Using MFRM and SEM in the Validation of Analytic Rating Scales of an English Speaking Assessment

Jinsong Fan and Trevor Bond

Introduction

In second language performance assessment, both holistic and analytic rating scales are often used to award scores to test candidates. Whereas holistic scales express an overall impression of a test candidate's ability in one score, analytic scales contain a number of criteria, usually 3–5, each of which has descriptors at the different levels of the scale (Luoma 2004). Compared with holistic scales which give only one score, analytic scales have several discernible advantages including, for example, providing rich information about test candidates' language ability (e.g., Kondo-Brown 2002), and improving rating accuracy through drawing raters' attention to specific criteria of language performance (Luoma 2004). Moreover, as pointed out by Sawaki (2007), analytic scales are consistent with the current view of the multidimensional nature of language ability (see also In'nami and Koizumi 2012; Sawaki et al. 2009). As such, analytic rating scales are extensively used in L2 performance assessment such as speaking and writing (e.g., Bachman et al. 1995; Lumley 2002; Shin and Ewert 2015), particularly in the contexts where testing is more closely aligned with teaching and learning, and where rich feedback information is deemed crucial to test candidates (e.g., Sasaki and Hirose 1999).

Fulcher (1996, p. 208) argued that rating scales tend to be "*a priori* measuring instruments" in the sense that the descriptors in the rating scales are usually constructed by an expert through his or her own intuitive judgment concerning the

J. Fan (✉)
Fudan University, Shanghai, People's Republic of China
e-mail: jinsongfan@fudan.edu.cn

J. Fan
The University of Melbourne, Melbourne, Australia

T. Bond
James Cook University, Townsville, Australia

Q. Zhang (ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*,
DOI 10.1007/978-981-10-1687-5_3

nature of language proficiency, or sometimes in consultation with a team of other language experts. Such an approach to scale development, as Fulcher (1996) continued to argue, inevitably leads to the lack of empirical underpinning. Therefore, after a rating scale has been constructed, *post hoc* validity studies are essential to verify that the descriptors are meaningful indicators of test candidates' proficiency in a specific language modality (see also Upshur and Turner 1995). This view resonates with that of Knoch (2011, p. 81) who argued that rating scales act as "the *de facto* test construct" in performance assessment. It follows therefore that construct validation of the rating scale is crucial to the establishment of the construct validity of a particular assessment. In response to this call for *post hoc* validity research of rating scales, an array of validation studies have been reported, most of which are in the domain of L2 writing assessment (e.g., Lumley 2002; Sasaki and Hirose 1999; Shin and Ewert 2015; Upshur and Turner 1999) with few focused on the assessment of L2 speaking ability (e.g., Sato 2012; Sawaki 2007; Upshur and Turner 1999).

A review of the existent research reveals that most studies have adopted either the Generalizability-theory (G-theory) which represents an extension of the Classical Test Theory (CTT) (e.g., Sato 2012; Shin and Ewert 2015) or the Many-Facets Rasch Model (MFRM), one of the Item Response Theory (IRT) models (e.g., Upshur and Turner 1999); few studies, however, have adopted a combination of two different yet complementary data analytic approaches. One exception is that of Lynch and McNamara (1998) who employed G-theory and MFRM in the development of a L2 speaking assessment for intending immigrants. As articulated by the two researchers, the G-theory is able to take all the various facets of a measurement procedure into account, and to differentiate their effects, via the estimated variance components, on the dependability of decisions or interpretations made from test scores. On the other hand, MFRM helps to identify particular elements within a facet that are problematic, or "misfitting." Through utilizing the potential of G-theory and MFRM, this study illustrated the complementary roles of these two methodologies in the validation of L2 performance assessment. In a later study, Sawaki (2007) examined the construct validity of the rating scale for a Spanish speaking assessment designed for student placement and diagnosis, using multivariate G-theory and confirmatory factor analysis (CFA) in Structural Equation Modeling (SEM). Similar to the Lynch and McNamara (1998) study, Sawaki articulated the complementary roles of the two methodologies, i.e., G-theory and CFA, in her investigation. She argued that while the G-theory could estimate and differentiate the effects of various aspects of a measurement procedure on the dependability of decisions, the CFA modeling of the rating data helped researchers examine the convergent and discriminant validity of the analytic rating scale, as well as the weighting of analytic ratings in the composite score.

These two studies clearly demonstrate how the potential of contrasting data analytic approaches might be harnessed in examining test validity. It is worth noting that such a research design also concurs with recent developments of test validity theory which advocate that multiple strands of validity evidence should be collected, evaluated, and synthesized into a validity argument to support test score interpretation and use (e.g., Chapelle et al. 2008; Kane 2012). Equipped with the evidence generated by two different methodologies, validation researchers should be placed in a more advantageous position to interrogate the plausibility and accuracy of the warrants which are crucial to test validity, as well as the rebuttals which might weaken or undermine that validity (Kane 2012). Following this line of argument, this present study seeks to use MFRM and MTMM CFA model in SEM to examine the construct validity of the analytic rating scale of an English speaking assessment developed and used within a research university. Drawing upon the theory of interpretive validity argument (e.g., Kane 2012), this preliminary study is aimed at examining, through utilizing both MFRM and SEM, three warrants (and their respective rebuttals) which are critical to the validity of the speaking assessment: (1) Raters demonstrate sufficiently high reliability and similar severity in using the rating scale to award scores to test candidates; (2) The category structure of the rating scale functions as intended, and can effectively distinguish between test candidates at different levels of speaking proficiency; (3) Since the criteria in the rating scale represent different aspects of test candidate's L2 speaking ability (e.g., pronunciation, vocabulary, grammar), dimensions representing these aspects should be correlated, but at the same time, be distinct enough from each other. To put in another way, the test should display both convergent and discriminant validity (Sawaki 2007). Correspondingly, the three rebuttals are: (1) Raters do not demonstrate sufficiently high reliability and the same level of severity; (2) The category structure of the rating scale does not function appropriately, and thereby fails to distinguish test candidates at different levels of speaking ability; and (3) The correlations between the ability dimensions are negligible, or cannot be neatly distinguishable from each other. In this study, MFRM and SEM are used to examine these three warrants (and their respective rebuttals). Consequently, evidence in favor of these three warrants would show support for the construct validity of the rating scale, and hence the validity of the speaking assessment (Knoch 2011); conversely, lack of such evidence, i.e., evidence in favor of the rebuttals, would weaken or undermine claims for the construct validity of the rating scale.

