to estimate the actual level of fit. In that sense the adjustment function may serve as a tool for determining whether the random samples are accurate or not.

Ideally, software developers will provide solutions allowing for simultaneously generating numerous random samples and calculating averaged tests of fit statistics, for larger numbers of random samples, providing estimates of fit analysis corresponding to the actual level of fit, also at the item level.

Even if the issue is not addressed specifically in this paper, situations where the analyst has too small data, rather than too large, constitutes a significant problem, causing too low power in order to conduct realistic statistical analyses. In these circumstances, there are few options in manipulating the data in order to facilitate statistical analysis. Naturally the random sample approach

would not be a solution. Nor is the adjustment function an optimal solution. However, technically it would be possible to adjust the sample size from a small sample to a larger effective samples size in the statistical analysis of fit. Nevertheless, there will be a problem with exaggerated results, but not with fit exaggerated but misfit, causing significant results in more cases than expected. However, by employing smaller levels of adjustments it may possible to reach an effective sample size large enough to conduct realistic statistical test of fit.

Conclusion

Based on the analyses presented here, the RUMM2030 adjust sample size function works well as a test of fit approximation, given an original sample size of 21,000 and adjustments to sample sizes of 5,000 individuals (equivalent to approximately 24 percent of original size) or more, compared to sets of random samples. However, when applying adjustments to lower sample sizes the adjustment function is not considered to be effective at approximating the Chi-square value for an actual random sample of the relevant size, as there is an evident risk of spuriously accepting misfitting data. Nevertheless, inferences based on *p*-values may be the same when adjustments are conducted to lower sample size levels. Working with large samples, the adjustment function may be used as a heuristic tool in the analysis of fit. By adjusting the sample to a smaller effective sample it is possible to identify the relationship between items in terms of best and worst fitting items, while the actual level of fit may be estimated using a random sample approach. Thus, the adjustment function may serve as a tool in determining whether the random samples are accurate or not.

Acknowledgements

I am very grateful to Professor David Andrich, Associate Professor Stephen Humphry, Dr. Joshua McGrane and Assistant Professor Ida Marais at the University of Western Australia, Graduate School of Education, and Professor Curt Hagquist at the Centre for Research on Child and Adolescent Mental Health, Karlstad University, Sweden, for providing invaluable manuscript comments and encouragements.

My Gratitude also goes to The University of Western Australia, Graduate School of Education, for providing office space and facilities during the postdoc period spent there, and which is when this paper was written.

The financial support received from Wenner-Gren Foundations is very much appreciated.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park, NJ: Sage.

Andrich, D., Sheridan, B., and Luo, G. (2009). *Interpreting RUMM2030 (Part I, Dichotomous data): Rasch unidimensional models for measurement.* Perth, WA, Australia: RUMM Laboratory.

Andrich, D., Sheridan, B., and Luo, G. (2013). RUMM2030: A Windows program for the Rasch unidimensional measurement model [Computer software]. Perth, WA, Australia: RUMM Laboratory.

Andrich, D., and Styles, I. (2011). Distractors with information in multiple choice items: A rationale based on the Rasch model. *Journal of Applied Measurement, 12*, 67-95.

Gustafsson, J.-E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33*, 205-233.

Lantz, B. (2013). The large sample size fallacy. *Scandinavian Journal of Caring Sciences, 27*, 487-492.

Marais, I., and Andrich, D. (2012). RUMMss. Rasch unidimensional measurement models simulations studies software. [Computer software].

Martin-Löf, P. (1973, May 7-12). *The notion of redundancy and its use as a quantitative measure of the deviation between a statistical*

hypothesis and a set of observational data. Paper presented at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observationtional data. *Scandinavian Journal of Statistics, 1*, 3-18.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Tennant, A., and Pallant, J. F. (2012). The root mean square error of approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions, 25*, 1348-1349.

Veldhuizen, I., Pasker-De Jong, P., and Atsma, F. (2012). Significance or relevance: What do you use in large samples? About *p* values, confidence intervals, and effect sizes. *Transfusion, 52*, 1169-1171.

Wright, B., D,, and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., and Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Measurement Transactions, 1990*, 84-85.