On the stability of the Rasch measure

Silvia Golia
Dipartimento Metodi Quantitativi
Università di Brescia
golia@eco.unibs.it

Abstract: *The paper aims at evaluating of the impact of the questionnaire size on the reliability and stability of the Rasch measure. A simulative study is used in order to deal with the issue.*

Keywords: Rasch measure, stability, simulation study

1. Introduction

The issue of determining a reliable and objective measurement of a latent trait of interest is a crucial problem in the analysis of social and economic phenomena. One of the proposed methods to deal with this issue is the so-called Rasch model which allows one to transform ordinal raw scores coming from a questionnaire into interval scale measures. The goodness of the obtained measures depends on meeting the model assumptions as well as the quality of the questionnaire used. The present paper studies the impact of the questionnaire size on the reliability and stability of the Rasch measure making use of simulated data.

2. Simulation study

The *Rasch model* (RM) (Rasch, 1960) is a family of measurement models which converts raw scores into linear and reproducible measurement. Its distinguishing characteristics are: separable person and item parameters, sufficient statistics for the parameters and conjoint additivity. It requires *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). 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).

According to the RM, the probability that a person *n* answers in a given way, say *x*, to the item *i* depends on subject ability and how difficult the item is to endorse. For polytomously scored items, that is when there are *m*+1 possible ordered response categories for each item (coded as *x* = 0, 1,…, *m*), following the *Rating Scale Model* (RSM) (Andrich, 1978), this probability is given by:

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

where $\tau_0=0$. β_n identifies the *ability* of person *n*, δ_i the *mean difficulty* of item *i* and τ_j, called *threshold*, is the point of equal probability of categories *j*-1 and *j*; thresholds add up to zero, i.e. $\sum_{j=1}^{m}\tau_j=0$.

The present simulation study wants to investigate the stability of the measures estimated from simulated data sets involving the RSM defined in (1) and different items and thresholds sets. The data are generated as follows. A sample of 1000 abilities was drawn from a standardized normal distribution; these abilities are used in the data simulation and represent the target or *real abilities* β_n. This sample size was chosen because it represents an appropriate size for this kind of studies. Then, the response given by the subject *n* to the item *i* is obtained as follows: for all the categories, the response probabilities of type (1) and their cumulative sum are computed. Then, a random number *rn* is chosen from a uniform distribution on the interval [0,1] and compared with the cumulative sum; the first category with cumulative sum larger than *rn* is assigned to the response. Table 1 reports the sets of the item mean difficulties δ_i used in the present study, whereas [-0.5 0.5] and [-1 -0.5 0.5 1] are the two sets of threshold parameters τ_j utilized. These two sets imply three and five response categories respectively.

4 item	**5 items**	**7 items**	**10 items**	**15items**	**20 items**
-1.7684	-1.7684	-1.7684	-1.7684	-1.7684	-1.7684
				-1.4726	-1.4726
					-1.2740
			-0.9496	-0.9496	-0.9496
-0.8373	-0.8373	-0.8373	-0.8373	-0.8373	-0.8373
					-0.6370
				-0.5323	-0.5323
		-0.3092	-0.3092	-0.3092	-0.3092
				-0.2662	-0.2662
			-0.1237	-0.1237	-0.1237
	0.0000	0.0000	0.0000	0.0000	0.0000
					0.1578
		0.3092	0.3092	0.3092	0.3092
					0.5078
0.7783	0.7783	0.7783	0.7783	0.7783	0.7783
				0.9029	0.9029
			1.0733	1.0733	1.0733
					1.2454
				1.3682	1.3682
1.8274	1.8274	1.8274	1.8274	1.8274	1.8274

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

For each combination of item mean difficulties δ_i and threshold set τ_j, 100 data sets were simulated and analyzed and 100 sets of estimated abilities 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[1].

Table 2 reports the mean values of the person reliability index and 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 able (level of estimated ability bounded by the first and ninth deciles) subjects.

The person reliability index is an estimate of the replicability 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 computed with missing values. The

[1] The data simulation was performed using Matlab 6.5 whereas the Rasch analysis using Winsteps 3.64.