MFRM and SEM

MFRM is a development of earlier Rasch models (e.g., dichotomous model, partial credit model) that incorporates multiple facets of the measurement procedure (Bond and Fox 2015). A facet of measurement is an aspect of the measurement procedure which the test developer claims might affect test scores, and hence needs to be investigated (Linacre 2013). Examples of such facets include the severity of rater judgments, task or item difficulty, and rating scale category options. All estimates of the measurement facets are calibrated on a single equal-interval scale (i.e., the logit scale), thereby creating a single frame of reference for interpreting the results of the analysis. Facets are estimated concurrently so they may be examined separately. Importantly, MFRM provides information about how well the performance of each individual examinee, rater, or task matches the expected values predicted by the strict mathematical model generated during the analysis. Therefore, MFRM can help researchers detect particular elements within any facet that are "misfitting", i.e., deviating from the expectations of the mathematical model. The "misfitting" element could be a rater who is unsystematically inconsistent in applying the ratings, a task that is unexpectedly difficult, or a person whose responses are inconsistent (Lynch and McNamara 1998). In MFRM analysis, the fit statistics are calculated from the item/person residuals and are reflected in Infit and Outfit Mean Square values, both with an expected value of 1.0 (Bond and Fox 2015).

In addition to fit statistics, the MFRM analysis also reports the reliability of separation index and the separation ratio. These statistics describe the amount of variability in the measures estimated by the Rasch model for the various elements in the specified facet relative to the precision by which these measures are estimated. The reliability of separation index for each facet ranges between 0 and 1.0, whereas the separation ratio ranges from 1 to infinity (Linacre 2013). The interpretation of these two statistics, however, is different for various facets. Low separation index for the examinee facet indicates lack of variability in the examinees' ability which might be symptomatic of central tendency errors, meaning that the raters do not distinguish the performance of test candidates at different ability levels. Conversely, low values of these two statistics for the rater facet are indicative of an unusually high degree of consistency in the measures for various elements of that facet. Once parameters of the model have been estimated, interaction effects, such as the interaction between raters and rating criteria, or between raters and examinees, can be detected by examining the standardized residuals (i.e., standardized differences between the observed and expected ratings) (Eckes 2011).

Thanks to its unique advantages, MFRM has been extensively used in the fields of language assessment, educational and psychological measurement, and across the health sciences (e.g., Bond and Fox 2015; McNamara 1996; McNamara and Knoch 2012). In the field of language assessment, MFRM typically is used in rater-mediated performance assessments such as speaking or writing assessments where a score is the result of the interaction between the rater, the task, the criteria, and the examinee (Batty 2015). In particular, this analytic approach has formed the cornerstone of the descriptor scales advanced by the Common European Framework of Reference (CEFR) (e.g., North 2000; North and Jones 2009). For example, *the Manual for Relating Language Examinations to the CEFR* clearly illustrates how to use MFRM to measure the severity (or leniency) of raters, assess the degree of rater consistency, correct examinee scores for rater severity differences, examine the functioning of the rating scale, and detect the interactions between facets in writing assessment data (Eckes 2011).

In comparison with MFRM, SEM has been more widely applied for various purposes in language assessment research. Also referred to as analysis of covariance structures and causal modeling (Kunnan 1998), SEM is a comprehensive statistical methodology that "takes a confirmatory (i.e., hypothesis-testing) approach" to the analysis of a structural theory bearing on some phenomenon (Byrne 2006, p. 3), and to test theoretical hypotheses about the relationships among observed and latent variables. It is a family of statistical techniques that includes confirmatory factor analysis, structural regression path, growth, multiple-groups, and MTMM models. The purpose of SEM is to examine whether the hypothesized relationships among variables are supported by empirical data. Usually, a model is specified *a priori* according to substantive theory, common sense, or a hypothesis to be tested. SEM is then used to estimate the discrepancy between the variance-covariance matrix as implied by the model and the observed variance-covariance matrix of the empirical data. The discrepancy is indicated by Chi-square statistics. The smaller the Chi-square value, the closer the data fit the model. In addition to the Chi-square test, a host of goodness of fit indices have been proposed to assess data/model fit, the most essential among which are CFI[1] (>0.90),[2] GFI (>0.90), SRMR (<0.05), and RMSEA (<0.05, with narrow 90 % confidence interval) (see e.g., Byrne 2006; In'nami and Koizumi 2011). When the fit is satisfactory, the model is considered to be an approximate representation of the relationships among the variables in the model. It represents one plausible explanation until future evidence falsifies this explanation (Xie and Andrews 2012).

As noted earlier, SEM techniques have been used in language assessment research for various purposes, including assessing the internal structure of a language test through structural modeling of the test data (e.g., Sawaki et al. 2009), assessing the effect of test methods on test performance (e.g., Llosa 2007), assessing equivalency of models for different populations (e.g., In'nami and Koizumi 2012), and understanding the effects of test tasks and strategy use on test performance (e.g., Kunnan 1995; Purpura 1999). SEM has also been used by language assessment researchers to investigate properties of questionnaires [see Kunnan (1998), and Ockey and Choi (2015) for a summary of the applications of SEM in language assessment research]. Despite the increasingly extensive applications of both MFRM and SEM in the field of language assessment, few attempts have been made to tap into the potential of these two different analytic approaches through combining them in test validation research. On the one hand, MFRM could function as "a magnifying glass," enabling researchers to examine closely the response patterns of individual examinees, raters, and tasks (e.g., Sawaki 2007, p. 357); SEM, on the other hand, allows researchers to hypothesize theoretical models which represent the factorial relationships between and among the variables under investigation, and to test the fit between the hypothesized model and the test data. Whereas MFRM analysis functions as the magnifying glass, SEM can provide validity evidence from a broader perspective through examining whether the hypothesized relationships between the various criteria in the rating scale are supported by the rating data, and whether such relationships are consistent with the substantive theory about language ability. Therefore, the evidence generated by

[1] CFI: Comparative Fit Index; GFI: Goodness of Fit Index; SRMR: Standardized Root Mean Residual; RMSEA: Root Mean Square Error of Approximation.

[2] The numbers in brackets are indicative of acceptable goodness of fit between the model and the empirical data.

both MFRM and SEM should be conducive to the construction of a more convincing validity argument for this speaking assessment, thus enabling us to provide a more compelling validity narrative.

Context of this Study

The Fudan English Test (FET)

In China, the College English Test (CET) has been recognized as a reliable and valid instrument in assessing university students' English language proficiency and achievement. However, recent years have witnessed the CET coming under heavy criticisms from some educators and researchers for its test format (e.g., heavy reliance on the multiple-choice questions), lack of alignment between the CET and the teaching curriculum developed within any particular university, and its rather negative washback effect on English teaching and learning at the tertiary level (e.g., Han et al. 2004). Though many of these criticisms might be seen as politically motivated or emotionally charged rather than empirically grounded, some high-ranking universities in China are attempting to develop their own English language tests in the hope of addressing the deficiencies of the CET and better aligning English testing with English teaching and learning within those university settings (see e.g., TOPE Project Team 2013; Tsinghua University Testing Team 2012). It is in this context that the FET project was initiated at Fudan University (FDU) in 2010 (see e.g., Fan and Ji 2014).

