Point to Point

A Critique to "Using Rasch Analysis to Create and Evaluate a Measurement Instrument for Foreign Language Classroom Speaking Anxiety"

Panayiotis Panayides
Secondary Education, Cyprus

Apple (2013) published a study in a commendable attempt at bringing to light the many advantages of the Rasch models over Classical Test Theory (CTT) and other Item Response Theory (IRT) models and how they can be productively used in contexts where researchers are measuring foreign language anxiety (FLA). Of special importance, and much to Apple's credit, is the detailed procedure followed for the establishment of the unidimensionality of the Foreign Language Classroom Speaking Anxiety Scale (FLCSAS).

Even though the Rasch model was originally developed for use in educational testing, measurement pioneers such as Wright (1967, 1977, 1983, 1997, 1999), Andrich (1978), Masters (1982) and Linacre (1992, 1996, 1998, 2006) have taken the field to a different level. Now the Rasch models can address every reasonable observational situation in the social sciences.

Unlike other statistically-oriented IRT models, the Rasch models provide a mathematical framework of ideal measurement, against which test developers can assess their data. Real data can, and always do to some extent, deviate from ideal measurement due to random measurement error.

One of the advantages of the Rasch models over other IRT models is that they are the only models that use the raw score as a sufficient statistic for estimating item difficulty or person ability. This means that the sufficient statistic for estimating item difficulty is simply the sum or count of the correct responses for an item over all persons. Similarly, for person ability, it is the sum or count of the correct responses for a person over all items. This ensures that, despite the fact that items in a test or scale have different difficulty estimates, the raw score ranking or order is maintained for both item difficulties and person abilities, and this is consistent with the widely-used practice for reporting results.

One flaw in Apple's study lies in exactly this feature of the Rasch models. Apple uses a 20-item scale (each item on a 6-point Likert scale) and when items are ranked by difficulty as estimated by the mean score, the order is quite different from when they are estimated by the Rasch Rating Scale model (RSM). For example, item 20 has the smallest mean score of 1.82 (thus it is the most difficult item), and the Rasch item difficulty estimate is -0.65, making item 20 the fourth easiest. Item 16 is the ninth most difficult in the mean difficulty ranking (2.26) but the most difficult in the Rasch item estimate ranking (0.90). Item 19 is the fourth easiest in the mean difficulty ranking (2.95) but the third most difficult in the Rasch item estimate ranking (0.57). Figure 1 shows a scatter plot of the Rasch item estimates against the item mean scores. The three aforementioned items are the outliers in the figure.

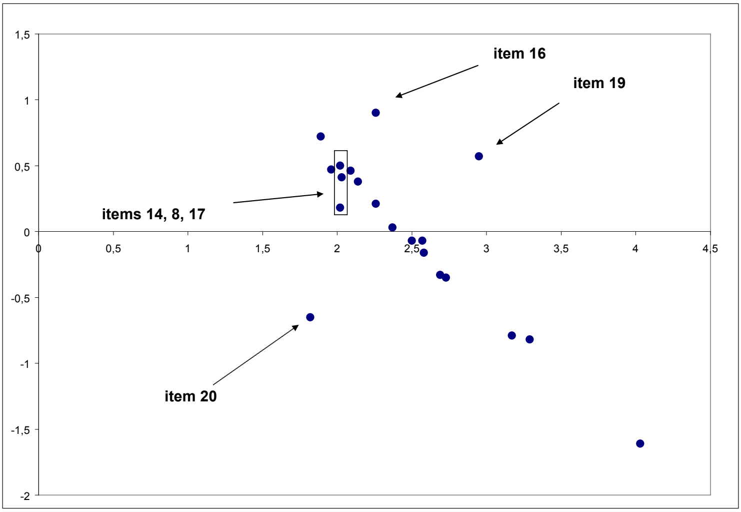

Figure 1. Scatter Plot of Rasch Item Estimates Against Item Mean Scores

The three items in the narrow rectangle (items 14, 8, and 17) are also problematic. Table 1 shows the details of these items. Even though they have essentially the same mean score, their Rasch estimates vary from 0.18 to 0.50 logits. The mean score shows that they are of the same difficulty, but in contrast the Rasch estimates show that item 17 is easier than item 8, which in turn is slightly easier than item 14.

