한국산학기술학회논문지
Vol. 12, No. 1 pp. 195-199, 2011

The Diversity in an English Oral Proficiency Test

Chung-yeol Park[1*]
[1]**Owens International College, Korea Nazarene University**

영어 능력평가를 위한 구술시험의 다양성

박정열[1*]
[1]나사렛대학교 오웬스국제대학

Abstract There are many causes for the variation of the result in oral proficiency test such as the examiner, the task, the theme of the interview, and the gender of the participants. Previous literature documents that the rater is an important variable influencing test scores of second language oral proficiency. Although much research in language testing has been conducted concerning rater effect on test scores, there has been little attention paid to the effect of potential rater variables in language testing on their rating process. There are noticeably different contents of the rating scales across different speaking tests developed in different context. Therefore, it would not be appropriate to apply the same rating criteria for various tasks. In conclusion, we need more subject protocol analyses and more thoughtful studies on rating processes. In other words, the oral proficiency test needs a more realistic and valid tool for the assessment of second language proficiency.

요 약 영어 능력평가를 위한 테스트는 평가자, 과제, 인터뷰의 주제, 그리고 평가 받는 사람의 성별 등 여러 가지 이유 때문에 그 평가 결과가 아주 다양할 수 있다. 이전의 자료는 평가자가 외국어로서의 영어구술 능력 평가 결과에 아주 중요한 영향을 끼친다는 것을 증명해 왔다. 이렇게 영어 능력 구술 평가에 영향을 미치는 평가자에 대한 많은 조사는 있었지만 평가 과정에 잠재적으로 영향을 줄 수 있는 평가자의 가변성에 대한 조사는 극히 드물었다. 시험 환경이 달라지면 그에 따라 구술 평가가 달라지고 그 평가 기준은 또 완전히 달라진다. 그러므로 다양한 시험 수행 과제에 대해 똑같은 평가 기준을 적용하는 것은 적당치 않다. 즉 평가 과정에 대한 더 많은 원안 분석과 심도 있는 연구가 필요하다. 제 2외국어로서 영어 능력평가를 위한 구술시험은 보다 더 현실적이고 효율적인 장치가 만들어져야 할 것이다.

Key Words : Raters, Test-takers, Rating Scales, Validity

1. Introduction

There are many researchers in the area of oral proficiency tests[2, 6, 7, 9, 14, 20, 23, 24]. The importance of the effects of test- taker characteristics on candidate performance in tests of language proficiency has emphasized[14]. I am interested in the effect of diversity of oral test-taker characteristics such as attributes to be rated, aspects of the discourse, and motivation.

Oral proficiency test increasingly calls for more performance-based tests. Performance-based tests require students to produce complex responses integrating various skills and knowledge for application to in their target language skills with regard to life-like situations. Such tests typically employ more than one test method and call for human raters' judgment. Consequently, these two

This work was supported by Korea Nazarene University
*Corresponding Author: Park, Chung-yeol(cyp4x4@kornu.ac.kr)
Received November 18, 2010 Revised (1st January 9, 2011, 2nd January 12, 2011) Accepted January 13, 2011

factors, the test method and the rater, have become integral components of performance-based tests that influence test scores.

Considering the potential influence of test methods and raters on test scores, what do total scores obtained from oral proficiency tests mean? As Bachman wrote, '... "What does this test really measure?" construct validation is called for'[3]. Validity researchers concur that the primary '...purpose of construct validity is to justify a particular interpretation of a test score by examining the behavior that the test score summarizes'[4]. Therefore, the fundamental issue in construct validation is to uncover the attributes of the constructs underlying test scores. In oral construct validation, therefore, researchers need to specify the attributes of the oral construct and minimize or explain factors such as test method and rater that might confound test score interpretation[10].

The investigation of variation in oral proficiency test is important for understanding. As Douglas and Selinker have pointed out, the abundant evidence linking interlanguage variation to features of the context in which it is elicited has important implications for the design of instruments that attempt to test and evaluate learners' proficiency in a second language[13]. In evaluating learners' proficiency, developers and users of language tests in effect make generalizations from a sample of interlanguage elicited in one context to learners' performance in other and different contexts[24]. Measuring ability in oral proficiency test is problematic because of the complexity both of the skills involved and the context in which these skills are to be elicited and assessed[1].

