УДК 372.881.111.22

Elena V. Likhacheva

Novosibirsk State University
Pirogov Str., 2, Novosibirsk, 630090, Russia
E-mail: orientlab@nsu.ru

SOME EFFECTIVE METHODS OF TESTING LANGUAGE COMPETENCE

This study investigated the underlying structure of language ability in relation to the levels of language proficiency and test methods. Teaching and testing are theoretically and practically interrelated. Theoretically, they have been influenced by the principles of linguistics, psychology and other language related fields. Practically, they are complementary because both aim at optimizing the efficiency of education. The main purpose of this study has been to investigate the factors which influence of the language testes in order to avoid errors of final estimation of their proficiency.

Keywords: diagnostic testing, placement test, achievement test.

Though being widely discussed, such important issue as testing students' knowledge hasn't proved to offer a universal and adequate solution so far. At least we, ESP (English for specific purposes) teachers, continue suffering from the diversity of materials chosen to check students and lack of criteria making us hostile what mark to put. This is especially true for testing students who enter our oriental language department with different levels of language competence but with a high motivation to improve it as English though not officially is recognized as the second state language in Japan, Korea and China. How can we prove our students and ourselves the objectivity of their mark? There is always a potential for errors. Just as the track on which the athletes will be competing should be prepared very thoroughly (its length, its smooth or rough surface, its straightness or curves) – all its features will reveal the best in the field. In the same way the testing of language abilities must be accurate, reliable and provide us with a set of measurements for a variety of purposes to be useful.

The types of tests depend on the purpose stated. Perhaps the most common use of language tests and educational test in general is to pinpoint strengths and weaknesses in the wanted abilities of a student. We may discover through testing that a given student has excellent pronunciation and fluency of speech but that she or he has a low level of reading comprehension which is caused by her or his too highly specialized vocabulary. We might recommend suitable approaches for vocabulary expansion. This kind of test is frequently termed **diagnostic testing** and it should make a learning process more efficient. Another important use of tests is to assist in the decision of who should be allowed to participate in a particular program or course (especially when there are more applicants than places available) This type of test is often called **screening** or **placement**. One more common use of tests, especially achievement tests, is to provide information about the progress a student has made obtaining a certain program or part of it. This type of test is especially close to us as we have to conduct it twice a year. **Achievement tests** are used to measure the extent of learning in a prescribed content domain often in accordance with the stated objectives of a learning program. These tests may be used for program evaluation as well as for certification of a learner's competence. Apart from this type a **proficiency test** or **mastery test** are more global measures of ability in a language. They are not necessarily developed with reference to some previously trained course. They are aimed at revealing the mastery level of all skills and abilities of the examinee.

Just as there are many purposes for which language tests are developed, so there are many types of language tests. As it has been noted, some tests serve a variety of purposes, while others are more restricted in the applicability. Let us have a brief view over the most common ones:

ISSN 1818-7919
Вестник НГУ. Серия: История, филология. 2012. Том 11, выпуск 4: Востоковедение

1. *Objective* / *subjective tests* – they are distinguished on the basis of the manner in which they are scored. An objective test is scored by comparing examinee responses with an established set of responses or scoring keys. A common example of it is a multiple-choice recognition test. A subjective test is based on the judgment on the part of the scorer, his expertise. An example might be the scoring of free written compositions with an element of the creativity presented but in a situation where no operational definition of creativity is provided and when there is only one rater. The latter type of test can be objectified through the use of precise rating schedules clearly specifying the kinds of errors to be quantified or through the use of multiple independent raters.

2. *Direct* / *indirect tests*. Direct tests are ratings of language use in real communicational situations which show language performance directly whereas other tests such as multiple choice recognition tests are indirectly touching true language performance and therefore are less valid for measuring language proficiency. Thus it can be recommended to locate them somewhere closer to natural situations. For example, an interview might be more direct for measuring overall language proficiency than an artificially thought of particular topic or contextualized vocabularies more natural and direct than a synonym-matching test.

3. *Discrete point* / *integrative*. Discrete point tests are designed to measure knowledge in very restricted areas of the target language. Thus a test of ability to use correctly the perfect tenses of English verbs or to supply correct prepositions in a passage may be termed as a discrete point test. Integrative tests, on the other hand, are to cover a greater variety of language abilities and even all language proficiency. Examples of integrative tests are oral interviews, dictations and oral imitation tasks. Frequently the attempts are made to achieve the best results through the use of discrete point subtests for diagnostic purposes but which provide a total score to reflect an overall language proficiency. A test of listening comprehension may check one of the four general skills in a discrete manner and have a limited value of language proficiency. On the other hand such a test may examine a broad range of lexical and grammatical structures and in this way be said to be integrative.