The FET is developed by the College English Center of FDU, one of the most prestigious institutions of higher learning in China. The test was formally launched in 2011, following a number of trials and pilot studies, and is currently administered once a year by FDU's Academic Affairs Office (AAO) to non-English major undergraduates. According to *the FET Test Syllabus* (FDU Testing Team 2014), the purpose of the FET is twofold: (1) to measure accurately students' English abilities and skills as reflected in the English teaching syllabus at FDU, and (2) to promote a more positive washback effect on English teaching and learning within FDU. Since September, 2011, all newly enrolled undergraduates at FDU have been required to take the FET, and to pass it within the four years of their Bachelor's program. A school-based English test notwithstanding, the FET is a reasonably high-stakes test because according to the AAO, the test is treated on a par with a compulsory English language course, which accounts for two credits in students' GPA calculations. The past few years since the inception of the FET have seen the number of test candidates increasing steadily. During the first FET administration in December 2011, 1337 students took the test,[3] and the number soared to 3575 during the most recent administration in December 2015.

[3] Typical annual undergraduate enrollment at FDU is around 3000.

Drawing upon recent models of communicative language ability and communicative language use (e.g., Bachman and Palmer 1996, 2010), the FET is designed to assess students' English language abilities in the four modalities of listening, writing, reading, and speaking, each accounting for 25 % of the test score (FDU Testing Team 2014). Previous research indicates that the FET is, on the whole, a reliable test, with internal consistency reliability coefficient reported at 0.83 (Fan and Ji 2013). Confirmatory factor analyses suggest that there is a higher-order general language competence factor and four first-order factors representing listening, reading, writing, and speaking, lending support to the construct validity of the test as well as its current score-reporting policy, i.e., reporting a composite score and four profile scores on the four subskills (Fan et al. 2014b). Furthermore, students were found to demonstrate a generally positive attitude toward the FET (Fan and Ji 2014; Fan et al. 2014a), in particular the listening, reading, and writing components. Previous research, however, has also indicated that students' attitude toward the speaking component tended to be more negative in light of the design, rating, and testing environment (Fan and Ji 2014). One concern voiced by test candidates, as previous research suggested, is that richer feedback information as to their speaking performance was lacking. An analytic scale was therefore developed to replace the holistic scale that had been used in the FET speaking component. This present study represents a preliminary attempt to validate this analytic scale developed for the FET speaking component.

The Speaking Component of the FET

The FET speaking component is a computer-mediated assessment of students' English speaking ability, administered in language laboratories. In this mode of speaking assessment, computers are used to present the tasks, and to capture students' speaking performance (Shohamy 1994). The FET speaking component comprises three tasks. In Task 1, students listen to an English passage of approximately 300 words, and respond to one or two questions based on the passage they have heard; in Task 2, students comment briefly on a topic which is mentioned in the input text; in Task 3, a graph or chart is presented on the computer screen, and students are required to describe and comment on the graph or chart. The speaking test takes about 14 min to complete. In light of the test purpose as well as previous research (e.g., Fan and Ji 2014; Fan et al. 2014a), an analytic rating scale was deemed to be more appropriate in this testing context.

The rating scale was developed on the basis of a comprehensive review of English speaking ability theory (e.g., Luoma 2004), as well as the English teaching and testing syllabus at FDU. The scale was designed to include the following four dimensions: (1) pronunciation; (2) content; (3) grammar; (4) vocabulary, all on a 4-point Likert-style scale (1-Very Poor; 2-Poor; 3-Moderate; 4-Good). Detailed descriptions accompanying each of those levels were drafted and provided. After the descriptors were drafted, they went through numerous content revisions based on the feedback through expert reviews and panel discussions. Given the centrality of *post hoc* validity research for rating scales (e.g., Fulcher 1996; Knoch 2011), this preliminary study was conducted to examine the validity of this rating scale, and to suggest directions for its improvement in the future.

Methodology

Participants

Due to the exploratory nature of this research, convenience random sampling was employed whereby emails were sent to the prospective participants of this study, calling for their participation. Consequently, a total of 74 students participated in this study on a voluntary basis with 35 males (47.3 %) and 39 females (52.7 %). Most participants had the experience of taking the FET at least once, and therefore understood the format of the FET speaking test. To ensure that each participant was familiar with the testing procedures, a package of testing materials was sent to each of the participants one month prior to the administration, including a brief introduction to the FET speaking test, sample test papers, and marking criteria. In addition, two FET certified raters were invited to participate in this study. The two raters were both very experienced in marking the FET speaking test, and had been directly involved in the development of the analytic rating scale.

Data Collection

We used test items from the FET item bank which were written by certified item writers, and had survived earlier moderation meetings and pilot studies. Students were arranged to take the test in two language laboratories. The testing procedures simulated, as closely as possible, an authentic FET speaking test. After all test takers had completed the recordings, the two raters rated students' performance, using the analytic scale developed for this study. For the sake of data connectivity, we followed Ecke's (2011) suggestion in rating design, wherein Rater 1 rated Examinees 1–45 and Rater 2 rated Examinees 30–74 (see also Linacre 2013). Each rater was required to rate students' performance on each of the three tasks (i.e., responding to question, short comment, and graph/chart description and comment) on each of the four language aspects (i.e., pronunciation, content, grammar, and vocabulary), generating a total of 12 scores for each student (i.e., 3 tasks × 4 aspects). In total, each rater awarded 540 scores (i.e., 45 examinees × 3 tasks × 4 aspects).

Data Analysis

In this study, MFRM was first of all utilized to analyze the rating data to examine the first two warrants in relation to rater reliability and severity, and the category structure of the rating scale. Given this research scenario, a four-facet Rasch model was used which included examinee ability (74 elements), task difficulty (3 elements), rater severity (2 elements), and the difficulty of the language aspects (4 elements). The mathematical expression of this four-facet Rasch model is presented below:

$$\log\left(P_{nijmk}/P_{nijm(k-1)}\right) = B_n - D_i - C_j - T_m - F_k$$

where $P_{nijmk}/P_{nijm(k-1)}$ is the probability of examinee n receiving a rating of k in relative to $k - 1$ from rater j on criterion i for task m; B_n is the ability of examinee n; D_i is the difficulty of criterion i; C_j is the severity of rater j; T_m is the difficulty of task m; and F_k is the difficulty of receiving a rating of category k relative to immediately lower category $k - 1$. FACETS 3.71.0 (Linacre 2013) was implemented to perform MFRM analysis in this study.

