

## Views on EFL Testing System



Naira Avakyan

A test or an examination is an **assessment**, often controlled on paper or on the computer, intended to measure the test-takers' or respondents' knowledge, skills and aptitudes. Tests are often used in education, professional qualifications, psychology and many other fields. The measurement that is the goal of testing is called a **test score**, and is "a summary of the evidence contained in an examinee's responses to the items of a test that are related to constructs being measured" (Thissen 2001:1). Test scores are interpreted with regards to a norm or criterion. The norm may be established independently or by statistical analysis of a large number of subjects. A standardized test is one that is administered and scored in a consistent matter to ensure legal defensibility.

The earliest known standardized tests including both practical and written works were the Chinese Imperial Examinations which began in 587. In Europe, traditionally school examinations were conducted orally. Students would have to answer questions formulated by teachers in Latin, and teachers would mark them on their answers. The first written exams were held at Cambridge University, England in 1792.

The main component of a test is an **item** stored in an item bank. Items are often colloquially referred to as "questions", but not every item is phrased as a question; it may be a true/false statement or a task to be performed.

Various item formats are available for test construction: multiple choice, free response, performance, etc. There is no best format to use; the applicability depends on the purpose and content of the test.

A common type of test item is a **multiple-choice question**; the author of the test offers several possible answers (usually four) from which the examinees must choose. There is one right answer represented by only one answer option.

Let us view the college/university entrance written exam test in English in Armenia. The test consists of 12 tasks checking the examinee's ability of text comprehension, the knowledge of grammar, vocabulary, logics and so on. Test authors generally create incorrect response options often referred to as distracters which correspond with likely errors, such as the use of *rather/quite/enough*, *besides/except*, *despite/in spite of/though*, *hard/hardly*, *late/lately*, *many/much/few/little/a few/a little*, *so/such*, *too/either*, etc. The stress is also put on common errors, those typical of Armenian learners of English as a foreign language. Here we have structures like *remember to do/doing*, *forget to do/doing*, *need to do/doing*, etc. which are not differentiated in their mother tongue.

A positive feature of multiple-choice questions is that they are particularly easy to score. Machines such as the Scantron and software grading of computer-based tests can be performed automatically and directly which is particularly valuable for situations

where there are not enough graders available to mark a large-scale standardized test. Multiple-choice tests are also valuable when the test sponsor desires to have immediate score reporting available to the examinee; it is impossible to provide a score at the end of the test if the items are not actually scored until several weeks later.

This system is not, however, appropriate for assessing all types of skills. Poorly written multiple-choice questions often create an overemphasis on simple memorization and deemphasize processes and comprehension. They also leave no opportunity for disagreement or alternate interpretation, making them particularly unsuitable for humanities such as literature and philosophy. Besides, “through specialized training on material and techniques specifically created to suit the test, students can be “coached” on the test to increase their scores without actually significantly increasing knowledge of the subject matter. However, research on the effects of coaching remains inconclusive, and the increase might be simply due to practice effects” (Domino 2006:340).

**Free response questions** evaluate the responses in a different way. Effective scoring involves reading the answer carefully and looking for specific features, such as clarity and logic which the item is designed to assess. Often, the best results are achieved by awarding scores according to explicitly ordered categories which reflect an increasing quality of response. This may involve the construction of marking criteria and support materials such as training materials for markers and samples of work which exemplify categories of responses. Typically, these questions are scored according to a uniform grading rubric for greater consistency and reliability.

At the other end of the spectrum, scores may be awarded according to superficial qualities of the response such as the presence of certain important terms. In this case it is easy for the examinees to “fool” scorers by writing a stream of generalizations that include the terms that the scorers are looking for. This, along with other factors that limit their reliability has caused the value of this item type to be questioned.

While free-response items have disadvantages, they are able to offer more differentiating power between examinees. However, this might be offset by the length of the item, such as if a free-response item provides twice as much measurement information as a multiple-choice item but takes as long as three multiple-choice items to complete.

Neither free-response nor multiple-choice questions can assess the knowledge of “how to do” something. It may be demonstrated only by a performance test. Art, music and language fall into this category, as well as non-academic disciplines such as sports and driving.

A practical examination may be controlled by an examiner (a person) or by means of an audio or video recording. It may be checked on its own or in combination with other types of questions. For instance, many driving tests in Armenia include a practical examination as well as a multiple-choice section regarding traffic rules.