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

	4 item	5 items	7 items	10 items	15 items	20 items
			3 categories			
Reliability	0.46 (0.023)	0.55 (0.015)	0.66 (0.012)	0.81 (0.007)	0.82 (0.005)	0.86 (0.003)
Least able	4.88 (0.567)	4.21 (0.556)	3.55 (0.530)	2.91 (0.514)	2.35 (0.598)	1.88 (0.379)
Most able	4.87 (0.610)	4.23 (0.540)	3.48 (0.536)	2.29 (0.521)	2.20 (0.463)	1.93 (0.449)
Able	4.79 (0.664)	3.82 (0.557)	2.87 (0.393)	2.20 (0.267)	1.74 (0.202)	1.47 (0.166)
			5 categories			
Reliability	0.71 (0.012)	0.75 (0.009)	0.81 (0.007)	0.86 (0.003)	0.90 (0.003)	0.93 (0.003)
Least able	3.92 (0.565)	3.22 (0.551)	2.58 (0.568)	1.93 (0.339)	1.55 (0.293)	1.29 (0.232)
Most able	3.88 (0.518)	3.20 (0.460)	2.50 (0.496)	1.99 (0.449)	1.53 (0.283)	1.27 (0.219)
Able	3.34 (0.374)	2.60 (0.301)	1.96 (0.231)	1.53 (0.173)	1.21 (0.130)	1.02 (0.104)

Table 2: The mean value of the person reliability index and the mean width of the empirical 95% confidence interval for the ability estimation computed considering the least able, the most able and the able subjects. Standard errors in brackets.

The width of the empirical 95% confidence interval for the ability estimation shows an inverse relation with the number of used items and 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.

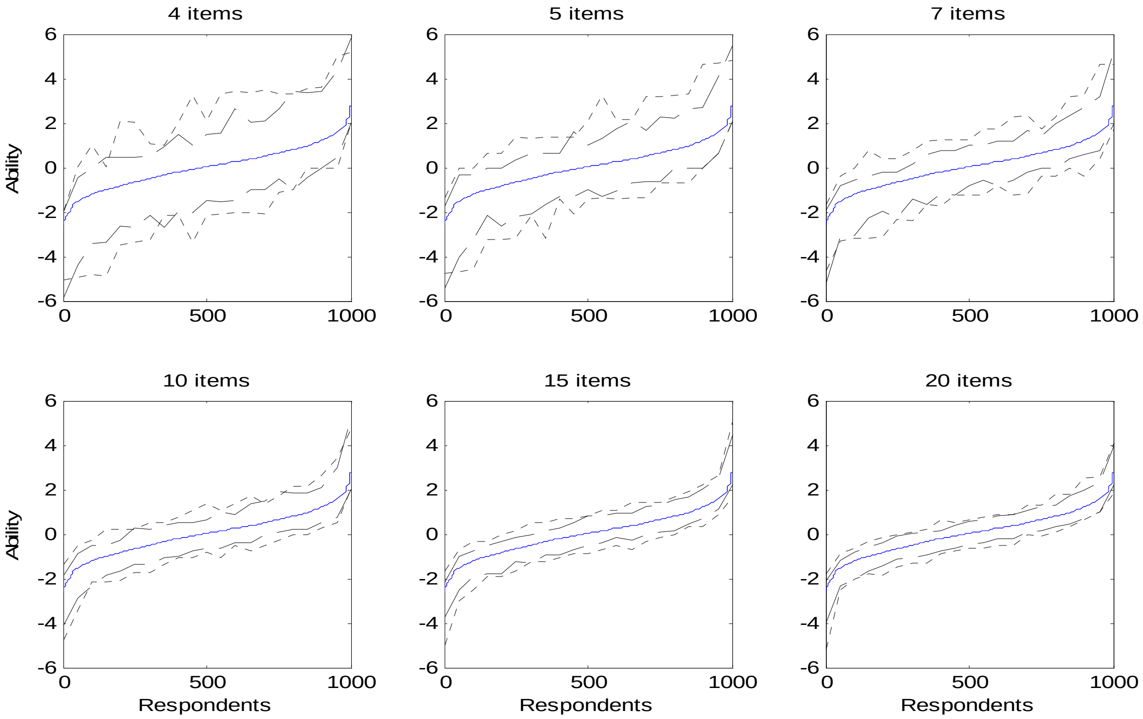

Figure 1: Graph of the real ability (solid line) and the empirical 95% confidence intervals smoothed using a cubic spline. Data obtained using 3 (dotted line) and 5 (dashed line) categories.

Figure 1 reports the graphs (ability versus respondents) of the empirical 95% confidence intervals, 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 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 seven, an high number of response categories is able to produce estimated measures $\hat{\beta}_n$ which are reasonably stable.

3. Conclusions

The paper deals with the evaluation of the impact of the questionnaire size on the reliability 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 stability of the estimated measures. The width of the empirical 95% confidence intervals decreases with the increase of the number of items and response categories. Moreover, 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.

Bibliography

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

Linacre J.M. (2006), *WINSTEPS Rasch measurement computer program*, Chicago: Winsteps.com

Rasch G. (1960), *Probabilistic models for some intelligence and attainment tests*, The Danish Institute of Educational Research, Copenhagen.