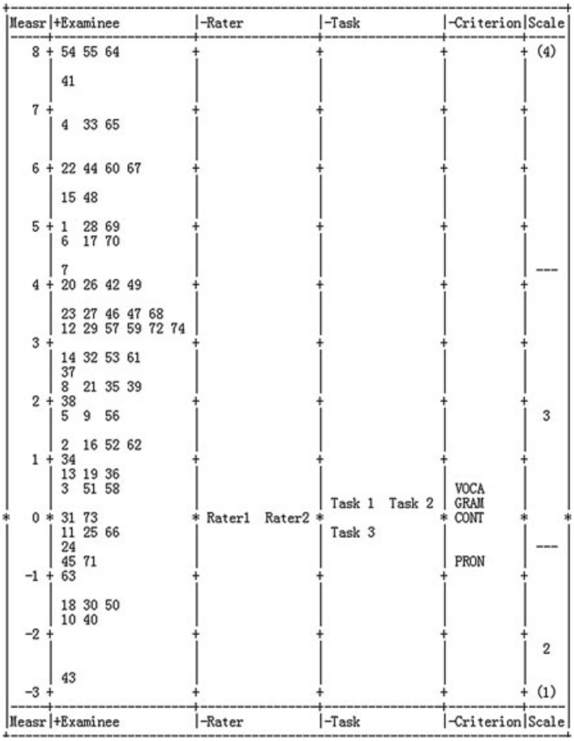
To examine the third warrant about the convergent and discriminant validity of the rating scale, the MTMM CFA model in SEM was utilized to model the test data. Based on Byrne (2006) and Kunnan (1998), the SEM analytic procedures followed three steps: (1) Model specification, i.e., specifying the hypothetical MTMM models; (2) Model evaluation, i.e., evaluating the fit between the hypothesized MTMM models and the test data; and (3) Model comparison, i.e., comparing the fit of the baseline MTMM model and the alternative competing models. The SEM analysis in this study was performed with EQS 6.3 (Bentler and Wu 2005).

Results and Discussion

MFRM Analysis

Heeding the warning that "lack of connectedness among elements of a particular facet would make it impossible to calibrate all elements of that facet on the same scale" (Eckes 2011, p. 110), we first of all examined the connectedness of the resulting data set in the FACETS output. The result suggested that the rater allocation design adopted in this study was unproblematic and provided for sufficient links between all facet elements. Next, we inspected the variable map generated by the FACETS analysis. The variable map, regarded as a distinctive advantage of Rasch analysis, can illustrate graphically the estimated locations of elements in each facet on the same interval-level measurement scale, containing a wealth of basic information that is central to Rasch measurement (Bond and Fox 2015). Figure 1 displays the variable map representing the calibrations of examinees, raters, tasks,

Fig. 1 The variable map



```
+------------------------------------------------------------------------------+
|Measr|+Examinee          |-Rater           |-Task            |-Criterion|Scale|
|-----+-------------------+-----------------+-----------------+----------+-----|
|   8 + 54 55 64          +                 +                 +          + (4) |
|     |                   |                 |                 |          |     |
|     | 41                |                 |                 |          |     |
|     |                   |                 |                 |          |     |
|   7 +                   +                 +                 +          +     |
|     | 4  33 65          |                 |                 |          |     |
|     |                   |                 |                 |          |     |
|     |                   |                 |                 |          |     |
|   6 + 22 44 60 67       +                 +                 +          +     |
|     |                   |                 |                 |          |     |
|     | 15 48             |                 |                 |          |     |
|     |                   |                 |                 |          |     |
|   5 + 1  28 69          +                 +                 +          +     |
|     | 6  17 70          |                 |                 |          |     |
|     |                   |                 |                 |          |     |
|     | 7                 |                 |                 |          | --- |
|   4 + 20 26 42 49       +                 +                 +          +     |
|     |                   |                 |                 |          |     |
|     | 23 27 46 47 68    |                 |                 |          |     |
|     | 12 29 57 59 72 74 |                 |                 |          |     |
|   3 +                   +                 +                 +          +     |
|     | 14 32 53 61       |                 |                 |          |     |
|     | 37                |                 |                 |          |     |
|     | 8  21 35 39       |                 |                 |          |     |
|   2 + 38                +                 +                 +          +     |
|     | 5  9  56          |                 |                 |          |  3  |
|     |                   |                 |                 |          |     |
|     | 2  16 52 62       |                 |                 |          |     |
|   1 + 34                +                 +                 +          +     |
|     | 13 19 36          |                 |                 |          |     |
|     | 3  51 58          |                 |                 | VOCA     |     |
|     |                   |                 | Task 1  Task 2  | GRAM     |     |
*   0 * 31 73             * Rater1  Rater2  *                 * CONT     *     *
|     | 11 25 66          |                 | Task 3          |          |     |
|     | 24                |                 |                 |          | --- |
|     | 45 71             |                 |                 | PRON     |     |
|  -1 + 63                +                 +                 +          +     |
|     |                   |                 |                 |          |     |
|     | 18 30 50          |                 |                 |          |     |
|     | 10 40             |                 |                 |          |     |
|  -2 +                   +                 +                 +          +     |
|     |                   |                 |                 |          |  2  |
|     |                   |                 |                 |          |     |
|     | 43                |                 |                 |          |     |
|  -3 +                   +                 +                 +          + (1) |
|-----+-------------------+-----------------+-----------------+----------+-----|
|Measr|+Examinee          |-Rater           |-Task            |-Criterion|Scale|
+------------------------------------------------------------------------------+
```

criteria, and the 4-point scale as raters used it to score examinees' performance on each language aspect. Summary statistics from the FACETS analysis for the four-facets are presented in Table 1.

Figure 1 indicated that there was a wide spread of examinees' ability with a range from −2.64 to +8.08 logits. The mean ability of examinees was 2.69 logits, with a standard error of 0.69 logits (see Table 1). The Chi-square test indicated that

Table 1 Summary statistics for the MFRM analysis

Statistics	Examinees	Raters	Tasks	Criteria
M Measure	2.69	0.01	0.26	0.60
M SE	0.69	0.00	0.00	0.00
x^2	1142.8^*	0.01	8.9^*	50.1^*
df	73	1	2	3
Separation index	3.64	0.00	1.88	4.03
Separation reliability	0.93	0.00	0.78	0.94

Note *Significant at the $p > 0.05$ level the examinees came from statistically distinct ability groups ($x^2 = 1142.8$, $df = 73$, $p < 0.01$). Figure 1 also revealed that most examinees were located above the difficulty of the three speaking tasks in the variable map. The mean ability of examinees (2.69 logits) was substantially higher than the mean task difficulty (0.26 logits), suggesting that, on average, the three tasks were quite easy for this group of test candidates. It should be noted, however, that all participants in this study were volunteers, and students at higher ability levels might be more motivated to participate in such a study. That said, the results indicate that more difficult tasks might be developed in the future to tap into test candidates' speaking ability. The satisfactorily high reliability (0.93) indicated the reproducibility of the measures, suggesting that the same number of statistically distinct levels of proficiency could be expected if we repeated the same data collection (Linacre 2013).