Tests of the sciences may include laboratory experiments to make sure that the student has learned not only the body of knowledge comprising the science but also the experimental methods through which it has been developed. Again, the use of explicit criteria is generally valuable in marking practical examinations.

General aptitude tests are used in certain countries as a basis for entrance into col-

leges and universities. A criticism associated with this use of such tests is that they do not necessarily assess the accumulated learning of students during their schooling years. However, the goal of these tests is not to assess accumulated learning; they are designed to measure aptitude, not achievement.

“Similarly, college entrance exams are criticized for not accurately predicting first-year university grade point average (GPA) as well as high school GPA. However, the intent is for test scores to be used along with other measures in university selection; large-scale test scores are only one aspect of the university selection process. Universities are free to place more emphasis on high school GPA or extracurricular activities. Any criticism might be better directed to a university than the test itself, which most people consider fair” (Domino 2006:342).

The content of the exam might not correspond with its intended use or representation. As an extreme and unrealistic example, a mathematics exam may ask only about the names, birthdates, and country of origin of various mathematicians when such knowledge is of little importance in a mathematics program. Back again to college/university entrance exams in Armenia, this time the oral exam in English. The examinee is checked in the ability of text comprehension (Item 1) and communicating skills (Item 2). Item 1 includes an unknown text in the original followed by a number of questions. The examinee is supposed to read the text, translate a passage and answer the questions on the text. Item 2 includes topics like “My family/friend/flat/school/day off,” etc. and some topics on land study (the UK, London and major cities, the USA, etc) for which the examinee has to get ready beforehand trying to memorize some facts, geographical names and historical events. A good memory is very important in this case. Such knowledge is of little importance, like in the above mentioned case with mathematics. Apart from comprehension check there should be free communicative skills check in the form of questions on various topics which the examinee has to answer on the spot without getting ready beforehand. The applicants should prepare for the entrance exams not by getting trained mechanically but by reading books enriching their word-stock. In early 1980s in Armenia the college/university entrance oral exam in English included oral grammar check – sentence analysis. This is the best option for grammar check. Nowadays the knowledge of grammar is checked mainly by multiple-choice questions during the written exam. The test-takers sometimes tick the right answer at random without a deep knowledge of the subject.

Another problem with the assessment is the psychological factor. People are variously inclined to stress. Some are practically unaffected and excel on tests, while in extreme cases individuals can become very nervous and forget a substantial part of exam material. To balance this, often teachers and professors don't mark their students on tests alone paying considerable attention to homework, attendance, in-class discussion activity, and laboratory investigations (where applicable). On the other hand, in some cases anxiety factor stimulates examinees to rise to meet the exam's high expectations.

Finally, the greatest problem with the assessment is the justice factor. Test organizers attempt to prevent injustice and impose strict penalties for it; academic dishonesty can be used to obtain an advantage over other test-takers. On a multiple-choice test lists of

answers may be obtained beforehand. On a free-response test the questions may be obtained in advance or the examinee may write an answer that creates the illusion of knowledge. If students sit close to one another, they are likely to copy answers off other students. Despite such issues tests are less disposed to cheating than other tools of learning evaluation. Laboratory results can be fabricated and homework can be done by one student and copied mechanically by others. We can guard against cheating by the presence of a responsible test administrator. Thus, considerable steps should be taken to ensure objective assessment.

### References:

1. Cangelosi, J. (1990) *Designing Tests for Evaluating Student Achievement*. NY: Addison-Wesley.
2. Domino, G., Domino, M. (2006) *Psychological Testing: An Introduction*. Cambridge: Cambridge University Press.
3. Douglas, D. (2000) *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.
4. Haladyna, T. (2004) *Developing and Validating Multiple-Choice Test Items*. Mahwah, NJ: Lawrence Erlbaum Associates.
5. Thissen, D., Wainer, H. (2001) *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

### Անգլերենի՝ որպես օտար լեզվի թեստային համակարգի շուրջ

Հոդվածը քննարկում է անգլերենի՝ որպես օտար լեզվի թեստային համակարգի երկու առավել տարածված թեստերի տեսակները: Դրանք են՝ բազմակի ընտրության և ազատ հարց ու պատասխանի թեստերը: Թեստերի երկու տեսակներն էլ ունեն առավելություններ և թերություններ: Քննական գործընթացում կարևոր են հոգեբանական և արդարության գործոնները: Հոդվածը ներկայացնում է անգլերենի բուհական ընդունելության բանավոր քննությանը առնչվող հեղինակի կողմից առաջարկվող որոշ բարեփոխումներ: