Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 11 (4): 707-712, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.707.712

The Impact of Questionnaire Size on the Accuracy of the Rasch Measure

Silvia Golia
Department of Quantitative Methods, University of Brescia, C.da S.Chiara, 50 25122 Brescia, Italy

Abstract: The study aims at evaluating the impact of the questionnaire size on the accuracy and stability of the Rasch measure. The Rasch measurement model is used to obtain a reliable and objective measurement of a latent trait of interest. A simulation study is performed in order to deal with the issue.

Key words: Rasch model, stability, simulation study, number of items

INTRODUCTION

The issue of determining a reliable and objective measurement of a complex concept not directly observed, or latent trait, is a crucial problem in the analysis of social and economic phenomena. A latent trait refers to a latent continuum, or dimension, which all individuals are mapped on, based on their pattern of responses on a set of categorical variables. These categorical variables result from the submission of questionnaires with items referring to the different aspects of the concept being measured. Responses usually indicate the degree of agreement with each statement, with higher scores reflecting greater agreement.

A very simple tool to assess subjective attitudes is the summated rating scale, also referred to as raw score. However raw score has little inferential value, is neither interval nor ratio measure and is affected by missing values; this means it cannot be compared for conclusions about subject latent trait. Hence, raw score can only be an indication of a possible measure of the latent trait.

One of the methods proposed to deal with the issue to identify an objective measurement of the latent trait underlying a multiple-item scale is the so-called Rasch model, which allows one to transform ordinal raw scores into interval scale measures. The model has been successfully applied to many contexts which include the psychological, educational, medical and socio-economic (such as the evaluation of customer or job satisfaction) field (Kubinger, 2005; Waugh *et al*., 2000; Tesio, 2003; King and Bond, 2003; Brentari and Golia, 2008).

The goodness of the obtained measures depends on meeting the model assumptions as well as the quality of the questionnaire used. The present research studies the impact of the questionnaire size on the accuracy and stability of the Rasch measure making use of simulated data. Increasing the questionnaire length, the goodness of the estimated measures increases, as expected. Nevertheless, the improvement is significantly high when the questionnaire size is small and more responsive to the increasing number of response categories than items. In the empirical study, questionnaires of small size are quite common and it is interesting to study the goodness of the estimated measures.

SIMULATION STUDY

The Rasch Model (RM) (Rasch, 1960) is a measurement model which converts raw scores into linear and reproducible measurement. It is built around the idea that the probability of a certain answer when a person is confronted with an item, is described as a function of the person's position on the latent trait under study and the parameter characterizing that particular item. Under the hypotheses of unidimensionality (all items forming the questionnaire measure only a single construct, i.e., the latent trait under study) and local independence (conditional to the latent trait, the response to a given item is independent from the responses to the other items in the questionnaire) the person and item parameters enter in the response probability as a linear combination. The mathematical form of the RM provides the separation of item and person parameters. A concomitant of separability is minimally sufficient statistics for person as well as item parameters; the person raw score is a sufficient statistic for the unknown person parameter and the item sum score across persons is a sufficient statistic for the unknown item parameter (Wright and Master, 1982). If the data fit the model, then the measures produced applying the RM to the sample data are objective and expressed in logits (logarithm of odds); the logit scale is an interval scale.

The RM introduced by Rasch (1960) can be used to deal with dichotomous data; if data come from polytomously scored items, that is when there are more than two possible ordered response categories for each item, extensions of the RM, such as the Rating Scale Model (RSM) (Andrich, 1978), must be taken into account. Assuming that there are m+1 possible ordered

response categories for each item, coded as x = 0, 1,…, m, following the RSM, the probability that the person n answers x to the item i is given by:

$$P(X_{ni} = x) = \frac{\exp\left\{x(\beta_n - \delta_i) - \sum_{j=0}^{x}\tau_j\right\}}{\sum_{k=0}^{m}\exp\left\{k(\beta_n - \delta_i) - \sum_{j=0}^{k}\tau_j\right\}}, \quad x = 0,1,...,m \qquad (1)$$

Therefore, it depends on the subject ability, or level of latent trait, β_n and how difficult the item is to endorse, identified by its mean difficulty δ_i and the thresholds τ_j; τ_j is the point of equal probability of categories j-1 and j. Thresholds add up to zero, i.e.,

$$\sum_{j=1}^{m}\tau_j = 0 \text{ and } \tau_0 = 0$$

The present simulation study wants to investigate the stability of the measures estimated from simulated data sets involving the RSM defined in Eq. 1 and different items and thresholds sets. The data are generated as follows.

