Available at jtpsol.journals.ikiu.ac.ir

Journal of Teaching Persian to Speakers of Other Languages

Randomness among Teacher Raters Rating Language Learners' Essays: A FACETS Approach

Rajab Esfandiari*

Assistant Professor, Imam Khomeini International University

ABSTRACT

Inconsistency of the ratings of raters may invalidate test results, adversely affecting the decisions made about the placement of language learners to a higher level of education. In the present study, the researcher used the many-facet Rasch measurement model to examine how consistently teacher raters rated the essays written by language learners in their writing classes at Imam Khomeini International University. The teacher raters each rated 56 essays, using a researcher-made, 5-point analytic rating scale. Using FACETS, the Rasch-based computer programme for rating data, the researcher analysed the data. The results of FACETS analysis, including separation indices and fit values, showed that teacher raters were self-consistent in rating the essays language learners wrote. The results of single rater-rest of the raters revealed that each teacher rater's ratings were consistent with those of other raters. These findings may carry implications for research and pedagogy, shedding light on rater training.

Keywords: randomness, rating, rating scale, self-consistency

* Received on: 3/4/2016 Accepted on: 21/6/2016

E-mail: esfandiari@hum.ikiu.ac.ir

1. Introduction

Communicative language testing requires that test developers use more open-ended test methods, including essays, to measure test takers' written abilities (Douglas, 2010). These constructed-response assessments (see e.g. Brown & Hudson, 1998) require that test takers produce more extended stretches of language. As such, they are more aligned with more recent approaches to test development. The rationale for using these techniques is predicated on the assumption that they are more authentic, more performance-based, and more direct, thereby reflecting the test takers' underlying written abilities, increasing the validity of the test results, and contributing to the generalizability of the test scores (Hughes, 2003).

One of the major advantages which constructed-response assessments offer relates to the elimination of guessing factor (Brown & Abeywickrama, 2010). However, these assessments are relatively subjective, because language learners' written performance should be subjectively evaluated. Test owners usually employ raters to rate language learners' performance, and the raters award ratings to these products, using a scale with some guidelines about how to use the scale for rating purposes.

When raters rate language learners' performance, they may not necessarily use the scale properly and may exert their own idiosyncrasies. As a result, errors may occur. These errors are regarded construct-irrelevant variance and have nothing to do with language learners' performance. Physical and mental characteristics of raters may affect test takers' language abilities positively or negatively. Therefore, language ability is confounded with irrelevant factors, and these extraneous factors may endanger the validity of test scores, thereby leading to unfair decisions for language learners.

One of the errors which may be manifested while raters rate a piece of writing relates to randomness. Raters may not be consistent in the ratings they award; their ratings may not be consistent with those of other raters either. This systematic rater error is very detrimental if left undetected. The present researcher examined randomness among native-Persian raters who rated 56 essays non-native Persian speakers wrote in their writing classes.

2. Literature Review

In this part of the paper, the researcher first introduces rater errors, defines them, and lists a catalogue of rater errors; next, he explains many-facet Rasch measurement; finally, he summarises the findings of studies on randomness.

2.1. Rater Effects

Researchers use different terms to refer to the errors raters introduce. In the literature, rater effects, rater errors, and/or rater biases are used interchangeably (Myford & Wolfe, 2003). A rater effect refers to an error which a rater introduces into a rating setting and does not relate to a language learner's language ability. In the words of Scullen, Mount, and Goff (2000), rater effects refer to a "broad category of effects [resulting in] systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the rate" (957). This succinct definition is very informative and is used throughout this study.

Rater effects fall into two main categories: classic psychometric errors and lesser-known effects (Myford & Wolfe, 2003). Classic psychometric errors include severity/leniency, halo, central tendency, restriction of range, differential severity/leniency, and randomness (Engelhard, 2002). These effects are more traditional, and they have been extensively empirically investigated. The lesser-known effects include inaccuracy, logical error, contrast error, influences of rater biases, beliefs, attitudes, and personality characteristics, influences of rater/ratee background characteristics, proximity error, recency (or primacy) error, and order effects (Myford & Wolfe, 2003). These are more recent errors, and much less is known about them.

