Papers in Language Testing and Assessment, 2019, Volume 8, Issue 2, 117-142

Fairness in language assessment: What can the Rasch model offer?

Jason Fan & Ute Knoch
Language Testing Research Centre, University of Melbourne

Drawing upon discussions of fairness in the field of language assessment, this systematic review study explores how the Rasch model has been used to investigate and enhance fairness in language assessment. To that end, we collected and systematically reviewed the empirical studies that used the Rasch model, published in four leading journals in the field from 2000 to 2018. A total of 139 articles were collected and subsequently coded in NVivo 11, using the open coding method. In addition, matrix coding analysis was implemented to explore the relationship between the topics that were identified and the language constructs that constituted the focus of the collected articles. Five broad themes were extracted from the coding process, including: 1) rater effects; 2) language test design and evaluation; 3) differential group performance; 4) evaluation of rating criteria, and 5) standard setting. Representative studies under each category were used to illustrate how the Rasch model was utilised to investigate test fairness. Findings of this study have important implications for language assessment development and evaluation. In addition, the findings also identified a few avenues in the application of the Rasch model which language assessment researchers should explore in future studies.

Key words: The Rasch model, test fairness, language assessment, systematic review

Introduction

In recent years, language assessments are playing an increasingly significant role in education and social life in general, as evidenced, for instance, by the fact that an array of high-stakes decisions are made based on language assessment results, ranging from admissions into higher education to immigration and citizenship (e.g., McNamara,

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

2005; Taylor, 2009). Fairness is the perennial concern for language assessment research and practice, with considerable ramifications for test takers, test users, and other stakeholders such as teachers and parents. This is particularly the case when assessment results are used to make high-stakes decisions. In this paper, we explore how the Rasch model has been used to investigate and enhance fairness of language assessments. To this end, we collected and systematically analysed the research articles that utilised the Rasch model, published in the four leading international journals in the field of language assessment, including *Language Testing*, *Language Assessment Quarterly*, *Assessing Writing*, and *Papers in Language Testing and Assessment* from 2000 onwards when the Rasch wars were essentially over (McNamara & Knoch, 2012), and when language assessment researchers started to focus on the practical benefits of utilising the Rasch model (Bachman, 2000).

Test fairness

Despite the primacy of fairness in language assessment, test fairness has proven to be an elusive concept which is challenging to unpack. Indeed, it was not until quite recently that language assessment researchers started to conceptualise it more systematically. Kunnan (2004) spearheaded the discussions on fairness through presenting the Test Fairness Framework in which test fairness was believed to consist of five aspects: validity, absence of bias, access, administration, and social consequences. This framework was subsequently expanded in Kunnan (2008) where a complementary Test Context Framework was added such that the wider context of language assessments could be analysed. Xi (2010), in her discussions of test fairness, argued that previous approaches to fairness such as Kunnan (2004, 2008) were difficult to operationalise in practice as they failed to provide a mechanism to integrate evidence into a fairness argument; nor did they offer a means to plan research and set priorities. She observed that a crucial difference among the various approaches to fairness lay in their understanding of the relationship between fairness and validity. She summarised three approaches which were commonly adopted by language assessment researchers to view test fairness vis-à-vis validity: 1) fairness as a test quality which is independent of test validity; 2) fairness as an all-encompassing test quality which includes test validity; and 3) fairness as a fundamental test quality which is linked directly to validity. In her discussions, she defined test fairness as 'comparable validity for *identifiable* and *relevant* groups across all stages of assessment, from assessment conceptualization to the use of assessment results' (p. 154, italics in the original). Based on this definition, she proposed a new approach to investigating fairness, i.e., a fairness argument in a validity argument, which, in her opinion, 'offers

a principled approach to evaluating the soundness of the overall fairness argument and setting research priorities' (p. 167).

Drawing on the recent developments in the theory of argument-based validation (Chapelle, Enright, & Jamieson, 2008) and her working definition of fairness, Xi mapped out the warrants and assumptions in the validity argument for each interference, including domain description, evaluation, generalisation, explanation, extrapolation, and utilisation, together with the counter-arguments that would weaken the fairness argument for each inference. For example, in the case of TOEFL iBT, she argued that the counter-arguments for the inference of domain description might include: 1) assessment tasks are not equally representative of the academic domain for different groups; 2) critical English language skills, knowledge, and processes required for some sub-domains are not accessed; and 3) varieties of English included in the test are not representative of the domain (p. 159). Arguably, Xi's approach to test fairness provides a workable mechanism to collect and integrate evidence to investigate the fairness of a language assessment in practice.

McNamara and Ryan (2011) critiqued the different conceptualisations of test fairness in the field by arguing that 'the discussion of fairness and justice in language testing has emphasised the quality of testing procedures, while not engaging as critically or as productively with the question of the values in test constructs' (p. 166). Drawing on Messick's (1989) seminal theory on validity, they proposed a useful distinction between *fairness* and *justice* in language assessment. They reasoned that *fairness* is internal to a language assessment, referring to the technical quality of a language assessment; whereas *justice* is external to a language assessment, concerning the values in the test constructs, and the use, impact, and consequences of a language assessment. In the well-known validity matrix advanced by Messick (1989), *fairness* relates to the evidential basis of test validity, while *justice* to the consequential basis – the lower row of the matrix. Referring back to Xi's (2010) approach to test fairness, McNamara and Ryan argued that 'Xi elides the complexity of the highly contested values dimensions of tests, which need to be addressed in a different way, through direct political and ethical argumentation, discussions which will necessarily be open-ended, given that they are arguments about values' (pp. 166-167).