A sample of 1000 abilities was drawn from a standard normal distribution; these abilities are used in the data simulation and represent the target or real abilities β_n. The response given by the subject n with ability β_n to the item i, which has difficulty δ_i, is obtained as follows. For each category, the corresponding response probability is computed making use of Eq. 1. Then the response probability cumulative sum is calculated and compared with a random number rn drawn from a uniform distribution on the interval [0,1]. The response category corresponding to the first element of the cumulative sum which is equal or larger than rn is assigned as the response of the subject n to the item i. This procedure is repeated for all the items in order to simulate the response record of each of the 1000 subjects forming the simulated sample.

Table 1: The sets of the item mean difficulties used in the simulation study

5 items	7 items	10 items	20 items
-1.7684	-1.7684	-1.7684	-1.7684
			-1.4726
			-1.2740
		-0.9496	-0.9496
-0.8373	-0.8373	-0.8373	-0.8373
			-0.6370
			-0.5323
	-0.3092	-0.3092	-0.3092
			-0.2662
		-0.1237	-0.1237
0.0000	0.0000	0.0000	0.0000
			0.1578
	0.3092	0.3092	0.3092
			0.5078
0.7783	0.7783	0.7783	0.7783
			0.9029
		1.0733	1.0733
			1.2454
			1.3682
1.8274	1.8274	1.8274	1.8274

Table 1 reports the sets of the item mean difficulties δ_i used in the present study. Each set of the difficulty parameters is drawn from a continuous uniform distribution on the interval from -1.9 to 1.9 so that the parameters sum is equal to zero, as required by the calibration procedure and each set includes the previous one.

The two sets of threshold parameters τ_j utilized are [-0.5 0.5] and [-1 -0.5 0.5 1]; they imply three and five response categories, respectively.

For each combination of item mean difficulties δ_i and threshold set τ_j, 200 data sets were simulated and analyzed and 200 sets of estimated abilities and item difficulties were computed.

In the calibration procedure the analysis was performed by setting the mean of item difficulty estimates to 0.0 logits and by using the (unconditional) maximum likelihood estimation method. The data simulation was performed using Matlab 6.5 whereas the Rasch analysis using Winsteps 3.65 (Linacre, 2006).

RESULTS AND DISCUSSION

Table 2 reports the mean value of the person reliability index and the rejection percentages of the null hypothesis underlying the Jarque-Bera test for normality, the two-sample Kolmogorov-Smirnov test and the t-test for zero mean.

The person reliability index (Bond and Fox, 2007) is an estimate of the reproducibility of people placement that can be expected if the same sample of respondents was to be given another set of items measuring the same latent construct. It is bounded by 0 and 1 and can also be

Table 2: The mean value of the person reliability index (standard errors in brackets) and the rejection percentages of the null hypothesis underlying the Jarque-Bera, Kolmogorov-Smirnov and t-tests (significance level 5%)

Items	Reliab.	JB test	KS test	t-test
3 categories				
5	0.55 (0.018)	69.00%	100%	0%
7	0.66 (0.011)	98.00%	100%	0%
10	0.74 (0.008)	96.50%	100%	0%
20	0.86 (0.003)	53.50%	18%	0%
5 categories				
5	0.75 (0.009)	80.50%	100%	0%
7	0.81 (0.006)	78.50%	100%	0%
10	0.86 (0.004)	48.00%	84.50%	0%
20	0.93 (0.004)	5.50%	0%	0%

