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

Matthew T. Apple
Ritsumeikan University

In his response to my paper concerning the use of the Rasch model for creating and evaluating foreign language classroom speaking anxiety (Apple, 2013), Dr. Panayides makes some interesting observations; however, there also appear to be several points of misinterpretation of the study results. The initial issue is his opening assertion that my paper was designed to show advantages of the Rasch model over classical test theory (CTT) models as well as item response theory (IRT) models. In fact, the paper was designed only to demonstrate the advantages of the Rasch model for Japan-based classroom teachers of English. I made no mention whatsoever of other IRT models. I also did not set the Rasch model against all CTT methods; I merely demonstrated that simple descriptive statistics were not as informative or useful as Rasch analysis when creating and evaluating questionnaires.

The main argument of Panayides's critique is that, because the correlation of mean raw scores and Rasch logits was not the "expected" 1.0, I had somehow miscalculated the raw scores or had not used Rasch analysis properly. As a point of clarification, I did fail to clearly indicate the difference in *N*-size between the mean scores in Table 1 (p. 14), which was produced to compare raw score results, and the Rasch item analysis in Table 2 (p. 15). Whereas Table 1 with traditional mean scores showed the original *N*-size of 172, Table 2 with Rasch item analysis showed the reduced *N*-size of 152, following the removal of 20 persons whose responses systematically misfit the model. The descriptive statistics were meant to show what a researcher with no knowledge of the Rasch model would have done. The researcher would not have known that 20 persons' responses misfit the model, because merely summing up raw scores from questionnaire items and then averaging them does not provide a measure of person fit.

The issue in the correlation-based argument is the assertion that, because both traditional mean scores based on classical test theory (CTT) and the Rasch model use the raw score as a sufficient statistic for the estimation of item difficulty, raw scores from Likert-type category data ought to correlate highly with Rasch logits. As Panayides states, for a typical Rasch model-based analysis of testing data, "[the] sufficient statistic for estimating item difficulty is simply the sum or count of the correct responses for an item over all persons." However, there are two important points to be made regarding the use of raw scores and the Rasch model.

First, there is a crucial distinction between CTT and the Rasch model, which Bond and Fox (2007) have made explicit:

> *To the extent that the data fit the Rasch model's specifications for measurement* [emphasis added], then *N* is the sufficient statistic. (Bond & Fox, 2007, p. 267)

Data obtained from a well-established questionnaire with a large *N*-size ($N > 1000$) may indeed show good item fit. However, the data in my paper were obtained from a newly created measurement instrument, which had many misfitting items; the *N*-size, while adequate, was not large and the participants were not well targeted by the items, which may have introduced measurement error. Additionally, data from a test, which has correct answers, and a Likert category scale, which has no correct answers, are necessarily different. The frequency of responses to the Likert response categories for each item may adversely affect item fit and thus measurement

of the construct (Linacre, 1999). Traditional raw scores take neither fit nor measurement error into account, nor do they consider differences in Likert category utility.

Second, the relatively low correlation of traditional mean scores to Rasch logits demonstrates that Rasch logits based on Rasch model analysis are not simply another type of descriptive statistics. The Rasch model is not a model of observed responses, but a model of the probability of the observed responses (Wilson, 2005, p. 90). In other words, the Rasch model attempts to predict the likelihood of a questionnaire respondent to give the same answer to similar items on a future iteration of the questionnaire. Raw scores are descriptions that do not take item or person fit, measurement error, or Likert category scale functioning into account. Rasch logits are the result of attempting to model probabilistic item responses as a function of the level of endorsability of the construct for both respondents and items.

As I mentioned in my discussion on the drawbacks of traditional statistics (pp. 7-8), there are two problematic assumptions with averaging or adding raw scores such as participant responses to Likert-scale items on a questionnaire in order to create an item mean score. First, Likert-type category data are not true interval data. With interval data, the distances between each adjoining pair of data points are required to be equal. Although the distances between points on a Likert scale may look equal on the surface, they may actually vary from person to person and item to item. Second, because Likert-type category data are ordinal and not interval, such data are not additive. Different questionnaire respondents may have very different perceptions of the distinction between a *strongly agree* and an *agree* for one item. Adding a 1 from one person's response to a 2 to another person's response doesn't really equal 3. Averaging the two responses doesn't really equal 1.5, either. Likewise, a response of a 3 on one item by one person is not necessarily the same as a 3 on another item by the same person. The use of mathematical averages ignores the possibility that responses from different people on the same items or the same person on different items may represent very different perceptions of the intended Likert categories. Thus, any correlation of means, based on raw scores, to Rasch logits of data from a Likert category scale seems of questionable value.

Additionally, I stated in the conclusion (p. 23) that not only did several items on the questionnaire need revision, but that, indeed, several studies had already used revised versions of the anxiety questionnaire. Each use of the questionnaire with new participant samples required new Rasch analysis for validation; this construct has proved valid and reliable for Japan-

based samples. Readers are invited to review Apple (2011) and Hill, Falout, and Apple (2013) for further information.

Finally, Panayides claims that I implied that "item order change is common practice in using the Rasch models." I did not imply this; I did, however, suggest that reliance on raw scores to judge item difficulties in a questionnaire may lead to erroneous conclusions about which items are more difficult than others to endorse, because raw scores do not take item fit or measurement error into account. I thank Dr. Panayides, and I thank the editors of *JALT Journal* for giving me this opportunity to respond to the issues raised and give clarifications. I hope this exchange of ideas will encourage *JALT Journal* readers to learn more about the use of Rasch model analysis for second and foreign language teaching and research.

References

Apple, M. (2011). *The Big Five personality traits and foreign language speaking confidence among Japanese EFL students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3457819)

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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Hill, G., Falout, J., & Apple, M. (2013). Possible L2 selves for students of science and engineering. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings* (pp. 210-220). Tokyo: JALT.

Linacre, J. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103-122.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.