While aware of the different conceptualisations of test fairness in the field, we adopted McNamara and Ryan's (2011) distinction between fairness and justice in this paper primarily because of the focus of this study, that is, exploring how the Rasch model has been used to investigate and enhance test fairness. As a measurement model and a statistical technique (Bond & Fox, 2015), the Rasch model has considerable potential

which could be harnessed by language assessment researchers to investigate the technical quality of language assessment - the way in which test fairness is defined by McNamara and Ryan (2011). Having said that, we will discuss the major findings that emerge from this study in connection with other conceptualisations of fairness in the field.

It should be noted, however, that technical quality of a language assessment is reflected in multiple aspects. For a language assessment which does not involve human raters, technical quality might refer to the psychometric or statistical properties of items, tasks, or the whole test, which may encompass such aspects as difficulty level, reliability, dimensionality, or differential item or test functioning. In the case of rater-mediated language assessment, technical quality concerns the various facets in a measurement context, including, for example, rater severity, intra- and inter-rater reliability, task difficulty, and the functioning of rating scales (McNamara, 1996). Among these measurement facets, raters play a particularly vital role in the scores that test takers receive; as such, implications of raters for test fairness should be highlighted. If a test taker's performance happens to be rated by a more severe as opposed to by a more lenient rater, it is likely that the scores that are awarded fail to reflect their standing on the latent trait being measured, as rater severity (or leniency) has introduced variance in the test scores, thus causing construct contamination (AERA, APA, & NCME, 2014) or construct-irrelevant variance (Messick, 1989). In consequence, test fairness is compromised.

The Rasch model

The Rasch model is a probabilistic mathematical model which calibrates person ability and item difficulty on the same interval scale (Bond & Fox, 2015). The fundamental rationale underlying the Rasch model is that 'a person having a greater ability than another person should have the greater probability of solving any item of the type in question'; similarly, 'one item being more difficult than another means that for any person the probability of solving the second item is the greater one' (Rasch, 1960, p. 117).

The Rasch model is an overarching umbrella term which includes a family of models: the basic (or dichotomous) Rasch model, the Rating Scale Model (RSM), the Partial Credit Model (PCM), and the Many-Facets Rasch Model (MFRM) (McNamara, 1996; McNamara, Knoch, & Fan, 2019). The basic Rasch model was developed to analyse dichotomous data (i.e. responses that are either correct or incorrect) (Rasch, 1960). This model was later expanded to the RSM (Andrich, 1978) which could be used to analyse

Likert-scale type data, and the PCM (Masters, 1982) which is used to analyse responses where partial credit schemes have been applied. An even more exciting development is the invention of the Many-Facets Rasch Model (MFRM) by Linacre (1989) which is best suited to analyse data from rater-mediated performance assessments.

The Rasch model provides powerful means to investigate the technical quality of language assessments. The basic Rach model, RSM, and PCM have been typically used to examine the technical quality of non-rater-mediated assessments. A host of Rasch analyses can be implemented to generate evidence pertaining to the technical quality of language assessments, including model-data fit analysis, dimensionality analysis, and differential item functioning (DIF) analysis, to name but a few. Winsteps (Linacre, 2013) is the computer program which has been most frequently used by language assessment researchers to implement these analyses (McNamara & Knoch, 2012). Researchers can easily assess model-data fit through the infit and outfit mean square values as well as their standardised Z values in Winsteps output. Misfitting items need to be removed from score reporting as they contaminate the effective measurement of the latent constructs. Dimensionality analysis is another important function that Rasch analysis has to offer, which sheds light on the construct validity of a language assessment. In Winsteps, dimensionality analysis involves a factor analysis of the Rasch residuals to ascertain whether a meaningful secondary dimension exists, in addition to the primary dimension explained by the Rasch measure (Fan & Bond, 2019). The Rasch model has also proven to be effective in the investigation of DIF which concerns whether a test item functions equivalently on different groups of test takers with the same ability on the latent trait. Winsteps can provide the local difficulties of each item for test taker groups of interest and compare them through the Welsh t test.

The MFRM is the model which is usually utilised to research rater effects in performance assessment (Eckes, 2011). The MFRM calibrates the facets of interest in a measurement context (e.g., rater severity, rating criteria difficulty, test takers' ability on the latent trait) onto the same equal-interval scale. By so doing, the MFRM creates a single frame of reference, thus making it much easier to understand rater effects, or the effects of any other facets in the measurement context. MFRM analysis is most typically implemented in FACETS (Linacre, 2013), which generates a wealth of evidence relating to a particular measurement facet (McNamara & Knoch, 2012). In the case of rater facet, a chi-square test is provided as part of rater measurement report. A significant result is indicative that at least two raters have applied differential severity levels in their ratings. In addition, rater separation or strata statistics, as well as separation reliability estimates, are reported to provide additional information in

interpreting rater severity (Linacre, 2017). Rater consistency can be evaluated by: 1) infit and outfit mean square statistics of each rater; 2) the number of large standard residuals between observed scores and the scores expected by the Rasch model; 3) the correlation between a single rater and the rest of the raters (Myford and Wolfe, 2000). The 'fair scores' generated by the MFRM analysis compensate for rater severity or leniency by computing the scores that a test taker would receive if his or her performance is rated by a rater who is neither too severe nor too lenient. Finally, the bias or interaction analysis provides valuable information about whether different measurement facets (e.g., raters, rating criteria, test takers) interact in significant ways, thus impinging on the ratings and test fairness.