Researchers use a variety of methods to study rater effects. Myford and Wolfe (2004) neatly summarized these methods. These procedures comprise classical test score theory, including the means and standard deviations of trait ratings, inter-correlations among ratings across traits, confirmatory factor analysis, analysis of variance, and regression-based procedures; generalizability theory; item response theory, including two-parameter logistic model and three parameter-parameter logistic model and the variants of these two models; and many-facet Rasch measurement procedures.

Researchers also use a variety of strategies to minimise rater effects. Although the adoption of strategies depends on the type of rater effect, research has shown that rater training has proved promising and can be regarded as the most common technique to apply to reducing almost all rater effects (McNamara, 1996). The findings of studies on rater training have revealed that although it can reduce rater variability, rater training does not necessarily eliminate it.

2.2. Randomness

Randomness (Myford & Wolfe, 2004) is commonly known as inconsistency (Knoch, Read, & von Randow, 2007). In Myford and Wolfe's words (2004), "the randomness effect is defined as a rater's tendency to apply one or more trait scales in a manner inconsistent with the way in which the other raters apply the same scales" (p. 206). Similarly, building upon this definition, Knoch et. al. (2007) defined randomness as "a tendency of a rater to apply one or more rating scale categories in a way that is inconsistent with the way in which other raters apply the same scale" (p. 27).

According to these two definitions, randomness is manifested in two different ways. Raters may award ratings which may show much variability. In this case, raters are not self-consistent in the ratings they assign. Alternatively, raters may award ratings which do not correlate with those of other raters. In this case, a rater will rank language learners in a different order than will the other raters (Myford & Wolfe, 2004).

Many factors may lead to inconsistency. The first main reason for inconsistency relates to the raters' insufficient background, or expertise, to make fine discriminations between the scale categories (Lumely & McNamara, 1995). Raters may not have developed a solid understanding of the sale categories; as a result, they may not rate the examinees in a reliable fashion, tending to use different categories in an indiscriminate fashion. The second main reason for inconsistency concerns the raters' characteristics. Raters may feel exhausted as the rating session proceeds. Therefore, fatigue may lead to raters' inattention to rating criteria, making them inconsistent in the ratings they award (McNamara, 1996).

The studies using rater training to examine randomness have yielded in some very interesting points. Research has shown that "rater training is successful in making raters more self-consistent …. Without this self-consistency, no orderly process of measurement can be conducted" (Lumely & McNamara, 1995, p. 57). Similarly, McNamara (1996) concluded that "rater training is successful in making raters more self-consistent. That is, the main effect of training is to reduce the random error in rater judgments" (p. 126). Esfandiari and Myford (2013) discussed that "rater training can reduce but not eliminate differences in rater severity, help raters to become more self-consistent, and reduce some biases that individual raters may display" (p. 117).

Some researchers have used many-facet Rasch measurement model to empirically study inconsistency. Lumley and McNamara (1995) used two groups of raters (Group A, n = 13; Group B, n = 4) to rate 83 tapes on three different times on the speaking subtest of Occupational English Test in Australia. The authors were interested in whether the raters were consistent in their ratings over time. The raters used a 6-point analytic rating scale to rate the tapes on six linguistic dimensions. FACETS was used to analyse the spoken rating data. The results of FACETS analysis revealed that the raters were not necessarily rating the tapes consistently over time, and much variation was found among the raters' ratings from one occasion to another occasion.

In a longitudinal study spanning four and a half years at three time periods, Lim (2011) used 11 new, inexperienced raters to rate the writing section of the Michigan English Language Assessment Battery (MELAB) (n = 20,662 ratings). Lim used FACETS to analyse the ratings. The overarching goal in the study was to make sure these new raters were able to rate consistently over time. The findings showed that as raters gained in rating experience, they remain consistent in their ratings over time. Lim neatly summarized the findings as follows:

> new raters may or may not be inconsistent when they begin rating, and those who are do not stay that way for very long. With one brief exception, experienced raters all began and stayed within acceptable bounds of fit throughout the time periods tracked. (p. 555)

Lim attributed the consistency of experienced raters to the frequency of ratings.

The findings of the above empirical studies and other studies suggest that consistency may be affected by many factors. These factors may include training, experience, expertise, language proficiency, feedback, language background, and type of rater. Unfortunately, many of these factors remain unexplored, and new empirical studies are required to shed light on rater consistency.