Of particular interest to this current study is the rater facet. The interpretation of the statistics for raters in Table 1, however, is decidedly different from that for the other three-facets. When raters within a group exercised a highly similar degree of severity, rater separation reliability will be close to 0 (Eckes 2011). This is exactly what happened in this study where the two raters were found to demonstrate highly similar patterns in their rating behavior. This could be first observed from Column 3 of the variable map, as shown in Fig. 1. In addition, this was also indicated by the insignificant Chi-square test, as well as the extremely low separation index and reliability (see Table 1). Rater fit statistics present statistical indicators of the degree to which raters used the rating scale in a consistent manner (Eckes 2011). The Infit and Outfit Mean Square were 1.03 and 1.09 for Rater 1, and 0.94 and 0.92 for Rater 2, all approximating the ideal value of 1. These statistics suggested that both raters were consistent in their ratings. Such a finding is unlikely when multiple raters are used, and is at odds with previous research which tended to identify significant rater effects (e.g., Eckes 2005; Lynch and McNamara 1998). As a preliminary study, only two raters were involved; both raters, as noted earlier, were very experienced in rating the FET speaking test, and were directly involved in the construction of the rating scale. As such, caution needed to be exercised in overinterpreting the results emanating from this part of the research. A larger and more representative sample of raters should be included in a future investigation. On the basis of this preliminary study, it seems reasonable to conclude that the first warrant, i.e., the two raters demonstrate sufficiently high reliability and similar level of severity, was supported.

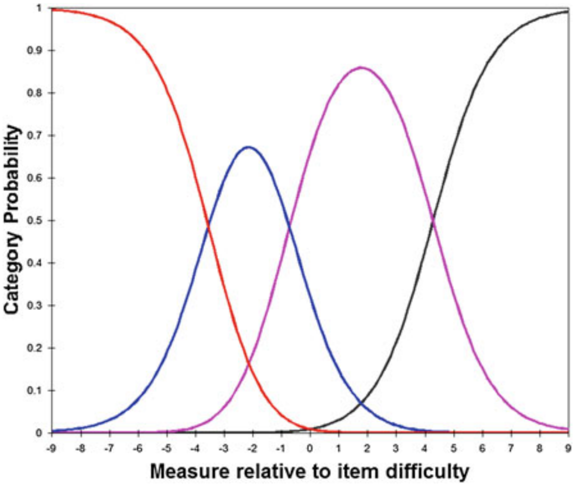
The second warrant is pertinent to the utility of the category structure of the rating scale. To verify the functioning of each response category, Linacre's (2004) criteria were applied, including: (1) A minimum of 10 observations is needed for each category; (2) Average category measures must increase monotonically with categories; (3) Outfit Mean Square statistics should be less than 2.00; (4) The category threshold should increase monotonically with categories; (5) Category thresholds should be at least 1.4–5-logits apart, and (6) The shape of the probability curves should peak for each category (cited in Oon and Subramaniam 2011, p. 125). Summary of category structure of the 4-point scale is presented in Table 2. As shown in this table, though all categories were used by raters, the first category (i.e., Very Poor) was substantially under-used with only 1 % frequency, suggesting

Table 2 Summary of category structure of the 4-point rating scale

Category	Observed count (%)	Average measure	Outfit MnSq	Threshold calibration
1. Very Poor	14 (1 %)	−1.99	0.80	None
2. Poor	137 (14 %)	−0.65	0.90	−3.55
3. Moderate	538 (57 %)	2.10	1.00	−0.73
4. Good	259 (27 %)	4.96	1.10	4.18

that this category should be removed or collapsed with its adjacent category (Bond and Fox 2015). It should be noted that this finding concurred with our earlier observation that the three tasks in this speaking test were, on average, too easy for this sample of test candidates. The "Very Poor" category should be attempted on a larger and more representative sample of test candidates in the future to further examine the functioning of this category. Table 2 also showed that average category measures increased monotonically from −1.99 to 4.96, suggesting that these categories were used as expected by the raters. Outfit Mean Square values ranged from 0.80 to 1.10, suggesting that these categories did not introduce noise into the measurement process. An inspection of the distance between two adjacent categories showed that the required range of 1.4–5-logits was met, suggesting that the four categories defined distinct positions on the latent variable. The category probability curves (displayed in Fig. 2) further revealed that each category emerged as a peak. The analysis of the category structure only partly supported the second warrant, and suggests that the category structure should be revised in the future through either removing the redundant category or collapsing it with its adjacent category.

Fig. 2 Category probability curves for the 4-point rating scale



SEM Analysis

The SEM analysis in this study followed the procedures outlined by Byrne (2006). In this analysis, a MTMM design was adopted by which multiple traits are measured by multiple methods. The four language performance aspects (i.e., pronunciation, content, grammar, and vocabulary) were specified as the trait factors in the MTMM model, whereas the three tasks (i.e., responding to questions, short comment, and graph/chart description and comment) were the method factors. According to Campbell and Fiske (1959), convergent validity refers to the extent to which different assessment methods concur in their measurement of the same trait, whereas discriminant validity refers to the extent to which independent assessment methods diverge in their measurement of different traits. The convergent and discriminant validity of the rating scale could be examined at both the matrix and parameter level, as advised by Byrne (2006). Specifically, four MTMM CFA models were specified in this study, including Correlated Traits/Correlated Methods Model (Model 1), No Traits/Correlated Methods Model (Model 2), Perfectly Correlated Traits/Freely Correlated Methods Model (Model 3), and Freely Correlated Traits/Uncorrelated Methods Model (Model 4). Readers are referred to Figs. 3, 4, 5, and 6 for the graphic representations of these four MTMM models. It is worthnoting that: (1) the variances of latent factors in the four models were set to be 1.0 for model identification purposes; (2) the only difference between Model 1 and Model 3 (displayed in Figs. 3 and 5 respectively) lies in that the correlations between the trait factors in Model 1 were freely estimated, but fixed to 1 in Model 3, as indicated by the dotted lines in the figure.