The Rasch model started to be used in the field in the 1980s, and experienced exponential growth from the 1990s. It is worth noting, however, that the rise of the Rasch model was not without controversy. In fact, considerable debates emerged concerning the application of the Rasch model, described by McNamara and Knoch (2012) as 'the Rasch wars'. According to McNamara and Knoch (2012), the Rasch wars were fought on several fronts, including the dimensionality debate (i.e. whether the Rasch model is suitable for analysing language assessment data) and the relationship between the Rasch model and the one-, two-, and three-parameter IRT models (i.e. whether discrimination and guessing should be incorporated in the model as legitimate parameters). In this paper, we will focus on the research that was published from 2000 onwards, when the Rasch wars were essentially over (McNamara & Knoch, 2012) and when researchers started to focus on the practical benefits of utilising the Rasch model (Bachman, 2000). Specifically, we will explore how the Rasch model has been utilised to investigate and enhance test fairness by investigating the following two research questions:

- What fairness topics were investigated by the Rasch model in the published research?
- How was the Rasch model used to investigate and enhance test fairness?

Method

Article collection

We collected the papers that were published in four international journals in the field of language assessment, including *Language Testing*, *Language Assessment Quarterly*

(LAQ), *Assessing Writing*, and *Papers in Language Testing and Assessment* (PLTA)[2] from 2000 to 2018. We targeted these journals because they are considered as high-impact, leading journals in the field. In addition, Assessing Writing has a special focus on rater-mediated language assessment. Keywords were used to search for the articles, including 'Rasch model' and 'Rasch measurement theory' as well as the different models in the Rasch model family, such as 'basic Rasch model'[3],'partial credit model', 'rating scale model', and 'many-facets Rasch model'. As a result, a total of 147 articles were collected. The search results were verified through going through all the issues published in the four journals during this period manually. Eight articles were found to be either position or review papers, focused on different topics. For example, McNamara and Knoch (2012) provide an engaging narrative on the rise of Rasch measurement in language assessment; Wind and Peterson (2018) report a systematic review of methods for evaluating rating quality in language assessment among which the Rasch model was mentioned as one of the methods. While these review papers offer insights into the role of the Rasch model in language assessment, they are not directly relevant to the present study. Therefore, these eight articles were excluded from analysis, yielding a total of 139 articles in our collection.

Analyses

The collected articles were analysed inductively, using qualitative research methods where themes were extracted through coding the data (Richards, 2014). Due to the exploratory nature of this study, open coding method was employed which means a coding scheme was not specified *a priori* but was generated through the coding process. One researcher coded the collected articles in NVivo 11 (QSR, 2012) and generated the themes. Another researcher then applied the themes to code 30 articles (21.6%) which were selected randomly from the collection. A high exact agreement percentage (93.3%) was reached between the two researchers. Discrepancies or disagreement about the themes were resolved through discussion.

[2] LAQ made its inception in 2004. Therefore, all articles that were collected in this journal were published from 2004 to 2018. PLTA was officially established in 2011, and before that, it was published as *Melbourne Papers in Language Testing* (MPLT). As such, the papers that were collected from PLTA include those published in MPLT from 2000 to 2011 and in PLTA from 2011 onwards.

[3] We are aware that authors might use variants of the same Rasch model. For example, instead of using 'basic Rasch model', some authors prefer to use 'dichotomous Rasch model'; similarly, instead of using 'many-facets Rasch model', some authors might use 'multi-facet Rasch model', 'multi-faceted Rasch model', or 'many-facet Rasch model'. All these variant terms were included in the search terms.

Five broad themes were extracted from the coding of the articles which include: 1) rater effects; 2) test design and validation; 3) differential group performance; 4) evaluation of rating criteria; and 5) standard setting (see Table 1). These themes represent the broad fairness topics that the Rasch model was utilised to explore or investigate in the published research. Matrix coding analysis in NVivo was then implemented to explore the relationships between the topics that were extracted and the language constructs that constituted the focus of the collected articles.

Table 1. Matrix coding results of topics and language constructs (n = 139)

Topics	Listening (n = 11)	Reading (n = 6)	Voc & Gram (n = 19)	Writing (n = 53)	Speaking (n = 29)	Translation (n = 2)	Multiple (n = 12)	Others (n = 7)
Rater effect (n = 64)	0	0	0	38	20	0	4	2
Test design and validation (n = 56)	7	5	18	11	8	0	5	4
Differential group performance (n = 20)	3	0	1	9	5	0	1	1
Evaluation of rating criteria (n = 12)	0	0	0	5	5	0	1	1
Standard setting (n = 7)	1	1	0	0	1	2	2	0

Notes. 1) The totals in some columns are larger than the frequency statistics on the top row because some papers were classified into more than one category. 2) 'Multiple' means more than one language skill is involved in the test being investigated; 3) 'Others' means the article focuses on abilities or skills which are not included in this table, including, for example, sign language (Bochner et al., 2016) and pragmatic ability (e.g., Roever, 2007; Youn, 2015).

Results

In this section, we will report the results in accordance with the five broad themes that were extracted through the coding process. We will use some illustrative studies under each category to demonstrate how the Rasch model was used to investigate and enhance test fairness.

Rater effects

As indicated in Table 1, the most prominently featured theme and skill combination were rater effects in writing and speaking assessments. The important topics under this category include rater severity and consistency, the effect of raters' background on their rating performance, and rater training. Each of these are further described below through illustrative studies.

Rater severity and consistency

A study by Eckes (2005) serves as an excellent example which demonstrates how the Rasch model can be used to explore rater severity and consistency. The study was conducted in the context of the writing and speaking sections of TestDaF, a high-stakes German test for those who apply to study in higher education in Germany. Using the MFRM as the primary method of data analysis, he investigated several questions which were considered crucial to the fairness of the writing and speaking tests in focus, including raters' severity and consistency, their use of the rating scale, and whether they applied differential severity across male and female test takers (an effect known as 'differential rater functioning').