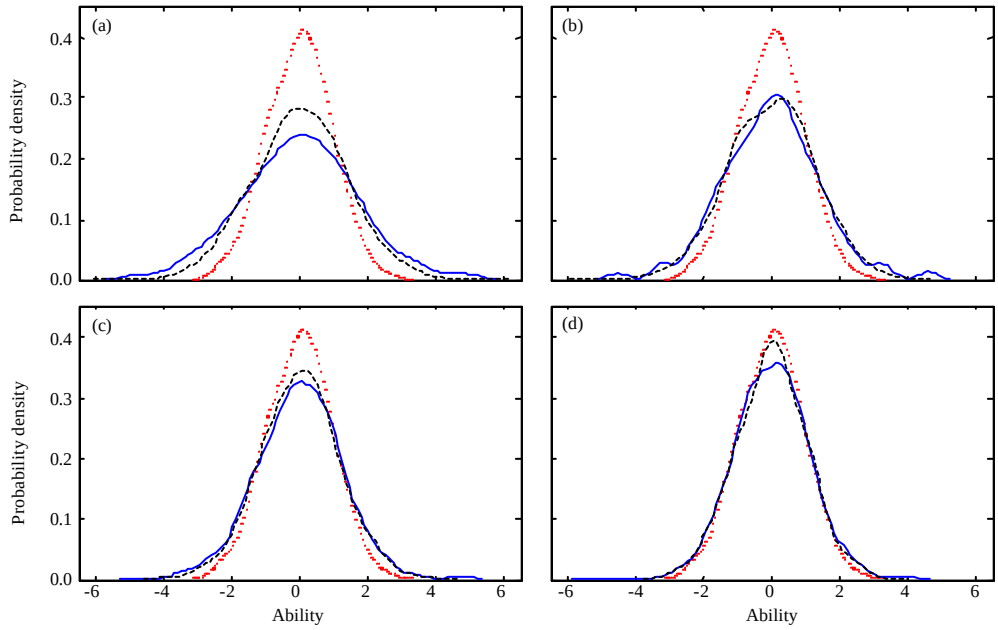

Fig. 1: Graph of the kernel probability density of the real (dotted line) and an estimated ability (3 categories: solid line, 5 categories: dashed line). (a) 5 items, (b) 7 items, (c) 10 items and (d) 20 items

computed with missing values. The values of the person reliability index are sufficiently high if there are 20 items with three categories or at least seven items with five categories; increasing the number of items and thresholds, the people placement in the ability scale is more reliable.

The Jarque-Bera test for normality (Bera and Jarque, 1980) has been performed for each estimated ability set to check if the null hypothesis of normality is a reasonable assumption regarding the population distribution. If three response categories are used, the rejection percentage of the null hypothesis is higher than 5% for all the cases; the distribution of the estimated abilities is not normal. If the questionnaire admits five response categories, the rejection percentage of the null hypothesis is consistent with the fixed significant level if 20 items are used.

The two-sample Kolmogorov-Smirnov test (Massey, 1951) is used to verify if two independent random samples (the real and estimated abilities) are drawn from the same underlying continuous population. The rejection percentage of the null hypothesis is consistent with the fixed significant level if five categories and 20 items are used.

The t-test is used to verify if the mean estimated ability is equal to zero. In all the eight cases the null hypothesis is accepted.

We can conclude that all the eight types of questionnaires are able to reproduce ability estimates with zero mean, as the real one, but only one type (five categories and 20 items) is able to produce estimates coming from a normal distribution, as in the real case.