Among the four hypothesized models, Model 1 was the least restrictive model, and therefore served as the baseline model against which the alternative MTMM

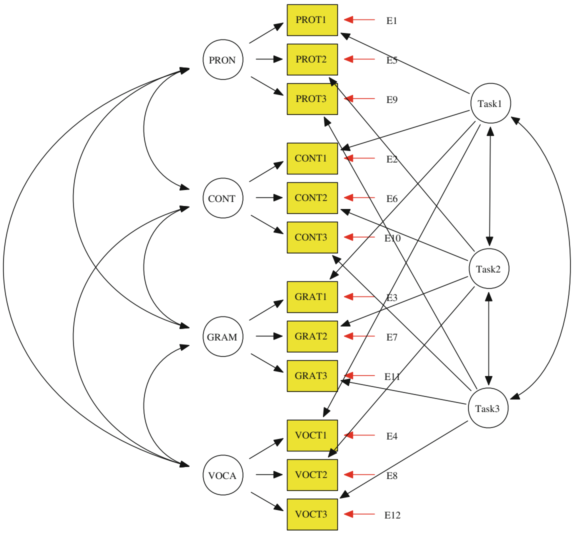
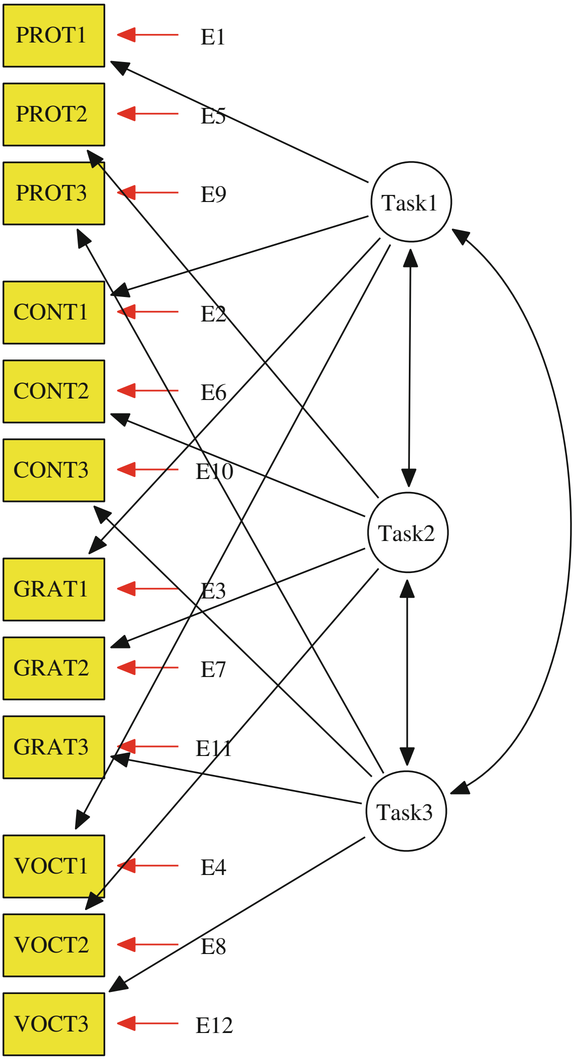


Fig. 3 Correlated Traits/Correlated Methods Model (Model 1)

Fig. 4 No Traits/Correlated Methods Model (Model 2)



models were compared. Since the other three models were nested models of Model 1, the Chi-square difference test was used to compare whether the difference between the fit of the alternative models and the baseline model was statistically significant. A non-significant Chi-square value would indicate that the difference was negligible. In addition to the Chi-square difference test, Cheung and Rensvold (2002) recommended that if the CFI difference values did not exceed 0.01, then the difference between the fit of the two models would be of minimal practical significance. While aware that the CFA format allows for an assessment of construct validity at both matrix and individual parameter levels (Byrne 2006), this

Fig. 5 Perfectly Correlated Traits/Correlated Methods Model (Model 3)

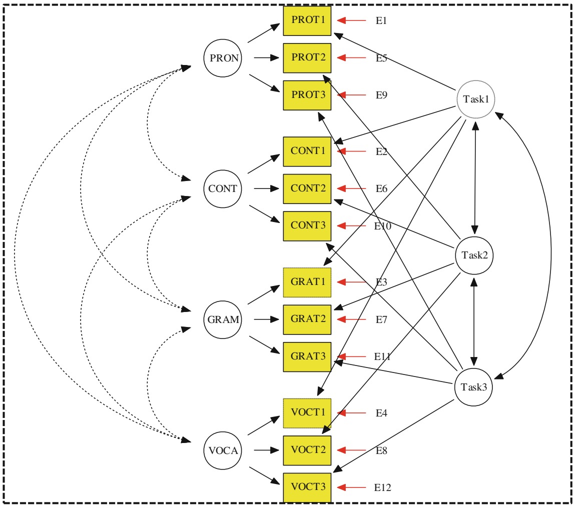
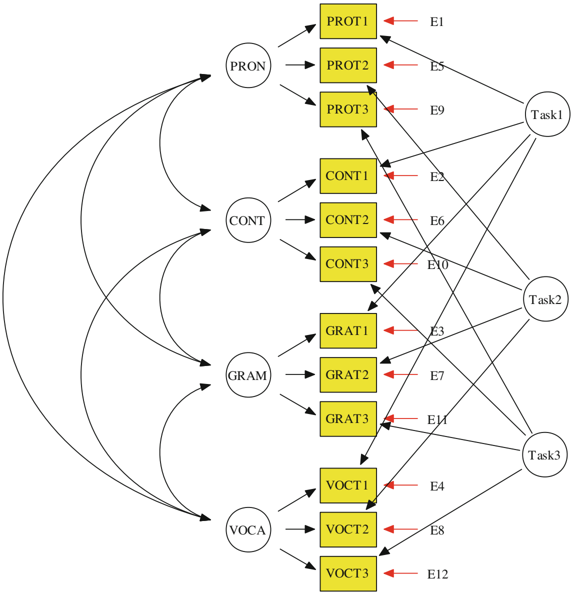


Fig. 6 Freely Correlated Traits/Uncorrelated Methods Model (Model 4)



preliminary study tested for evidence in relation to convergent and discriminant validity primarily at the matrix level.

