Evaluating the effectiveness of the training program on direct and semi-direct oral proficiency assessment: A case of multifaceted Rasch analysis

Houman Bijani*

Abstract: An Oral Proficiency Interview (OPI) may be evaluated either during the interview procedure (direct-method) or from a tape-made oral interaction (semi-direct method). Such variety in methods of assessment can influence test takers' scores extensively. However, it is not conclusive whether such differences are due to test format or something else. Besides, most previous studies have investigated the application of multifaceted Rasch measurement (MFRM) on only one or two facets and few studies have used a pre- and post-training design. 20 English as a foreign language (EFL) teachers rated the oral performances produced by 200 test takers before and after a training program. The tasks were implemented via two methods of task delivery, direct and semi-direct. The findings indicated that test takers found semi-direct oral tests harder and more stressful than direct oral tests since they are more difficult due to having more complicated linguistic and communicative features and contained more lexically complex structures. Therefore, direct oral tests are more appropriate for low-level test takers, and semi-direct tests for higher ability ones. Data analyses demonstrated no significant difference between the ratings of direct and semi-direct oral assessment by the raters. Consequently, semi-direct oral tests can be regarded as a reliable substitute for direct oral tests.

ABOUT THE AUTHOR
Houman Bijani is an assistant professor in Applied Linguistics, teaching English as a foreign language (TEFL) at Islamic Azad University, Zanjan Branch, Iran. He got his PhD in TEFL from Tehran Islamic Azad University, Science and Research Branch and his MA in TEFL from Allameh Tabatabai University as a top student. He is also an English language teacher and supervisor at Iran Language Institute (ILI). He is a CELTA holder awarded by LTTB Center in Brussels, Belgium. He has published several research papers in scholarly national and international language teaching and assessment journals. His areas of interest include quantitative assessment, teacher education, and language research. The current research paper is a part of a wider study entitled oral performance assessment: the use of FACETS in raters' rating biasedness which was conducted for the fulfillment of the requirement of PhD study.

PUBLIC INTEREST STATEMENT
Second/Foreign language speaking can be evaluated through various formats. The two most popular formats are direct and semi-direct oral assessments where test takers are interviewed directly in the former and through a tape-made oral interaction in the latter format. However, there is paucity of research findings indicating whether test takers’ oral score variation is due to the test format or other factors. In this study, 20 English as a foreign language (EFL) teachers scored the oral tasks produced by 200 test takers before and after a training program implemented via two methods of task delivery, direct and semi-direct. The findings indicated that test takers found semi-direct oral tests harder and more stressful. Consequently, direct oral tests are more appropriate for low-level test takers, and semi-direct tests for higher ability ones. The results also displayed that semi-direct oral tests can be regarded as a reliable substitute for direct oral tests.
a reliable substitute for direct oral tests. Finally, the outcomes showed that although training can result in higher levels of interrater consistency and reduced levels of severity/leniency, biasedness and inconsistency, it cannot turn rater into duplicates of one another.

Subjects: Language & Linguistics; Language & Education; Language Teaching & Learning

Keywords: Bias; direct oral assessment; interrater consistency; intrarater consistency; Multifaceted Rasch Measurement (MFRM); rater training; semi-direct oral assessment

1. Introduction

Oral Proficiency Interview (OPI), in which raters and test takers participate in a non-written discussion, has been a common method of assessing speaking proficiency. Because such test is free from conventional testing features, it allows test takers to demonstrate their interactional capabilities (Mirzaei, Hashemian, & Farsani, 2016). Although OPIs are considered non-written and rather unstructured, the rater is provided with some guidelines suggesting various topics which differ for each test taker. Moreover, OPI has shown a high amount of reliability and validity. Kyle, Crossley, and McNamara (2016) stated such this test is reliable because it is rated by trained raters who assign almost the same score to both live and taped interviews. They further added that OPI is a face valid test because it demands that test takers speak in a realistic conversational situation. By the use of trained raters, the content validity of OPI tests is high in various interview settings. Along with the OPI, special guidelines are given to raters for evaluation in order to provide them with descriptions of various proficiency levels ranging from novice to advanced.

A major problem with OPI tests is that they require fully trained raters and that they are time-consuming; however, the essentiality of oral competence assessment justifies for this drawback. According to Kuiken and Vedder (2014), the OPI test is a commonly used assessment measure which is typically a part of a battery of tests used for diverse purposes. OPIs have been a popular method of oral language proficiency assessment since the emergence of modern oral proficiency testing (Maria-Ducasse & Brown, 2009). An OPI may be tested meanwhile it is in progress during the interview process, or later on from a tape-made OPI. Second language oral assessment increasingly requires more performance-oriented tests. Performance-oriented tests demand that test takers produce complicated responses using their knowledge and expertise and apply their target language skills to similar real-life situations. Such tests typically require more than one test method and human rater judgments. Therefore, these two factors, the test method and the rater have become incorporated elements of performance-based tests that affect test scores (Kim, 2015).

2. Literature review

Second language oral tests are greatly influenced by the method used to measure them (Kim, 2015). Studies by Foster, Tonkyn, and Wigglesworth (2000) demonstrated that the use of various test methods can influence the construct to a great extent, affect test takers’ scores differently, tap various aspects of test takers’ second language oral proficiency, and identify test takers’ various attitudes. An oral interaction may be evaluated by whether they are in progress during the interview procedure or from a tape-made oral interaction. Clark (1978) provides the basis for distinguishing two different types of oral proficiency tests including direct and semi-direct oral language assessment methods.

2.1. Direct oral assessment

Direct tests of speaking proficiency have been developed by involving an assessment setting in which the test taker and one or more human raters get involved in a communicative oral interaction (Clark, 1978). According to him, once the need for an active test of speaking ability has been established, the main issue becomes selecting which test (direct vs. semi-direct) procedure best fulfills the test users and other stakeholders’ requirements.
OPI along with its rating scales has been in use since 1956. Even though this test has been revised and refined several times since then, its main body format has been left intact. According to Stansfield and Kenyon (1992), several scholars, since the 1970s, have identified that this model did not include some facets of oral language proficiency including pragmatic, contextual and strategic facets of language proficiency. Clark (1978) further argued that OPI suffers from two main limitations as a result of its disability in reflecting real-life conversational settings. Firstly, there is the problem of the rater, i.e., in the interview situation. The test taker is certainly aware that s/he is talking to a rater and not a common person in the society. Secondly, the language which is elicited in an interview hardly reflects the real-life conversational discourse. However, in spite of these criticisms, according to Clark (1978), a face-to-face interview seems to possess the highest degree of validity as a measure of global oral proficiency and is thus superior to both semi-direct and indirect speaking tests.