Results revealed that the raters' severity was far from homogenous for both writing and speaking assessment, though in both cases raters' internal consistency was found to be satisfactory. The findings resonate with numerous previous studies on rater effects in performance assessment (e.g., Engelhard & Myford, 2003), suggesting that rater variability may exist even after raters are rigorously trained or certified. Given the high-stakes nature of this test, it is essential then to understand the extent to which raters' heterogeneous severity levels affect the decisions that were made of test takers, in this case, the levels into which they were assigned. By comparing the fair and observed scores, the author was able to show that 13.5% of the sample (n = 184) were misclassified in the writing section, and 17.1% (n = 208) in the case of speaking assessment. Arguably, these

classification decisions could have significant fairness ramifications for test takers, particularly those who were misclassified into a lower level. By employing multiple types of Rasch analysis, this study clearly demonstrates the significant role of the Rasch model in exploring rater effects in performance assessment. Similar Rasch-based studies focusing on rater effects can be found in Bonk and Ockey (2003), Goodwin (2016), and Kondo-Brown (2002), among many others in our collected articles.

In addition to high-stakes language assessments, the Rasch model has also been utilised to explore fairness issues surrounding classroom assessment procedures, such as students' self- and peer-assessment. When students rate their own performance, or the performance of their fellow students, are they more severe or lenient as compared with their teachers? Can they rate consistently? Investigating these questions has implications for the fairness of classroom-based assessment (CBA), especially in the contexts where assessment results are used to make decisions which are important to students (Bachman & Dambock, 2018). Such questions have been explored with traditional data analysis methods (e.g., Butler & Lee, 2010; Suzuki, 2015); however, the Rasch model, primarily the MFRM, has been increasingly utilised to investigate these questions.

Matsuno (2009) reported a study which was conducted in a second language (L2) writing classroom in Japan's higher education. The focus of this study was the comparison between three assessor types, i.e., self-, peer-, and teacher-assessor in the L2 writing classroom, in terms of their severity and consistency in rating writing performance. Results of MFRM analysis indicated that self-assessors, especially the high-achieving ones, tended to rate their own writing more harshly and their ratings were somewhat idiosyncratic. Peer-assessors did not display much variance in their ratings and were found to be internally consistent, irrespective of their writing proficiency levels. Another interesting observation was that teacher-assessors, though featuring satisfactory internal consistency, exhibited quite significant variance in their ratings. The findings led the author to conclude that peer-assessment could play a useful role in L2 writing classrooms.

A somewhat different picture, however, emerged from another study conducted by Esfandiari and Myford (2013) who compared the severity and consistency of self-, peer-, and teacher-assessors in rating essays in Iranian universities. Similar to Matsuno (2009), the MFRM was employed as the primary data analysis method. Results of this study indicated that the three types of assessors applied differential severity levels in rating the L2 essays. Teacher- and peer-assessors were found to be significantly harsher than self-

assessors; no significant difference, however, was identified between the average severity of teacher- and peer-assessors. In view of the similar average severity levels of teacher- and peer-assessors, the study suggested that peer-assessment could be potentially used as a useful summative procedure in L2 writing instruction. Despite the different findings, these studies demonstrate that the Rasch model could be fruitfully applied in CBA contexts, with considerable implications for the fairness of CBA procedures.

The effect of rater background and rater training

The Rasch model has been used to explore the effect of raters' backgrounds on their rating performance. Two background variables that have captured most attention from the field are raters' language background and their experience. When it comes to raters' language background, both their L1 and L2 have been the focus of investigation. Research focusing on raters' language background is significant in the sense that large-scale international language assessment programs such as the Test of English as a Foreign Language (TOEFL) are expanding their rater pools through recruiting raters from diverse L1 backgrounds (Xi & Mollaun, 2011), and that research on English as a Lingua France (ELF) communication is gaining momentum (Jenkins & Leung, 2014). Raters' backgrounds may introduce construct-irrelevant variance into their ratings, hence impinging on test fairness.

As part of a rater evaluation study, Yan (2014) investigated whether raters of different L1 backgrounds rated test taker subgroups defined by their L1s in the same way. The study was conducted in the context of the Oral English Proficiency Test (OEPT), a local English oral test used to select prospective international teaching assistants in a North American university. The study focused on raters with English and Chinese as their L1s and test takers with Chinese, Korean, and Indian as their L1s. The bias or interaction analysis in FACETS is typically implemented to investigate research questions of such a nature. Results indicated that L1 Chinese raters were significantly more lenient toward Chinese test takers, but more severe towards Indian test takers; L1 English raters were significantly more lenient toward Indian test takers. Reassuringly, the effect sizes in both cases were small, suggesting that the effects of such interactions tend to be peripheral on ratings.

Another interesting question emerges: does raters' L2 background constitute a potential source of bias in their ratings? A study by Winke, Gass, and Myford (2013) investigated

whether raters' L2 backgrounds interacted with test takers' L1 backgrounds in significant ways, thus causing bias in their ratings. In their study, the participating raters had L2 backgrounds in Spanish, Chinese, and Korean, whereas test takers were from Spanish, Korean, and Chinese L1 backgrounds. The findings indicated that L2 Spanish raters were significantly more lenient with L1 Spanish test takers; the same trend could be observed of L2 Chinese raters with L1 Chinese test takers. Resonating with Yan's (2014) study, the effect sizes of these interactions were found to be small, suggesting that they constituted an insignificant source of variability in ratings. Several other studies can be found in our collection, focusing on similar topics, including Wei and Llosa (2015), and Zhang and Elder (2011). Despite the reassuring findings, it should be noted that these studies have important implications for language assessment programs because even a small amount of variability might affect decision-making based on test scores. Therefore, it is recommended that assessment programs consider including specific modules in their rater training programs, which aim to sensitise raters to potential sources of construct-irrelevant variance in ratings, such as accent familiarity (Winke et al., 2013, p. 247).