Given that Mardia's normalized estimate was 3.74, the data were considered normally distributed, and hence the default estimation method in EQS, i.e., the maximum likelihood method, was used for parameter estimation purposes (Bentler and Wu 2005). The goodness-of-fit indexes of the four MTMM models are

Table 3 Summary of Goodness-of-Fit Indexes for MTMM Models

Model	x^2	*df*	CFI	GFI	SRMR	RMSEA (90 % C.I.)
Model 1	35.93	33	0.995	0.934	0.038	0.035 (0.000–0.094)
Model 2	142.59*	51	0.858	0.756	0.083	0.187 (0.126–0.186)
Model 3	59.91*	39	0.968	0.890	0.039	0.086 (0.037–0.126)
Model 4	42.44	36	0.990	0.912	0.047	0.050 (0.000–0.101)

Notes Model 1: Correlated Traits/Correlated Methods Model; Model 2: No Traits/Correlated Methods Model; Model 3: Perfectly Correlated Traits/Correlated Methods Model; Model 4: Freely Correlated Traits/Uncorrelated Methods Model. $^*p < 0.05$

presented in Table 3. As could be seen from this table, the baseline model, i.e., the Correlated Trait/Correlated Method model fits the data well [$x^2 = 35.93$, $df = 33$, $p > 0.05$, CFI = 0.995, RSMEA = 0.035 (90 % C.I., 0.000–0.094)]. In comparison, Model 2, i.e., No Traits/Correlated Methods Model, displayed extremely poor fit to the data [$x^2 = 142.59$, $df = 51$, $p < 0.05$, CFI = 0.858, RSMEA = 0.187 (90 % C.I., 0.126–0.186)]. A Chi-square difference test was performed to compare the fit of these two models, yielding a highly significant result ($\Delta x^2 = 106.66$, $df = 18$, $p < 0.01$). Furthermore, the CFI difference value was 0.137, well above the criterion value of 0.01 recommended by Cheung and Rensvold (2002). The results supported the convergent validity of the rating scale which required correlations between independent measures of the same trait (e.g., the ratings of pronunciation on Task 1, Task 2 and Task 3) that should be substantial and statistically significant.

Discriminant validity, on the other hand, is assessed in terms of both traits and methods. In testing for evidence of discriminant validity among traits, a model in which trait factors were posited to be freely estimated (Model 1) was compared with one in which they were perfectly correlated (Model 3). An inspection of the goodness-of-fit indexes in Table 3 revealed that Model 3 did not fit the data well [$x^2 = 59.91$, $df = 39$, $p < 0.05$, CFI = 0.968, RSMEA = 0.086 (90 % C.I., 0.037–0.126)]. The comparison between Model 1 and Model 3 yielded a statistically significant result ($\Delta x^2 = 23.98$, $df = 6$, $p < 0.05$), and the difference in practical fit was quite large (ΔCFI = 0.03). This result supported discriminant validity among traits, and suggested that although the traits were substantially correlated, they were still distinguishable from each other. In the field of language assessment, numerous factor analytic studies have supported the notion that language ability is a complex construct with multiple dimensions, though the research community has not reached an agreement regarding the nature of the constituents, or on the manner in which they interact (e.g., Gu 2014; In'nami and Koizumi 2012; Sawaki et al. 2009). The tenability of Model 1 and rejection of Model 3 lends further support to this view, suggesting that not only is general language ability multidimensional, but any single language modality such as speaking ability might also have multiple constituents.

Based on the same logic, when testing for the evidence of discriminant validity related to method effects, a model in which method factors were posited to be freely estimated (Model 1) was compared with one in which method factors were specified to be uncorrelated (Model 4). The goodness-of-fit indexes in Table 3 indicated that Model 4 was a reasonably satisfactory fit to the data [$\Delta x^2 = 42.44$, $df = 36$, $p > 0.05$, CFI = 0.912, RSMEA = 0.050 (90 % C.I., 0.000–0.101)]. A comparison of this model with the baseline model (i.e. Model 1) yielded a Δx^2 value which was statistically not significant ($x^2 = 6.51$, $df = 3$, $p > 0.05$) with negligible difference in CFI values (ΔCFI = 0.01). A large Δx^2 or substantial ΔCFI argued for the lack of discriminant validity, thereby suggesting common method bias. Given that this analysis yielded a non-significant Chi-square test and the ΔCFI was minimal, it was reasonable to conclude that the scale displayed evidence of discriminant validity related to methods.

An inspection of the factor loadings in the baseline model, i.e., the Correlated Traits/Correlated Methods Model revealed that the path coefficients of the observed ratings to the corresponding four trait factors in the model were high and significantly different from zero, ranging from 0.48 to 0.89. The results indicated strong linear relationships between the trait factors and the observed ratings. In addition, the correlations between the four trait factors were significant, ranging from 0.49 to 0.93. The substantial path and correlation coefficients again support the convergent validity of the rating scale related to the traits. However, the correlations between the three method factors were found to be reasonably high. For example, the correlation between Task 2 and Task 3 was 0.56. The high correlation coefficients between the methods argued against discriminant validity in relation to the methods, and suggested common method bias in measurement (Byrne 2006). These results recommended that the FET speaking test designers should adopt elicitation methods which are distinct enough so as to avoid common method bias. Taken together, the MTMM modeling of the test data lent reasonably strong support to the third warrant, i.e., the rating scale displays both convergent and discriminant validity. However, the strength of this warrant was somewhat weakened by the identification of common method bias which should be addressed in future revisions of this speaking assessment.

Conclusions, Limitations, and Implications

This preliminary validation research demonstrated how MFRM and SEM could be used in tandem in the interrogation of the construct validity of the analytic rating scale developed for a school-based English speaking assessment. Through harnessing the potential of both research methodologies, this study examined the plausibility and accuracy of three warrants (and their respective rebuttals) which were deemed crucial to the construct validity of the rating scale, and hence to this speaking assessment. Specifically, a four-facet MFRM model was utilized to calibrate examinee ability, rater severity, task difficulty, and the difficulty of ability dimensions on the same interval measurement scale. MFRM analysis of the rating data lent support to the first warrant, i.e., raters displayed high reliability and the same level of severity in awarding scores to test candidates. The second warrant, i.e., the category structure of the rating scale functioned appropriately, was partly supported by the MFRM analysis. The lowest category was found to be substantially under-used, thereby weakening the strength of this warrant. The third warrant regarding the convergent and discriminant validity of the rating scale was examined through using the MTMM CFA design to model the rating data. Four MTMM models were specified, evaluated, and compared. As it turned out, the SEM analysis partly supported the third warrant, suggesting that the rating scale displayed convergent and discriminant validity in relation to both traits and methods. The high correlations between the method factors, however, argued against the discriminant validity about methods, and suggested common method bias. This finding somewhat weakened the strength of the third warrant, and should be addressed in future test revisions.

In this research, the Rasch model and SEM have been used on the same set of data, but separately. Bond and Fox (2015) are much more direct about using the models collaboratively, in conjunction: "For those who are more thoughtfully wedded to SEM, our advice would be spread over two steps: First, that Rasch analysis should be adopted to guide the construction and quality control of measurement scales for each of the variables that feature in the research. Second, the interval-level person Rasch measures and their standard errors (SEs) that derive from each of those instruments should be imputed into the SEM software for the calculation of the relationships between those variable measures" (p. 240). In defense of the current analyses and results, we assert that this speaking test has almost satisfied Rasch model requirements for the production of interval-level measurement. To the extent that the data fit the Rasch model, total scores are the necessary statistic for parameter estimation, so using raw ordinal-level data in the SEM analyses is likely to be unproblematic in this case. For researchers who wish to explore the applicability of further developments of the Rasch model for answering the questions broached in this research, a generalized form of the Rasch model, the mixed coefficients multinomial logit model (MCMLM; Adams et al. 1997) might be applied. Such multidimensional item response model analyses combine the response information for different tests according to the size of the correlations between the latent variables. When the correlations are high but not perfect, as in this case, the MIRM uses information from all tests to estimate performances on each of the latent traits (after Bond and Fox 2015, pp. 291–292).

