LANGUAGE TESTING BETWEEN STANDARDIZATION AND CLASS ROUTINE

Lect. Speranza TOMESCU[*], Ph.D.
"Carol I" National Defence University

Language testing, like most assessment, irrespective of the domain in which tests apply, can be either summative or formative, according to the framework in which it is administered: during a course to check the acquisition of the taught material within a specific period of time (in the case of formative assessment) or at any moment of one's life to pass an entrance or final examination or simply to prove one's abilities in a domain for various reasons or purposes (in the case of summative assessment). The current trend in language assessment, and generally as well, is to standardize the tests to be administered, especially those that fall under the heading of summative, but not solely. The present paper will analyze how standardized language testing is performed in Romania and in the English-speaking environment nowadays, starting from a general theoretical perspective.

Keywords: *language testing; summative assessment; formative assessment; standardization.*

While performing our duties as language teachers, more precisely English language teachers, to Romanian students, we know we have to develop and apply periodical tests to check how well our students have internalized the knowledge taught and have developed the skills and competencies we strove to form along the language course. These tests verify *acquisition*. This is the so-called *formative assessment* of linguistic abilities and it is more or less up to each teacher, as well as according to each type of course and level of study, which eventually decide whether this test or that one should be a standard or an ad-hoc one.

Sometimes, the language teachers have to also keep in mind that their students prepare for specific exams, and under such circumstances teachers are under the obligation to help their students perform on a standardized pattern imposed by the institution where the assessment is organized. These

[*] e-mail: *speranzat@yahoo.com*

tests verify *performance*. In such cases, training for standardized tests is no longer ad-hoc; standard formats of language assessments have to be used in order to ensure the success of our students in such tests as: Test of English as a Foreign Language, Michigan English Language Assessment Battery, International English Language Testing System, Graduate Management Admission Test, CAE or STANAG-6001 in the case of the military (for NATO-member countries).

Nonetheless, whether formative or summative, whether administered on children, teenagers or adults, evaluation creates anxiety and the prospect of unsatisfactory grades/scores/percentages induces resentment. It is the teacher's role not only to train students in their field of expertise, but also to appease their students' negative emotions regarding evaluation, to boost the students' morale regarding the process of evaluation in a coherent effort to reconcile what happens during regular/routine classes to the idea of assessment.

The purpose of evaluation – between theory and practice

During one's lifetime of formal education, training is performed based on three essential pillars: *teaching* (a process through which the teacher facilitates the students access to knowledge), *learning* (a process through which the students internalize and retain knowledge that has been taught), and *evaluation* (a process through which internalization and retention of knowledge is verified). These three basic elements are interconnected in very complex ways, but put in simple words the relationship between them states this: if *teaching* and *learning* have been performed correctly, then *evaluation* will show it, and the whole educational process turns to good account.

Nonetheless, as simple and logical as the whole process may seem, and as hard as teachers and students may work, evaluation is constantly feared. Indeed, as H. Douglas Brown stated (2004, p. 3) tests should be positive experiences, build a person's confidence and become learning experiences, they should bring out the best in students. Unfortunately, the prospect of getting a grade that measures the students' knowledge may, and it often does, inhibit personal progress in the studied domain. According to Kohn (2011, p. 29), "the more students are led to focus on *how well* they're doing, the less engaged they tend to be with *what* they're doing", which is to say that the fear of performing badly in tests and of getting poor grades paradoxically discourages students to learn, as it makes them focus on getting good grades for the specific areas they know will be verified by the coming test, driving them away from the pleasure of learning itself of analyzing data, investigating facts, discovering phenomena. Thus, students become result-centered and lose their interest in learning as such. As Kohn notes (2011, p.28-29), "educational

psychologists systematically studied the effects of grades " and "the research supports three robust conclusions: [g]rades tend to diminish students' interest in whatever they are learning(…); (…) create a preference for the easiest possible task (…) and (…) tend to reduce the quality of students' thinking". In his endeavor to make a clear point out of the fact the fear of grades in students undermines deep thinking and their desire to study thoroughly, Kohn also suggest that "[r]eplacing letter and number grades with narrative assessments or student-teacher conferences – qualitative summaries of student progress offered in writing or as part of a conversation – is not a utopian fantasy. It has been done successfully in many (…) schools" in the United States of America (2011, p.32). But, in the long run, probably the most detrimental effect of the system based on grades is the fact that students tend to memorize facts for the sole purpose of passing a test and later on they forget the information altogether, which underlines the futility of the grade.