2.3. Many-Facet Rasch Measurement Model

The many-facet Rasch measurement model (also commonly known as multi-faceted or many-faceted Rasch measurement (Engelhard, 1994; McNamara, 1996), many-faceted conjoint measurement (Linacre, Engelhard, Tatum, & Myford, 1994), or multifacet Rasch modeling (Lunz & Linacre, 1998)) is a family of Rash models (Eckes, 2011).

The many-facet Rasch measurement model was introduced by Linacre (1989). It is an extension of the basic dichotomous Rasch model. It enables researchers to estimate the difficulty of items, ability of test takers, and severity of raters. The many-facet Rasch measurement makes it possible to estimate simultaneously, but independently, the effects of many facets on an equal-interval scale. In other words, it transforms the ordinal data into continuous data, making it possible to use parametric tests.

Many-facet Rasch measurement model is best suited for polytomous data for which there is no correct or incorrect answer. Many-facet Rasch measurement model transforms ordered data into equal-interval scale, making it possible to analyse more than two facets onto a single frame of reference.

Knock (2007) summarised the many facet Rasch measurement model as

> a generalization of Wright and Master's (1982) partial credit model that makes possible the analysis of data from assessments that have more than the traditional two facets associated with multiple-choice tests (i.e., items and examinees). In the many-facet Rasch model, each facet of the assessment situation (e.g., candidates, raters, traits) is represented by one parameter. The model states that the likelihood of a particular rating on a given rating scale from a particular rater for a particular student can be predicted mathematically from the proficiency of the student and the severity of the rater. The advantage of using multi-faceted Rasch measurement is that it models all facets in the analysis onto a common logit scale, which is an interval scale. Because of this, it becomes possible to establish not only the relative difficulty of items, ability of candidates and severity of raters as well as the scale step difficulty, but also how large these differences are. Multi-faceted Rasch measurement is particularly useful in rating scale validation as it provides a number of useful measures such as rating scale discrimination, rater agreement and severity statistics and information with respect to the functioning of the different band levels in a scale. (p. 116)

The examination of the literature shows that randomness is a pervasive rater effect, which may have detrimental effects on the validity of the ratings the raters award. The studies on randomness have been limited to the English-speaking raters rating EFL ratings. The present study is the first to examine the ratings Persian raters award nonnative Persian-speaking students learning Persian as a second language. Therefore, the present study addresses the following two research questions:

1. To what extent are Persian raters self-consistent when they rate the essays of nonnative Persian speaking students?

2. Are the Persian speaking raters consistent in the ratings they assign to the ESL essays?

3. Method

In this section, the participants are fully described, data collection methods are explained, and data analysis procedures are described.

3.1. Participants

Two groups of participants, as described below, took part in this study. The writers included 28 male and female advanced nonnative Persian-speaking students, who were majoring in humanities and engineering, learning Persian as a second language at Persian Language Center at Imam Khomeini International University. Their ages ranged between 18 and 30, and their average age was 22.52. Seven (25%) writers did not specify their age. The number of years they were learning Persian ranged from 3 to 24 months, and the average learning experience was 7.26 months.

In addition to the writers, two native Persian speaking raters participated in this study. The raters were MA holders of teaching Persian to nonnative Persian speaking language learners. They had three years of teaching experience. One rater had one year of teaching writing experience, and the second rater had two years of teaching writing experience to Persian to nonnative Persian speaking language learners. They did not have any rating experience.

Table 1 shows the number of participants, their mother tongue, and their nationality.

3.2. Data Collection

The rating data for the present study were collected during winter 2015. Data were collected in two different phases. The first phase of data collection began in November 2015. During the first phase of data collection, 28 language learners in four different writing classes were asked to write a 15-20 line essay on the following topic: What is the effect of the Internet and

Television on personal and social relations? Please, explain as fully as possible.

