Higher Education Studies; Vol. 3, No. 2; 2013
ISSN 1925-4741 E-ISSN 1925-475X
Published by Canadian Center of Science and Education

# Teacher Interpretation of Test Scores and Feedback to Students in EFL Classrooms: A Comparison of Two Rating Methods

Mu-hsuan Chou[1]

[1] Department of Foreign Language Instruction, Wenzao Ursuline College of Languages, Kaohsiung, Taiwan

Correspondence: Mu-hsuan Chou, Department of Foreign Language Instruction, Wenzao Ursuline College of Languages, 900 Mintsu 1st Road Kaohsiung 807, Taiwan. Tel: 886-7-342-6031 ext. 5221. E-mail: mhchou@gmail.com

Received: February 13, 2013 Accepted: February 28, 2013 Online Published: March 25, 2013

doi:10.5539/hes.v3n2p86 URL: http://dx.doi.org/10.5539/hes.v3n2p86

**Abstract**

Rating scales have been used as a major assessment instrument to measure language performance on oral tasks. The present research concerned, under the same rating criteria, how far teachers interpreted students' speaking scores by using two different types of rating method, and how far students could benefit from the feedback of the description of the two rating methods. Fifteen English teachers and 300 college students in Taiwan participated in the study. Under the same rating criteria, the two types of rating method, one with the level descriptors and the other with a checklist, were used by the teachers to assess an English speaking test. It was discovered that the rating method had a noticeable impact on how teachers judged students' performance and interpreted their scores. The majority of the students considered feedback from the verbally more detailed method more helpful for improving language ability.

**Keywords:** rating scale, rating checklist, role-play, speaking, reflection

## 1. Introduction

### *1.1 Overview*

Performance rating scales have long been used to provide information regarding test candidates' performance abilities in speaking or writing. The aim of using rating scales to interpret candidates' language ability is to diminish the drawback of low reliability in holistic scoring, by incorporating a number of relevant linguistic aspects to help reduce the problem of biased or unfair judgment by scorers (Hughes, 2003). According to my teaching experience in Taiwan, however, many university teachers have regarded using rating scales with detailed descriptors in speaking tests as a waste of time. They often turned to holistic scoring, where they simply gave students single scores based on their overall speaking performance, but later they discovered it was hard to interpret students' scores after marking and to know how far the students had achieved the teaching and learning objectives. This resulted in problems with verbalizing students' performance based on their scores and offering informative feedback to students, the teachers themselves, and other relevant stakeholders. Rating methods have been developed and revised to use under different assessment circumstances. The present study examines Taiwanese university teachers' perceptions of using two different rating methods to interpret their students' oral scores in role-play and simulation tasks in an English course. The aim was to discover firstly, whether using the same test criteria, the formats of the two rating methods influenced the teachers' marking, and if so, how much, and secondly, which type of rating they thought would be useful for them to reflect the teaching. In addition to teachers' usage of rating methods, the students were asked which rating method helped them better reflect on their speaking performance after receiving feedback from both. It is hoped that the results of the present study can provide teachers and educational researchers with guidance on whether to employ either rating method in the context of classroom assessment and feedback to students.

### *1.2 Rating Scale for Assessing Language Performance*

Rating scales for performance assessment have been designed in various forms depending on whether the research interest relates to the student's underlying ability, or the purpose of the test users (Hamp-Lyons, 1991; Alderson, 1991). The most common form of rating scale is called a "performance data-driven scale", where a separate score for each of a number of language aspects of a task, say, grammar, vocabulary, or fluency, is given

to a student and the score from each linguistic component is added up as a total score to represent the student's performance on the task. When constructing the scale, samples of learner performance undertaking test tasks in specific language contexts need to be collected, following the transcription, and identification of key performance features by discourse analysis (Fulcher, 2003; Fulcher, Davidson, & Kemp, 2011). Descriptors for a performance data-driven approach are thus generated from the key discourse features of observed language use, such as grammar or vocabulary. The major disadvantage of the approach is that it is complicated and difficult to use in real-time ratings (Fulcher, 2003; Hughes, 2003), when raters need to observe the performance, read numerous detailed descriptors, and mark everything in a limited period of time. Assessing interactive ability with small groups in a classroom is always exceedingly hard, due to the difficulty of measuring communicative performance at different levels and on different tasks at the same time. However, as Nunn (2000) has pointed out, the task is made even harder if a performance data-driven scale needs to be used, and rather than operate the multiple detailed descriptors involved, many classroom teachers simply avoid the problem (and do not measure interactivity at all).

