

Task-based Performance Assessment for Teachers: Key Issues to Consider

Hyunjoo Kim

Teachers College, Columbia University

In the last two decades or so, the tenets of communicative language teaching with their strong emphasis on students' ability to use language in real-life situations have taken hold in foreign and second language classrooms. Accordingly, task-based language instruction that employs communicative tasks as the basic unit of analysis for motivating syllabus design and L2 classroom activities has received increasing recognition. As the primary goal in language instruction is shifted from an object of study to a system of communication, the need to assess students' ability to use the language communicatively has been raised, and performance assessment, including task-based assessment, has become more and more popular.

Performance assessment refers to any assessment procedure that involves either the observation of behavior in the real world or a simulation of a real-life activity with raters to evaluate the performance (Bachman, 2002; Norris, Brown, Hudson, & Yoshioka, 1998; Norris, Hudson, & Bonk, 2002; Weigle, 2002). Performance assessment thus differs from traditional paper-and-pencil tests in that the primary focus is to get an accurate picture of students' communicative abilities and to generalize about students' ability beyond the learning/testing situation to real-life communication. In order to provide test-takers with the opportunity to display more language and have relatively more control over the language produced, open tasks have been widely used (Chalhoub-Deville, 2001). In addition, assessment utilizing open tasks is claimed to provide test-takers the opportunity to utilize their background knowledge and experiences in the testing situation as well as to be active and autonomous. However, using tasks for assessment does not simply mean replicating real-life activities and then asking test-takers to perform. There are several key issues to take into account when designing and scoring task-based assessment.

First, despite the prevalent use of the term *task*, its definition varies (e.g., Crookes, 1986; Doyle, 1983; Long, 1985; Nunan, 1989; Skehan, 1998). Norris et al. (1998) defined tasks as real-world activities "that people do in everyday life and which require language for their accomplishment" (p. 33). In this definition, a test task is a real-world activity. On the other hand, Bachman and Palmer (1996) consider tasks as "an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation" (p. 44). Their definition is broader as it encompasses tasks specifically designed for assessment and instruction as well as real-world activities.

No matter which definition is adopted, as with all tests, teachers or test developers must start with a clear purpose of the test. In other words, they must state what it is that they want to find out based on the test scores. It could be the level of mastery in tense and aspect in English that they have been teaching in class. Or it could be to find out who can advance to the next level or who should repeat the same level. The purpose of the particular testing occasion will decide the subsequent steps in test development.

In addition, the purpose of utilizing tasks needs to be clearly specified in advance. There are currently two approaches. One is to use tasks as a means of eliciting knowledge and skills of interest, and the other is to view them as the ultimate goal of testing. This distinction corresponds to construct-based and task-centered approaches to test design (Bachman, 2002; Messick, 1994;

Nichols & Sugrue, 1999). The reason why this is a critical issue in task-based assessment is because of its direct bearing on what type of inferences the teachers can make from the test scores. In the construct-based approach, the principal inference is about test-takers' language ability elicited through the task, whereas in the task-centered approach, it is the degree of success in completing a particular task using language. In an English class specifically designed to train air traffic controllers, their ability to accurately direct airplanes using the target language may be critical and so needs to be judged according to the real-world criteria. However, when it comes to a typical ESL classroom where students come from all over the world with diverse purposes of learning English, our goal as language teachers and testers should be to make inferences about the state of test-takers' language ability and their ability to use language in situations that resemble real-life.

Regardless of which approach is to be taken, appropriate tasks for the assessment need to be carefully selected and their characteristics thoroughly described. Authenticity of the tasks is a critical quality in task-based assessment. It might be assumed that the closer the relationship between the test tasks and real-life situations, the more accurate the generalization of test scores to non-testing situations will be. However, the use of real-life activities as test tasks may be impractical to administer and/or inappropriate or unfair for certain test-takers since they might presuppose prior knowledge or experience that the test-takers may not possess. Furthermore, given the complex nature of real-life tasks, the issue of task comparability is often raised. In other words, the criteria to select an assessment task among many different real-life tasks become a major concern. For this, teachers will need to have a well-specified target language use domain defined as "a set of specific language use tasks that the test-taker is likely to encounter outside the test itself, and to which we want our inferences about language ability to generalize" (Bachman & Palmer, 1996, p. 44). The TLU domain could be a syllabus or a textbook with specific instructional goals or results from needs analysis when there is no well-specified course content.

In addition to the procedure of identifying TLU domains and TLU tasks from each TLU domain and of selecting test tasks among the TLU tasks, scoring criteria need to be developed. Scoring criteria should be tightly aligned with the types of inferences to be made discussed earlier. If the inference is about how successfully a candidate completes a task, the test is more task-centered and toward a "strong" sense of performance tests (McNamara, 1996, p. 43). Subsequently, performance will be judged on real-world criteria, and adequate real-world criteria need to be identified (e.g., the ability to control airplane traffic efficiently). In this case, the focus of assessment is not language ability but rather the successful completion of the task. Thus, the definition of successful completion is necessarily task-dependent, which would be likely to limit the generalizability of test scores. On the other hand, if the inference is about students' ability to use the language in the situation, the test is more construct-centered and closer to a "weak" sense of performance tests. Its scoring criteria, therefore, must reflect the definition of construct of interest. In either case, explicit scoring criteria are essential since they will function as standards of excellence and self-evaluation criteria for students, and as a consistent, unbiased, and accurate scoring guide for teachers and raters (Herman, Aschbacher, & Winters, 1992).

In summary, it has also been argued that task-based assessment can be very useful for meeting actual inferential demands in language classrooms as it employs complex, integrative, and open-ended tasks. Furthermore, its high degree of authenticity may be beneficial in achieving the intended consequences of assessment by bridging the gap between what the students face in the world and the way they are tested (Delandshere & Petrosky, 1998; Eisner,

1999; Khattri, Reeve, & Kane, 1998; Wiggins, 1993). However, designing and scoring performance assessment is not a simple task of copying a real-world activity to a test. By its nature, task-based assessment involves a lot more variables affecting test scores and, subsequently, the interpretation of those test scores. Thus, the key issues raised in the present commentary must be taken into account if fair and valid tests are to be created for students.

REFERENCES

- Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210-228). Essex, UK: Pearson Education.
- Crookes, G. (1986). *Task classification: A cross-disciplinary review* (Technical Report 4). Honolulu: University of Hawaii.
- Delandshere, G., & Petrosky, A. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*, 14-24.
- Doyle, W. (1983). Academic work. *Review of Educational Research, 53*, 159-199.
- Eisner, E. (1999). The uses and limits of performance assessment. *Phi Delta Kappan, 80*, 658-660.
- Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Khattri, N., Reeve, A., & Kane, M. (1998). *Principles and practices of performance assessment*. Mahwah, NJ: Lawrence Erlbaum.
- Long, M. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77-99). Clevedon, UK: Multilingual Matters.
- McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-23.
- Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18*, 18-29.
- Norris, J., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (Technical Report 18). Honolulu: University of Hawaii.
- Norris, J., Hudson, B., Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*, 395-418.
- Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge, UK: Cambridge University Press.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.
- Weigle, S. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*.
San Francisco: Jossey-Bass.

Hyunjoo Kim is a doctoral student in the Applied Linguistics program at Teachers College, Columbia University. Her research interests include second language assessment and acquisition. She is currently working on her dissertation proposal, defining and measuring speaking ability.