In addition to language background, rater experience is another background variable which has attracted researchers' attention. Lim (2011) investigated the development and maintenance of rating quality in L2 writing assessment on a longitudinal basis. The study was conducted in the context of the writing section of the Michigan English Language Assessment Battery (MELAB), which is an international assessment of English proficiency. Using rater severity and consistency as the two indictors of rater performance, he compared the performance of novice and experienced raters over three time periods of 12 to 21 months. MFRM was used in this study as the primary data analysis method. The findings are encouraging. Novice raters might struggle with meeting the quality standards in the beginning, as compared with their more experienced counterparts; however, they rated consistently within a short period of time. Once their rating quality was aligned with the standards of experienced raters, they could maintain their level of performance over time. This study has important implications for rater training and certification, suggesting that 'the idea of an expert rater is potentially legitimate' (p. 557). The findings are largely consistent with Davis (2016) who observed negligible changes occurring to raters' severity and consistency after they had obtained a certain amount of rating experience.

It has been argued that training helps to ensure the reliability and validity of scores in performance language assessment (e.g., Fulcher, 2014; Luoma, 2004), and hence is crucial to its fairness. However, empirical research on the effects of training has yielded inconclusive findings. Several studies have been conducted to investigate the effects of training on rater performance using the Rasch model as the primary data analysis method. A longitudinal study by Knoch (2011a) examined the effectiveness of individualised feedback to rating behavior. The study was conducted in the writing and speaking sections of an English for Specific Purposes (ESP) test for health professionals over eight administrations. Several indicators were selected to represent rater behavior, including rater severity, consistency, and use of rating scales. The MFRM was used to model the rating data across the eight administrations. The findings were rather disappointing – raters who received individualised feedback did not perform any better than those who had not received feedback. It was further revealed through a follow-up qualitative study that raters seemed to have a positive attitude towards the feedback that was provided; however, no relationship was found to exist between perceptions of the feedback and its effect on rater behavior. The topic of rater training has been approached from different perspectives, using different methods. It can be envisaged that the Rasch model will continue to play a potent role in exploring the various avenues of rater training research, such as understanding the content, format, and duration of training, the timing of providing feedback, etc. Crucially, the continued efforts to research rater effects will have significant repercussions on the fairness of performance assessment practices.

Language test design and validation

Table 1 indicates that language test design and validation is the second important topic that was examined through the use of the Rasch model. This is reflected by the application of the Rasch model to investigate the statistical properties of language assessments, the difficulty levels of assessment tasks, the comparability between different assessment forms, and the effect of assessment method on test takers' performance. The Rasch model has been frequently used to examine the statistical properties of language assessments as part of their validation research. A Rasch-based validation study by Beglar (2010) serves as an excellent example to demonstrate the powerful functions of the Rasch model to interrogate the different aspects of test validity. The study focused on the Vocabulary Size Test (VST), a test of written receptive vocabulary size from the first 1,000 to the fourteenth 1,000-word families of English. To examine the validity of the VST,

various Rasch analyses were implemented in Winsteps, including item fit analysis, reliability analysis, dimensionality analysis, and DIF analysis. Results of these analyses provide compelling evidence in support of the validity of the VST.

In addition to generating evidence concerning test validity, the Rasch model has also been used to improve and refine test design. This is illustrated by a study by Lee-Ellis (2009) who developed a Korean C-test, and subsequently validated it through the Rasch model and revised it based on Rasch analysis results. Similar to Beglar (2010), multiple Rasch analyses were performed to interrogate the validity of the C-test in focus. Based on the analysis results, the misfitting items were removed through filling in the missing parts in the words which were mutilated. Furthermore, the C-test was streamlined to maximise its efficiency through removing a redundant C-test passage. This was accomplished through comparing the difficulty levels of each passage generated by the Rasch analysis which treated each passage as a super-item.

The Rasch model has also been utilised to compare test forms and to explore the effects of test method on test takers' performance. A recent study by Batty (2015) investigated whether the use of audio and video in listening assessment affects its difficulty level. The MFRM, which is typically used to investigate the quality of performance assessment, was adopted in this study. Results revealed that the two test formats, namely, audio- and video-mediated listening assessment, were at almost the same difficulty level, suggesting that the format had negligible effect on test takers' performance. In view of the findings, it was argued that test developers could decide on their own whether to use video or audio in listening assessment. Different from the studies focusing on performance assessment, this study shows that the MFRM could be usefully applied to explore the comparability of different test forms in non-rater-mediated assessments.

With the increasing use of computer technology in language assessment, an important fairness question is: what extent is a computer-mediated semi-direct speaking test comparable to a traditional direct face-to-face test? Exploring this question has clear implications for test fairness in the context where the semi-direct test is available to only part of the test takers. Kiddle and Kormos (2011) investigated this question, using the MFRM as one of their research methods. Rasch analysis results revealed that only minimal difference existed between the difficulty levels of the two test forms, thus suggesting that mode of administration did not affect the difficulty of the test. The studies that we reviewed in this section demonstrate that the Rasch model could play a vitally

important role in language test design and validation with significant implications for test fairness.

