with performance-based assessment, as fully discussed by Norris et al. (1998), concern the practicality issues and its consequences. Performance-based tests are usually difficult to design and administrate. This is in itself a major problem, but, for the very same reason, the number of test tasks designed in a single test will be relatively small, and, by the passage of time, this can endanger the test security too because the test tasks would lose their originality, and testees would become too familiar with them. Another by-product of this limitation could be the insufficient coverage of the content, which stands out as a threat to the content validity of the test and its generalizability, as generalizations would not be easily possible from one single test to the whole gamut of real life contexts. Performance-based tests also cost very much because of the equipment (such as voice recorders or cameras) usually needed to record the performances or the time and money needed for hiring or training raters.

Conclusion

When first introduced seriously in the field of language testing in the 1970s, performance-based assessment received widespread attention, and many considered it as a revolution. Morrow (1979, p. 144) even called it "the Promised Land," and many people had high hopes for this new approach; however, this "Promised Land" was never truly reached. Just like almost any other novelties, performance-based assessment was welcomed enthusiastically at first, but, by the passage of time, its popularity faded as its weaknesses gradually came to be better known after putting this approach into practice. Nonetheless, the contributions of this approach to the field of language testing should not be underestimated. It was, after all, a move forward, and although it never proved to be the Promised Land for language testers, it did pave the way towards this Land – if, of course, such a land exists at all.

References

Bachman, L. F. (1995). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, E. (1995). *Validity and equity issues in educational assessment*. Paper presented at 17th annual Language Testing Research Colloquium, Long Beach CA, March.

Brown, J. D. & T. Hudson. (1998). The alternatives in language assessment. *TESOL Quarterly,* 32, 653 – 675.

Farhady, H., Jafarpur, A., & Birjandi, P. (2006). *Testing language skills*: From theory to practice. Tehran: SAMT.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*, 28, 483-497.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGrawHill.

Heaton, J. B. (1990). *Writing English language tests*. London: Longman Group UK Limited.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing*, 13(1), 53-61.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: CUP.

Jones, R 1. (1985). Second language performance testing: an overview. In Hauptman, P. Le Blanc, R., and Wesche, M. (Eds.). *Second language performance testing*. Ottawa: University of Ottawa Press: 15-24.

Larsen-Freeman, D. (1986). *Techniques and principles in language teaching.* Oxford: Oxford University Press.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Addison Wiesley Longman.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice*, 12, 9-15.

Morrow, K. (1979). Communicative language testing: revolution of evolution? In Brumfit, C.K., Johnson, K. (Eds.). *The communicative approach to language teaching*. Oxford University Press: Oxford: 143-159.

Moss, P. A. (1992). Shifting concepts of validity in educational measurements: Implications for performance-based assessment. *Review of Educational Research*, 62(3), 229-258.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawaii's Press.

Richards, J. C. and R. Schmidt. (2002). *Longman dictionary of language teaching and applied linguistics*. Essex: Pearson Education.

Robert, L. L., Eva, L. B., & Dunbar, S. B. (1991). Complex, performance-based assessment: *Expectations and validation criteria.* Educational Researcher, 20, 15-21.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 188-211.

Underhill, N. (1987). T*esting spoken language*. Cambridge University Press: Cambridge.

Weir, C. J. (1990). *Communicative language testing*. Hertfordshire: Prentice Hall International.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi *Delta Kappan*, 70, 703-713.

interlocutor/assessor."

The problem of reliability seems, at least to some extent, inescapable because an effort to enhance validity naturally leads to a relative reduction of reliability. The opposite of this holds true as well: In order to make a test more reliable, the validity of that test might be threatened (Heaton, 1990). This seems to be a give-and-take equation, determined by the laws of nature, and one should decide in advance which issue should receive the primary attention. According to Hughes (2003, p. 50), "[T]here will always be some tension between reliability and validity. The tester has to balance gains in one against losses in the other." This continuing tension between reliability and validity has been the point of much argument and concern in the field, and many other scholars have also dealt with it (e.g., Morrow, 1979 or Underhill, 1982).

Unlike many traditional types of tests, such as multiple-choice questions, in which scoring is highly consistent and very easily done, rating in performance-based assessment becomes a major concern because raters deal with real-world performances, not the simple tests of factual knowledge (Shohamy, 1995)

Hughes (2003) also provides an example of a composition writing test to clarify the tension between reliability and validity. In order to make such a test more reliable, the test writer needs to take measures to decrease the potential variability this test could cause in testees' performance on this test. One way to reach this goal would be to set the instructions in a way that testees will be restricted in terms of what they can write about or how they can write it. Think of a writing test in which testees are required to use, for example, the past simple tense to write their composition. This limits testees' freedom of choice, and less freedom results in more reliability, but would such a test be a really valid test of real-world writing? Does it reflect the realities of how we write in our daily lives? Do we feel forced to use a specific verb tense when writing something in the real world? The answer is a definite "No."