Besides the lack of dependability that the grade system proves, another drawback that the practice of testing often meets is given by the fact that not all teachers use *washback*, as H. Douglas Brown writes (2004) or *backwash*, as Hughes calls it (2003). Either term signifies the same concept, a concept that should be central to testing as a classroom fact: "the effect that tests have on learning and teaching" (Hughes, 2003, p. 53). Since testing, "though an essential component of any sound language curriculum, is only **part** of the curriculum" as J.D. Brown states (1996, p. 288), it should most definitely have an impact on the other parts of the curriculum. The results the students got on a test should influence not only the way the students will prepare for the following test, but also the way the teacher helps them prepare. The teacher acts as a coach, suggesting strategies for success, methods of improving particular elements of the students' performance that turned out badly or not satisfactorily in the previous test. Sometimes, the whole syllabus of a course has to be reconsidered in order to put the educational process on the right track. Just as Hughes contends (2003, p. 1), "if the test content and the testing techniques are at variance with the objectives of the course, there is likely to be harmful backwash". In other words, as mentioned later by the same author, "if the syllabus is badly designed, or the books and other materials are badly chosen, the results of a test can be very misleading" (Hughes, 2003, p.13). Backwash is, on the contrary, beneficial if it leads to re-designing the syllabus in order to fit the aims of the training, which, in its turn, will bring about good test results (Hughes, 2003, p. 2). Other authors put this truth in shorter terms: "When a test becomes a learning experience, it achieves washback" (H. Douglas Brown, 2004, p. 63). By learning experience, we must read a real feedback benefit for the student, but also for the teacher who should draw realistic conclusions regarding their way of

structuring tests (so as to challenge both weaker and stronger students, without actually overwhelming either of them), as well as regarding the way of re-designing the syllabus in order to attain reasonable objectives as a result of following an adequate program of study.

Starting with the mid-90's, another big step forward in the realm of language evaluation has been accomplished – language testers "were prodded to cautiously combat the potential tyranny of 'objectivity' and its accompanying impersonal approach" and to "test interpersonal, creative, communicative, interactive skills, and in doing so to place some trust in our subjectivity and intuition" (H. Douglas Brown, 2004, p. 13). In other words, in the practice of language evaluation, although we strive for the implementation of standardized tests, we also allow for some personal touch in testing, paradoxical as it may seem at a theoretical level.

Testing versus assessment

Although we have used both terms in the article so far, mention should be made of the fact that **testing** does not exactly cover the same area of meaning as **assessment**, according to most of the theorists in the field, point that we support as well. Hughes says that testing is just *one* form of assessment (2003, p. 5). Similarly, H. Douglas Brown states that "tests are one of a number of possible types of assessment" (2004, p. 251), just "a subset of assessment" (2004, p. 4). He also defines a **test** in simple and easy to understand terms as a method by which teachers can *measure* the student's ability/knowledge/performance in the domain under study, mentioning that, for the purpose of *measuring* (as much as the author may resent the term, the **test** "must be explicit and structured" (H. Douglas Brown, p. 3).

This definition sets **tests** apart from **assessment**, the latter being seen as "an ongoing process that encompasses a much wider domain" (H. Douglas Brown, p. 4). Thus, we may consider that assessment never ceases to manifest itself in a classroom, be it physical or virtual, as teachers constantly ask their students questions, or to perform various tasks such as use new words in context to check comprehension of the meaning and so on and so forth, the natural consequence of these simple acts of classroom interactions being the fact that the teachers subconsciously make assessments of their students as the latter offer answers to the teachers' stimuli. This is also called **alternative assessment**, a proposal "to assemble additional measures of students – portfolios, journals, observations, self-assessments, peer-assessments, and the like, in an effort to triangulate data about students" (H. Douglas Brown, p. 251). We can conclude by saying that teachers never cease to assess, but they only test at particular moments, according to the syllabus.

Various types of assessment

As we have already noted, **assessment** is much more complex a term than **tests**. Thus, several dichotomies have been established in the realm of **assessment**. One of the basic dichotomies is the classical **formative** vs. **summative assessment** (H. Douglas Brown, 2004; Hughes, 2003, and others). Indeed, most of the assessment that takes place in a classroom is **formative assessment**, and it is used by teachers to verify the progress undergone by their students in mastering what they were supposed to learn. This kind of assessment is a necessary tool for teachers as it provides "information to modify their future teaching plans" (Hughes, 2003, p.5). As opposed to this, **summative assessment** aims at summarizing what students have grasped during a term/semester or year of study. Therefore, "[f]inal exams in a course and general proficiency exams are examples of **summative assessment**" (H. Douglas Brown, p.6).

Another such dichotomy is the one between **traditional** and **alternative assessment** (Armstrong, 1994 and Bailey, 1998, as quoted in H. Douglas Brown, p. 13). Nowadays, many forms of assessment are either in between the two types or a wise and profitable combination of both. This situation occurs as a result of a current tendency in language testing to supplement the traditional test formats and designs with alternatives of what is largely regarded to be considerably more authentic, comprehensible and meaningful elicitation techniques that lead to the production of genuine samples of language. What is traditional has not been removed or replaced, but rather adjusted to new formats or elicitation techniques. One good example of this is the computer-adaptive test in which a database of multiple-choice language test items is loaded on the computer, the test-taker starts answering the questions, and according to the number of subsequent correct answers, the computer selects the next question and the next level of difficulty for the questions to be addressed.

Many recent authors have used the term **authentic assessment** instead of the one indicated above – **alternative assessment** (see, among others, Kohn, 2011; Mathur & Murray, 2006; Muller, 2003, etc.). One good reason why they are called *authentic* is that they observe and measure the behavior of students in very natural or real-life formats of language samples: interviews, projects, portfolios, journals, rubrics, blogs and wikis, discussions, self-testing, peer-testing, etc. Most of these forms of assessment also have the merit of alleviating the stress of the students concerning grades, especially because they are creative, not predictably and implacably right or wrong; they ask the students to project themselves into the sample of language or the artifact that they produce, to be original and creative rather than to focus on what is the correct form of a word or phrase, or if grammar or spelling are accurate.

Various types of tests

When it comes to language testing, the wide gamut of possible **test types** is stunning – for one thing because there is the matter of the four basic skills tested; and for each skill, there are many varieties of test formats. Beyond the issue of the **specific tests for each basic skill**, though, several classifications of tests have been mentioned in the specialized literature. There is **direct testing** vs. **indirect testing** (H. Douglas Brown, 2004, p. 23). Most often, language teachers confront their students with either **achievement tests** or **proficiency tests**. Mentions are also made by many authors of **placement tests** and **diagnostic tests**.

More traditional reference books in the field of testing classified tests at large in two categories: **norm-referenced tests** and **criterion-referenced tests** (see, among others, J. D. Brown, 1996). By **norm-referenced tests** the theorists mean tests that help administrators and teachers to make *program-level decisions* and they are "designed to measure global language abilities" (J. D. Brown, 1996, p. 2), while **criterion-referenced tests** are those that help teachers to make classroom-level decisions and are "usually produced to measure well-defined and fairly specific objectives" (J. D. Brown, 1996, p. 2). The author further divides these two classes of tests into the two more classes each – **norm-referenced test** can be of the following types: **proficiency** and **placement tests**; and **criterion-referenced tests** can be either **diagnostic** or **achievement tests**.

Proficiency tests are designed to measure a person's ability in a specific language irrespective of the fact that the person has had any formal training in that particular language or not. All proficiency tests are designed based on very clear and concrete specifications of what the test-taker has to be able to do in the tested language if the person is to be considered proficient in that particular language. As Hughes states, a proficiency test shows if the test-taker has "sufficient command of the language for a particular purpose" (2003, p. 11).

According to H. Douglas Brown, "[c]ertain proficiency tests can act in the role of **placement** tests, the purpose of which is to place a student into a particular level or section of a language curriculum or school" (2004, p. 45). However, there are many possible different formats of **placement tests**, depending on the needs and type of the program of instruction it applies to (written tests, oral production tests, multiple choice reading/listening tests, closes, open-ended questions, etc.). The **placement tests** that work best are those especially devised for each particular program, in order to check exactly if the candidate for the course will find the material of the course adequately challenging, that is to say not too difficult to be inhibiting for progress, but not too easy either as not to stimulate interest and not lead to progress.

Diagnostic tests represent a method of measurement limited to a particular and specified aspect of a studied language; they may be meant to evaluate only pronunciation, or only a certain area of the vocabulary, etc. This is why they are called *criterion-referenced tests*.

Another type of *criterion-referenced tests* is the **achievement tests**. These are strictly related to each stage of the program of instruction in turn. **Achievement tests** are somewhat limited because they are meant to verify to what extent the students have internalized the material taught during a stage of the curriculum. This stage may be a lesson, a chapter, a module or even the entire curriculum, which means, on the one hand, that **achievement tests** are oriented on the objectives of the lesson/chapter/module/course, and on the other hand, that an **achievement test** may be either *formative* (progress test) or *summative* (final test), according to how much of the curriculum it tests and when it does it.