Differential group performance

The Rasch model has proven highly effective in exploring whether test taker subgroups with the same ability on the latent trait being measured have a differential probability of getting an item correct – a phenomenon known as DIF. Clearly, test items exhibiting DIF constitute a threat to test fairness, as the probability of getting an item correct is related to test takers' group membership, as opposed to their ability on the language construct. Aryadoust, Goh, and Kim (2011) investigated gender-based DIF in the Michigan English Language Assessment Battery (MELAB) listening test. A series of analysis techniques in the Rasch model were employed to detect the items that exhibit DIF because male and female test takers with the same ability had unequal probabilities of answering them correctly. Prior to the DIF analysis, the authors explored whether the data was suitable for Rasch analysis through analysing the fit of items to the Rasch model and test dimensionality as well as checking the local independence assumption. In this study, both uniform and non-uniform DIF analyses were implemented. While the former concerns whether an item favours male or female test takers, the latter further divides subgroups into high- and low-ability subsections and explores whether an item functions differentially on these subsections of test takers. As a typical Rasch-based DIF investigation, this study demonstrates the power of the Rasch model in DIF analysis, which has clear implications for test fairness.

Takala and Kaftandjieva (2000) argued that DIF analysis should not stop at the item level; rather, it should investigate whether and to what extent the items that exhibit DIF affect the total test scores, that is, at the whole test level. As part of a larger item bank construction project, the study investigated the potential gender effect on an L2 English vocabulary test. Similar to Aryadoust et al. (2011), this study used the Rasch model as the primary data analysis technique. To investigate uniform DIF, the study compared the item difficulty parameters which were estimated separately for female and male test taker subgroups. The study also examined the effect of the items that displayed DIF on total test scores. As revealed by the results, significant differences existed in the passive vocabulary of female and male test takers, which should be taken into consideration in test construction and the measurement of language proficiency, in the interest of fair score interpretations and use. It was also argued that the advice of excluding items that

exhibit DIF from item bank construction may be too restrictive, as their impact on total test scores need to be empirically verified.

In addition to DIF, the other subcategory under differential group performance is the analysis of raters' bias when rating test takers' language performance. This is typically implemented through the bias or interaction analysis function that is available in FACETS. Several studies explored the interactions between raters and test takers (e.g., Schaefer, 2008; Winke, Gass, & Myford, 2013; Yan, 2014), raters and rating criteria (e.g., Di Gennaro, 2009; Kondo-Brown, 2002), and raters and assessment tasks (e.g., Eckes, 2005; Kim, 2009). The findings of these interaction analyses, needless to say, have significant implications for understanding the technical quality of performance assessment, which is how fairness is interpreted in this paper. It can be anticipated that the Rasch model will continue to play a prominent role in researching differential group performance in language assessment, and this contributes substantially to our understanding of test fairness.

Evaluation of rating criteria

Rating criteria or rubrics are often used in the scoring process, and are seen to represent 'the *de facto* test construct' of performance assessment (Knoch, 2011b, p. 81). As such, it is essential to ensure that the rating criteria are properly developed and rigorously evaluated. Table 1 indicates that the Rasch model has been frequently used to evaluate the technical quality of rating criteria, mostly in writing and speaking assessment. Knoch (2009) represents a typical study in this regard. Her study was aiming to compare two rating scales for the writing assessment in an English for Academic Purposes (EAP) context: one scale with less specific descriptors whereas the other with detailed level descriptors which was therefore considered as more useful for diagnostic purposes. The MFRM analysis was implemented in FACETS to compare the functioning of the two rating scales in terms of their ability to discriminate test takers' writing ability, rater spread and agreement, and variability in the ratings. Analysis results indicated that the new scales, that is, the one with detailed level descriptors, featured a higher discrimination among test takers, smaller differences in raters' severity, greater rater reliability, and less variability in the ratings. The study has important implications for the development and use of rating scales for diagnostic purposes. For example, it suggests that not all analytical scales can be assumed to have the diagnostic functions; therefore, scale developers and assessment users should be mindful when selecting or developing scales for diagnostic assessment.

Janssen, Meier, and Trace (2015) serves as another example of using the Rasch model to evaluate rating criteria or rubrics. Employing the MFRM as one of the analysis techniques, their study evaluated the functioning of a well-known rating rubric, which was adapted for the writing component of a local English placement test. As pointed out by the authors, the rubric was also used by PhD program directors for different course level placements, which, in turn, were used as one admission criterion to determine students' entrance into the university's PhD programs – an apparently high-stakes use of the rubric. In their study, the MFRM was used to compare the functioning of the categories in the original and adapted rubric. The comparison focused on such aspects as understanding the fit and difficulty of the categories, identifying redundant scores, and determining whether the distances between adjacent categories were appropriate so that raters could distinguish them when rating. The multiple analyses brought the researchers to the conclusion that the rubric category scales contained too many possible scores. In consequence, the rubric was revised accordingly. Similar evaluation studies using the MFRM can be found in Bonk and Ockey (2003), and Youn (2015), among others.

Standard setting

As a procedure aimed at establishing cut points or scores, standard setting has implications for test fairness in the sense that it affects the classification of individuals into different performance levels (Kenyon & Romhild, 2014). Some of these classification decisions may affect test takers significantly, depending on the purpose of the assessment. In the standard setting process, a group of panelists is typically recruited to participate in workshops to review the performance standards and test takers' performance on language tests, and then allocate them into different performance levels. As such, the standard setting process involves panelists' subjective judgements, the quality of which could have appreciable impact on the establishment of the cut points, and consequently, the classification of individuals. The Rasch model can be used to examine whether and to what extent panelists agree with each other in their judgements.