The issue of rating and raters has received a great deal of attention in performance-based assessment, and this is considered by many (such as Norris et al., 1998) as one of the main drawbacks of this approach to language testing. Unlike many traditional types of tests, such as multiple-choice questions, in which scoring is highly consistent and very easily done, rating in performance-based assessment becomes a major concern because raters deal with real-world performances, not the simple tests of factual knowledge (Shohamy, 1995). This is the very thing that causes trouble: the complexity of human behavior in real life. No matter how hard the raters might try to be objective in their ratings, they would inevitably engage in some subjective judging process as well. When objective scoring is not present, as Henning (1996) points out, inter-rater reliability estimates should be calculated. According to him, even when high inter-rater estimates are obtained, they cannot be totally dependable (See Henning, 1996 for a discussion of how this is possible).

Many other problems often associated

Backwash effect is undoubtedly among the most crucial topics studied by many scholars of the field, and, as Hughes (2003, p.53) states, "an entire issue of Language Testing has been devoted to the study of the ways of achieving beneficial backwash effect." This is where performance-based assessment is very much valued, as it not only is effective in eliminating the negative backwash effects caused by traditional tests, but it also initiates positive backwash. In case of multiple-choice questions, as an example of more traditional tests, the negative backwash is that the testees would try to learn the strategies of doing better in answering this format of tests rather than try to master the language. As for the performance-based assessment, however, the testees need to work on their language proficiency if they wish to get good grades, which is considered a totally beneficial backwash effect.

What Are Some of the Problems with Performance-Based Assessment?

Performance-based assessment, just like any other approach to language testing, is not without its drawbacks. Bachman (1995), for example, in a separate chapter on test methods, points to a number of problems involved in assessing performance. The effect of test methods, as one of the widely discussed issues in language testing, is a source of concern, as people simply may differ in their reactions to specific testing methods. For instance, some people might find it really difficult to enter a conversation with native speakers, but they might be very good at giving a lecture, or vice versa. Similarly, the test setting (such as personnel's behavior) could also affect the assessment of a testee's performance. Furthermore, the personality factors of an individual could influence his/her performance on a language test. As an example, people with different social backgrounds or learning styles might perform differently on the same test.

Although the problems mentioned above might be present in many other approaches to language testing other than performance-based assessment, there are other disadvantages that could be argued to more specifically concern this approach. As discussed by McNamara (1996), the relevance of the test tasks to the real world is a crucial factor. Since it is not always possible to assess performance in real-world contexts,, there always exists a major concern of not being able to match the real world sufficiently and/or appropriately. Therefore, despite the fact that performance-based assessment enjoys much higher levels of validity in comparison to the traditional tests of knowledge of the language, the validity of such an assessment should not be blindly overestimated or easily taken for granted. Presenting a detailed survey of the problems with establishing different types of validity for performance-based assessment is beyond the scope of this introductory paper. The interested readers may refer to McNamara (1996) for a more detailed discussion.

Another point of concern in performance-based assessment is the issue of reliability. Due to the nature of the real world, its unpredictability, and the lack of scientific control, tests dealing with real-world tasks generally fail to maintain high levels of reliability. As Fulcher (2000, p. 484) states, "performance would be judged subjectively, qualitatively and impressionistically, by a sympathetic

real-life situations (Wiggins, 1989), and this provides support for the construct validity - as well as the face validity - of such tests. For example, having a short conversation with a stranger to get some information about a specific issue or to ask for directions to a specific destination could constitute a typical form of test task in a performance-based test. Such a task would, to a great extent, resemble a real-life situation that many people may encounter when going to a foreign country. This resemblance of test tasks to real life tasks, i.e. the authenticity of test tasks, renders performance-based tests more valid. Using a mock job interview as a test task could be a good example here.