The necessary features of an effective language proficiency test

Most language testing theorists are in agreement concerning the issue of the necessary features of a language proficiency test that is effective to administer: **practicality**, **reliability**, **validity**, **authenticity** and **washback** (see, among others, H. Douglas Brown, 2004, Hughes, 2003, and J.D.Brown, 1996).

Practicality is the feature that all standardized language proficiency tests have, especially as it is painstaking and time-consuming to create a test (which implies a very thorough needs analysis, drawing up the test specifications for each level that you expect to be testing, devising test tasks – including multiple-choice items for reading comprehension and listening comprehension, following to pilot the items and to finally validate them). Multiple-choice items are not only **practical**, but also **reliable**. There are some drawbacks in using them nevertheless: they only test recognition knowledge, but not production; as we have already mentioned – they are difficult to conceive in a successful way; besides, guessing has a big impact on test scores, not to mention that cheating is facilitated.

Test reliability is defined as "the extent to which the results can be considered consistent or stable" (J. D. Brown, 1996, p.192). Again, standardized tests meet the requirements, they are considered reliable, since test items are always piloted and selected to meet specifications before they are released on the market, and the scoring procedures are specified and consistent.

Test validity has two basic components: **content validity** and **face validity**. **Content validity** is an intrinsic feature of a test that can be used, in the sense that the things that a test is designed to measure and verify must represent rigorously what has been actually covered during the course as part of the curriculum. The consistency of the content of the test with the objectives of the course as they are projected in the class activities guarantees

a representative sample that can be used as an efficient test. This is why theorists advise teachers and administrators to "base [their] assessment on accomplished class work" (H. Douglas Brown, 2004, p. 32) and to bear lesson objectives in mind when they conceive the test specifications (idem, p. 33). **Face validity**, on the other hand, is an extrinsic feature of an efficient test as it addresses the manner in which the test is being perceived by the test-takers. By this, we mean the administrative aspect of the test that elicits a certain level of performance from the test-takers: the clear phrasing of directions, the logical organization of the test structure, the appropriateness of the level of its difficulty, etc. In Hughes' terms, **face validity** is ensured if the test "looks as if it measures what it is supposed to measure" (Hughes, 2003, p 33).

Authenticity is a natural prerequisite of any material used in the language classroom, so consequently tests are efficient and consistent with everything else that takes place in the language class only if the language used in the test is as natural as possible, tasks are as close to real-life situations as possible, topics that the tasks approach are relevant and appealing to test-takers, and test items are contextualized, not isolated.

Washback, or **backwash**, as mentioned some pages above, is probably the most important factor of an effective test as it represents the concrete effect that *testing* has on both *teaching* and *learning* and which does by no means restrict to communicating scores to the test-takers, but it means giving details about the students' individual test performance and advice on how to approach similar issues in the future, as well as it also means that the teachers may need to re-design the syllabus of the course or at least to approach certain remedial training measures.

Conclusions

As shown above, it is extremely difficult, time-consuming and resource-demanding to design a complex language test that is efficient as well; this is why language teachers are recommended to use commercially produced tests available on the market, the result of a complex process of creation, piloting, validation by trial and error techniques instead of creating their own overall proficiency test (H. Douglas Brown, 2004, p. 45).

Nonetheless, organizations, such as the Romanian military (to name just one among all the NATO-member militaries that deal with language testing) may engage in the endeavor of creating batteries of standardized tests for each basic skill, harnessed by a large group of specialists, observing strict and rigorously defined level descriptors and skill specifications. Such batteries, although subject to periodical re-designing, and rigorous upgrading, may be used repeatedly on a large scale.

BIBLIOGRAPHY

Brown H. Douglas, *Language Assessment. Principles and Classroom Practices*, Pearson Longman, 2004.

Brown James Dean, *Testing in Language Programs*, Upper Saddle River, New Jersey, Prentice Hall Regents, 1996.

Hughes Arthur, *Testing for Language Teachers*. Second edition, Cambridge Language Teaching Library, Cambridge University Press, 2003.

Kohn Alfie, *The Case against Standardized Testing*. Westport, CT, Heinemann, 2000.

Kohn Alfie, *The Case against Grades*. in *Effective Grading Practices*, November 2011/Volume 69/Number 3, 2011.

Mathur S. & Murray T., *Authentic Assessment Online: A Practical and Theoretical Challenge in Higher Education*. in *Online Assessment, Measurement and Evaluation. Emerging Practices*. William, D., Howell, S.L., Hricko, M. (ed). Information Science Publishing, 2006.

Muller J., *Authentic assessment toolbox*. Retrieved June 18, 2004, http://jonathan,muller.faculty/nocrtl.edu/toolbox/whatisi.htm