Hsieh (2013) evaluated the quality of the panelists' judgements in a standard setting study in the context of an English assessment for elementary students. This study shows that the MFRM can be used to provide multiple strands of evidence concerning the reliability and validity of panelists' judgements in a standard setting procedure. For example, the variable map generated by FACETS can be used to evaluate panelists' levels of severity or leniency; the separation indices and their reliability estimates provide

evidence about variability in panelists' judgements. In addition, large infit and outfit statistics help to identify the less consistent panelists; the standardised residual plot can be used to detect aberrant decisions that are made by each panelist. These quality control indices help assessment users understand the recommended performance standards, which leads to more informed and fairer decisions. Similar studies can be found in Pill and McNamara (2016) and Kozaki (2004). These studies demonstrate that the Rasch model can be effectively applied to examining the reliability and validity of judgements made by panelists in standard setting process, thus enhancing the fairness in using the cut scores derived from standard setting procedures.

Discussion and conclusions

In this study, we explored how the Rasch model has been used to investigate and enhance test fairness through a systematic review of the research which was published in four leading international journals from 2000 to 2018, when the Rasch wars were essentially over (McNamara & Knoch, 2012). Using the open coding method, five broad themes emerged from the articles we included in the study, including rater effects, test design and validation, differential group performance, evaluation of rating criteria, and standard setting. We then used a few illustrative studies under each major category to demonstrate how the Rasch model was utilised in language assessment to investigate and enhance test fairness, defined as technical quality of language assessments in this study (McNamara and Ryan, 2011).

Our analysis results indicate that the Rasch model has been used extensively by language assessment researchers to explore a variety of research questions which are germane to test fairness. The trend is particularly pronounced in the case of performance assessment such as writing and speaking. As a matter of fact, it is not an exaggeration to claim that the Rasch model has indeed become one of the default methods or analysis techniques to examine the technical quality of performance assessments. It should be noted that compared with speaking assessment (n = 20), the Rasch model seemed to be more frequently used to investigate rater effects in writing assessment (n = 38), as indicated by the matrix coding results in Table 1. This may be related to the fact that writing is a language skill that is more likely to be included in language assessments than speaking, though admittedly, the latter is in recent years routinely included in major language

assessments. The finding may also be related to the journals that this study targeted. One of the four journals, namely, *Assessing Writing*, focuses on writing research only.

Two categories that were extracted from our review are related to rater and rating in performance assessment: rater effect (n = 64) and the evaluation of rating criteria (n = 12) (see Table 1). Three topics under the category of rater effect were most prominent, including rater severity and consistency, the effect of rater background on their rating performance, and the effect of rater training. Another category that pertains to human judgement is standard setting (n = 7) for both performance and non-rater-mediated assessments. These findings can be interpreted in connection with other conceptualisations of test fairness in the field, such as Xi (2010). Her approach to test fairness, as mentioned previously, is couched in the argument-based validation framework, where the topics that were identified about rater and rating are primarily related to the two inferences of evaluation and generalisation (Knoch & Chapelle, 2018). Following Xi (2010) and Knoch and Chapelle (2018), the counter-arguments that would weaken the test fairness argument relating to rater effect might include: 1) raters apply differential severities on identifiable groups of test takers; 2) raters rate inconsistently at task and test level; 3) detectable rater characteristics introduce construct-irrelevant variance into the rating process; and 4) raters are not thoroughly or regularly trained. As far as the rating criteria are concerned, the counter-arguments might include: 1) scale criteria are found to tap into different constructs than hypothesised; and 2) scale steps are inconsistent with the levels that appear in the scale. As shown in the illustrative studies, the Rasch model can be utilised to investigate the plausibility of each of the counter-arguments as set above. This is best evidenced by Eckes (2005), Kondo-Brown (2002), Winke, et al. (2013), and Esfandiari and Myford (2013), among many others.

In addition to rater-mediated performance assessment, the matrix coding results reveal that the Rasch model has also been used quite extensively to investigate the technical quality of the assessments which do not typically involve rater judgement, targeting such language abilities and skills as listening, reading, and vocabulary and grammar. This is manifested by several important topics under the second category, that is, test design and validation, including statistical properties of language assessments, the comparability between assessment forms, and the effect of test method on test takers' performance. In the argument-based validation framework, the studies along these lines are largely related to the inferences of evaluation, generalisation, and explanation (Chapelle, et al.,

2008). Applying Xi's (2010) approach to test fairness, some counter-arguments that would weaken the fairness argument might include: 1) test takers' responses to the items do not fit the expectations of the Rasch model; 2) a meaningful secondary dimension is identified through analysing the Rasch residuals; 3) test takers' performance on different assessment forms is not comparable; and 4) test method introduces construct-irrelevant variance into the assessment of test takers' language abilities. The illustrative studies such as Beglar (2010), Lee-Ellis (2009), and Batty (2015) serve as excellent examples demonstrating that the Rasch model can be effectively used to investigate the plausibility of these potential rebuttals.

