Application of Rasch measurement theory in language assessment: Using measurement to enhance language assessment research and practice

Jason Fan[1], Ute Knoch[1] & Trevor Bond[2]
[1]Language Testing Research Centre, University of Melbourne
[2]James Cook University

Language assessment is a hybrid discipline which draws on expertise from both applied linguistics and measurement (Bachman, 1990; McNamara, 2011). While research and theories in applied linguistics inform the conceptualisations of language ability, expertise in measurement enables language assessment researchers and practitioners to ascertain the extent to which the target language constructs have been assessed in a reliable and valid manner. In fact, such knowledge and expertise have long been recognised as essential components of language assessment literacy (Taylor, 2013) and incorporated into the training of pre-service language testers in different contexts (e.g., Brown & Bailey, 2008; Jin, 2010). The Rasch model for measurement, an overarching umbrella term which encompasses a family of related models, is a data analysis technique which has been widely used in the field of language assessment for interrogating the technical quality of language assessment, and for enhancing test design in the interest of test validity and fairness (e.g., McNamara & Knoch, 2012; McNamara, Knoch, & Fan, 2019).

The Rasch model has been used in the field since the 1980s, and its application grew quite significantly in the 1990s, following the advent of the many-facets Rasch model (MFRM) (Linacre, 1989) which focussed on rater-mediated performance assessments. In addition, the publication of *Measuring Second Language Performance* by McNamara (1996) provided a catalyst for this trend, thanks to its accessible introduction to the MFRM in researching rater effects in performance language assessment such as writing and speaking (Bond, 2016). Nonetheless, the initial uptake of the Rasch model generated considerable controversy in the field at the time. As described by McNamara and Knoch (2012), 'the Rasch wars' were fought for a lengthy period and on several fronts. The focus of the debates includes the relationship between the Rasch model and the other Item Response Theory (IRT) models (i.e. 2PL and 3PL models) and whether it was at all appropriate for the analysis of language assessment data (i.e. 'the unidimensionality debate'). By the beginning of this millennium, 'the Rasch wars' were essentially over and language

Email address for correspondence: jinsong.fan@unimelb.edu.au

assessment researchers focussed on the practical benefits of utilising the Rasch model to investigate an array of questions which are germane to test reliability, validity and fairness (Bachman, 2000).

Featuring five full-length research papers and a book review of *Fairness, justice and language assessment* (McNamara et al., 2019), this special issue brings together researchers working across diverse language assessment contexts. The papers in this special issue were carefully selected from presentations delivered at the Pacific Rim Objective Measurement Symposium (PROMS) held in Shanghai, China in July 2018, with the theme of *Application of Rasch measurement theory in language assessment and across the human sciences*.

The first paper, by Koizumi, Kaneko, Setoguchi, In'nami and Naganuma, used the Rasch model in conjunction with generalizability theory (G-theory) to evaluate the validity of spoken interaction tasks based on the adapted Japanese version of the Common European Framework of Reference (CEFR-J). The study has two purposes: 1) to examine the measurement properties of the spoken interaction tasks; and 2) to examine whether the difficulty levels of these tasks align with the CEFR-J levels. MFRM analysis results indicate that the measurement quality of the tasks was positive overall, though some problems were also identified, such as the relatively large step calibrations in the rating scale. The results also suggest that the difficulty levels of the interaction tasks were largely consistent with the CEFR-J levels. G-theory analysis results provide insights into the number of tasks that were required to achieve a satisfactory level of reliability. This study has implications for the development of language assessment tasks based on language proficiency frameworks such as the CEFR; it also demonstrates how MFRM and G-theory can be combined for interrogating the technical quality and validity of performance language assessment.