In an attempt to overcome problems with reliability and validity of rating scales, Upshur and Turner (1995) designed a different type of performance rating scale – the empirically derived, binary-choice, boundary definition (EBB) scale, which did not contain detailed description like a performance data-driven scale, but only binary choices. An EBB scale requires raters to make a series of hierarchical binary choices about the features of student performance that define boundaries between score levels. Taking Upshur and Turner's example of responses to a writing test, the rater begins by asking the first level question: 'Did the student write ten lines or more?' If the answer is No, the rater asks a second level question: 'Did the student write on the assigned topic?' If the answer to this question is also No, the writing sample is scored1; if the answer is Yes, the score is 2. If, however, the answer to the first question (Ten lines or more?) is Yes, the rater asks the other second level question: 'Was everything clearly understood?' A No answer to this question would result in a score of 3; a Yes answer would yield a score of 4. Fulcher *et al.* (2011), combining the advantages of the descriptive richness of the performance data-driven approach and the simplicity of decision making of the EBB scale, devised a 'performance decision tree' (PDT) for service encounters for L1 examinees. A PDT comprises a series of questions describing performance in service encounters and two options. For example, the rater reads questions such as 'Is there clear identification of purpose in the discourse?' If it is 'explicit', the candidate receives 2 points. If it is 'implicit' (deemed less desirable in this context), he or she receives 1 point.). The problem of a time-consuming rating is largely eliminated by using a PDT. One problem Fulcher *et al.* (2011) found with PDT scales was that each was developed for a particular task and a single population. Furthermore, the feature of dichotomous choices in EBB and PDT scales casts doubt on the accuracy of interpretation of learner performance, because the score interpretation is simply based on whether or not learners achieved fixed criteria, rather than on the degrees of learner skill or knowledge. While rating research has focused on the theoretical usability of scale types, this does not answer the question of how useful they prove to teachers in classroom contexts trying to establish what their students can and cannot do well, and create helpful formative feedback on what they should work on in the immediate future.

### *1.3 Interpretation of Scores and Feedback to Students*

The accuracy and appropriateness of the interpretation of test scores requires raters' fair and unbiased judgment on learners' language ability. Purpura (2009), summarizing from Bachman (1990), Frederiksen and Collins (1989), and Messick (1989), notes that the decision teachers make on the basis of test score interpretations can have a significant impact on individuals, groups and institutions. However, Taylor (2009) argues that decisions teachers made while scoring may be influenced by socio-cultural sensitivities and subject to prejudices, and this can threaten the validity, reliability, and the social impact of the tests. Oscarson and Apelgren (2011), a study on Swedish secondary school language teachers' conceptions of student assessment and their grading process, discovered that teachers frequently experienced difficulties in the grading of their students. Specifically, they found that 'there is a conceptual conflict between the assessment, testing and grading being carried out by the teachers in their classroom' where there is no criterial guideline for assessment (ibid, p. 14). This resulted in discrepant interpretations of learner performance among the teachers. Oscarson and Apelgren's study supported Orr's earlier (2002) finding that different raters paid attention to different aspects of assessment criteria, in the case of the Cambridge First Certificate in English speaking test. Even though raters gave the same score to a participant, they still perceived the aspects of student performance differently, and moreover, noticed many aspects of the performance which were irrelevant to the assessment criteria. In short, the lack of consistent ratings based on agreed criteria resulted in misinterpretations of language ability and unfairness to learners, and generally threatened the overall validity of the test.

Recent research has emphasized the importance of including student participation in assessment (Shohamy, 2001; Lynch, 2003), suggesting that students can benefit from feedback on their performance, by allowing them to monitor and improve their language ability. Oscarson and Apelgren (2011) found, moreover, that when feedback was given, learning was genuinely enhanced. Nunn (2000, p. 171) states that the most important thing is that a conclusion drawn from a rating scale should be able to help candidates understand their strengths and weaknesses. However, the methods used to score oral performance may influence how teachers (as raters) perceive and decide on students' speaking ability. Thus, the present study investigates teachers' perceptions of how far they could judge students' performance and interpret their students' test scores in the speaking test when they used two different rating methods but the same scoring criteria. It also explores how far the feedback from the two rating methods helps students reflect on and improve their oral performance. The study reported here is a case study which addressed these points via three research questions:

*(1) How far did college teachers think that the two rating methods (a performance data-driven scale and a rating checklist) helped them interpret students' scores in a university speaking test?*

*(2) How far did the formats of the two rating methods influence the scoring process?*

*(3) How far did students think the feedback from the two rating methods helped reflect on their oral performance?*

## 2. Method

The research was undertaken with fifteen course instructors and 300 first-year undergraduate students taking a compulsory general English course at a university in southern Taiwan. The oral test took the form of a role-play and simulation task, where paired participants performed tasks in a simulated context relevant to the content-based topics taught in class. This kind of activity had been the major form of oral test on this course and was practiced regularly (approximately one hour per week) by the participants in class. Each participant was paired up randomly by drawing lots. Each pair was given ten minutes to prepare a two-minute conversation based on a task and topical context from a predetermined list and they were asked to perform the required grammar point (e.g., present perfect) and conversation strategy that had been taught in class in the conversation.

The performance of the students was either video- or audio-recorded with the consent of the course instructors and the students themselves. Each teacher randomly chose ten pairs in his or her own class to rate. The course instructors first used a performance data-driven rating scale (see Table 1). The rating scale was designed, piloted, and modified by the researcher, who collected and analyzed data of oral performance from tasks and students that were similar to those in the present study. The scale was divided into five categories – 'topic relevance', 'pronunciation and fluency', 'grammar', 'communication strategy', and 'pragmatic competence' – which were considered highly relevant to the characteristics of language use in oral task performance (Cohen & Dörnyei, 2002; Bachman & Palmer, 1996). Messick (1989; 1996) suggests that testers first need to establish the criteria for the assessment that can provide adequate and appropriate interpretations for the intended test score use and decision-making, and then design assessment items based on the criteria. In the present study, the rating scale designed for the role-play and simulation task was based on the course objective that the participants needed to successfully apply communication strategies and specific grammar to various simulated conversational contexts.

Table 1. Rating scale for the role-play and simulation speaking test

| | Range | Description of Criteria |
|---|---|---|
| **Topic Relevance** | 4 | The discourse content was related to the topic. The student managed the topic without problems. Ideas and opinions were logically presented. |
| | 2-3 | The discourse was partially related to the topic. Changes of topic or jumping suddenly between ideas in the middle of interaction. Some ideas and opinions were not logically connected. |
| | 0-1 | The discourse was not related to the topic. Ideas and opinions were disconnected and there was jumping due to frequent changes of topic. |
| **Pronunciation & Fluency** | 4 | The pronunciation was correct. The student's spoken discourse was fluent and natural. |
| | 2-3 | The pronunciation was sometimes incorrect, or the student was hesitant or uncertain while expressing opinions or ideas. |
| | 0-1 | The pronunciation was frequently incorrect, and the student had difficulty in expressing his or her message completely. |
| **Grammar** | 4 | The required grammar items were used correctly and appropriately. |
| | 2-3 | The required grammar items were used in incorrect forms, in inappropriate, or seldom used. In general, the student had some difficulty in managing grammar in a timed short conversation. |
| | 0-1 | The required grammar items were not used. In general, the student could not use grammar correctly and appropriately in conversation. |
| **Communication Strategy** | 4 | The required strategies were used appropriately and other relevant communication strategies were appropriately adopted to manage the conversation. |
| | 2-3 | The required strategies were partially used or used inappropriately. The student had some difficulties in applying other relevant communication strategies appropriately. |
| | 0-1 | The required strategies were not used. Other relevant communication strategies were not obviously applied to facilitate interaction. |
| **Pragmatic Competence** | 4 | The utterances were appropriate to the communicative goals of the specific language use setting. The student responded correctly and appropriately to the interlocutor's utterances. |
| | 2-3 | The utterances were at times inappropriate to the specific language use setting. The student had some difficulties in responding correctly and appropriately. |
| | 0-1 | The utterances were not related or appropriate to the communicative goals in the specific language use setting. The student failed frequently to respond correctly and appropriately. |

One week later, the teachers rated the same performance again, but this time using a rating checklist (see Table 2), which was devised by the researcher. The concept of a rating checklist was adapted from EBB scales (Upshur & Turner, 1995) and the PDT (Fulcher *et al.*, 2011). The elements in the checklist for this study were almost the same as the criteria in the rating scale (see Table 1), but with trichotomous instead of binary choices. The reason for using trichotomous choices was because, in the process of piloting, it was discovered that it was difficult to decide on some participants' language performance; sometimes teachers needed to indicate where certain linguistic skills were partially shown or not shown. To avoid being influenced by the first ratings, the fifteen teachers agreed that they would not refer back to their first rating while doing their second rating.

Table 2. Rating checklist for the role-play and simulation speaking test

| | | Yes | Partially | No |
|---|---|---|---|---|
| **Topic** | 1. Are required topical elements covered? | ☐2 | ☐1 | ☐0 |
| | 2. Are ideas/opinions logically presented? | ☐2 | ☐1 | ☐0 |
| **Fluency** | 3. Is the language fluent? | ☐2 | ☐1 | ☐0 |
| **Pronunciation** | 4. Is the pronunciation correct? | ☐2 | ☐1 | ☐0 |
| **Grammar** | 5. Is the required grammar used correctly? | ☐2 | ☐1 | ☐0 |
| | 6. In general, is grammar used correctly? | ☐2 | ☐1 | ☐0 |
| **Pragmatic Competence** | 7. Are the utterances appropriate to the features of the specific language use setting? | ☐2 | ☐1 | ☐0 |
| | 8. Is the participant responding correctly & appropriately to the interlocutor's utterances? | ☐2 | ☐1 | ☐0 |
| **Communication Strategy** | 9. Are the required communication strategies used appropriately? | ☐2 | ☐1 | ☐0 |
| | 10. Is the participant using other relevant communication strategies to manage conversation? | ☐2 | ☐1 | ☐0 |

After marking the students' performance, each teacher gave feedback to their students according to the descriptions of the two rating methods. The participants then filled out a questionnaire to explore their opinions towards the feedback and reflections on their oral performance (Appendix 1). Semi-structured interviews were conducted with the teachers after completing the rating. One-third of the participants who received the feedback and answered the questionnaire (i.e., 64 students) were also interviewed.

## 3. Results

### *3.1 Interpretation of Student Performance on the Two Rating Methods*

Although the criteria in the performance data-driven rating scale and the rating checklist were the same, the fifteen teachers reported differing interpretations of student performance on the role-play/simulation tasks. Eleven out of fifteen (73.3%) considered that the rating scale, with its more detailed descriptors, offered more comprehensive information of students' speaking ability and skills than the rating checklist. Seven said that the scale was effective in terms of helping them decide whether their students were able to use what they had learned in class and whether they used it correctly or not. Four teachers reported that they tended to focus more on what had *not* been done by the students in the task than pay attention to what *had* been done, so the rating scale with detailed descriptors helped them focus on positive aspects of student performance. The rest of the teachers said they found it easier to use the 'Topic', 'Communication Strategies' and 'Pragmatic Competence' sections, while there was limited room for personalized feedback on students' 'Pronunciation' and 'Grammar'. They indicated that it was not possible for them to tell each student which part of, say, grammar, was incorrect, unless they took notes while scoring, which was not feasible due to the time constraints. In other words, the rating scale could provide teachers with a general description of the linguistic abilities in the tasks. But, individual linguistic deficiencies could not be easily interpreted from the scale. Inevitably, all but six teachers noted that the scoring was rushed for pair work; four specifically indicated that the scale might have worked more efficiently for individual oral tests.

Unlike the rating scale, the rating checklist provided the teachers with a more efficient approach to scoring. All but five teachers (66.7%) stated that it was relatively quick to use the checklists to score. However, when it came to the interpretation, only four teachers (26.7%) regarded the descriptions of the checklist as effective, straightforward and easy to understand. The other seven reported that the checklist was efficient and user friendly, but not effective, because the descriptions provided a less comprehensive and objective view of students' speaking ability that did not help them better interpret the participants' scores. From the standpoint of the teachers, the majority said that they preferred the rating checklist when they did not need to provide detailed feedback to students or test users. However, when the teachers wanted to know more about their students' performance, the majority thought the detailed level descriptors satisfied their needs. Specifically, three teachers mentioned that the score range in the checklists was too small (0 to 2 for each question), which meant even where the two students clearly differed in their performances; their scores would not actually vary markedly. However, the statistics showed the opposite to the teachers' conjecture. The standard deviations of the scores