There are many causes for the variation of the result in oral proficiency test such as the examiner, the task, the theme of the interview, and the gender of the participants. Previous literature documents that the rater is an important variable influencing test scores of second language oral proficiency. Although much research in language testing has been conducted concerning rater effect on test scores, there has been little attention paid to the effect of potential rater variables in language testing on their rating process. This paper concentrates on how the rater can make variability of the oral proficiency test result.

2. Literature Review

Just as important and critical as the selection of tasks used in oral performance tests is the choice of rates. Diverse rater groups may differ in judging learners' second language ability depending on their background and the set criteria with which they operate[9, 22]. Researchers have access to techniques such as Facets that investigate rater fit and adjust for rater severity. Nevertheless, the validity issue remains because raters' perceptions and their impact on scores have not been addressed. According to Brindley, because 'different judges may operate with their own personalized constructs irrespective of the criteria they are given, it would be a mistake to assume that high inter-rater reliability constitutes evidence of the construct validity of the scales or performance descriptors that are used'[9].

Validity is a primary concern in a testing operation. However, the concept of validity has shifted over time. Traditionally, validity was classified by different types, such as content, criterion, and construct validity. However, one of the most influential psychometricians of the second half of the 20th century challenges this traditional validity. He believes that validity is a unitary concept. There are not different sorts of validity; rather, there are many sorts of evidence that can be presented to help document validity. In other words, the different types of validity traditionally discussed are all relevant to help establish a validity argument. Another noticeable change in validity is Messick's emphasis on exploring test score interpretation and use. Therefore, research is needed that focuses on the interpretation and use of the test scores of language oral proficiency.

The rules of speaking continually change with time and place. Although Bachman's Communicative Language Ability (CLA) model has been regarded as the best depiction of language test performance[2], several researchers have raised questions about other factors influencing test scores[10, 18].

In terms of performance assessment, McNamara pointed out that it was necessary to consider 'rating is a result of a host factors interacting with each other'[18]. He interpreted the rating as an end-product of an interaction among task, test-taker, testing performance, rating criteria, rater, and interlocutor. As McNamara argues, test scores

are closely linked to tasks, raters, and rating scales[18]. It is essential to identify sources of variability in the assessment process and estimate the magnitude of their effects in test scores. If the effect is sizeable, the interpretation of test scores will be problematic[2].

English language oral proficiency is usually evaluated by human raters, mostly native speakers[8]. Raters play a major role in the assessment process and influence the quality and meaning of scores obtained. Douglas writes: to attempt to isolate any single component of language ability may be fruitless. We need to know more about how raters arrive at judgments. What aspects of the discourse they attend to in making their ratings, and how different arrive at similar ratings for perhaps very different reasons[12].

There have been some previous studies of the relationship between raters and test scores of L2 oral performance assessments[5, 7, 15, 17, 23, 25]. All these studies have found significant differences among raters. Brown has assumed that different ratings can be controlled by rater training with explicit assessment criteria and samples of performance at different levels[8].

Douglas strongly argued that raters could not arrive at the same rating for the same reasons[12]. He included six examinees on a tape-recorded test consisting of five tasks: answering three unscored warm-up questions, completing 10 partial sentences, answering questions about a picture, responding to two open-ended questions, and describing a diagram. He analyzed the test scores based on five rating criteria; grammar, comprehensibility, vocabulary, fluency, and organization. The results of his study demonstrated that similarly proficient students had different scores for different rating components. His qualitative (interview data) results showed that similar quantitative scores represent qualitatively different speakers' performances with differences identified in degree of value on rating criteria. These results imply that raters are influenced by features of test performance that are not included in the scoring rubric and called for more through research on understanding the basis on which raters make decisions about speaking ability. Chalhoub-Deville investigated the impact of raters as well as tasks on construct validity. She called for further research to derive rating scales based on empirical evidence from a variety of tasks and raters[9].

Lumley compared the extent of agreement in the ratings of candidates' overall language proficiency on twenty audio-recordings of role plays from the Occupational English Test of speaking given by 10 trained ESL raters and nine medical doctors[16]. Each of two rater groups represents (1) language-trained specialists and (2) representatives of the medical profession. Lumley found that there was considerable variation in levels of agreement of holistic ratings within and between the two groups of raters and that the ESL raters were harsher than the doctors. However, he also found that there were broad similarities in judgments between the two groups. The significance of his study is that he observed considerable variation in ratings not between rater groups but between individuals of each group. Therefore, it points to the need for further studies on raters such as the present study.

O'Loughlin examined the effect of raters' gender on test scores[19]. Because oral tests such as the International English Language Testing System (IELTS) has an interviewer as a rater, he has questioned whether gender affects the rating decision of the candidate's oral proficiency level. Sixteen students(8males and 8 females) had a practice IELTS interview on two different occasions, once with a female and once with a male interviewer. All 32 interviews were tape-recorded and reevaluated by 4 raters (2males and 2females) and then analyzed using multifaceted faceted Rasch bias analyses. O'Loughlin found that gender did not have a significant impact on the IELTS ratings.

Bonk and Ockey also used FACETS many-facet Rasch analysis software to examine to what extent variables such as examinee, prompt, rater and rating scales influence score variance[6]. In their study Japanese EFL students viewed a video in their first language which explains what to do in a group oral test and after one minute of preparation time they discussed the assigned topic for 10 minutes in a group of three or four people. Two raters outside the group assigned scores to students independently on pronunciation, fluency, grammar, vocabulary/content, and communication skills/strategies. Bonk and Ockey found that rater differences in terms of severity/leniency were generally large and that these differences were not stable over time for each rater[6]. For example, returning raters tended to move toward greater severity and consistency while new raters showed much more inconsistency. Therefore, they argued that

rater misfit may be a serious threat to general test validity and may lead to inaccurate interpretation of test scores.

The literature review on rater effect on test scores shows that some rater characteristics such as teaching and rating experiences, residing places, and exposure to non-native speakers' English influence the rating performances[9]. On the other hand, other variables such as occupation, gender, and rater training did not influence test scores [7, 15, 19]. This implies that there might be unexamined potential variables of individual raters that influence test groups in their rating of L2 oral tests[12, 16]. Additionally, rater misfit will cause a serious problem in test score interpretation and use, which is the most important issue in validity in language testing[14]. The present study attempts to find these potential variables that impact individual raters' rating performance when they evaluate non-native speakers' English language oral proficiency and that eventually influence test scores.

The literature review on tasks and rating scales shows that these variables in the assessment process of L2 oral tests influence test scores significantly[9, 11, 25]. Therefore, it is difficult to generalize the ratings of one task or by one rating scale to other tasks and rating scales. Another issue to consider is the research which documents that different tasks require different rating criteria [9]. Although much research has been done on this issue, there is no consensus on rating criteria for various tasks[12]. As a result, there are noticeably different contents of the rating scales across different speaking tests developed in different context. Therefore, it would not be appropriate to apply the same rating criteria for various tasks.

3. Conclusion

A number of suggestions are provided for both language testing research and practice. However, there are remarkably few descriptive linguistic studies of oral proficiency test discourse. One of the major criticisms of oral proficiency tests is that the unique roles and meaning created during the course of an interaction seriously compromise reliability.

The analysis of discourse features of the oral proficiency test suggests that there are a number of phenomena useful for providing ancillary criteria for assessing proficiency[21]. Second language research has documented variability in language performance across different tasks/test methods. Performance variability may be attributed to some extent to the different demands the test places on the linguistic and cognitive processes of the subjects, thus influencing their performance. Because both tests and raters affect learners' second language oral scores, researchers might reconsider employing generic component scales[9]. The results help inform second language test developers and researchers about the extent to which rater variables in language testing affect test scores of language oral proficiency.

In conclusion, we need more subject protocol analyses and more thoughtful studies on rating processes. Just like Douglas suggests strongly, I wish to use raters' judgments about learner performance as evidence of underlying language ability. To do this, we need to understand more thoroughly the bases upon which the raters are making their judgments[12]. The results of my research make the oral proficiency test a more realistic and valid tool for the assessment of second language proficiency.

References

[1] Bachman, L. F., Problems in examining the validity of the ACTFL oral proficiency interview. Studies in Second Language Acquisition, 10(2), pp.149-164, 1988.

[2] Bachman, L. F., Fundamental Considerations in Language Testing. Oxford: Oxford University Press, 1990.

[3] Bachman, L. F., Some reflections on task-based language performance assessment. Language Testing, 19(4), pp.452-47, 2002.

[4] Bachman, L. F., & Palmer, A. S., Language testing practice. Cambridge: Cambridge University Press, 1996.

[5] Bachman, L. F., & Savignon, S. J., The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. The Modern Language Journal, 70(4), pp.380-390, 1986.

[6] Bonk, W. J. & Ockey, G. J., A Many-Facet Rasch Analysis of the Second Language Group Oral Discussion Task. Language Testing 20(1), pp.89-110,

2003.

[7] Brown, A., The effect of rater variables in the development of an occupation-specific language performance test. Language Testing, 12(1), pp.1-15, 1995.

[8] Brown, A., Iwashita N., & McNamara, T., An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks, TOEFL Monograph Series, 29, Educational Testing Service, 2005.

[9] Chalhoub-Deville, M., Deriving oral assessment scales across different tests and rater groups. Language Testing, 12, pp.16-33, 1995.

[10] Chalhoub-Deville, M., Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, M. Swain (Eds.), Researching pedagogic tasks: Second language learning, teaching and testing, Harlow, England: Pearson Education Limited, pp.210-228, 2001.

[11] Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I., Second Language Fluency: Judgments on Different Tasks, Language Testing, 54(4), pp. 655-679, 2004.

[12] Douglas, D., Quantity and quality in speaking test performance. Language Testing. 11, pp.125-143, 1994.

[13] Douglas, D., & Selinker, L., Analyzing oral proficiency test performance in general and specific purpose contexts. System 20, pp.317-328, 1992.

[14] Joo, M., The need for an alternative Approach to Oral Testing. The English Teachers Association in Korea, 14(1), pp.1-20, 2008.

[15] Lumley, T., Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. English for Specific Purposes 17(4), pp.347-367, 1998.

[16] Lumley, T., & McNamara, T. F., Rater characteristics and rater bias: Implications for training. Language Testing, 12(1), pp.54-71, 1995.

[17] Lynch, B. K., & McNamara, T. F., 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), pp.158-180, 1998.

[18] McNamara, T. F., Language Testing. Oxford: Oxford University Press, 2000.

[19] O'Loughlin, K., The impact of gender in oral proficiency testing. Language Testing, 19(2), pp.169-192, 2002.

[20] O'Sullivan, B., Learner acquaintanceship and oral proficiency test pair-task performance, Language Testing, 19, pp.277-275, 2002.

[21] Ross, S., Accommodative questions in oral proficiency interviews, Language Testing, pp.173-186, 1992.

[22] Shohamy, E., The stability of oral proficiency assessment on the oral interview testing procedures. Language Learning, 33, pp.527-540, 1983.

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

[24] Young, R. & Milanovic, M., Discourse variation in oral proficiency interview., Studies in Second Language Acquisition, 14, pp.403-424, 1992.

[25] Wigglesworth, G., An investigation of planning time and proficiency level on oral test discourse. Language Testing, 14, pp.85-106, 1997.

Chung-yeol Park [Regular member]

- Feb. 1992 : Sookmyung Women's Univ., English Literature, MS
- Dec. 2004 : Oklahoma City Univ., TESOL, MS
- Feb. 1999 : Sookmyung Women's Univ., English Literature, PhD
- Jan. 2007 ~ Dec. 2008 : California State Univ., Dept. of Education, Visiting Professor
- Aug. 2009 ~ current : Korea Nazarene Univ., Owens International College, Professor

<Research Interests>
English Drama, Language Education