Performance-based tests also generally enjoy high levels of predictive validity, which is regarded as a very important factor when we wish to predict the future performance of our testees as in, for example, entrance examinations of many universities that foreign students are usually required to take before being admitted. Norris, Brown, Hudson, and Yoshioka (1998) rightly observe:

[…] unlike other types of tests, performance assessments can be used to approximate the conditions of a real task in a real-life situation. As a result, performance assessments have value in that their scores can be used to predict students' abilities in future, real-world situations, unlike other tests where scores are only very indirect predictors of ability to perform a real-life language task. We suggest that this potential for predicting or generalizing to future, real-world language use is one of the key contributions that performance assessment might make as an alternative for language assessment. (p.14)

After all, the most significant advantage of performance-based assessment in comparison with the more traditional approaches to language testing is the positive backwash effect that it causes.

to make a distinction between "weak" and "strong" versions of performance-based tests with regard to the extent a test includes nonlinguistic assessment (1996, p. 8). The work sample approach (or the stronger version) is best applied in English for Occupational Purposes contexts, where the nonlinguistic side of the performance of people is more easily observable. The cognitive approach (or the weaker version), however, might better fit the contexts in which only the linguistic performance of people could be tested. This limitation in the scope of assessment, of course, reduces the reality element of performance-based assessment, which is considered a disadvantage, as it is in mild opposition with the rationale of performance-based assessment itself, that is, assessing real life language use. Jones (1985) presents a similar categorization of performance-based assessment by naming "direct assessment," "work sample method," and "simulation techniques." This categorization, just like McNamara's, is done in a strong-to-weak order in terms of the reality of the testing process. An example of such an assessment could be a record of the extracts of language someone has produced over a period of time in his/her workplace.

Why Is Performance-Based Assessment Valued?

As discussed above, in the section about the history of performance-based assessment, it should now be clear that this approach to language testing, or testing in general (see Robert, Eva, and Dunbar, 1991, for a variety of subject matters in which performance-based assessment is used), was basically an attempt to satisfy the needs of the governments or universities that were seeking to find ways to have a valid measure of people's real abilities, rather than their mere knowledge. From this perspective, at least, performance-based assessment is valuable because it has been a step forward in responding to the emerged needs of the time. Aside from how successful or unsuccessful this approach proved to be later, it was nevertheless a change for the better, and many (such as Miller and Legg, 1993; Moss, 1992; Wiggins, 1989) started to support it.

In case of multiple-choice questions, as an example of more traditional tests, the negative backwash is that the testees would try to learn the strategies of doing better in answering this format of tests rather than try to master the language. As for the performance-based assessment, however, the testees need to work on their language proficiency if they wish to get good grades, which is considered a totally beneficial backwash effect

Performance-based assessment is a more valid type of assessment when it comes to assessing testees' language proficiency compared with the older tests of language knowledge or translation. In a performance-based test, the testees are required to use language in a way that they will most likely need to use it later in

The advent of Communicative Language Teaching and its popularity focused attention on communication as a "process" which demanded the application of knowledge of target language forms, meanings, and functions in meaning negotiation (Larsen-Freeman, 1986). This further supported the introduction of performance-based testing into the field, as it was the communicative competence (and, subsequently, the realization of such competence) that rested at the heart of both. The increasing number of the foreign students entering British and American universities constituted the practical reason for embracing performance-based assessment, and the flourishing of Communicative Language Teaching is considered as the theoretical basis of this new approach to testing.

What Is Performance-Based Assessment?

According to Longman Dictionary of Language Teaching and Applied Linguistics (Richards and Schmidt, 2002, p. 392), performance-based assessment is "an approach to assessment that seeks to measure student learning based on how well the learner can perform on a practical real task." The examples brought for this approach to assessment in the same book include essay writing or doing conversations, as opposed to unreal tasks, such as multiple-choice questions or gap-filling ones. These are, of course, the more common examples of performance-based tests; however, there are other types of such tests with even higher levels of resemblance to real-life language and context. McNamara (1996) states:

> *[There are] two main approaches to second language performance assessment: (1) work sample approach, which has its origins in general and vocational education and in personnel selection, and has influenced both general purpose and specific purpose assessment in second languages; and (2) a more cognitive and distinctively linguistic approach in which attention is focused less on the task, which may be relatively unrealistic in real-world terms, but on the qualities of execution in the performance, and/or the evidence it provides about the candidates' control of the underlying linguistic system. (p. 2)*

McNamara goes even further

discrete-point tests, integrative tests were developed, and they received a great deal of support from many scholars. Integrative tests differ from discrete-point tests in assessing two or more skills at a time, instead of testing each item separately. Cloze and dictation are the typical examples of such tests. As Heaton (1990, p.16) rightly observes, integrative tests could also be viewed as an improvement on discrete-point tests by "testing of language in context." Nonetheless, it should not be forgotten that integrative tests involve "functional language but not the use of functional language". That is, they do not tap the socio-cultural, socio-linguistic, or communicative performance of the learners and do not consider the interaction between interlocutors (Farhady, et al., 2006).

The failure of discrete-point and integrative tests to measure communicative competence along with the new policies of the governments of English-speaking countries set the ground for the advent of performance-based assessment. In order to be able to give admission to students with the minimum level of language proficiency required for living and studying in their countries and universities, British and American universities felt the need for more authentic tests that could assess the communicative competence of the applicants rather than their knowledge of English language (Baker, 1995).

The main criticism of multiple-choice items, on the other hand, concerns the strong claim that the ability to answer discrete items of a language test correctly does not equal proficiency in that language

Introduction

The importance of testing, as discussed by Farhady, Jafarpur, and Birjandi (2006) and Hughes (2003), makes us act more meticulously when making a decision about the testing approach that best fits our objectives and purposes. Throughout the history of language testing, which goes hand in hand with the history of language teaching, different theories, approaches, methods, and techniques, have been proposed and utilized in various contexts (Brown and Hudson, 1998). Now the question is: Why has the field of language testing undergone so many changes? The most straightforward answer to this question could be that different needs arise at different times, and so new approaches would be introduced to satisfy those needs. Here follows a short description of the needs that gave rise to performance-based assessment.

How Performance-Based Assessment Came into Vogue

1950s witnessed the introduction of scientific approaches in the field of testing. Science deals with numbers and calls for precision, so objectivity lies at the heart of scientific analyses of all kinds. The result of the application of scientific practices in language testing during this time led to a new approach of constructing tests, widely known as discrete-point approach (Weir, 1990). As the name itself implies, discrete-point testing breaks the language into distinct segments. Hughes (2003, p. 19) defines discrete-point testing as "the testing of one element at a time, item by item". Such breaking of language into smaller parts and testing each part separately provides the test taker with more control on the process of testing. To better grasp the idea of discrete-point testing and its strengths and weaknesses, we can refer to multiple-choice items, which very often serve as the typical example of discrete-point tests (although there are many other test formats falling in the same category, such as true-false questions).

performance-based assessment is "an approach to assessment that seeks to measure student learning based on how well the learner can perform on a practical real task

Harris (1969) points out a number of the advantages of multiple-choice items. One of the significant benefits of multiple-choice items is that the test writer can include many multiple-choice items in a single test, as they do not take much time to answer, which could result in an increase in the content validity of the test. The other eye-catching point is that scoring multiple-choice items is easy, fast, and objective, which would add to the reliability of the test. The main criticism of multiple-choice items, on the other hand, concerns the strong claim that the ability to answer discrete items of a language test correctly does not equal proficiency in that language. As Brown and Hudson (1998, p.659) put it, "real-life language is not multiple choice." In other words, having good knowledge of language elements (linguistic competence), such as grammar or vocabulary, is one thing, while being able to use that language communicatively (communicative competence) is quite a different thing (Farhady et al., 2006).

To compensate for the deficiencies of

Knowledge Improvement

Performance-Based Assessment: What We Should Know about It

Mehrdad Yousefpoori-Naeim
PhD Candidate, TEFL, Shahid Beheshti University
Email: m_yousefpoori@sbu.ac.ir

چکیده

با افزایش محبوبیت رویکرد ارتباطی در آموزش زبان، حوزهٔ سنجش و ارزشیابی زبان نیز دستخوش تغییرات چشمگیری شد که نتیجهٔ آن ظهور سنجش عملکرد ـ محور در این حوزه بود. طولی نکشید که این نوع جدید سنجش مورد توجه مدرسان و پژوهشگران قرار گرفت و به محبوبیتی دست یافت که تا امروز از آن کاسته نشده است. مقالهٔ حاضر تلاشی است برای معرفی سنجش عملکرد ـ محور، چگونگی شکل‌گیری این نوع سنجش و قوت‌ها و ضعف‌های آن. در پایان نیز چنین نتیجه گرفته می‌شود که با وجود دستاوردهای فراوان سنجش عملکرد ـ محور در حوزهٔ سنجش و ارزشیابی زبان، این نوع سنجش نتوانسته است انتظاراتی را که در ابتدای مسیر رشد آن وجود داشت، براورده کند.

کلیدواژه‌ها: رویکرد ارتباطی آموزش زبان، سنجش و ارزشیابی زبان، سنجش عملکردـمحور

Abstract

With the rise of the Communicative Language Teaching, the field of language testing also witnessed outstanding changes, which resulted in the advent of the performance-based assessment. It did not take long for this new approach to language assessment to gain popularity among teachers and researchers, and its passion has not petered out to this day. The present paper makes an attempt to provide readers with a concise overview of what performance-based assessment is, how it was developed, and how it should be viewed in terms of its strengths and weaknesses. It is concluded that despite the numerous benefits the field of language testing has gained from performance-based assessment, it still falls short of the expectations raised at its early stage of development.

Key Words: communicative language teaching, language testing, performance-based assessment