Table 1
Demographic Information of Language Learners

Nationality			Mother tongue		
Nationality	Absolute frequency	Relative frequency	Mother tongue	Absolute frequency	Relative frequency
Iraqi	2	7.1	Arabic	14	50
Tajikistani	6	21.4	Kurdish	2	7.1
Sudanese	1	3.6	Chinese	1	3.6
Ivory coast	1	3.6	Korean	1	3.6
Kazakhstani	1	3.6	Tajikistani	6	21.4
Syrian	5	17.9	French	1	3.6
Chinese	1	3.6	English	1	3.6
Korean	1	3.6	Kazakhstani	1	3.6
Lebanese	3	10.7	Other	1	3.6
Yemeni	3	10.7			
Palestinian	2	7.1			
Nigerian	1	3.6			
Other	1	3.6			
Total	28	100	Total	28	100

The second phase of data collection began one week later in December 2015. The same language learners in the first phase wrote a second 15-20 line essay on the following topic: Compare the urban life with rural life and explain the advantages and disadvantages. Language learners were allotted 50 minutes to write each essay.

3.3. Data analysis

The researcher used Facets (version 3.68.1), the Rasch-based computer programme for rating data, to analyse the data for this study. Three facets were specified: raters, essays, and items. The mathematical formula for the measurement model used in this study is as follows:

$$Log\ (P_{nirk} / P_{nir}\ (k-1))\ B_n - D_i - C_j - F_k$$

where: P_{nijk} is the probability that an essay n will receive a rating of k on item i from rater j, P_{nij} (k−1) is the probability that an essay n will receive a rating of k−1 on item i from rater j, B_n is the level of proficiency shown in an essay n, D_i is the difficulty of $item_i$, Cj is the severity of rater j, and F_k is the difficulty of scale category k, relative to scale category k−1.

3.4. Instrument

The raters used a researcher-made, 5-point analytic rating scale to rate the 56 essays the 28 writers wrote for this study. The scale included five items: grammar, vocabulary, coherence, content and development, and mechanics. The scale categories ranged from 1 (very bad), 2 (bad), 3 (average), 4 (good), and 5 (very good). The items were all weighted.

The raters were instructed how to rate the essays. Before they began rating, they were first given the instrument to get familiarized with the scale, the items, and the categories. They were then given some benchmark essays to understand how ratings were assigned to the essays according to the items. Finally, they were given the essays, were asked to rate them at home, and return them to the researcher two weeks later.

4. Results and Discussion

In this section, I first report on the functionality of the scale used in this study; then I present the findings; and finally, I discuss the findings.

4.1. Proper Functioning of the Scale

To examine the proper functioning of the rating scale, I used the category statistics as implemented in Facets. Table 2 presents the information related to the reliable functioning of the scale.

Table 2
Category Statistics

Category score	Response category	Absolute frequency	Relative frequency	Average measure	Outfit Mnsq	Most probable from
1	Very poor	15	3%	-.88	1.1	Low
2	Poor	114	20%	-.50	1.0	-2.94
3	Average	227	41%	.29	1.0	-.80
4	Good	146	26%	1.08	1.0	1.07
5	Very good	58	10%	2.07	.9	2.66

Note. Mnsq = Mean square

The first two columns show the number and the name of the scale categories. Column three shows the total number of ratings raters assigned scale categories across essays on items. In column four, the percentage of ratings is shown. The information in column five shows that as the scale categories increase, the ability of the writers on essays increases as well. Outfit Msq in column six shows quality control, and it has an expected value of 1. The last column shows how flat or peaked the response categories are.

According to Linacre (2004), in order for a rating scale to perform effectively, a number of guidelines should be met: (a) there should be at least 10 ratings in each category; (b) average measures should advance monotonically with counts; (c) outfit mean-square values should be less than 2; (d) step difficulties (or scale calibrations) should advance monotonically, signifying that each category is the most probable one for raters to assign to essays that are located in a particular portion of the writer proficiency continuum; and (e) step difficulties (or scale calibrations) should increase by 1.4, but less than 5 logits. Table 2 shows that the rating scale I used met all these guidelines except for the last one. Fortunately, as Linacre (2004) noted, “this degree of rating scale refinement is usually not required in order for valid and inferentially useful measures to be constructed from rating scale observations” (p.274).

Figure 1 schematically shows the scale categories of the rating scale. As the figure shows, the categories are the most probable, implying that they have distinct peaks.