2.2. Semi-direct oral assessment

The term semi-direct oral test was first coined by Clark (1978) to describe the tests which elicit active speech from test takers through tape recordings, printed text booklets, or any other non-human elicitation procedure, rather than through a face-to-face conversation with a present interlocutor. Tape-based oral tests may be the result of a direct test in which the test takers are interviewed in a face-to-face situation, or a semi-direct test in which the test takers do the test in a language laboratory giving responses to tape prompts (May, 2006). The Simulated Oral Proficiency Interview (SOPI) was created as an alternative to direct OPI for the sake of having more feasibility of administration. Semi-direct oral tests were developed to ensure reliability and validity without having the burden of direct testing features. They represent an attempt to standardize the assessment of speaking while keeping the principles of direct testing as well (Leaper & Riazi, 2014). Stansfield (1991) states that SOPI provides raters with an economical alternative to OPI. Since SOPI is tape-mediated, it does not require a trained rater for test administration. Moreover, SOPI could be performed simultaneously to a group of test-takers, whereas OPI should be performed individually.

Regarding the validity issue, Stansfield (1991) argues that one critical problem with regard to OPI is that test takers' performance is determined by the skill of the rater to a great extent, whereas SOPI enjoys the same language input quality of each test taker. Consequently, the decision of which test method to use depends on the purpose of the test. In other words, OPI may be more appropriate for placement and course curriculum evaluation purposes, whereas SOPI is more suitable when making important decisions.

There has been some research on the relationship between direct and semi-direct oral language testing with impressive outcomes. In a study by Clark and Li (1986), four types of oral proficiency interviews were administered in China which were then performed live the second time. The result represented a high correlation of 0.93. Shohamy (1994) in a comparison of her test takers' performances on both test methods found the concurrent validity of the two types of tests were high. She also pointed out that these tests were different from OPI test; however, the scores were highly comparable. She also concluded that OPI test was more suitable for low-level test takers, whereas SOPI for high-level ones. Besides, she found that test takers used more self-correction and paraphrasing in SOPI while they resorted to first language (L1) more in OPI.

The degree to which such tests are a valid and reliable alternative to direct oral tests was investigated by Stansfield (1991). Stansfield (1991) in a comparative study of OPI and SOPI tests argues that SOPI has shown itself to be a valid and reliable substitute for OPI. On behalf of the comparison of scores on the two kinds of tests, he reports Pearson correlations between 0.89 and 0.95. A large majority of test takers (86%) who preferred the live test felt more nervous in the taped test. Moreover, a majority of them (90%) found the taped test more difficult. On behalf of the comparison of scores on the two kinds of tests, he reports Pearson correlations between 0.89 and 0.95. Stansfield (1991) then justifies that the reason why OPI and SOPI are highly correlated is perhaps because both tests do not
let test takers represent their interactive skills fully. He argues that, even in the OPI test, both the rater and the test taker believe that it is the test taker’s responsibility to perform the talking. Thus, the kind of spoken language is not the mirror manifestation of natural talk in a real speech context. However, he argues that still there is more interactive talking in OPI than SOPI. Unlike Stansfield (1991), Shohamy (1994), regarding the comparative statistical scores of the two types of tests, argues that high correlations between scores on the two different tests provide necessary but inadequate evidence for the substitution of either for the other. In other words, according to her, these two tests may not be measuring identical things. She further argues that it is required to measure the validity of them from various perspectives, not just through a simple correlation of their outcome scores.

Kenyon and Tschirner (2000) in a quantitative comparison of Spanish OPI and SOPI oral performances found that SOPI test takers use significantly more turns, quotes, speech acts, switches to L1 than those on OPI. Ahmadi and Sadeghi (2016) found rather similar findings but with regard to lexical density. They compared OPI and SOPI of 20 test takers and found significant differences between the two in a way that SOPI produces more lexically dense type of language. Shohamy (1994), in her own study, found that SOPI elicits more limited range of language functions than OPI. That is, most of the language functions used in the SOPI, as she found, were self-correction, repetition of phrases and paraphrasing. There were also more prosodic features in SOPI performance including hesitation and silence, and the language was more formal with more cohesive devices. She further added that SOPI tests produced language which was more literate than OPI ones especially with regard to lexical density. She, on the basis of these results, claimed that these two tests do not measure the same thing.

Qian (2009) compared live and audio-recorded interviews of 27 IELTS test takers rated by three raters. The raters tended to rate the recorded interviews lower than the live ones and the results demonstrated that some raters were influenced by extralinguistic, paralinguistic and nonlinguistic features. In another empirical study, Khabbazbashi (2017) investigated oral test scores using the multi-faceted Rasch program, FACETS. The research was on the basis of the data obtained from 83 candidates on a direct and semi-direct test. They reported that the ability estimates of candidates obtained using the FACETS program were highly correlated at r = 0.92. Jeong and Hashizume (2011) analyzed the 83 test takers’ attitudes and performance over a live and tape-based OPI test. The FACETS outcomes revealed a preference over the direct (live) method although the results indicated both methods to be valid enough. Moreover, female test takers, also, found the taped method more difficult than the live method.

However, most of the studies conducted so far have investigated the application of FACETS on only one or two facets. For example, the study of rater’s severity/leniency on specific test takers (Lynch & McNamara, 1998), on task types (In’nami & Koizumi, 2016), and on certain rating time (Lumley & McNamara, 1995). Thus, no study, so far, has included the facets of test takers’ ability, raters’ severity, task difficulty, group expertise, and test method all in a single study along with their bilateral effects. Besides, while a few studies have looked at the differences between trained and untrained raters in speaking assessment (Bijani, 2010; Elder, Iwashita, & McNamara, 2002; Gan, 2010; Kim, 2011) few, if any, studies have used a pre- and post-training design. Also, no research could be found showing that the change and alteration of various elicitation technique in oral testing prompts may affect test takers’ output and hence their scores. In this respect, although the results of some studies (e.g., Stansfield & Kenyon, 1992) suggest different test performances on the direct OPI and SOPI by test takers, it is not conclusive whether such differences are due to test format or something else. Therefore, this research is aimed to investigate the above-mentioned shortcomings through taking a meticulous analytical approach investigating the five mentioned facets using a pre, post-training design to investigate raters’ change of rating behavior for the direct and semi-direct oral assessment tests. Besides, any possible difference between the two methods of direct and semi-direct oral assessment test with regard to raters’ severity, bias and consistency differences for both groups of rater expertise is explored. This
will explore which test method evaluates test takers’ oral proficiency more reliably and validly. Therefore, the following research question can be formed:

RQ1: Is there a difference between the direct and semi-direct oral language proficiency assessment in terms of scoring quality? What particular micro-linguistic and communicative features play a role in this respect?

RQ2: Is there any significant difference in raters’ measures of severity, bias and consistency in scoring for the direct and semi-direct test methods before and after the training program?