from the rating scale were smaller (for all but T14) than those from the rating checklist (see Table 3). This contradicted the teachers' feedback that there was less variation on the scoring of the checklist. Although the scores of the two rating methods in each category were exactly the same, it appeared that the display of rating criteria had some influence on the perception, interpretation and judgment of teachers' scoring, even though they might not have noticed it. Because the statements of the level descriptors were changed to a list of direct questions with definite 'answers', i.e., 'Yes', 'Partially', or 'No,' the itemized format allowed the teachers to select the answer which best described the performance. Nonetheless, the level descriptors of the performance data-driven rating scale provided teachers with ranges of scores under each criterion, which left more freedom of marking. The results suggest that the more descriptions of speaking performance and ranges of scores were provided, the less variation of scores between the participants there was. In addition, four teachers noted that not just the rating scale but also the rating checklist had a common problem of not giving specific feedback on grammatical and pronunciation mistakes to individual students.

Table 3. Means and standard deviations of student scores

| Teacher | Rating Scale (Mean) | Rating Scale (SD) | Rating Checklist (Mean) | Rating Checklist (SD) | Correlation coefficient (*r*) |
|---|---|---|---|---|---|
| **T1** | 12.9 | 1.89 | 14.8 | 2.78 | .78 (p<.000) |
| **T2** | 13.1 | 1.96 | 12.8 | 2.18 | .83 (p<.000) |
| **T3** | 15.3 | 1.85 | 14.9 | 1.91 | .82 (p<.000) |
| **T4** | 14.7 | 2.34 | 13.6 | 2.52 | .80 (p<.000) |
| **T5** | 13.3 | 2.70 | 13.2 | 2.79 | .94 (p<.000) |
| **T6** | 13.6 | 2.42 | 12.9 | 2.48 | .79 (p<.000) |
| **T7** | 13.7 | 1.94 | 12.9 | 2.26 | .81 (p<.000) |
| **T8** | 14.8 | 1.76 | 14.2 | 2.02 | .90 (p<.000) |
| **T9** | 14.5 | 1.87 | 13.2 | 2.04 | .84 (p<.000) |
| **T10** | 12.8 | 1.26 | 12.6 | 2.17 | .83 (p<.000) |
| **T11** | 12.6 | 2.01 | 12.3 | 2.21 | .91 (p<.000) |
| **T12** | 14.3 | 2.17 | 13.9 | 2.53 | .85 (p<.000) |
| **T13** | 14.7 | 1.99 | 13.9 | 2.39 | .82 (p<.000) |
| **T14** | 13.2 | 2.47 | 14.8 | 1.96 | .79 (p<.000) |
| **T15** | 14.1 | 1.89 | 14.9 | 2.12 | .83 (p<.000) |

### *3.2 Feedback from Rating Scales to Improve Speaking Ability*

The participants' opinions about the feedback based on the descriptions of the two rating methods concerned whether or not the descriptions of their oral performance helped them (1) understand their language problems, and (2) reflect on and improve their speaking skills. Of the 300 participants in the present study, 192 agreed to fill out the questionnaire, and one third of them (i.e., 64 students) agreed to be interviewed.

In general, the majority of the participants agreed that they understood their speaking problems better in the test from the feedback of the rating scale than the checklist (see Table 4). More than three quarters of the interviewees said that the descriptions on the scale being more detailed, appeared to provide them with more information regarding their speaking performance, and they felt it was useful to know which aspect of language performance needed to be improved. As Interviewees 12 and 45 said:

*After finishing the oral conversations, it was very useful to know how well or poorly I had spoken. I think the first form (rating scale) was more detailed in terms of describing my oral ability. The information from the second form (checklist) was understandable, but not as informative as the first one.* (Interviewee 12; trans.)

*I preferred the first form, because it was apparent to see how far….I mean the degree…I had done in the conversation. When I read the description at other level, I knew if I did better or worse in this oral activity.* (Interviewee 45; trans.)

In the aspects of 'Conversation Strategy' and 'Pragmatic Competence,' both two rating methods were highly comprehensible for the majority of the participants to understand their task performance. Particularly, twenty nine interviewees said that managing interaction with partners in the conversational tasks was not considered a problem for them because they could pick up the topic and try to continue the conversation without difficulty when the communication broke down. For example:

*I think it was sometimes hard to use the required conversation effectively in the conversation because it was only two minutes, however, I would try to use other strategies to pick up the conversation quickly. I think it pretty obvious to read whether I had use the appropriate ways to communicate from both forms.* (Interviewee 32; trans.)

Table 4. Student reactions to feedback via the two types of rating method

| | Topic Relevance (%) | | Pronunciation & Fluency (%) | | Grammar (%) | | Communication Strategy (%) | | Pragmatic Competence (%) | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Type** | **RS** | **RC** | **RS** | **RC** | **RS** | **RC** | **RS** | **RC** | **RS** | **RC** |
| **Comprehension** | 91.7 | 72.9 | 87.5 | 68.8 | 77.1 | 60.4 | 85.4 | 70.8 | 93.8 | 81.3 |
| **Reflection** | 87.5 | 56.3 | 65.9 | 51.0 | 48.2 | 33.6 | 75.0 | 60.3 | 83.3 | 72.1 |

Note: 'RS' means 'Rating Scale', and 'RC' 'Rating Checklist'

However, in all five categories a lower percentage of the participants agreed that the feedback helped them reflect on and improve speaking ability, due to its limited verbal descriptions (see Table 4). However, one criterion on which both rating methods failed to give useful feedback was grammar, where fewer than half of the participants (48.2%) reported they benefited from the feedback, particularly in the case of the checklist where only one-third considered the feedback useful for reflection. There are two reasons why the participants did not think that both rating methods helped them reflect on their grammar performance. First, twenty-three interviewees said that it was challenging to fit the required grammar into, as well as use correct grammar in, the role-play tasks, because in the real-life situation, they did not pay attention to the grammar they used in spoken English. Two students said:

*To be honest, I seldom paid attention to my grammar in the conversation. I knew grammar was important but I think the purpose of speaking was for communication. I didn't think I could correct my grammar mistake from both rating forms because I didn't pay attention to what mistakes I had made* (Interviewee 51; trans.).

*When I talked to my partners in the oral activities, I let the conversation flow. So I think it was sometimes hard for me to use the required grammar in the specific tasks. Even though I knew I had to use it in the activities or tests, I found it difficult* (Interviewee 27; trans.).

Also, the other seven made the more general point that in order to speak fluently, grammatical accuracy was sometimes sacrificed. Ellis (2002) notes that frequent exposure to patterns or morphosyntax can facilitate learners' understanding of grammar. However, using and fitting required grammar points into different specific contexts in oral tasks could be challenging for foreign language learners, owing to the disparate contexts of the simulated tasks. This was also in line with the situation that the teachers were unable to point out specific grammatical mistakes when marking. In addition to grammar, fluency was the other difficult problem for the participants to reflect on and improve. For example:

*When I spoke English, I couldn't concentrate on both grammar and fluency. So when I need to finish the conversation fluently in time, I couldn't pay attention to my grammar. I hope that the teacher could tell me what mistakes I had made, because I had no idea of the mistakes I had made from the rating form, let alone correct them.* (Interviewee 14; trans.)

Sixteen interviewees said that they needed more opportunities to create an English environment for practicing speaking English after class. The other eight reported that nervousness was the main reason why they found it hard to improve in the speaking test and why they thought the feedback from both rating methods did not actually help them reflect; this corresponded to Bachman and Palmer's (1996) statement that personal characteristics have considerable influence on language performance.

## 4. Discussion

From the teachers' standpoints, most regarded the rating scale as a more comprehensive and informative means of interpreting students' oral scores from the simulation and role-play task than the rating checklist. As such, it appears to have encouraged them to focus more on positive aspects of their students' oral performance.

The rating scale also led to better understanding students own oral performance and reflecting on what they needed to improve. Although both rating methods were based on the test construct that the teachers (on this particular course) wished to measure, the rating scale involved more comprehensive descriptions of the underlying construct which made it easier for the teachers to relate test scores to the participants' speaking ability. In contrast, the rating checklist provided most teachers with an efficient rating process that largely reduced the burden of reading long descriptors in real-time rating. However, the scoring with the checklist was more subjective than it with the rating scale, due to its trichotomous format. Although the criteria were the same in both cases, the format affected the perception of rating for most teachers, who thought it harder to distinguish between students when using the checklist. However, contrary to their belief, the standard deviation of scores showed that the variation in scores from the rating scale was less than that from the checklist. Although the rating scale was considered a more detailed approach to assessing students' speaking performance and giving them feedback, the difference in oral performance between paired students was less marked.