Table 3: The mean width of the empirical 95% confidence interval for the ability estimation computed considering the least able, the mid-able and the most able subject (standard errors in brackets)

	5 items	7 items	10 items	20 items
3 categories				
Least able	4.31 (0.467)	3.55 (0.443)	2.94 (0.446)	1.91 (0.348)
Mid-able	3.84 (0.438)	2.88 (0.327)	2.21 (0.218)	1.47 (0.119)
Most able	4.33 (0.451)	3.47 (0.474)	2.93 (0.473)	1.92 (0.415)
5 categories				
Least able	3.29 (0.481)	2.58 (0.492)	1.99 (0.378)	1.30 (0.213)
Mid-able	2.60 (0.301)	1.97 (0.231)	1.53 (0.132)	1.03 (0.081)
Most able	2.61 (0.223)	2.52 (0.174)	1.96 (0.368)	1.28 (0.197)

If the probability density function, estimated using a kernel smoothing method based on a normal kernel function, is considered, as shown in Fig. 1a-d, it is possible to observe that the extreme subjects, that is the respondents with higher or lower level of ability, are the most difficult to estimate with accuracy.

Table 3 reports the mean values of the width of the empirical 95% confidence interval for the ability estimation $\hat{\beta}_n$ computed considering the least able (level of estimated ability lower than the first decile), the most able (level of estimated ability higher than the ninth decile) and the mid-able (level of estimated ability bounded by the first and ninth deciles) subjects. The width of the empirical 95% confidence interval for the ability estimation shows an inverse relation with the number of used items and

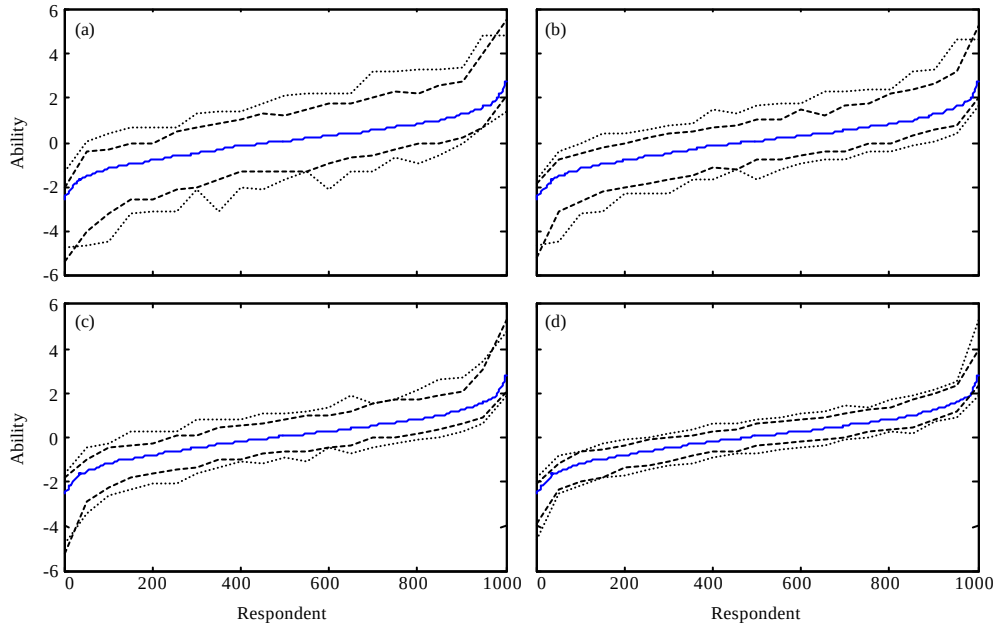

Fig. 2: Graph of the real ability (solid line) and the empirical 95% confidence bands smoothed using a cubic spline. Data obtained using 3 (dotted line) and 5 (dashed line) categories. (a) 5 items, (b) 7 items, (c) 10 items and (d) 20 items

Table 4: Variations in the mean width of the empirical 95% confidence interval for the ability estimation between contiguous questionnaire lengths

Items	Least able	Mid-able	Most able
3 categories			
5	-	-	-
7	-0.1761	-0.2498	-0.1978
10	-0.1725	-0.2316	-0.1568
20	-0.3503	-0.3353	-0.3450
5 categories			
5	-	-	-
7	-0.2147	-0.2473	-0.2167
10	-0.2296	-0.2213	-0.2231
20	-0.3485	-0.3290	-0.3464

Table 5: Mean correlation coefficient between estimated and true measures (standard errors in brackets)

Items	3 categories	5 categories
5	0.777 (0.010)	0.869 (0.006)
7	0.830 (0.008)	0.904 (0.005)
10	0.872 (0.005)	0.931 (0.003)
20	0.930 (0.003)	0.965 (0.002)

thresholds; increasing this number, the width decreases and the estimation is more stable. Moreover, a larger and asymmetric empirical confidence interval corresponds to the least and the most able subjects; the estimation is more complex for extreme respondents.