The Park and Yan study investigated a theoretically and practically intriguing topic in the field of L2 performance assessment, namely, how rater behavior is influenced by different rating scales. The study was conducted in the context of a university-level English as a Second Language (ESL) placement test in the United States. Park and Yan compared raters' performance on two types of rating scales: a holistic scale and a binary, analytic scale, using the MFRM together with qualitative research methods, including semi-structured retrospective interviews and think-aloud protocols. MFRM analysis generated evidence concerning intra- and inter-rater consistency, indicated by raters' fit statistics and exact agreement percentages on the two types of rating scales. Qualitative data, on the other hand, provided a window into raters' rating processes. Results indicate that raters had lower rater agreement with the holistic rating scale and had a reduced cognitive burden when using the binary, analytic scale. The authors concluded that a binary, analytic scale might be used in a complementary manner with a holistic scale to improve rater reliability, and to provide more diagnostic feedback on students' writing performance. This paper serves as an excellent example of how the Rasch model could be used together with qualitative methods in L2 rating research to explore both the rating quality and the processes that raters engage with while assigning scores.

In the next paper, Zhu and Aryadoust examined gender-based differential item functioning (DIF) in the Pearson Test of English – Academic (PTE-A) reading test. They employed two methods in DIF detection: a Rasch-based DIF method using the partial credit model, and MIMIC (multiple indicators, multiple causes) method in the framework of structural equation modeling (SEM). They argued in their paper that this was probably the only study to date that employed both a Rasch and MIMIC method in DIF detection in the field of language assessment. The Rasch analysis was implemented to explore both uniform and non-uniform DIF in the reading test items, whereas SEM analysis was used to further investigate whether DIF existed through regressing the latent variable (i.e. reading ability in this case) and test items on the variable of interest (i.e. test takers' gender). Rasch analysis results were then compared with MIMIC results to ascertain whether measurement invariance could be established for the test items. Findings of this study indicate that the reading test functioned basically equivalently on male and female test takers, hence lending support to the fairness of this high-stakes English reading test. While demonstrating the power of Rasch analysis in DIF detection, this study also showcases the advantages of using the Rasch model and SEM in examining DIF from different perspectives.

The paper by Wang and Luo explored the rater effects on the writing subtest English Test for International Communication (ETIC), a relatively new English language test in China designed to assess English ability in the international workplace. Like Koizumi et al. (2019, this issue), they approached rater effects using both the MFRM and G-theory. Specifically, their investigation focussed on raters' consistency in assigning scores, including inter- and intra-rater consistency as well as consistency across different test takers and categories in the rating criteria. A generalizability study (G-study) and a decision study (D-study) under the G-theory framework were both implemented to investigate rater reliability and to inform rating designs. This was followed by MFRM analysis which was used to pinpoint inconsistent raters; in addition, the bias analysis functions were employed to explore raters' consistency across test takers and the different categories in the rating scale. This paper highlights the crucial roles that quantitative analytical techniques such as the MFRM and G-theory could play in revealing the technical qualities of large-scale performance assessment, and in enhancing test design in a context where large-scale language assessments play a crucial role in language education (Cheng & Curtis, 2010).

The last paper by Fan and Knoch features a systematic review of the published research that utilised the Rasch model in the field of language assessment and explored how the Rasch model has been used to investigate and enhance test fairness. Adopting the useful distinction between fairness and justice made by McNamara and Ryan (2011), *fairness* was conceptualised as the technical quality of a language assessment, whereas *justice* concerns the values in the test constructs, and the use, impact, and consequences of a language assessment. A total of 139 papers were collected from four high-impact international journals in the field. A qualitative research method (i.e. thematic coding) was applied to analyse the collected papers. Five prominent themes emerged from the data, representing the topics that were most frequently investigated by language assessment researchers using the Rasch model. The findings mirror previous reviews of Rasch-based studies in the field (e.g., McNamara & Knoch, 2012), signalling the crucial role that the Rasch model could play in enhancing the fairness of language assessment research and practice.