3. Methodology

3.1. Participants

200 adult Iranian students of English as a Foreign Language (EFL), including 100 males and 100 females, ranging in age from 17 to 44, participated as test takers. The students were selected from, upper-intermediate, and advanced levels (50 students of each level) studying at the Iran Language Institute (ILI). 20 Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 participated in this study as raters.

3.2. Instruments

3.2.1. The speaking tests

The elicitation of test takers’ oral proficiency was done through the use of five different tasks including description, narration, summarizing, role-play and exposition tasks. Task 1 (Description Task) is an independent-skill task which reflects test takers’ personal experience or background knowledge to respond in a way that no input is provided for it. On the other hand, tasks 3 (Summarizing Task) and 4 (Role-play Task) reflect test takers’ use of their listening skills to respond orally. For tasks 2 (Narration Task) and 5 (Exposition Task), the test takers were required to respond to pictorial prompts including sequences of pictures, graphs and tables. The tasks were implemented via two methods of task delivery: (1) direct and (2) semi-direct. The direct test was designed for use in an individual face-to-face method (i.e., a single test taker speaking to an interlocutor _ here a rater), whereas the semi-direct test was designed for use in a language laboratory setting. Although it is evident that test takers might differ in terms of their background knowledge about the given topics for both test prompts, what matters in their productions is the data analysis of linguistic features.

3.2.1.1. Direct speaking tasks. The direct speaking tasks were administered through face-to-face oral interaction between the test takers and the raters which took around 15 to 20 minutes. Table 1 demonstrates the range of tasks and their criteria used in the direct method.

3.2.1.2. Semi-direct speaking tasks. The semi-direct format requires test takers to respond to pre-recorded prompts on a handout in a language laboratory setting. After getting the task prompt, the test takers were given 60 seconds time to prepare themselves before recording their responses. Then, they were given 5 minutes to respond. Table 2 demonstrates the range of tasks and their criteria used in the semi-direct method.

3.2.2. The scoring rubric

For both methods of the test, each test taker’s task performance was assessed using the Educational Testing Service (ETS, 2001) analytic rating scale. In ETS (2001) scoring rubric, individual tasks are assessed using appropriate criteria including fluency, grammar, vocabulary, intelligibility, cohesion and comprehension.
| Task No. | Task Form | Language Function | Task Type | Topic | Preparation Time | Response Time |
|---------|-----------|-------------------|-----------|-------|------------------|---------------|
| 1       | Description | Describing | Open Monologic Unplanned Divergent Perspective (both) Immediacy (both) | Educational setting | | |
| 2       | Narration | Narrating, explaining | Closed Monologic Planned Divergent Perspective (both) Immediacy (both) | Car breakdown | 1 minute | |
| 3       | Summarizing | Summarizing | Closed Monologic Planned Divergent Perspective (both) Immediacy (both) | Looking for new ways to fight malaria | 1 minute | 1 minute |
| 3       | Role play | Explaining, apologizing, making excuses | Closed Dialogic Planned Convergent Perspective (both) Immediacy (both) | Jobs | 1 minute | |
| 4       | Exposition | Describing, comparing, analyzing, interpreting | Closed Monologic Planned Divergent Perspective (both) Immediacy (both) | Mobile use and fast foods | 1 minute | |
| Task No. | Task Form | Language Function | Task Type | Topic | Preparation Time | Response Time |
|----------|-----------|-------------------|-----------|-------|------------------|---------------|
| 1        | Description | Describing | Open Monologic Unplanned Divergent Perspective (both) Immediacy (both) | Work setting | 1 minute | 2 minutes |
| 2        | Narration | Narrating, explaining | Closed Monologic Planned Divergent Perspective (both) Immediacy (both) | Renting a flat | 1 minute | 2 minutes |
| 3        | Summarizing | Summarizing | Closed Monologic Planned Divergent Perspective (both) Immediacy (both) | Looking for new ways to fight malaria | 1 minute | 1 minute |
| 4        | Role play | Explaining, apologizing, making excuses | Closed Dialogic Planned Divergent Perspective (both) Immediacy (both) | Party invitation | 1 minute | _____ |
| 5        | Exposition | Describing, comparing, analyzing, interpreting | Closed Monologic Planned Divergent Perspective (both) Immediacy (both) | Staff employment | 1 minute | 2 minutes |
3.3. Procedures

3.3.1. Pre-training phase
The 200 test takers participating as data providers were divided randomly into two groups in a way that each group took part in each stage of the study (pre-, and post-training). From each group, half of the test takers took the direct and the remaining half took the semi-direct test method. The reason for not having all the participants perform both methods of the test was due to the fact that performance in one method would most certainly affect their performance on the other method through enabling them to get used to the typology of the questions and this would invalidate the findings of the study. The rating design was in a way that each test taker on each method of the test was rated by all raters participating in the study so that the data matrix was fully crossed; however, for practical reasons, the raters were classified into two groups in a way that half of them scored the direct method first and then the semi-direct method, whereas for the other group, they scored the semi-direct method first and then moved to the direct method.

3.3.2. Rater training program
After the pre-training scoring stage, the raters participated in a training (norming) session in which the speaking tasks and the rating scale were introduced and time was given to practice the instructed material with some sample responses. The training program consisted of rater norming and feedback on previous rating behavior and was conducted in two separate norming sessions, each lasting for about six hours, with an interval of one week. Regarding feedback on raters' biases, the raters having Z-scores beyond ±2 were considered to have a significant bias and were reminded individually to mind the issue accordingly. For feedback on raters' consistency, the raters having infit mean squares beyond the acceptable range of 0.6 to 1.4, as suggested by Wright and Linacre (1994), were considered as misfitting in a way that the raters with an infit mean square value below 0.6 are referred to as too consistent (overfit the model) and those with an infit mean square value of above 1.4 as inconsistent (underfit the model). Therefore, the raters were pointed out individually on the issue if they were identified as misfitting.

3.3.3. Post-training phase
After the training program, the tasks of both methods of the test were run. The second half of the test takers (including 100 students) was used from whom to elicit data. All the raters participating in this study were given one week to submit their scorings. The rating design was in a way that each test taker on each method of the test was rated by all raters participating in the study so that the data matrix was fully crossed; however, for practical reasons, the raters were classified into two groups in a way that half of them scored the direct method first and then the semi-direct method, whereas for the other group, they scored the semi-direct method first and then moved to the direct method. Moreover, the selection of 100 oral performance data of both methods at each stage was done randomly for each rater. Randomization was done in order to counteract the influence of sequencing the performances on the raters' behaviors so that they could not remember how many data at a particular score were rated by them. Table 3 demonstrates the scoring procedure at the pre-training and post-training data collection stage.