Table 4 reports the variations in the mean width of the empirical 95% confidence interval for the ability estimation observed when the questionnaire size is n_i instead of n_{i-1}; for example -0.1761 is the variation in the mean width of the least able respondents when a questionnaire with three categories and seven items, instead of one with five items, is used. The reduction of the width is almost constant until 10 items (around 20%). When the questionnaire size doubles, going from 10 to 20 items, the improvement in the reduction of the width is higher but not so high as one could expect doubling the number of items, in comparison with the previous sizes. The impact of increasing the questionnaire size is stronger when this size is small. If mid-able subjects are considered, it can be noted that the number of categories does not affect the magnitude of the improvement.

Table 5 reports the mean correlation coefficients between the true and estimated ability measures. In all cases the values are significantly high, showing a strong linear relation between the measures. The linear link becomes stronger as the length of the questionnaire increases.

Figure 2a-d display the graphs (ability versus respondents) of the empirical 95% confidence bands, smoothed using a cubic spline and the real abilities. The shapes of the confidence intervals are less precise when a small number of items and categories is used and the estimated measures are less stable. In all the cases the estimation procedure underestimates the real abilities β_n of the least able subjects and overestimates the real abilities β_n of the most able subjects, highlighting difficulties in estimating the ability of extreme respondents. The empirical confidence interval is centred

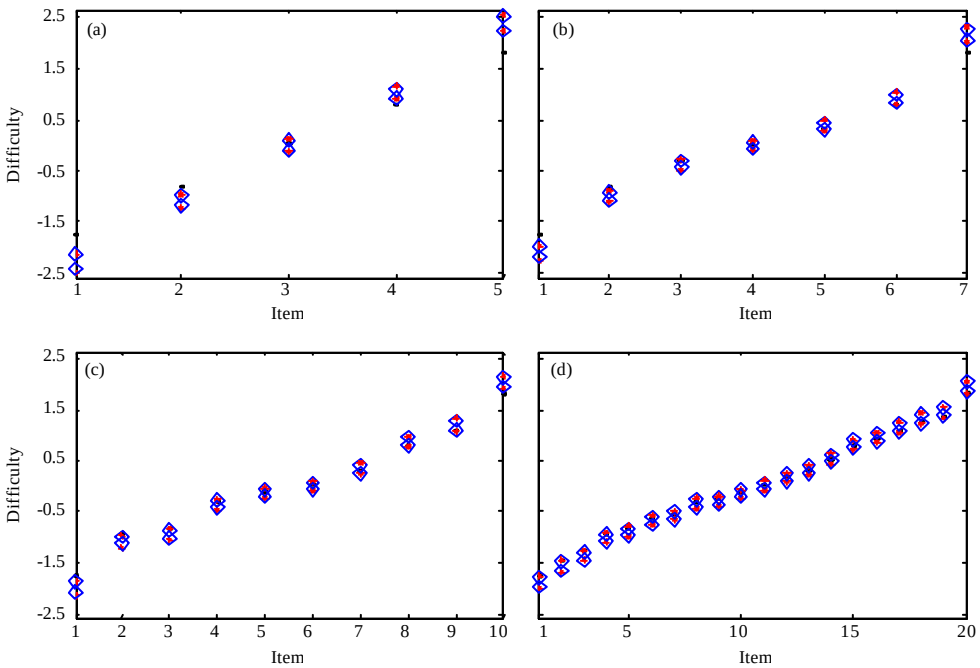

Fig. 3: Graph of the real item mean difficulty (dots) and the empirical 95% confidence intervals obtained using 3 (star) and 5 (diamond) categories. (a) 5 items, (b) 7 items, (c) 10 items and (d) 20 items

for almost all the non-extreme subjects. Moreover, the effect on the confidence interval and the stability of the estimated measure $\hat{\beta}_n$ due to the number of the response categories is stronger than the effect obtained increasing the items number. Even if the questionnaire is made by a small number of items, for example 10, a high number of response categories is able to produce estimated measures $\hat{\beta}_n$ which are reasonably stable.