There are several salient features that make the studies reported in this issue unique. First, the collection of papers highlights the usefulness of the MFRM in researching rater effects in performance language assessment, resonating with the finding of Fan and Knoch (2019, this issue) that the topic of rater effects stood out as the most prominent theme from their coding of the published research. Among the four empirical papers in this special issue, three focus on performance assessment (two on writing and one speaking), all using the MFRM in their analyses. This further underscores the conclusion that 'the Rasch model has indeed become one of the default methods or analysis techniques to examine the technical quality of performance assessments' (Fan & Knoch, 2019, p. 136). Second, all four empirical studies reported in this special issue use the Rasch model together with other research methods, including qualitative methods such as think-aloud protocols (Park & Yan, 2019, this issue) and other quantitative methods, such as G-theory (Koizumi, et al., 2019, this issue; Wang & Luo, 2019, this issue) and SEM (Zhu & Aryadoust, 2019, this issue). These research methods help to elucidate a phenomenon from different perspectives, thus strengthening the validity of the research findings. This is particularly true for the collaborative application of the Rasch model and G-theory in researching the technical quality of performance assessment, as has been evidenced by previous studies (e.g., Lynch & McNamara, 1998; Sudweeks, Reeve, & Bradshaw, 2004). On a final note, these studies also showcase the applicability of the Rasch model in diverse assessment contexts, including classroom or school-based assessment (see Koizumi, et al., 2019 and

Park & Yan, 2019 in this issue) and large-scale standardised assessment (see Wang & Luo, 2019, and Zhu & Aryadoust, 2019 in this issue).

However, several areas still remain which could be considered for future studies by language assessment researchers. First, more complex Rasch models, though highly valuable, have not been utilized by language assessment researchers. For example, the mixed coefficients multinomial logit model (MCMLM), advanced by Adams, Wilson, and Wang (1997), takes into consideration the correlations between the latent traits, and could be applied to language assessments to address multidimensionality issues; the Rasch testlet model (Wang & Wilson, 2005) could be used to analyse data from testlet-based language assessments. Second, though the Rasch model has been used in combination with SEM (e.g., Zhu & Aryadoust, 2019), few attempts have been made by language assessment researchers to use the two data analysis methods sequentially and collaboratively as advised by Bond and Fox (2015), namely, using the Rasch model for quality control of instruments and imputing the person measures generated by Rasch analysis into a subsequent SEM-based path analysis (see also Fan & Knoch, 2019). Finally, it would be both timely and worthwhile to develop a set of best practice principles for applying and reporting Rasch-based research for language assessment researchers. This could be accomplished through learning from previous endeavors to develop best practice for other research methods or data analysis techniques, such as Ockey and Choi (2015) in the case of SEM, and Plonsky and Gonulal (2015), in the case of exploratory factor analysis. Given the increasing application of the Rasch model in the field, such an endeavor would help to improve the rigour and transparency of Rasch-based studies, and by so doing, language assessment researchers can utilise the Rasch model in a more appropriate and productive manner.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bond, T. G. (2016). Enhancing the capacity of English language teachers to develop English language testing: Lessons from the Orient. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The wiew from the Middle East and the Pacific Rim* (pp. 574-594). New Castle, UK: Cambridge Scholars Publishing.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.).* New York: Routledge.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing, 25*(3), 349-384.

Cheng, L., & Curtis, A. (Eds.). (2010). *English language assessment and the Chinese learner*. New York & London: Routledge, Taylor and Francis Group.

Fan, J., & Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? *Papers in Language Testing and Assessment 8*(2), 117-142.

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555-584.

Koizumi, R., Kaneko, E., Setoguchi, R., In'nami, Y., & Naganuma, N. (2019). Examination of CEFR-J spoken interaction tasks using many-facet Rasch measurement and generalizability theory. *Papers in Language Testing and Assessment 8*(2), 1-33.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

McNamara, T. (1996). Measuring second language proficiency. London: Longman.

McNamara, T. (2011). Applied linguistics and measurement: A dialogue. *Language Testing, 28*(4), 435-440.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 553-574.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and langauge assessment*. Oxford: Oxford University Press.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*(2), 161-178.

Ockey, G. J., & Choi, I. (2015). Structural Equation Modeling reporting practices for language assessment. *Language Assessment Quarterly, 12*(3), 305-319.

Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment 8*(2), 34-64.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*(S1), 9-36.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403-412.

Wang, J., & Luo, K. (2019). Evaluating rater judgments on ETIC Advanced writing tasks: An application of generalizability theory and Many-Facets Rasch Model. *Papers in Language Testing and Assessment 8*(2), 91-116.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149.

Zhu, X., & Aryadoust, V. (2019). Examining Test Fairness across Gender in a Computerized Reading Test: A Comparison between the Rasch-based DIF Technique and MIMIC. *Papers in Language Testing and Assessment 8*(2), 65-90.