3.4. Data analysis
In order to investigate the research questions, the researcher employed a pre-post, research design using a quantitative approach to investigate the raters’ development with regard to rating second language (L2) speaking performance (Cohen, Manion, & Morrison, 2007). Quantitative data (i.e., raters' scores based on an analytic scoring rubric) were collected and analyzed with a Multifaceted Rasch Model (MFRM) during two scoring sessions including the facets of takers, rater, rater group, task, and test method and their interactions to investigate variations in rater behavior and rater biasedness. Subsequent bias analyses were also performed to investigate the interaction between raters and the two test methods, i.e. direct and semi-direct. Moreover, the comparative difficulty of the two test methods was investigated in a way that the degree of item difficulty obtained from the FACETS analysis was mapped using both test methods.
4. Results

RQ1: Is there a difference between the direct and semi-direct oral language proficiency assessment in terms of scoring quality? What particular micro-linguistic and communicative features play a role in this respect?

Measuring **linguistic features** was done through focusing on the number of errors made within the domains of morphology, syntax and lexicon. In detail, it involved micro-features of word order, tenses, verb structure, pronouns, grammatical gender, singularity/plurality, prepositions, articles and lexical accuracy. Frequency of errors for every form was counted. Accordingly, the number of errors for each form was added and then divided by the total number of words produced by the test takers on each test. A comparative analysis of the ratio between the two types of tests (direct and semi-direct) was performed using a Mann–Whitney U test (Nonparametric t-test). The means standard deviations, t-values and significance levels are displayed in Table 4. The outcomes demonstrated a significant difference in pronouns, tenses and verb structure between the two tests. For the rest of the linguistic features, no significant difference was observed.

Measuring **communicative strategies** was done through focusing on issues of hesitation, self-correction, paraphrasing, and switching to L1. Like before, the frequency of occurrence of the above-mentioned issues was measured. The means of the frequencies, obtained through dividing the frequencies by the total number of words produced by the test takers on each test method, was compared using a Mann–Whitney U test. The means standard deviations, t-values and significance levels are displayed in Table 5. The results showed a significant difference for paraphrasing and self-correction. The frequency of self-correction and paraphrasing instances of semi-direct test were more than those of direct test.

RQ2: Is there any significant difference in raters’ measures of severity, bias and consistency in scoring for the direct and semi-direct test methods before and after the training program?

A bias analysis was performed to analyze raters’ behavior with respect to their severity, biasedness and consistency in scoring both test methods at the pre-training phase (Table 6). FACETS is capable of calculating raters' biases in various testing contexts—in particular here test methods (direct and semi-direct tests) by comparing the expected and observed values in a set of data and then reporting the outcome in a form of residuals. Later on, through converting residuals into Z-scores, the bias value is obtained. This Z-score shows any significant difference from what was expected from that particular rater allowing for routine and acceptable score variation. Finally, a Z-score between ±2 is regarded as a rater's normal scoring behavior thus acceptable range of biasedness.

*Column one (Oral test method) displays the oral test methods used in the study, i.e., direct and semi-direct test method.*
Columns two (Observed average score) displays the average observed scores given by the raters to test takers’ oral performance on each test method.

Column three (Fair average) demonstrates the extent to which the mean ratings of raters on each test method differ. For instance, here, the mean rating of the direct test method was 22.12 and her fair average was 22.90. Similarly, the mean rating of the semi-direct test method was 18.74 and his fair average was 19.82. These data show that the two direct test methods were 0.78 raw-score points apart when comparing their mean ratings and 1.08 raw scores apart when comparing their fair averages. According to Winke, Gass, and Myford (2012) both values demonstrate severity spread; however, the difference is that fair average is a better estimate when not all raters scored all the tasks. Eckes (2015) further reiterated that when fair average is greater than 1 point, then this shows a significant high difference between the severest and the most lenient raters in the use of scoring scale.

Column four (Obs-Exp score in logits) displays the total observed score for all the 100 test takers participating at the pre-training phase on each test method minus the total expected score for the test takers on the same test method. Since there were five tasks in the study for each test method and for each task and the possible score range for each task was between 1 and 7, there would be the possibility of scoring each test taker a score of 5 to 35.

Column five (Bias logit) demonstrates the bias value, representing raters’ severity/leniency (in each test method) in the performance assessment of test takers in that test method. Positive values represent severity, whereas negative ones represent leniency. Here, the outcome shows

| Micro-linguistic features | Semi-direct | Direct | t-value | Sig. |
|---------------------------|-------------|--------|---------|------|
|                           | X           | Sd     | X       | Sd   |     |
| Word order                | 16.34       | 0.98   | 16.84   | 1.08 | 0.56 | 0.63|
| Tenses                    | 22.34       | 1.09   | 26.28   | 1.79 | 3.57 | 0.033|
| Verb structure            | 26.84       | 1.68   | 33.34   | 2.15 | 4.46 | 0.015|
| Pronouns                  | 19.78       | 1.13   | 15.93   | 0.84 | 3.84 | 0.021|
| Grammatical gender        | 14.75       | 2.14   | 15.03   | 2.56 | 0.45 | 0.24|
| Singularity/Plurality     | 24.41       | 2.87   | 24.05   | 2.64 | 1.76 | 0.37|
| Prepositions              | 33.07       | 1.17   | 32.59   | 1.02 | 2.12 | 0.14|
| Articles                  | 37.18       | 2.85   | 36.79   | 2.52 | 2.37 | 0.51|
| Lexical accuracy          | 18.24       | 2.45   | 18.62   | 3.01 | 0.69 | 0.49|

| Communicative features    | Semi-direct | Direct | t-value | Sig. |
|---------------------------|-------------|--------|---------|------|
|                           | X           | Sd     | X       | Sd   |     |
| Hesitation                | 13.73       | 0.81   | 12.22   | 0.77 | 0.88 | 0.094|
| Self-correction           | 31.62       | 2.71   | 22.19   | 2.03 | 4.08 | 0.027|
| Paraphrasing              | 19.39       | 2.12   | 13.15   | 1.98 | 3.75 | 0.042|
| Switching to L1           | 9.41        | 0.84   | 11.29   | 1.56 | 2.19 | 0.170|

*Table 4. Frequency of error means, standard deviations, T-values and significance for linguistic features of direct and semi-direct tests*