Even though feedback based on the two rating methods was given to the participants, who found the rating scale more helpful for self-reflection, there were two linguistic elements – grammar and fluency – which they found difficult to improve. On the other hand, adopting appropriate communication strategies for the conversations was not reported as being a problem. This is in line with studies on communication strategy instruction and oral testing (involving either individual performance or learner-learner interaction), which have repeatedly shown that strategy instruction can have a significant positive effect on the development of speaking skills and the improvement of strategy use (Bejarano, Levine, Olshtain, & Steiner, 1997; Dörnyei, 1995). In addition, the participants' presentation of and reflection on pragmatic competence in terms of interacting with their partner was generally managed well. Davies (2009), in his case study on the influence of interlocutor proficiency in a paired oral assessment, discovered that the proficiency level of an examinee's partner in a paired test had little influence on scores, even though the pairing type appeared to affect language quantity or interaction characteristics in some conditions. Thus, although Nakatani and Goh (2007) argued that using other students as a communicative partner in a test may introduce intervening variables that can affect the test result, in the present case, improving the interaction between partners was not reported a problem for the participants. The variation of topical contexts of the task and the linguistic elements had a greater influence on determining whether or not the students could successfully adopt and integrate the communication strategies into the topic contexts.

## 5. Conclusion

The conventional performance data-driven rating scale is frequently considered an appropriate assessment instrument for teachers to interpret oral test scores, allowing practical feedback to the participants in paired (or potentially individual) speaking tasks. Muñoz and Álvarez (2010) stress the point that positive washback from a test may also be fostered by informing students explicitly about the assessment procedures and scoring scales used. The present study sheds light on how far Taiwanese EFL learners thought feedback on their oral test results which involved showing them two types of assessment actually served to help them reflect on their performance. It was discovered that, under the same rating criteria, the formats affected score variation.

Although the rating checklist was efficient, practical, easy to use, and appeared to be better at distinguishing the difference of paired performance, the majority of teachers and the participants still thought that detailed descriptions of the rating scales were generally more informative in terms of explaining language performance to students or test users.

Rating scales have limitations regarding the interpretation of scores; as Luoma (2004, p. 60) notes, 'scales are difficult to write, both because of the scarcity of solid evidence about language learning and because of the need to summarize descriptors into short statements to make them easy to use.' Finding a balance between informative interpretation of learner scores and efficient scoring is challenging. Also, the learning value of both rating methods regarding helping students reflect on their own performance was limited to certain aspects of L2 performance, in the sense that it was narrower than is desirable for personalized classroom feedback. Thus any scheme that is realistic for teachers is unlikely to be of great formative value to students.

However, considering that full diagnostic profiling is rarely possible in the real world, a compromise regarding detailed score interpretation of learners' performance needs to be made, and a simplified rating checklist can serve the purposes of efficient scoring and comprehensible description of learner performance.

**References**


Alderson, J. C. (1991). Bands and scores. In J. C. Alderson, & B. North (Eds.), *Language Testing in the 1990s*. London: Modern English Publications and the British Council.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press. http://dx.doi.org/10.1177/026553229601300201

Bejarano, Y., Levine, T., Olshtain, E., & Steiner, J. (1997). The skilled use of interaction strategies: Creating a framework for improved small-group communicative interaction in the language classroom. *System*, *25*(2), 203-214. http://dx.doi.org/10.1016/S0346-251X(97)00009-2

Cohen, A. D., & Dörnyei, Z. (2002). Focus on the language learner: Motivation, styles, and strategies. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 170-190). London: Arnold.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367-396. http://dx.doi.org/10.1177/0265532209104667

Dörnyei, Z. (1995). On the teachability of communication strategies. *TESOL Quarterly*, *29*(1), 55-85. http://dx.doi.org/10.2307/3587805

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188. http://dx.doi.org/10.1017/S0272263102002024

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32. http://dx.doi.org/10.3102/0013189X018009027

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5-29. http://dx.doi.org/10.1177/0265532209359514

Hamp-Lyons, L. (1991). Scoring procedures for ESL context. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 241-276). Victoria, Australia: Deakin University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511733017

Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256. http://dx.doi.org/10.1177/026553229601300302

Muñoz, A. P., & Álvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing*, *27*(1), 33-49. http://dx.doi.org/10.1177/0265532209347148

Nakatani, Y., & Goh, C. (2007). A review of oral communication strategies: Focus on interactionist and psycholinguistic perspectives. In A. D. Cohen, & E. Macaro (Eds.), *Language learner strategies* (pp. 207-227). Oxford: Oxford University Press.

