Issues of Rating Scales in Speaking Performance Assessment

Hyun Jung Kim
Teachers College, Columbia University

As with any other area of language assessment, the fundamental issues to be considered in a speaking assessment are: (a) whether or not the test is used as intended, and (b) what its consequences may be (Bachman & Purpura, in press). To ensure that the uses and consequences of a speaking test are fair, the operational definition of speaking ability in the testing context should be examined, since the definition of speaking ability varies with respect to the targeted use and the decisions made. One way to elicit the construct of speaking ability for a certain context is through a scoring rubric which informs test users what a test aims to measure (Luoma, 2004). However, a scoring rubric can affect the speaking assessment, as there may be an interaction effect between the rating criteria and examinees' performance (Luoma, 2004; McNamara, 1996). Different interpretations of the construct may cause biased effects on test-takers' performance, leading to unfairness in scoring and test use. Thus, careful examination of how rating scales interact with speaking performance needs to be considered to determine the fairness of the speaking assessment.

The first issue in examining rating scales is whether the scores given based on the rating scale truly reflect the quality of the test-taker's speaking performance. Douglas (1994) hypothesized that quantitatively similar scores may not necessarily guarantee qualitatively similar speaking performance. In order to test this hypothesis, the performance of six test-takers in a semi-direct speaking test was rated for (a) grammar, (b) vocabulary, (c) fluency, and (d) content and rhetorical organization. The taped responses of test-takers were transcribed for qualitative analysis, where the actual language produced by the test-takers was described in terms of four rating criteria. Both quantitative and qualitative analyses of test-takers' performance revealed a weak relationship between their quantitative scores based on the ratings and their language production analyzed qualitatively. Meiron and Schick (2000) also found that similar quantitative scores represented qualitatively different performance in a role-play simulation task. In their study, the pre- and post-speaking performance of 25 participants in an EFL teacher training program was scored based on a five-category rubric (topic control, pronunciation, grammatical control, lexical control, and conversational control). Close examination of the performance of two test-takers, one whose scores increased considerably from pre- to post-test, and the other who exhibited a very small increase, showed that their performances were very different qualitatively, despite similar quantitative scores on their post-test performance. For example, although these two examinees received the same score on conversational control, one examinee's performance showed more of "an academic approach to rhetorical control" while the other's performance exhibited more of "a dialogic approach to conversational control" (p. 166). The mismatch between examinees' quantitative scores and their qualitative performances, which was found in both of the cited studies, raises questions about the reliability and validity of the testing scores. Thus, for better estimation of test-takers' speaking ability, rating scales should be designed to accurately reflect the operational definition of speaking ability (Meiron & Schick, 2000). This step can prevent different raters from attending to different features in a test-taker's discourse.

What should be considered before deciding on rating scales that ensure the validity of interpretations of test-takers' speaking performance? Alderson and Banerjee (2002) divided rating scales into two categories. The first category are "generic scales" (p. 95), which refer to scales that are constructed in advance by proclaimed experts and that are used to evaluate test-takers' performance on any type of task. The second category includes rating scales designed to target specific tasks. Rating scales and tasks are thus directly linked because the scales describe the kinds of speaking skills that the tasks elicit (Luoma, 2004). Generic scales have the potential to present inappropriate criteria in measuring the intended ability, a concern related to the issue of validity. Different interpretations of descriptors also lead to problems of reliability (Upshur & Turner, 1995). Thus, rating scales developed for particular tasks are more desirable and preferred since they should have greater validity and reliability, particularly those based partially or wholly on a sample of test-takers' performance (Fulcher, 1987; Upshur & Turner, 1995, 1999).

Another consideration in deciding on rating criteria involves what the speaking test intends to measure. That is, it should be clear what speaking ability means in a given task or test and whether or not defined aspects or features of speaking ability are appropriate for the purposes of the test. Based on criteria used in assessing performance, McNamara (1996) distinguished between strong and weak language performance tests. Strong performance tests evaluate test-takers' performance based on real-world criteria where how well test-takers perform on a given task is the main interest. On the other hand, weak performance tests focus more on the language itself. Such tests attempt to elicit a sample of the test-takers' language for evaluation through simulated and artificial tasks, where success of the task is less important than the language elicited.

Although this dichotomy should be understood on a continuum rather than as two separate extremes, McNamara (1996) claimed that most general purpose language performance tests are weak in nature. Douglas and Myers (2000) questioned what appropriate rating criteria are necessary in a language testing context that has a specific purpose. In their study, they reviewed veterinary students' recorded performances in simulated patient/client interviews. The researchers found out that proficiency was judged according to three different criteria. Participants who were professional veterinarians focused on the test-takers' professional relationship with the client and content knowledge, applied linguists concentrated on framework of language use and measurement construct, and student participants used their own knowledge base and the authenticity of the test format. In conclusion, Douglas and Myers (2000) argued that raters should blend criteria from different perspectives. Rating criteria derived from task-specific and real-world concerns might not be useful beyond a certain context. Nevertheless, knowledge of indigenous criteria employed in a real-world situation makes it possible to better understand speaking test performance in relation to the situation at hand (Douglas & Myers, 2000).

In summary, in order to ensure validity and reliability of a speaking performance test, attention needs to be paid to the quality of the speaking performance along with scoring that is based on criteria specific to that particular testing context. Efforts to ensure high validity and reliability can help guarantee fairness in the speaking assessment. Ultimately, "the point is to get test developers to be clearer about what they are requiring of test takers and raters, and to think through the consequence of such requirements" (McNamara, 1996, p. 45).

REFERENCES

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, *35*, 79-113.

Bachman, L. F., & Purpura, J. E. (in press). Language assessments: Gate-keepers or door openers? In B. M. Spolsky & F. M. Hult (Eds.), *Blackwell handbook of educational linguistics*. Oxford, UK: Blackwell Publishing.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*, 125-44.

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60-81). Cambridge, UK: Cambridge University Press.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal*, *41*, 287-91.

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

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

Meiron, B., & Schick, L. (2000). Ratings, raters and test performance: An exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 153-176). Cambridge, UK: Cambridge University Press.

Upshur, J., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*, 3-12.

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

Hyun Jung Kim is a doctoral student in Applied Linguistics at Teachers College, Columbia University. Her research interests include second language assessment, especially speaking assessment, and applied psychometrics.