*Table 5. Means, Standard Deviations, T-values and significance levels for communicative strategies of direct and semi-direct tests*
Table 6. Direct and semi-direct test method measurement report (pre-training)

| Oral test method | Observed average score | Fair average | Obs-Exp score (logit) | Bias Logit | SE | Infit MnSq. | Z | Outfit MnSq. | Z | Point biserial |
|------------------|------------------------|-------------|-----------------------|-------------|----|------------|---|-------------|---|---------------|
| Direct           | 22.12                  | 22.90       | 0.75                  | -0.41       | 0.05| 1.2        | -0.81| 1.1          | -0.77| 0.26          |
| Semi-direct      | 18.74                  | 19.82       | -0.31                 | 0.23        | 0.04| 1.3        | 0.66 | 1.4          | 0.43 | 0.38          |
| Mean             | 20.43                  | 21.36       | 0.22                  | -0.09       | 0.04| 1.25       | -0.07| 1.25         | -0.17| 0.32          |
| SD               | 2.39                   | 2.17        | 0.74                  | 0.45        | 0.00| 0.07       | 1.03 | 0.21         | 0.84 | 0.08          |

Fixed (all same) chi-square: 87.64, df = 1, p < 0.00
Separation index (Direct): 2.73 Reliability (Direct): 0.91
Separation index (Semi-direct): 2.44 Reliability (Semi-direct): 0.94
that the raters in the direct test method were rather lenient towards the test takers with the leniency of (−0.41 logits). However, in the semi-direct test, they were severe with the severity of (0.17 logits). The mean bias value (in logits) measured −0.09, thus the raters in either test method displaying more than half a logit value above or below the mean logit value (between −0.59 and 0.41) would be considered as either too severe or lenient (McNamara, 1996; Wright & Linacre, 1994). In this respect, no significant severity/leniency was observed on the ratings of the either direct or semi-direct test method.

Column six (SE) displays the standard error of bias estimation. The small amount of SE provides evidence for the high precision of measurement.

Columns seven and nine (Infit and outfit mean square) display the fit statistics which show to what extent the data fit the Rasch model. An observed score is the one given by a rater to a test taker on one criterion for a task, and an expected score is the one predicted by the model considering the facets involved (Wright & Linacre, 1994). In other words, fit statistics simply is used to determine within-rater consistency (Intra-rater consistency) which indicates the extent to which each rater ranks the test takers consistent with his/her true ability. Fit statistics is categorized into two subparts entitled infit and outfit statistics and most researchers employ them because they are said to be less sensitive to sample size and that they are commonly weighted on the information provided by the responses. Infit is the weighted mean square statistic which is weighted towards expected responses and thus sensitive to unexpected responses near the point where the decision is made. Outfit is the same as above but it is unweighted and is more sensitive to sample size, outliers and extreme ratings (Bonk & Ockey, 2003). Fit statistics has the expected value of 1 and a range of zero to infinity; however, there is no straightforward rule or universally definite range for interpreting fit statistics value; therefore, the acceptability of fit is done on a judgmental basis. The acceptable range of fit statistics, although various among statisticians, according to Wright and Linacre (1994), is within 0.6 to 1.4 logit values. Therefore, in order to investigate the fit statistics value. The raters (of each test method) who are placed below this range are overfit or too consistent, and those above this range are underfit (misfit) or too inconsistent. The infit mean square for the direct test method measured 1.2 and for the semi-direct method 1.3. This finding demonstrates that the ratings of both test methods, according to Wright and Linacre (1994), are at the acceptable fit statistics range showing relative consistency before training, however, for the semi-direct test method, through considering the outfit mean square value, they were spotted on the borderline of consistency.

Also, columns eight and ten (Z-scores) which are sometimes called standardized infit statistics display test method rater bias estimate at this phase. Bias is the difference between expected and observed ratings of the obtained data which is then divided by its standard error to achieve the Z-score (Stahl & Lunz, 1992). The most preferable amount of Z value is 0 which indicates that the data match the expected model, thus there exists no bias on the side of raters. According to McNamara (1996) Z values between ±2 are considered as the acceptable range of biasedness, thus any values above or below the given score are considered to be either to positively or too negatively biased. Accordingly, the raters in both test methods were considered as having non-significant biasedness but to opposite directions, i.e., the raters of the direct test method had the tendency towards leniency ($Z_{Direct} = -0.81$) while the raters of the semi-direct test towards severity ($Z_{Semi-direct} = 0.66$). Although the ratings of both test methods were within the acceptable range of biasedness, the result indicates that the amount of biasedness for the ratings of the direct test method at the pre-training phase was more than that of the semi-direct test.

However, the logit severity estimates do not themselves tell us whether the differences in severity/leniency estimates are meaningful or not; consequently, FACETS also provides us with several indications of the reliability of differences among the elements of each facet. The most helpful ones to study are Separation index, Reliability and Fixed chi-square which can be found at the bottom of the table.
The separation index is the measure of the spread of the estimates related to their precision. In other words, it is the ratio of the corrected standard deviation of element measures to the root-mean-square estimation error (RMSE) which shows the number of statistically distinct levels of severity among the raters. Adequate separation is important in situations in which a test produces scores that test users use to separate test takers into categories defined by their performance (Eckes, 2015). In case the raters were equally severe, the standard deviation of the rater severity estimates should be equal to or smaller than the mean estimation error of the entire data set which results in a separation index of 1.00 or even less (if there is a total agreement among raters, the separation index should be 0.00). In the case of this phase of the study, the separation index of 2.73 for the direct test method and 2.44 for the semi-direct test indicate that the variance among the raters, of each test method, is more than the error of estimates and shows that the raters of each test method were not equally severe. The reliability demonstrates to what extent or how well the analysis distinguishes among the facet elements with respect to various levels of severity/leniency. The high amount of reliability in this phase of the study, \( r = 0.91 \) for the direct test method and \( r = 0.94 \) for the semi-direct test method, indicate that the analysis could reliably separate the raters of each test method into different levels of severity. Fixed chi-square tests the null hypothesis to check whether all elements of the facet are equal or not. The fixed chi-square value for all the 20 raters rating the test takers' oral performance of each test method was measured. The chi-square value indicates whether there was a significant difference in raters' level of severity (\( X^2(1, N=20) = 87.64, p < 0.00 \)). Here, the high value of chi-square indicates that the ratings of the two test methods did not share the same on a parameter (e.g., severity). Consequently, the outcome suggested that the raters of either test method are not at the same level of severity.

As it was already indicated above, the raters' separation index which was measured 2.73 and 2.44 for the direct and semi-direct test methods, respectively, indicates that there were almost three statistically distinct levels of severity. Statistically distinct levels are defined as those separation indices that are three standard errors apart, centered on the mean of the sample (Davis, 2016). The reliability of this rater separation index was 0.91 and 0.94 for the direct and semi-direct test methods, respectively, showing that the raters were reliably separated with respect to their level of severity and the analysis was reliable. As explained by Winke et al. (2012) separation reliability indices close to zero show that raters did not differ significantly in terms of their levels of severity and that they had rather similar levels of severity; whereas the separation reliability indices close to 1.0 demonstrate that the raters were very reliably separated with respect to their severity levels. Here, the rater separation reliability of 0.91 and 0.94 for direct and semi-direct test methods represents that the raters differed with regard to their severity variation in scoring the examinees oral performance.