It is interesting to observe if there is an impact of the questionnaire size on the accuracy of the estimated mean difficulties $\hat{\delta}_i$. It is well-known that the goodness of δ_i estimates strongly depends on the number of respondents involved in the survey; in this simulation study this number (1000) is fairly high.

Figure 3a-d display the empirical 95% confidence intervals for the estimated mean difficulties. As the questionnaire length increases, the real mean difficulty δ_i comes closer to the empirical confidence interval. The estimates of the easiest and most difficult items are the ones which are mostly influenced by the size of the questionnaire; the δ_i of the easiest items are overestimated whereas the difficult parameters of the most difficult items are underestimated. Moreover, the number of categories slightly affects the goodness of the estimations; the bias is almost the same when three or five response categories are used.

CONCLUSIONS

The study deals with the evaluation of the impact of the questionnaire size on the accuracy and stability of the Rasch measure making use of simulated data. The quality of the obtained measures depends on meeting the hypothesis underlying the RM as well as the quality of the questionnaire used in terms of number of items and response categories.

The results obtained show an inverse relationship between, on one side, the length of the questionnaire and the number of categories and, on the other side, the accuracy and stability of the estimated measures.

All the types of questionnaires considered in the study are able to reproduce ability estimates with zero mean, as the real one, but only one type (five categories and 20 items) is able to produce estimates coming from a normal distribution, as in the real case.

The width of the empirical 95% confidence intervals decreases with the increase of the number of items and response categories, nevertheless the impact of increasing the questionnaire size is stronger when this size is small. Moreover, it is important to underline that the effect on the confidence interval and the stability of the estimated measure due to the number of the response categories is stronger than the effect obtained increasing the items number.

ACKNOWLEDGMENT

The author wishes to thank M. Carpita for useful and valuable discussions and the two anonymous reviewers for their useful comments.

REFERENCES

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

Bera, A.K. and C.M. Jarque, 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ. Lett., 6: 255-259.

Bond, T.G. and C.M. Fox, 2007. Applying the Rasch Model. Fundamental Measurement in the Human Sciences. 2nd Edn., Lawrence Erlbaum Associates Publishers, New Jersey. ISBN: 0805854622.

Brentari, E. and S. Golia, 2008. Measuring job satisfaction in the social service sector with the Rasch model. J. Applied Measurement, 9: 45-56.

King, J.A. and T.G. Bond, 2003. Measuring client satisfaction with public education I: meeting competing demands in establishing state-wide benchmarks. J. Applied Measurement, 4: 111-123.

Kubinger, K.D., 2005. Psychological test calibration using the rasch model-some critical suggestions on traditional approaches. Int. J. Testing, 5: 377-394.

Linacre, J.M., 2006. Winsteps: Rasch Measurement Computer Program. Version 3.60, Winsteps.com, Chicago.

Massey, Jr. F.J., 1951. The kolmogorov-smirnov test for goodness of fit. J. Am. Stat. Assoc., 46: 68-78.

Rasch, G., 1960. Probabilistic Models for some Intelligence and Attainment Tests. University of Chicago Press, Chicago, IL USA.

Tesio, L., 2003. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J. Rehab. Med., 35: 105-115.

Waugh, R.F., T.K. Hii and A. Islam, 2000. An approach to studying scale for students in higher education: A rasch measurement model analysis. J. Applied Measurement, 1: 44-62.

Wright, B.D. and G.N. Masters, 1982. Rating Scale Analysis. MESA Press, Chicago. ISBN: 0941938018.