4.2. Self-consistency of Raters

To answer the first research question, I used Facets. Facets generates group-level and individual-level statistics. When raters show randomness in their ratings, they differ very little in their levels of performance, indicating that they it would be difficult for them to distinguish reliably among writers. The following group-level statistical indicators were used to detect whether the raters were self-consistent in their ratings. A fixed-effect chi-square tests the hypothesis that all writers show the same level of performance after accounting for measurement error. A non-significant chi-square test implies raters are not self-consistent. Writer separation index connotes the number of statistically distinct levels of performance among writers. A low writer separation index suggests self-inconsistency. Reliability of the writer separation index shows how reliably raters are able to distinguish writers in terms of their performance.

A low reliability of the writer separation index shows self-inconsistency.

For individual-level statistics, fit indices were used. Facets produces mean-square fit statistics for each rater. Rater infit is an estimate of how each rater is self-consistent in his or her ratings across essays and items. Infit stands for "information weighted" and has an expected value of one. Mean square infit

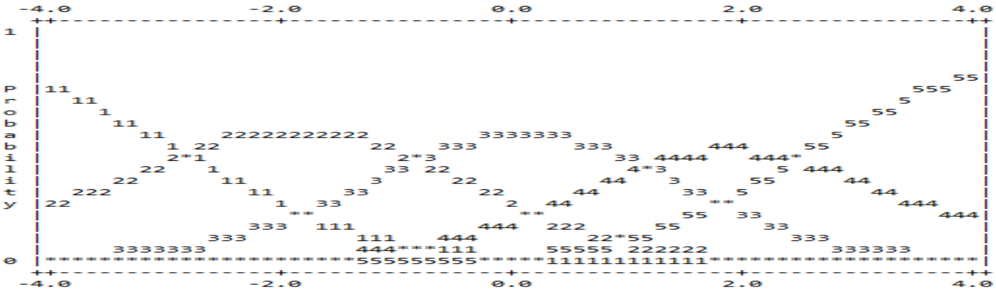

Figure 1. Probability curves

values larger than one show more variation than expected, and infit values lower than one show less variation than expected. Rater outfit is sensitive to highly unexpected, surprising ratings. Outfit stands for "outlier-sensitive fit statistic". Rater outfit is more problematic than rater infit. Rater outfit should be treated before attending to rater infit.

There are no hard-and-fast rules regarding the upper-limit and lower-limit for rater fit indices. Different assessment programmes use different ranges, depending on the nature of the tests and the decisions to be made. In the present study, the upper-limit 1.20 and the lower-limit .80 were set.

Table 3 includes group-level and individual-level statistics regarding whether the raters were self-consistent.

On a group level, the fixed-effect chi-square was statistically significant, $\chi^2(27) = 208.7$, $p<.05$, suggesting that writers were statistically significantly different in their levels of performance. Writer separation index was 3.91, implying that there were approximately four statistically distinct levels of performance among writers. Reliability of writer separation index was .99, showing that raters were able to reliably distinguish levels of performance among writers. On an individual level, rater infit mean square and rater outfit mean square were with acceptable limits set in this study.

Table 3
Group-level and individual-level statistical indicators for raters

Raters	Rater fit statistics	
	Infit Mnsq	Outfit Mnsq
1	.87	.87
2	1.12	1.12
Chi-square: 208.7, df = 27		
Writer separation index: 3.91		
Reliability of writer separation index: .99		
Significance: .00		

Note. Mnsq = mean square

4.3. Consistency of Raters' Ratings in Relation to Those of Other Raters

To answer the second research question, I used single rater–rest of the raters (SR/ROR) correlation as implemented in Facets (Myford & Wolfe, 2003). Also known as point biserial correlation, SR/ROR summarises the extent to which the ratings of each rater are consistent with those of other raters, comparing the ratings of an individual rater with those of all the other raters included in the analysis. SR/ROR is meaningful when it is interpreted in relation to rater fit indices.

Table 4 provides useful information on consistency of the ratings. In column one, the number of raters is given. In the next two columns, rater infit and rater outfit values are presented. They are the same values as presented in Table 3. Column four shows how the ratings of the two raters are consistent. When these values are significantly different, it may indicate the ratings of the raters are inconsistent. However, these values are close to each other.