The research described herein has several limitations. First, as a preliminary study, convenience sampling was adopted. Consequently, it cannot be held that the sample of test candidates used in this study was representative of the test population for which this test was designed. Moreover, SEM is a large-sample analytic technique (e.g., In'nami and Koizumi 2011; Kline 2005). Given the complexity of the MTMM models specified in this study, a larger sample of test candidates is essential for ensuring the viability of parameter estimations. Also, in view of the central role that raters played in performance assessment, a larger and more representative rater sample should be attempted in future validation research. Second, some essential features of the MFRM analysis, such as the interaction or bias analysis (e.g., Eckes 2011; Linacre 2013) were not included in this research. Investigations into the potential interactions between raters and the criteria in the rating scale could be particularly meaningful to such a study. Finally, the MTMM CFA format allows the researchers to investigate the convergent and discriminant validity at both matrix and parameter levels (Byrne 2006). This preliminary study, however, was primarily focused on evidence at the matrix level. A closer examination of convergent and discriminant validity at the parameter level could be very revealing, and should therefore be attempted in the future. Meanwhile, SEM allows researchers to hypothesize and evaluate a host of different theoretical models. Some alternative MTMM models could therefore be subsequently specified and evaluated, such as the Higher-Order Trait Model (e.g., Sawaki 2007) with a view to understanding more clearly the relationships between and among the observed and latent variables in this study.

This research has implications for the future revision and improvement of the analytic rating scale, as well as the speaking assessment under study. First, the category structure of this rating scale warrants adjustment. The redundant category, as discussed earlier, could be either removed or collapsed with its adjacent category. Second, the tasks in the speaking assessment could be redesigned. Given the common method bias identified by SEM analysis, testing methods which are sufficiently distinguishable from each other could be adopted in the future. In the current test format, both Task 2 and Task 3 in the FET speaking component require test candidates to give comments on a certain topic; such a design is very likely to cause common method bias. Other task formats such as reading aloud or listening to summarize could be attempted in the future development of this speaking test (see e.g., Fan 2014). Finally, this study has methodological implications for language assessment researchers, in particular the developers and validators of performance assessments (e.g., speaking, writing). Echoing the view of Bond and Fox (2015) regarding the collaborative use of Rasch model and SEM in conjunction, future researchers may consider using the MFRM to examine the quality of the rating scale, and revise it accordingly before utilizing the SEM to either test the tenability of specific theoretical models or examine the convergent and discriminant validity at both matrix and parameter levels. By doing so, the potential of the two methodologies could be harnessed more adequately.

Acknowledgments The study reported in this chapter was supported by the National Social Sciences Fund of the People's Republic of China under the project title of "Development and Validation of Standards in Language Testing" (Grant No: 13CYY032), and the Research Project of National Foreign Language Teaching in Higher Education under the project title of "Teacher-, Peer-, and Self-assessment in Translation Teaching: A Many-Facets Rasch Modeling Approach" (Grant No: 2014SH0008A). Part of this research was published in the third issue of *Foreign Language Education in China* (*Quarterly*) in 2015.

References

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–24.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257.

Bachman, L. F., & Palmer, A. S. (1996). *Language assessment in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

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

Bentler, P. M., & Wu, E. J. (2005). *EQS 6.1 for Windows*. Encino, CA: Multivariate Software.

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

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, New Jersey: Psychology Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

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.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

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

Eckes, T. (2011). *Introduction to many-facets Rasch measurement*. Frankfurt: Peter Lang.

Fan, J. (2014). Chinese test takers' attitudes towards the Versant English Test: A mixed-methods approach. *Language Testing in Asia, 4*(6), 1–17.

Fan, J., & Ji, P. (2013). Exploring the validity of the Fudan English Test (FET): Test data analysis. *Foreign Language Testing and Teaching, 3*(2), 45–53.

Fan, J., & Ji, P. (2014). Test candidates' attitudes and their test performance: The case of the Fudan English Test. *University of Sydney Papers in TESOL, 9*, 1–35.

Fan, J., Ji, P., & Song, X. (2014a). Washback of university-based English language tests on students' learning: A case study. *The Asian Journal of Applied Linguistics, 1*(2), 178–192.

Fan, J., Ji, P., & Yu, L. (2014b). Another perspective on language test validation: The factor structure of language tests. *Theory and Practice in Foreign Language Teaching, 4*, 34–40.

FDU Testing Team. (2014). *The FET Test Syllabus*. Shanghai: Fudan University Press.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31*(1), 111–133.

Han, B., Dan, M., & Yang, L. (2004). Problems with College English Test as emerged from a survey. *Foreign Languages and Their Teaching, 179*(2), 17–23.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEFL test: A multi-sample analysis. *Language Testing, 29*(1), 131–152.

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8*(3), 250–276.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Knoch, U. (2011). 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.

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

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach* Cambridge: Cambridge University Press.

Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*(3), 295–332.

Linacre, M. (2013). *A user's guide to FACETS (3.71.0)*. Chicago: MESA Press.

Linacre, M. (2004). *Optimal rating scale category effectiveness*. In E. V. Smith & R. M. Smith (Eds.), Introduction to Rasch measurement (pp. 258–278). Maple Grove, MN: JAM Press.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489–515.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

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

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facets 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., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. Language Testing, *29*(4), 553–574.

North, B. (2000). *The development of common framework scale of language proficiency*. New York: Peter Lang.

North, B., & Jones, N. (2009). *Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling*. Strasbourg: Language Policy Division.

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

Oon, P. T., & Subramaniam, R. (2011). Rasch modelling of a scale that explores the take-up of Physics among school students from the perspective of teachers. In R. F. Cavanaugh & R. F. Waugh (Eds.), *Applications of Rasch measurement in learning environments research* (pp. 119–139). Netherlands: Sense Publishers.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16*(4), 457–478.

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223–241.

Sawaki, Y. (2007). Construct validation of analytic rating scale in speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355–390.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5–30.

Shin, S.-Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing, 32*(2), 259–281.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99–123.

TOPE Project Team. (2013). *Syllabus for Test of Oral Proficiency in English (TOPE)*. Beijing: China Renming University Press.

Tsinghua University Testing Team. (2012). *Syllabus for Tsinghua English Proficiency Test (TEPT)*. Beijing: Tsinghua University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing, 16*(1), 82–111.

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49–70.