Nunn, R. (2000). Designing rating scales for small-group interaction. *ELT Journal*, *54*(2), 169-178. http://dx.doi.org/10.1093/elt/54.2.169

O'Malley, J. M., Chamot, A. U., Stewner-Manzanares, G., Russo, R. P., & Küpper, L. (1985). Learning strategy application with students of English as a second language. *TESOL Quarterly*, *19*(3), 557-584. http://dx.doi.org/10.2307/3586278

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154. http://dx.doi.org/10.1016/S0346-251X(02)00002-7

Oscarson, M., & Apelgren, B. M. (2011). Mapping language teachers' conceptions of student assessment procedures in relation to grading: A two-stage empirical inquiry. *System*, *39*(1), 2–16.http://dx.doi.org/10.1016/j.system.2011.01.014

Purpura, J. E. (2009). The impact of large-scale and classroom-based language assessments on the individual. In

L. Taylor, & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 301-325). Cambridge: Cambridge University Press.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373-391. http://dx.doi.org/10.1177/026553220101800404

Taylor, L. (2009). Setting language standards for teaching and assessment: A matter of principle, politics, or prejudice? In L. Taylor, & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp.139-157). Cambridge: Cambridge University Press.

Upshur, J., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3-12. http://dx.doi.org/10.1093/elt/49.1.3

**Appendix**

**Questionnaire on the Comprehension and Reflection Provided by the Two Rating Methods**

| | | Yes | No |
|---|---|---|---|
| **Rating Scale** | 1. The feedback from the rating scale helped me understand how far the content I said was related to the topic. | ☐ | ☐ |
| | 2. I know how to improve the topical content I said in the conversation based on the feedback of the rating scale. | ☐ | ☐ |
| | 3. The feedback from the rating scale helped me understand how well my pronunciation and fluency were. | ☐ | ☐ |
| | 4. I know how to improve my pronunciation and to speak fluently in the conversation based on the feedback of the rating scale. | ☐ | ☐ |
| | 5. The feedback from the rating scale helped me understand how well my grammar was in the conversation. | ☐ | ☐ |
| | 6. I know how to improve my grammar in the conversation based on the feedback of the rating scale. | ☐ | ☐ |
| | 7. The feedback from the rating scale helped me understand how far I used the conversation strategies taught in the textbook. | ☐ | ☐ |
| | 8. I know how to improve the usage of conversation strategies in the conversation based on the feedback of the rating scale. | ☐ | ☐ |
| | 9. The feedback from the rating scale helped me understand the degree of my pragmatic competence appropriately in the conversation. | ☐ | ☐ |
| | 10. I know how to improve my pragmatic competence in the conversation based on the feedback of the rating scale. | ☐ | ☐ |
| **Rating Checklist** | 11. The feedback from the rating checklist helped me understand how far the content I said was related to the topic. | ☐ | ☐ |
| | 12. I know how to improve the topical content I said in the conversation based on the feedback of the rating checklist. | ☐ | ☐ |
| | 13. The feedback from the rating checklist helped me understand how well my pronunciation and fluency were. | ☐ | ☐ |
| | 14. I know how to improve my pronunciation and to speak fluently in the conversation based on the feedback of the rating checklist. | ☐ | ☐ |
| | 15. The feedback from the rating checklist helped me understand how well my grammar was in the conversation. | ☐ | ☐ |
| | 16. I know how to improve my grammar in the conversation based on the feedback of the rating checklist. | ☐ | ☐ |
| | 17. The feedback from the rating checklist helped me understand how far I used the conversation strategies taught in the textbook. | ☐ | ☐ |
| | 18. I know how to improve the usage of conversation strategies in the conversation based on the feedback of the rating checklist. | ☐ | ☐ |
| | 19. The feedback from the rating checklist helped me understand the degree of my pragmatic competence appropriately in the conversation. | ☐ | ☐ |
| | 20. I know how to improve my pragmatic competence in the conversation based on the feedback of the rating checklist. | ☐ | ☐ |