PERFORMANCE ASSESSMENT IN LANGUAGE TESTING

By

MOHAMMAD ALI SALMANI-NODOUSHAN*

ABSTRACT

Over the past few decades, educators in general, and language teachers in specific, were more inclined towards using testing techniques that resembled real life-language performance. Unlike traditional paper-and-pencil language tests that required test-takers to attempt tests that were based on artificial and contrived language content, performance tests are authentic so that the test-taker is asked to perform language tasks that he or she will need to perform in real-life interactions. A very valuable type of performance test is called portfolio assessment in which a record of students' performance across a wide range of language tasks over a logical period of time is kept so that a profile of performance can be obtained for the evaluation of achievement. This article defines performance assessment, trace its origins and development, explain how performance tests can be constructed, and describes the nature and advantages of portfolios.

Keywords: Testing, Assessment, Measurement, Performance Assessment, Portfolio Assessment, Simulation, Work-Place Assessment.

INTRODUCTION

Some ten or fifteen years ago, few people questioned the widespread use of the standardized achievement tests. After all, standardized achievement tests take relatively little time to administer and are inexpensive. In addition, the results are simple to report and understand. Often a single score is reported for each student, and aggregate scores are reported for a classroom. Finally, and very significantly, standardized achievement tests are promoted as "objective" measures of achievement, meaning that the results are not affected by the personal values or biases of the person who scores the test.

For the past few years, however, language testing scholars have called for dramatic changes in how we assess what students know and are able to do. They have directed most of their criticism at the widespread use of standardized achievement tests. However, many teacher-made tests and tests found in textbooks have similar weaknesses and limitations. Those who propose changes in assessment rest their argument on the premise that what we assess, and how we assess it, affects both what is taught and the way it is taught. Critics of current assessment practices argue that the ultimate goal of assessment should be to have students who can create, reflect, solve problems, collect and use information, and formulate interesting and worthwhile questions. They therefore argue that our assessments must measure the extent to which students have mastered these types of knowledge and skills. They propose what is commonly called Performance Assessment (PA) or, as Flynn (2008) calls it, Performance-Based Assessment (PBA). Performance Assessment may also be taken as synonymous to what, in education literature, has been called Curriculum-based measurement (CBM) (Deno, 2003).

Performance-based assessment utilizes tasks conducted by students that enable them to demonstrate what they know about a given topic. The difference between PBA and the more traditional methods of testing is that, in PBA, students are given the opportunity to better communicate what they have already learnt (Flynn, 2008). CBM is an approach for assessing the growth of students in basic skills that originated uniquely in special education. A substantial research literature has developed to demonstrate that CBM can be used effectively to gather student performance data to support such a wide range of educational decisions as screening to identify, evaluating prereferral interventions,

determining eligibility for and placement in remedial and special education programs, formatively evaluating instruction, and evaluating reintegration and inclusion of students in mainstream programs (Deno, 2003). To provide an accurate reading of students' and schools' rates of progress, and to provide cues for instruction, assessment at every level should be connected to explicit learning goals and standards. (Niemi, Baker, and Sylvester, 2007).

The idea behind performance assessment is not to say that concepts, facts, definitions, dates, names, and locations have no place in education. However, as the critics of traditional assessment practices point out, many of our assessment practices place too much emphasis on assessing content and give far too little attention to the skills and knowledge. They also argue that we must no longer treat assessment as fundamentally separate from instruction. If curriculum, instruction, and assessment are integrated, the assessment itself becomes a valuable learning experience. They conclude that, by requiring students to complete high quality performance tasks, we have the potential to bring about significant and positive changes in instruction and learning. This article provides a useful review of performance assessment in language programs.

1. Background

Language testing has always followed linguistic theories of the time. Thus, the communicative era in the 1970s generated a wave of criticism of the traditional non-communicative tests. These tests were seen as being limited in their concept and as producing artificial language, as opposed to the language normally produced by human beings. For example, the kind of tests used for testing oral language included mostly mechanical repetition of words and sentences and the supplying of pattern answers to pattern questions. In subsequent years there was a shift in language testing towards the development and use of tests that resembled features of real language use and that required test takers to perform language that was authentic, direct, communicative, and performance-based. Such tests, it was believed, would reflect better 'real life' language use as they would tap a broader construct of 'what it means to know a language. A number of terms were used along with these types of tests. Clark (1975) referred to 'direct tests' in which both the testing format and the procedure duplicate, as closely as possible, 'the setting and operation of real life' situations in which language proficiency is normally demonstrated. Jones (1977) proposed performance tests in which test takers provide information on functional language ability. Morrow (1977) recommended few tests that would offer test takers the opportunity for spontaneous language use in authentic settings and activities which the candidate would recognize as relevant. Canale and Swain (1980) referred to performance-based communicative tests which required test takers to perform language while considering criteria such as saying the right thing, at the right time, to the right person. The Foreign Service Institute (FSI) Oral Interview (OI) test was the most relevant example of such a direct, performance-based test (Clark, 1975; Jones, 1977), requiring test takers to use language in a face-to-face oral interaction. The tester asked questions on a variety of topics, and the test taker provided the oral language sample which was then evaluated by the tester with the aid of a rating scale.

In this way, 'performance' became one feature among a number of others, such as 'direct,' 'functional,' and 'authentic,' all of which characterized communicative tests of that era. The unique aspect of the 'performance' feature was that test-takers were expected to replicate, as much as possible, the type of language used in non-testing situations (Bachman, 1990; Bailey, 1985). Thus, performance testing referred to tests where a test taker is tested on what s/he can do in the second language in situations similar to 'real life.' Jones (1985) specified that such tests also required the application of prior learning experiences in an actual or simulated setting where either the test stimulus, the desired response, or both were intended to lend a high degree of realism to the test situation.

The above description characterized features of performance tests in the 1970s. In the 1980s, performance testing became associated more with

specific tasks and contexts of professional preparation and certification, mostly in the workplace (Wesche, 1992). In this context, performance testing borrowed from the field of vocational testing in which a test taker needs to carry out realistic tasks applying language skills in actual or simulated settings (Carroll and Hall, 1985). The criteria used to evaluate the performance was an approximation of the way performance would be judged in the specific and actual target circumstances, including adequate fulfillment of tasks. Wesche (1992) notes that these tests tap both second language ability and the ability to fulfill nonlinguistic requirements of the given tasks. With these types of tests, the main psychometric feature is that of predictive validity; the tests predict how well a test taker will perform under real conditions in a specific context (Jones, 1985). The underlying assumptions with those type of performance tests is that nonlinguistic factors are present in any language performance; consequently, it is important to understand their role and channel their influence on language performance.

In this regard, McNamara (1996) has proposed a distinction between strong and weak hypotheses on performance tests. In the strong sense, knowledge of the second language is a necessary but not a sufficient condition for success on the performance-test tasks: success is measured in terms of performance on the task, and not only in terms of knowledge of language. In the weak sense, knowledge of the second language is the most important, and sometimes the one factor, relevant for success on the test. The specific contexts in which performance testing is used involves a clientele (students, employees, etc.) with certain shared second language needs that can be identified and described, and that can subsequently be translated into test tasks and overall test design. Performance testing, therefore, is associated with a specific context and its strongest requirement will be a detailed description of that context and the language performances associated with it (Sajavaara, 1992; Wesche, 1992).

Jones (1985) distinguished among three types of performance tests according to the degrees that the tasks require actual performances: (a) Direct Assessment, (b) Work-Place Assessment, and (c) Simulation. In a 'direct' assessment, the examinee is placed in the actual target context, and the second language performance is assessed in response to the naturally evolving situation. In the 'work sample' type, there is a realistic task which is generally set in the target context: this type enables control of the elicitation task and a comparison of the performance of different examinees while simultaneously retaining contextual realism. The 'simulation' type creates simulation settings and tasks in such a way that they represent what are thought to be pertinent aspects of the real-life context. 'Role playing' is frequently used as a simulation technique where both the examiner and the examinee play roles. There have also been a number of efforts to use devices such as video, audio recorders, and telephones. For all these types, however, it should be clear that it is never possible to satisfy all the conditions of performance communication and contextual grounding since testing is not really a normal activity. Recognizing this fact, more recent techniques utilize a variety of non-testing procedures that reflect the real performance context: these include record reviews, portfolios, self assessment, participant and non-participant observations, and external indicators.

Wesche (1992) differentiated between performance testing in the work-place and in the instructional context. In the work-place context, tests are used for job certification and for prediction of post-training behavior. In the instructional context, tests are used for washback, diagnostic feedback, and increasing students' motivation. Early introduction of performance tests can help communicate to learners the importance of language objectives, instructors expectations, and criteria for judging performances. Texts and tasks which are used in performance testing also make very good instructional tasks, and ratings obtained from performance tests can be translated to diagnostic feedback in the form of profile scores. Thus, performance tests can actually be introduced in the pre-instruction phase for placement, formative diagnosis, and achievement purposes; during the program itself, these tests can be used for achievement purposes, for

summative testing at the end of a program, and for certification purposes. In instructional situations where the goals are based on an analysis of large language needs, there is a place in the curriculum for an evaluation system which includes performance-type tasks.

2. Construction of performance tests

In constructing a performance test, a need analysis is conducted in order to provide a detailed description of the specific context and tasks which learners will need to perform, the specific conditions under which these tasks will be performed, and the criteria against which the performance can be judged. Then, the learners' performances can be judged over a range of tasks that need to be sampled, using a variety of instruments and procedures. The needs analysis will specify the context of the second language use, the type of interactions foreseen, the roles, discourse types, and language functions to be performed, and the basis on which successful fulfillment of the second language tasks is to be judged. It is with respect to these needs that the performance test is designed, texts and tasks are selected, and evaluation criteria are determined. These are then translated into appropriate test objectives and tasks, and later into actual test design and scoring. Performance tests are generally assessed with the aid of rating scales which describe what a person can do with the language in specific situations.

There are a number of questions that need to be addressed in constructing performance tests: How can the evaluation criteria reflect the kinds of judgments and consequences that the performance would entail? What relative weighting should be given to the different criteria? How can the scoring information be interpreted and presented so as to give maximum information back to the test users? There are also questions more generally related to the criteria by which the performance should be judged: What is the proportion of 'language' vs. 'domain knowledge' to be assessed? Who should be the judge to assess the performance - a native speaker, a domain specialist, or a teacher? Although most performance tests do use the native speaker as the top level of the scale (Emmett, 1985), this issue has been a topic of debate in the language testing literature for many years (Alderson, 1980; Bachman, 1990). Hamilton, *et al.* (1993) claim that performance on a test involves factors other than straight second language proficiency that cause an overlap in the performance of native and non-native speakers. Therefore, the reference to native speaker performance is unwarranted.

In the past few years, performance testing has become a common form of assessment in the educational research context. It is associated with any procedure not employing paper-and-pencil multiple choice items, and it includes a variety of assessment alternatives such as open ended responses, constructed responses, problem solving tasks, essays, hands-on science problems, computer simulations of real world problems, exhibits, and portfolios of students' work. (Linn, Baker, and Dunbar, 1991)

In its simplest terms, a performance assessment is one which requires students to demonstrate that they have mastered specific skills and competencies by performing or producing something. Advocates of performance assessment call for alternative tests that measure students' ability to perform specific tasks. Such tasks might include (a) designing and carrying out experiments, (b) writing essays, (c) working with other students, (d) writing term papers, and so on.

Advocates of performance assessments maintain that every task must have performance criteria for at least two reasons. On the one hand, the criteria define for students and others the type of behavior or attributes of a product which are expected. On the other hand, a well-defined scoring system allows the teacher, the students, and others to evaluate a performance or product as objectively as possible. If performance criteria are well defined, another person acting independently will award a student essentially the same score. Furthermore, well-written performance criteria will allow the teacher to be consistent in scoring over time. If a teacher fails to have a clear sense of the full dimensions of performance, ranging from poor or unacceptable to exemplary, he or she will not be able to teach students to perform at the highest levels or help students to evaluate their own

performance. As such, performance-based assessments require individuals to apply their knowledge and skills in context, not merely completing a task on cue (Brualdi, 2001).

In developing performance criteria, one must both define the attribute(s) being evaluated and also develop a performance continuum. For example, one attribute in the evaluation of writing might be writing mechanics, defined as the extent to which the student correctly uses proper grammar, punctuation, and spelling. As for the performance dimension, it can range from high quality (well-organized, good transitions with few errors) to low quality (so many errors that the paper is difficult to read and understand). Testers should keep in mind that the key to developing performance criteria is to place oneself in the hypothetical situation of having to give feedback to a student who has performed poorly on a task. Advocates of performance assessment suggest that a teacher should be able to tell the student exactly what must be done to receive a higher score. If performance criteria are well defined, the student then will understand what he or she must do to improve. It is possible, of course, to develop performance criteria for almost any of the characteristics or attributes of a performance or product. However, experts in developing performance criteria warn against evaluating those aspects of a performance or product which are easily measured. Ultimately, performances and products must be judged on those attributes which are most crucial.

Developing performance tasks or performance assessments seems reasonably straightforward, for the process consists of only three steps. The reality, however, is that quality performance tasks are difficult to develop. With this caveat in mind, the three steps include:

1. Listing the skills and knowledge the teacher wishes to have students learn as a result of completing a task. As tasks are designed, one should begin by identifying the types of knowledge and skills students are expected to learn and practice. These should be of high value, worth teaching to, and worth learning. In order to be authentic, they should be similar to those which are faced by adults in their daily lives and work;

2. Designing a performance task which requires the students to demonstrate these skills and knowledge. The performance tasks should motivate students. They should also be challenging, yet achievable. That is, they must be designed so that students are able to complete them successfully. In addition, one should seek to design tasks with sufficient depth and breadth so that valid generalizations about overall student competence can be made;

3. Developing explicit performance criteria which measure the extent to which students have mastered the skills and knowledge. It is recommended that there be a scoring system for each performance task. The performance criteria consist of a set of score points which define in explicit terms the range of student performance. Well-defined performance criteria will indicate students what sorts of processes and products are required to show mastery and also will provide the teacher with an "objective" scoring guide for evaluating student work. The performance criteria should be based on those attributes of a product or performance which are most critical in attaining mastery. It is also recommended that students could be provided with examples of high quality work, so that they can see what is expected of them.

3. Portfolios in performance assessment

Proponents of performance assessment also advocate the use of student portfolios. In doing so, they also remind us that a portfolio is more than a folder stuffed with student papers, video tapes, progress reports, or related materials. As such, portfolios provide the teacher with a source for the summative evaluation of the students. It must be a purposeful collection of student work that tells the story of a student's efforts, progress, or achievement in a given area over a period of time. If it is to be useful, specific design criteria also must be used to create and maintain a portfolio system.

Advocates of portfolios suggest two reasons for their use. The first reason reflects dissatisfaction with the kind of information typically provided to students, parents, teachers, and members of the community about what students have learned or are able to do. Secondly, it is

argued that a well-designed portfolio system, which requires students to participate in the selection process and to think about their work, can accomplish several important purposes. For instance, it can motivate students. It can provide explicit examples to parents, teachers, and others of what students know and are able to do. It allows students to chart their growth over time and to self-assess their progress. It encourages students to engage in self-reflection.

Proponents of portfolios argue that the primary worth of portfolios is that they allow students the opportunity to evaluate their work. Further, portfolio assessment offers students a way to take charge of their learning. In other words, portfolio assessment encourages ownership, pride, and high self-esteem. Language teachers and testers should keep in mind that several decisions must be addressed prior to establishing a portfolio system. There must be a physical and a conceptual structure. The physical structure refers to the actual arrangement of documents used to demonstrate student progress. The conceptual structure refers to the underlying goals for student learning. In this connection, numerous questions need to be addressed: Who is the intended audience for the portfolios? Parents? Administrators? or other teachers?, What will this audience want to know about student learning? Will the selected documents show aspects of student growth that test scores don't capture? What kinds of evidence will best show student progress toward the identified learning goals? Will the portfolio contain best work only, a progressive record of student growth, or both?

If portfolios are to be evaluated, the evaluation standards should be established before the portfolio system is established. As for the evaluation itself, portfolios can be evaluated in terms of standards of excellence or on growth demonstrated within an individual portfolio, rather than on comparisons made among different students' work. The final decision item has to do with what is done with portfolios at the end of the course. They could, of course, be turned over to students. However, there are advantages to keeping portfolios over a long period of time and sharing them with other teachers. Portfolios give the teacher opportunities to promote continuity in students' education. By passing a portfolio on to other teachers, a teacher can share important information with the student's next teacher. Portfolios should be kept for long periods of time several years, and they should act as a type of passport as a student moves from one level of instruction to another.

Conclusion

Performance assessment, although a somewhat recent approach in language testing, is gathering momentum and size in much the same way as a snowball would do when moving downhill. Nowadays, language educator do not question its importance and applicability in language programs. Its broad scope allows both teachers and students to envisage a clearer picture of success and achievement.

It is quite safe and sound to claim that the logical conclusion of using performance assessment in language programs is students' self evaluation of their own success. By providing students with the opportunity of performing in a wide range of situations and contexts and a wide range of tasks, over a long period of time, students' portfolios accumulate which can, then, be used as a pedestal upon which students' performance can be judged. It is, therefore, recommended that language teachers give more credence to performance assessment in their profession.

As Verhoeven and Nico (2002) rightly noticed, one point of caution with the implementation of performance assessment in education in general, and in language programs in specific, is that curricular innovations that are based on Performance Assessment might be represented in teachers' professional rhetoric, but not in teacher-made school examinations. This indicates that Performance Assessment may remain at theoretical level and may not turn up in classroom practice. It is therefore vital that curriculum developers should find ways for guaranteeing the practical side of performance assessment in curricula.

References

[1]. **Alderson, J. C., (1980).** Native and non-native speaker performance on cloze tests. *Language Learning*,

30, 59-76.

[2]. Bachman, L. F., (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

[3]. Bailey, K., (1985). If I had known then what I know now: Performance testing of foreign teaching assistants. In P. Hauptman, R. LeBlanc, & M. Wesche (Eds.), *Second language performance testing* (pp. 153-180). Ottawa: University of Ottawa Press.

[4]. Brualdi, A. C., (2001). Implementing performance assessment. *Ed at a Distance Journal*, *15*(4). http://www.usdla.org/html/journal/APR01_Issue/index.html, retrieved on May 10, 2008.

[5]. Canale, M., & Swain, M., (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 6, 67-84.

[6]. Carroll, B. J., & Hall, P., (1985). *Make your own language tests*. Oxford: Pergamon.

[7]. Clark, J. L. D., (1975). *Direct testing of speaking proficiency*: *Theory and practice*. Princeton, NJ: Educational Testing Service.

[8]. Deno, S. L., (2003). Developments in Curriculum-Based Measurement. *Journal of Special Education, 37*(3), 184-192.

[9]. Emmett, A., (1985). The Associated Examining Board's Test in English for Educational Purposes (TEEP) In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 131-151). Ottawa: Ottawa University Press.

[10]. Flynn, L. A., (2008). In Praise of Performance-Based Assessments. *Science and Children, 45*(8), 32-35.

[11]. Hamilton, J., Lopes, M., McNamara, T. F., & Sheridan, E., (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. *Melbourne Papers in Language Testing*, 2, 1-24.

[12]. Jones, R. L., (1977). Testing a vital connection. In J. Phillips (Ed.), *The language connection*: *From the classroom to the world* (pp. 237-265). Skokie, IL: National Textbook Company.

[13]. Jones, R. L., (1985). Second language performance testing. In P. C. Hauptman, R. LeBlanc & M. B. Wesche (Eds.), *Second language performance testing* (pp. 15-24). Ottawa: University of Ottawa Press. 15-24.

[14]. Linn, R. L., Baker, E., & Dunbar, S. B., (1991). Complex performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20(8), 15-24.

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

[16]. Morrow, K., (1977). Authentic tests and ESP. In S. Holden (Ed.), *English for specific purposes*. London: Modern English Publications.

[17]. Niemi, D., Baker, E. L., & Sylvester, R. M., (2007). Scaling up, scaling down: Seven years of performance assessment development in the nation's second largest school district. *Educational Assessment*, *12*(3-4), 195-214.

[18]. Sajavaara, K., (1992). Designing tests to match the needs of the workplace. In E. Shohamy, & A. Walton (Eds.), *Language assessment for feedback*: *Testing and other strategies* (pp. 123-144). Dubuque, IA: Kendall/Hunt.

[19]. Verhoeven, P., & Verloop, N., (2002). Identifying changes in teaching practice: Innovative curricular objectives in classical languages and the taught curriculum. *Journal of Curriculum Studies*, *34*(1), 91-102.

[20]. Wesche, M., (1992). Performance testing for work-related second language assessment. In E. Shohamy, & R. Walton (Eds.), *Language assessment for feedback*: *Testing and other strategies* (103-122). Kendall/Hunt Publishing Company.

ABOUT THE AUTHOR

** Assistant Professor of TEFL, University of Zanjan, Zanjan, Iran*

Mohammad Ali Salmani-Nodoushan is an Assistant Professor of TEFL at the English Department of University of Zanjan, Iran. His research interests include language testing in general, and testing English for Specific Purposes, Computer Adaptive Testing, and Performance Assessment in particular. He is currently a member of the editorial board of The Asian EFL Journal, The Linguistics Journal and i-manager's Journal of Educational Technology.