The results of fixed effect chi-square test, writer separation index, reliability of writer separation index, and fit indices showed that when Persian native raters rate nonnative Persian language learners' essays, they are self-consistent in the assignment of their ratings. The results of point biserial correlation showed that Persian raters' ratings were consistent across writers and items.

Table 4

Consistency of the Ratings of Raters

Raters	Infit MnSq[a]	Outfit MnSq	Corr.PtBis[b]
1	.87	.87	.30
2	1.12	1.12	.26

a. Mean square, b. Point biserial correlation

The findings from this study lend support to some of those from previous studies. Knoch, Read, and von Randow (2007) trained two groups of raters (*n* = 8 in each group) on line and face to face to rate 70 writing scripts using an 6-analytic rating scale on fluency, content, and form. The results of Facets analyses revealed that both groups rated the scripts consistently before and after training; however, the online group rated slightly more consistently. Knoch et al. concluded that "it seems that both rater training methods were effective. Online training seemed slightly more effective …. There were no differences between the two groups when inconsistencies …. were analyzed" (p. 42).

The findings also confirm those demonstrating that rater training makes raters self-consistent. Building on the findings of previous studies and summarizing the arguments supporting training, McNamar (1996) concluded that “rater training is successful in making raters more self-consistent. That is, the main effect of training is to reduce the random error in rater judgments” (p. 126).

Some factors may have contributed to the consistent ratings of the Persian raters rating nonnative Persian language learners’ essays. One cogent reason documented in the previous studies relates to the training these raters received before rating. The raters were trained for half an hour before they actually started rating. This brief training seems to have been effective in making the raters self-consistent. A second reason for consistent ratings of the raters may have to do with writing experience. Although they were inexperienced raters, they had the informal rating of the essays nonnative Persian speaking language learners wrote at Imam Khomeini Language centre. When rating was combined with this informal rating, this combination may have led to consistency of raters. The final possible explanation may concern their familiarity with the scale they used to rate the essay. The researcher-made analytic rating scale in this study used some of the criteria they were using to rate the essays of nonnative Persian speaking language learners. Prior familiarity with the rubrics has been reported to result in reducing variation, thereby leading to more consistency.

5. Conclusion and implications

Analysis of rating data using FACETS resulted in two findings in this study. The first finding was that Persian teacher raters were self-consistent in rating the essays nonnative Persian speaking language learners at Persian language Centre at Imam Khomeini International University wrote. The second finding was that the raters were consistent in the assignment of ratings when their ratings were analysed. The points to note are that self-consistency is related to the ratings of a single rater across writers and items, but consistency of ratings has to do with all the ratings of a single rater in relation to all the ratings of all other raters across writers and items.

The findings are significant in that this was the first small-scale study using the many-facet Rasch measurement model to analyse randomness among Persian teacher raters rating ESL essays. Randomness, or inconsistency, is a systematic rater effect which endangers the validity of the ratings, thereby invalidating the decisions to be made on language learners' promotion to a higher level of language education. In achievement testing, it is very important that the language learners' written performance reflect the language learners' underlying written abilities, and these abilities should not be affected unduly by some other irrelevant factors, including randomness.

The findings may carry some implications for research and pedagogy. From a research perspective, using the many-facet Rasch measurement model can be successful in making the Persian teacher raters more self-consistent in the assigning ratings. From a pedagogical point of view, when raters are more consistent in the assignment of the ratings, less variation creeps into their ratings. Therefore, more valid results are obtained, and fairer decisions are made regarding language learners.

In closing, it should be noted that this was a small-scale, cross-sectional study on the rating data using the many-facet Rasch measurement model, and whether self-consistency of the Persian teacher raters and the consistency of the ratings persists over time merits more longitudinal studies in which more Persian teacher raters rate the nonnative Persian speaking language learners.

References

Brown, J. D., & Hudson, T. (1998).The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675.

Douglas, D. (2010). *Understanding language testing*. Oxon: Hodder Education.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs* *for all students: Development, implementation, and analysis* (pp. 261-287). Mahway, NJ: Lawrence Erlbaum Associates.

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.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Knoch, U., Read, J., von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

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., Engelhard, G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*(6), 569–577.

Lumely, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M. E., & Linacre, J. M. (1998). Measurement designs using multifacet Rasch modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 47–77). Mahwah, NJ: Erlbaum.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press