Our coding results also reveal research gaps which language assessment researchers might consider exploring in the future. First, very few studies were found to be related to the two inferences of domain description and utilisation in the argument-based validation framework (Chapelle, et al., 2008; Xi, 2010). As noted previously, a few standard setting studies pertain to the utilisation inference, but the number is comparatively very small. Therefore, attempts might be made in the future to utilise the Rasch model to investigate the plausibility of the assumptions underlying these two inferences. For example, the RSM can be used to analyse the data of Likert-type questionnaires which are often used in the investigations related to these inferences. Secondly, the matrix coding results suggest that in terms of test design and validation, the Rasch model has been used more frequently in assessments of grammar and vocabulary (n = 18) than listening (n = 7) and reading (n = 5). In addition, it seems that the Rasch model has not been applied very often in the DIF research of non-rater-mediated assessments (see Table 1). We suggest that researchers consider using the Rasch model more frequently to improve the design and validation of listening and reading assessment, as well as for the DIF research in the future.

Another observation is that the Rasch model was used quite frequently together with other research methods, including qualitative methods or other quantitative methods. Having said that, the qualitative methods in most cases involved either interviews or linguistic analysis of test takers' performance. Future studies may attempt to use the Rasch model in combination with other qualitative methods supported by technology, such as eye-tracking (Conklin & Pellicer-Sánchez, 2016). Such combinational use may provide new insights into test validity and fairness. Researchers may also consider using the Rasch model collaboratively with other more advanced quantitative data analysis

techniques, such as structural equation modeling (SEM). In the case of SEM, Bond and Fox (2015) suggest that the Rasch model and SEM be used sequentially and in a complementary fashion to maximize the strengths of both methods. They advise that the Rasch model be used for quality control of the measurement instruments and to derive interval-level person ability estimates and their standard errors, which can be imported into subsequent SEM analysis. Such applications, however, are rarely seen in the field of language assessment. As a final note, very few studies in our collection applied the Rasch model to analyse data on a longitudinal basis (see Knoch, 2011a; Lim, 2011 for exceptions). The Rasch model, as noted by McNamara et al. (2019), is well suited to analyse data collected on multiple occasions. For example, it can be effectively used to map language ability growth or development longitudinally. Nonetheless, such mapping functions have been rarely attempted by language assessment researchers. By working along these lines, we believe that the powers of the Rasch model can be more fruitfully exploited by language assessment researchers in the interest of fairer assessment practices.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washtington, D.C.: AERA.

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Aryadoust, V., Goh, C. C., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8*(4), 361-385.

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.

Bachman, L. F., & Dambock, B. (2018). *Language assessment for classroom teachers.* Oxford: Oxford University Press.

Batty, A. O. (2015). A comparison of video-and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*(1), 3-20.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118.

Bochner, J. H., Samar, V. J., Hauser, P. C., Garrison, W. M., Searls, J. M., & Sanders, C. A. (2016). Validity of the American Sign Language Discrimination Test. *Language Testing, 33*(4), 473-495.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* New York: Routledge, Taylor & Francis Group.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5-31.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language.* New York and London: Routledge, Taylor & Francis Group.

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research, 32*(3), 453-467.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

Di Gennaro, K. (2009). Investigating differences in the writing performance of international and Generation 1.5 students. *Language Testing, 26*(4), 533-559.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2(3),* 197-221.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement*. Frankfurt: Peter Lang.

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model.* New York: College Board Research Report No. 2003-1.

Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*(2), 111-131.

Fan, J., & Bond, T. (2019). Unidimensionality and local independence. In V. Aryadoust & M. Rachelle (Eds.), *Quantitative data analysis for language assessment (Volume I): Fundamental techniques* (pp. 83-102). New York and London: Routledge.

Fulcher, G. (2014). *Testing second language speaking.* London: Routledge.

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing*, 30, 21-31.

Hsieh, M. (2013). An application of Multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing*, 30(4), 491-512.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, 26, 51-66.

Jenkins, J., & Leung, C. (2014). English as a lingua franca. In A. J. Kunnan (Ed.), *The companion to language assessment (*Vol. 4, pp. 1605-1616). Oxford, UK: John Wiley & Sons, Inc.

Kenyon, D., & Romhild, A. (2014). Standard setting in language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-14). Hoboken, NJ: John Wiley & Sons.

Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly, 8*(4), 342-360.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Knoch, U. (2011a). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing, 28*(2), 179-200.

Knoch, U. (2011b). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81-96.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477-499.

Kondo-Brown, K. (2002). A FACET analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing, 21*(1), 1-27.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. J. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27-48). Cambridge: Cambridge University Press.

Kunnan, A. J. (2008). Towards a model of test evaluation: Using the test fairness and wider context frameworks. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 229-251). Cambridge: Cambridge University Press.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing, 26*(2), 245-274.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

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

Linacre, J. M. (2013). *Facets computer program for many-facet Rasch measurement, version 3.71.4.* Beaverton, Oregon: Winsteps. com.

Linacre, J. M. (2017). *Facets computer program for many-facet Rasch measurement, version 3.80.0 user's guide*. Beaverton, Oregon: Winsteps.com.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 075-100.

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

McNamara, T. (2005). 21st century Shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 351-370.

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 language 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.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (Third ed., pp. 13-103). McMillan: American Council on Education.

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the Test of Spoken English assessment system (*TOEFL Research Report No. RR-65*). Princeton, NJ: Educational Testing Service.

Pill, J., & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing, 33*(2), 217-234.

QSR. (2012). *NVivo qualitative data analysis software*. Melbourne: QSR International Pty Ltd.

Rasch, G. (1960). *Probalistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Paedagogiske Institut.

Richards, L. (2014). *Handling qualitative data: A practical guide*. London: Sage.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165-189.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493.

Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing, 32*(1), 63-81.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*(3), 323-340.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21-36.

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT Speaking Tasks. *Language Assessment Quarterly, 12*(3), 283-304.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161-192.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 20*(10), 1-24.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501-527.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199-225.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50.