Table 1. Mean Score and Rasch Estimates for Three Items

Item #	Mean score	Rasch estimate
14	2.02	0.50
8	2.03	0.41
17	2.02	0.18

Furthermore, the correlation between these sets of item difficulties is -.746. The negative sign is expected because a higher mean signifies an easier item and thus a lower Rasch item estimate. However, a value much closer to -1 was expected.

Apple (2013) referred to this change of item order by citing only items 16 and 20 and concluded that "This demonstrates how reliance on mean scores to judge which items are the best indications of levels of a psychological variable … may be potentially misleading" (pp. 20-21). He implies that this item order change is common practice in using the Rasch models, which in fact it is not. This could only occur when other IRT models are used, which employ parameters other than person ability and item difficulty (discrimination and guessing). However, this results in intercepting item characteristic curves and does not constitute ideal measurement because the fundamental assumption that *a more difficult item will always have a smaller chance of being answered correctly than a less difficult item* is violated.

Concluding Remark

A mistake must have been made either in the calculation of the mean scores or in the application of the Rasch RSM. If indeed such a mistake has occurred in the latter, it is obvious that the validation process has been distorted and Apple's results cannot be reliable.

Readers interested in the basics of the Rasch model and its applications to language education are advised to read Sick (2008a, 2008b, 2009a, 2009b, 2010, 2011). Also, readers interested in a detailed application of the Rasch RSM to FLA may like to read Panayides and Walker (2013).

Panayiotis Panayides holds a BSc in Statistics with Mathematics, an MSc in Educational Testing and a PhD in Educational Measurement (Durham University, UK). He is currently an assistant headmaster and head of the Mathematics Department at the Lyceum of Polemidia, Limassol, Cyprus. His research interests include educational and psychological measurement.

References

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

Apple, M. T. (2013). Using Rasch analysis to create and evaluate a measurement instrument for foreign language classroom speaking anxiety. *JALT Journal, 35*, 5-28.

Linacre, J. M. (1992). *Many-facet Rasch measurement*. Chicago: Mesa Press.

Linacre, J. M. (1996). The Rasch model cannot be "disproved"! *Rasch Measurement Transactions*, *10*, 512-514.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266-283.

Linacre, J. M. (2006). WINSTEPS (3.61.2) [Computer software]. Chicago: Winsteps.com

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Panayides, P., & Walker, M. J. (2013). Evaluating the psychometric properties of the Foreign Language Classroom Anxiety Scale for Cypriot senior high school EFL students: The Rasch measurement approach. *Europe's Journal of Psychology, 9,* 493-516. Retrieved from http://ejop.psychopen.eu/article/view/611.

Sick, J. (2008a). Rasch measurement in language education, Part 1. *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*, 1-6.

Sick, J. (2008b). Rasch measurement in language education, Part 2: Measurement scales and invariance. *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*, 26-31.

Sick, J. (2009a). Rasch measurement in language education, Part 3: The family of Rasch models. *Shiken: JALT Testing and Evaluation SIG Newsletter, 13*, 4-10.

Sick, J. (2009b). Rasch measurement in language education, Part 4: Rasch analysis software programs. *Shiken: JALT Testing and Evaluation SIG Newsletter, 13*, 13-16.

Sick, J. (2010). Rasch measurement in language education, Part 5: Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing and Evaluation SIG Newsletter, 14*, 23-29.

Sick, J. (2011). Rasch measurement in language education, Part 6: Rasch measurement and factor analysis. *Shiken: JALT Testing and Evaluation SIG Newsletter, 15*, 15-17.

Wright, B. D. (1967). *Sample-free test calibration and person measurement*. Retrieved from http://www.rasch.org/memo1.htm

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-115. doi:10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D. (1983). *Fundamental measurement in social science and education*. Retrieved from http://www.rasch.org/memo33a.htm

Wright, B. D. (1997) *Measurement for social science and education: A history of social science measurement.* Retrieved from http://www.rasch.org/memo62.htm

Wright, B. D. (1999). *Fundamental measurement for psychology*. Retrieved from http://www.rasch.org/memo64.htm