Column eleven (Point biserial correlation) displays the correlation coefficient between each rater and the rest of the raters participating in this study in either test method. Here, the correlation coefficient for the direct test method was measured 0.26 (less than typical according to Cohen's table of effect size) and for the semi-direct method 0.38 (typical according to Cohen's table of effect size).

Table 7 displays the bias analysis of the ratings of both oral assessment methods at the post-training phase.

Column five (Bias logit) displays raters' severity/leniency measures in each testing method. On the first glance, it is clear that the raters of both test methods have modified rating behavior with respect to severity/leniency indices to a high extent. Similar to the pre-training phase, the outcome of the table revealed that still the ratings of the semi-direct test method were severer (logit value: 0.17) than those of the direct test method (logit value: −0.08). It should be noted that the ratings of the direct test method have reduced severity/leniency a lot more than those of the semi-direct test method. At the post-training phase, the mean bias value (in logits) measured 0.04; thus, the rater groups displaying more than half a logit value above or below the mean logit value (between
Table 7. Direct and semi-direct test method measurement report (post-training)

| Oral test method | Observed average score | Fair average | Obs-Exp score (logit) | Bias Logit | SE | Infit MnSq. | Z | Outfit MnSq. | Z | Correlation Point biserial |
|------------------|------------------------|--------------|-----------------------|------------|----|-------------|---|--------------|---|--------------------------|
| Direct           | 21.56                  | 20.28        | 0.14                  | -0.08      | 0.05 | 1.1         | -0.28 | 1.1          | -0.31 | 0.71                     |
| Semi-direct      | 19.30                  | 20.64        | -0.49                 | 0.17       | 0.06 | 1.2         | 0.55  | 1.1          | 0.59  | 0.65                     |
| Mean             | 20.43                  | 20.46        | -0.17                 | 0.04       | 0.05 | 1.15        | 0.13  | 1.10         | 0.14  | 0.68                     |
| SD               | 1.59                   | 0.25         | 0.44                  | 0.17       | 0.00 | 0.07        | 0.58  | 0.00         | 0.63  | 0.04                     |

Fixed (all same) chi-square: 11.16, df = 1, p < 0.01
Separation index (Direct): 1.82 Reliability (Direct): 0.81
Separation index (Semi-direct): 1.93 Reliability (Semi-direct): 0.85
−0.46 and +0.54) would be considered as either too severe or too lenient (McNamara, 1996; Wright & Linacre, 1994). In this respect, both rater groups of either test method were identified at the acceptable range of severity/leniency.

Columns seven and nine (Infit and outfit mean square) display whether the ratings of either test method (Direct and Semi-direct) were at the acceptable fit index or not. The infit mean square for the direct test method measured 1.10 and for the semi-direct method 1.20. This finding demonstrates that the ratings of both test methods have been modified to a considerable extent and have got closer to the absolute consistency at the post-training phase. This finding proves the true effectiveness of the training program in affecting the ratings of both test methods making the ratings of direct and semi-direct tests more consistent. It would be better to indicate that the ratings of the direct test method tended to get closer to consistency at the post-training phase which provides enough evidence to claim that direct ratings benefited more as a result of the administration of the training program than semi-direct ones.

Also, Z-scores (columns eight and ten) display test method bias estimate after training. The outcome of the table obviously displays that both groups of raters, at the post-training phase and similar to before training, were placed at the acceptable range of biasedness. Nevertheless, it is noteworthy to indicate that the ratings of the direct method (Z_{Direct} = −0.28) had less interactional influence than those of the semi-direct method (Z_{Semi-direct} = 0.55). This, once again, reflects that the training program was more beneficial for the ratings of the direct test method than those of the semi-direct. After all, both test methods, similar to the pre-training phase, had interactional tendency to opposite directions.

Besides, in order to examine to what extent the two test methods were similar to each other in ranking the test takers, a correlational analysis, column eleven (point-biserial correlation), was run. The results demonstrated that the ratings of the direct test method had a correlation index of 0.71 and for the semi-direct test method 0.65. This outcome shows a drastic shift of correlation coefficient compared to the pre-training phase. The ratings of the direct test method, at the pre-training phase, had a much less correlational index (0.26) than those of the semi-direct method (0.38), whereas at the post-training phase, direct test method ratings turned out to take the lead thus having a higher degree of correlation (0.71) than the semi-direct test method (0.65). This outcome definitely implies that ratings in the direct test method benefited much more from feedback and the training program than those of the semi-direct method and that it was more useful for direct oral-test assessment than the other test method.

The separation index of 1.82 and 1.93 for the direct and the semi-direct test methods, respectively, demonstrated that the rater groups could be classified into nearly two groups of severity measures. The reliability index of 0.81 and 0.85 for the direct and the semi-direct test methods, which are relatively lower than that of the pre-training phase, show that the separation of raters of either test method into various levels of severity was less distinguishable at the post-training phase. This is due to the fact that at the post-training phase, since the raters got more consistency and had less degree of severity and leniency and biasedness, it was rather difficult to clearly separate the raters into various levels of severity measures. This itself adds more evidence to the usefulness of the application of the training program in bringing consistency and reducing biasedness and severity among the raters of both test methods. Besides, the reduced amount of reliability in this phase of the study (r = 0.81 for the direct and 0.85 for the semi-direct test method) indicates that the analysis could separate the raters into different levels of severity with less precision—due to the establishment of more consistency among the raters. The fixed chi-square value for all the 20 raters rating the test takers’ oral performance of each test method was measured at the post-training phase (X^2_{1, N=20} = 11.16, p < 0.01). The outcome suggested that the ratings of the raters of either test method, at the post-training phase, were still different with respect to the degree of severity/leniency indices.
5. Discussion

With regard to the first research question dealing with the difference between the direct and semi-direct oral language proficiency assessment in terms of scoring quality, two issues were investigated—linguistic and communicative features. For linguistic features, involving micro-features of word order, tenses, verb structure, pronouns, grammatical gender, singularity/plurality, prepositions, articles and lexical accuracy, differences in the use of pronouns, tenses, and verb structure were observed. Errors in pronouns were more frequent for the semi-direct test than the direct test, whereas the errors in tenses and verb structure were more frequent for the direct test than the semi-direct test. Such finding is parallel with that of Shohamy (1994) who found that test takers used more limited linguistic structures in the direct test. Similarly, regarding communicative strategies, differences in paraphrasing and self-correction were revealed, too. Self-correction and paraphrasing occurred more frequently for the semi-direct test than the direct test. The more frequent use of self-correction and errors in verb structure, tenses and pronouns reveal that test takers pay more attention to linguistic accuracy when taking the semi-direct oral tests than they do on the direct oral tests. This is perhaps due to the fact that students consider semi-direct tests as a test in which they should pay more attention to linguistic accuracy and monitoring their output. This is similar to what Kenyon and Tscharner (2000), Qian (2009) and Shohamy (1994) found in their studies indicating that SOPI elicits more frequent use of language functions than the OPI. However, in Qian’s (2009) study, the test takers’ performances were highly influenced by raters’ biases to extralinguistic features.

This finding could be attributed to Tarone’s (1983) theory of Interlanguage Continuum in which he considers differences between elicited, monitored language and the vernacular one. In this respect, semi-direct oral tests might have the features of elicited speech, whereas direct oral tests seem to match the features of vernacular language. This finding suggests that the direct oral tests may represent more the qualities of a real language. On the direct oral tests, the test takers are more concerned about running the communication and transmitting information, that is why they are more capable of communicating effectively. A deeper analysis of the type of self-corrections revealed that the errors were mostly related to plurality and singularity. This again verifies that, in direct oral tests, test takers are more concerned with what to say than how to produce the right kind of language.

Also, test takers tended to paraphrase their output more frequently in semi-direct oral tests than in direct oral tests. This could be due to the fact that test takers cannot use the exact words appearing in the questions. Thus, they resort to paraphrasing. Paraphrasing could be used as a tool to go around the questions but express the same content. This is not applicable for direct oral tests since the questions are not as specific as those of semi-direct oral tests that even a partial response would suffice. A further reason for the differences in paraphrasing might be that the test takers, in semi-direct oral tests, do not receive immediate and direct feedback from their interlocutors which can highly influence their language output. If this is true, then paraphrasing could be regarded as a tool which emerges as a result of lack of corrective signals that a test taker receives from his/her rater. Consequently, paraphrasing is used as a strategy to ascertain that the message is transmitted.

Finally, in addition to the above-mentioned findings, for making decision of which test to use, according to Kenyon and Tscharner (2000), considerations must be regarded to issues of validity which are mainly related to test utility, feasibility and fairness. Utility shows whether a test provides practical information that students require. Feasibility deals with the fact that whether a test is practical to administer in various contexts and to what extent the cost of the test is justifiable. Fairness demonstrates to what extent the codes of ethics have been followed in test administration and whether or not an oral test is based on the materials that the test takers have already been exposed to. However, in general, the findings indicated that both oral tests enjoyed high stability, reliability and validity. That is, as long as designed oral tests reflect the real-life language tasks and objectives that test takers are likely to encounter in community up to a high extent, they are valid enough to be used. This finding is rather in contrast with that of Stansfield (1991) who found lack of stability and low reliability and validity in the oral interview test.
Besides, the data analysis of the speech production of both test methods displayed that semi-direct-oral-tests produced speeches incorporated more lexically and structurally complex language than direct-oral-tests-produced ones. Direct-oral-test speeches included rather simpler language complexity in both the application of lexicons and language grammar. This is parallel with what (Ahmadi & Sadeghi, 2016) found in a study comparing the two test methods resulting in less difficult lexical density produced by test takers when taking the direct oral test.

The second research question dealt with raters’ measures of severity, bias and consistency in scoring to the direct and semi-direct test methods before and after the training program. With respect to raters’ change of behavior in rating after the training program, it must be mentioned that raters showed to have been responsive to training and feedback and could incorporate it in their subsequent ratings in both testing methods. A majority of the raters who displayed significant bias before training, no longer displayed extreme bias with respect to any of the test methods after training. Regarding the FACETS tables, through considering the chi-square value, separation and the reliability index, the findings obtained from the data analysis indicated that both oral tests could properly discriminate test takers into various levels of language proficiency and had a high reliability as well. For some individual raters who did not reduce biases or even gained more bias after training, this phenomenon had better be viewed optimistically telling us that rating is a matter of personal characteristics relevant to raters’ individual differences. In other words, some raters must only be trained for rating face-to-face oral interview tests and some for taped-based ones. This finding is rather consistent with that of Khabbazbashi (2017) who observed some abnormal rating behavior from some raters scoring different oral test format. One of the very significant findings is that the outcomes help test administrators and decision makers to make up a profile containing various rater characteristics on the basis of which to decide over which raters to select, train and employ for what sort of test to achieve the maximum reliability and validity in assessment.

With respect to the effectiveness of the training program in establishing consistency among raters in scoring similarly those test takers taking both methods of the oral test, the results of data analysis were quite promising. The outcome displayed the usefulness of the training program in establishing consistency with respect to assigning rather similar scores to a test taker tasking both formats of the oral test. This result is consistent with the finding by Bijani (2018), Jeong and Hashizume (2011) and Qian (2009) who found higher consistency in the raters’ scoring after training. The remaining differences regarding the rating of the test takers on the two methods of the oral test might be due to the actual varieties of test takers’ performances on the two different test methods rather but not due to inconsistency and unsystematicity in scoring by the raters. This finding is parallel with that of Kenyon and Tschirner (2000) who found a change in raters’ scoring behavior in assigning closer scores to test takers on various oral test formats.

The analysis of data demonstrated no significant difference between the ratings of direct and semi-direct oral proficiency assessment by the raters after training. In other words, a close relationship between the ratings of raters was found in their ratings of test takers’ rating both test methods. It is noteworthy to indicate that such correlation was only found after the training program. Consequently, semi-direct oral tests can be regarded as a reliable substitute for direct oral tests. This finding is closely relevant to the one achieved by Clark and Li (1986), Jeong and Hashizume (2011), Khabbazbashi (2017), Shohamy (1994); and Stansfield and Kenyon (1992) who observed high correlation coefficients between the two test methods. Such finding must be generalized with caution since, according to Jeong and Hashizume (2011), the obtained high correlation between the two test methods could be due to the fact that both tests do not let test takers represent the interactive oral skills completely and neither is the manifestation of real-life conversation act in natural contexts. One important factor that test takers pointed out as a source of test difficulty in the semi-direct oral tests was preparation time. Therefore, it can be concluded that the provision of more preparation time will reduce stress on the side of test takers and increase test taking facility for them. This is the finding that
Jeong and Hashizume (2011) also found in a similar study on the comparison of direct and semi-direct oral assessment tests.

6. Conclusions
The outcome of the study showed that direct and semi-direct oral tests vary in terms of linguistic features including pronouns, tenses and verb structure. Similarly, some differences were observed for communicative features in paraphrasing and self correction. Besides, semi-direct oral tests speeches incorporated more lexically and structurally complex language. The outcome of the study showed that training can result in higher levels of interrater consistency and reduced levels of severity/leniency, biasedness and inconsistency. A majority of the raters who displayed significant bias before training, no longer displayed extreme bias with respect to any of the test methods after training. However, it cannot turn rater into duplicates of one another, but to make them more self-consistent (intrarater consistent). Rater effects can threaten the validity of decisions made for ratings. The findings of this study on the basis of statistical MFRM outcomes demonstrated the usefulness of the use of this analytical approach in detecting rater effects and demonstrating the consistency and variability in rater behavior aiming to evaluate the quality of rating. MFRM can provide raters with rapid feedback on their instability and thus to apply adjustments on raters’ behaviors based on that feedback. This study also represented that a valid test of oral language proficiency should consist of both direct and semi-direct tests in order to provide sufficient assessment evidence. The outcomes could highly help decision makers in selecting the right kind of test on the basis of a given context and purpose. Knowing exactly what each test type measures can provide testers with valuable information on test selection.

Both the direct and the semi-direct tests were shown to discriminate test takers into various levels of language proficiency and had a high reliability as well. However, since still some individuals did not reduce bias after training, this indicates that training had better be viewed optimistically telling us that rating is a matter of personal characteristics relevant to raters’ individual differences. The outcome displayed the usefulness of the training program in establishing consistency with respect to assigning rather similar scores to a test taker taking both formats of the oral test. The outcomes also demonstrated no significant difference between the two test formats in assessing test takers’ oral proficiency by raters specifically after training.

This study has a number of practical implications in language assessment. Since the findings showed that both test formats are highly correlated, decision makers can get benefit from semi-direct tests instead of direct tests in assessment for the sake of consuming less energy and cost of performance. However, since semi-direct tests are more difficult, care must be taken in using this test for low-level test takers. This study specified the significant importance of training in bringing raters into acceptable range of consistency and reduced levels of severity and bias no matter what the test format is; thus, decision makers had better use budget establishing training programs rather than getting concerned about rater expertise. This study investigated the difference between direct and semi-direct oral assessment tests focusing on linguistics and communicative measures. Further research could be done investigating the use of other language features in test takers’ oral performance assessment. Besides, this study focused on direct and semi-direct oral tests, future studies can incorporate the use of indirect oral tests as well.

Funding
The author received no direct funding for this research.

Author details
Houman Bijani
E-mail: houman.bijani@gmail.com
Department of English Language Teaching, Zanjan Branch, Islamic Azad University, Zanjan, Iran.

Citation information
Cite this article as: Evaluating the effectiveness of the training program on direct and semi-direct oral proficiency assessment: A case of multifaceted Rasch analysis, Houman Bijani,Cogent Education (2019), 6: 1670592.

References
Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners’ oral performance: A comparison of
monologue, interview, and group oral test. Language Assessment Quarterly, 13(4), 341–358. doi:10.1080/15434303.2016.1326797

Bijani, H. (2010). Raters’ perception and expertise in evaluating second language compositions. The Journal of Applied Linguistics, 32(2), 69–89.

Bijani, H. (2018). Effectiveness of a face-to-face training program on oral performance assessment: The analysis of tasks using the multifaceted Rasch analysis. Journal of Modern Research in English Language Studies, 5(4), 27–53.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. Language Testing, 20(1), 89–110. doi:10.1191/0265532203ht2450a

Clark, J. L. D. (1978). Direct testing of speaking proficiency: Theory and application. Princeton, NJ: Educational Testing Service.

Clark, J. L. D., & Li, Y. (1986). Development, validation, and dissemination of a proficiency based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages. Washington, DC: Center for Applied Linguistics.

Cohen, L., Manion, L., & Morrison, K. (2007). Research methods in education. London: Routledge.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. Language Testing, 33(1), 117–135. doi:10.1177/0265532215582882

Eckes, T. (2015). Introduction to many-facet Rasch measurement. Frankfurt: Peter Lang Edition.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? Language Testing, 19(4), 347–368. doi:10.11177/026553221503640a

ETS. (2001). ETS oral proficiency testing manual. Princeton, NJ: Educational Testing Service.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language. Applied Linguistics, 21(3), 354–375. doi:10.1093/applin/21.3.354

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher-and lower-scoring students. Language Testing, 27(4), 585–602. doi:10.1177/0265532210364049

Innami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. Language Testing, 33(3), 341–366. doi:10.1177/0265532215587390

Jeong, H., & Hashizume, H. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. Language Learning, 61(3), 675–699. doi:10.1111/j.1467-9922.2011.06835.x

Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. The Modern Language Journals, 84(1), 84–101.

Khabbazi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. Language Testing, 34(1), 23–48. doi:10.1177/0265532215595666

Kim, H. J. (2011). Investigating raters’ development of rating ability on a second language speaking assessment (Unpublished PhD thesis). University of Columbia.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. Language Assessment Quarterly, 12(3), 239–261. doi:10.1080/15434303.2015.1049353

Kuiken, F., & Verhoeven, L. (2014). Raters’ decisions, rating procedures and rating scales. Language Testing, 31(3), 279–284. doi:10.1177/0265532214526179

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL IBT speaking tasks: Insights from natural language processing. Language Testing, 33(3), 319–340. doi:10.1177/0265532215587391

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. Language Testing, 31(2), 177–204. doi:10.1177/0265532213498237

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. Language Testing, 12(1), 54–71. doi:10.1177/026553229501200104

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. Language Testing, 15(2), 158–180. doi:10.1177/026553229801500202

Maria-Duque, A., & Bonk, W. J. (2009). Assessing paired oral Raters’ orientation to interaction. Language Testing, 26(3), 423–443. doi:10.1177/0265532209104669

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. Melbourne Papers in Language Testing, 11(1), 29–51.

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

Mirzoei, A., Hashemian, M., & Farsani, M. (2016). Lexis-based instruction and IELTS candidates’ development of L2 speaking ability: Use of formulaity in monologic versus dialogic task. Journal of Teaching Language Skills, 35(2), 69–98.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. Language Assessment Quarterly, 6(2), 113–125. doi:10.1080/15434300902800059

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. Language Testing, 11(2), 99–123. doi:10.1177/026553229401100202

Stahl, J., & Lunz, M. (1992). Judge performance report. Paper presented at AERA, San Francisco, CA.

Stansfield, C. W., & Kenyon, D. M. (1992). On the comparability of the oral proficiency interview and the simulated oral proficiency interview. System, 20(3), 347–364. doi:10.1016/0346-251X(92)90068-5

Tarone, E. (1983). On the variability of interlanguage systems. Applied Linguistics, 4(2), 142–164. doi:10.1093/applin/4.2.142

Winke, P., Gass, S., & Myford, C. (2012). Raters’ L2 background as a potential source of bias in rating oral performance. Language Testing, 30(2), 231–252. doi:10.1177/0265532212456968

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 369–386.