4. *Criterion-referenced tests* are designed before the instruction (program) itself is designed. So the teachers tend to teach to the test. The criterion or score is set in advance (usually 80–90 % of the total possible score) and those who do not meet it are required to repeat the course. These tests have both strengths and weaknesses. On the positive side the process of developing criterion referenced tests is helpful in clarifying objectives. Such tests are useful in defining the degree to which objectives have been met both in ongoing and final evaluation. The security is considered less of a problem since students know in advance the precise content for which they are responsible in the test. For the same reason students' test anxiety is reduced with this type of the test. On the negative side, the objectives measured are often too limited and restrictive. Bright students, who easily attain the criterion level of mastery, may not be encouraged to reach higher standards.

5. *Norm-reference or standardized tests* are quite different from the previous group in a number of aspects. First, it must be conducted for a large group of people and acceptable standards can only be determined after the test has been developed and administered. Since a broad range of scores is desired, items at various levels of difficulty are included. For the purpose of language testing norm reference tests have also strengths and weaknesses. Among the strengths is the fact that comparison can be made among the larger group of students. Since examinees are spread on a wider range of results, more valid information is provided about their abilities.

6. *Speed tests* / *power tests*. A speed test is one in which items are so easy that every person taking this test must be expected to get every item correct, given enough time. But sufficient time is not provided, so examinees are compared on their speed of performance rather than on knowledge. Power tests are tests that allow sufficient time for every person to finish but that contain such difficult items, that few examinees are expected to get every item correct. Most tests should fall somewhere between the two extremes since knowledge rather than speed is a primary focus, but time limit is necessary since a weaker student may take unreasonably long periods of time to finish.

7. Not all the categories were mentioned above. , distinction can be made between **single – stage / multi-stage tests, language skills tests** (listening comprehension, oral production, reading, composition writing) and **language feature tests** (verb tenses, subject-verb agreement, comparatives-superlatives), t. i. between **production / recognition tests.** Whatever the type of the test may be, all types must contain the following features: test validity, test difficulty, test applicability, test relevance and test interpretability (scoring and reporting). What is test validity? The test is valid

when it adequately measures what is supposed to measure. Here we should ask ourselves "Is the content of the test consistent with the stated goal for which it was administered?" If it is an achievement test, for example, it must accurately and fully reflect the extent to which students have mastered and mustn't contain items which were not encountered by the students in the program. Very often the examinations are too easy or too difficult for the examinees. To avoid this, a test should be tried on a sample of persons. Inspections of difficulty level (the length and complexity of a reading passage, the familiarity of the vocabulary and other common indicators reveal whether the test is appropriate for the examinee. Every test is subjected to inaccuracies due to different factors, such as the position of the examinee in the classroom, psychological state of a student, various levels of experience of the raters themselves, t. i. environmental and psychological factors. To minimize measurement error is to ensure test reliability. Examinations that serve as admission criteria must be highly reliable for they define a lot in the whole life of an examinee. For standardized tests accompanying test manuals usually report the reliability levels measured.

In developing our own tests we should estimate reliability using a variety of methods, such as test-retest method, parallel form method, split-half reliability and some others. Another problem with language tests is that they may turn out to be more reliable and valid for persons from one particular language background than for those from another background. So it is important to consider characteristics of the sample on which the test was developed before it is blindly applied to the same sample from a different environment. In our situation the relevance of the test very often is a crucial factor, for the students often are aware of different realia of foreign life which prevents them from successful completion of the test.

Before the test is selected, developed or applied, it is highly important to understand how the test is to be scored, how the scores are to be interpreted. In general the scoring procedures require extensive training and experience. There exist specific formulas and coefficients which are too complicated for us, ESP teachers to follow in details. Still there are ways to objectify the measurements of different types of tests. Let us begin with evaluating creative writing which tends to be the most subjective as a rule, specifically with the test of prose passages, t. i. writing of precis, compositions, dictations, essays and so on. Dictation passages may be scored by treating each word as a separate item, counting it correct as present and incorrect as absent provided it is spelled correctly as well. With grammatical-lexical tests however the more common procedure is to allot a maximum score possible and then systematically subtract points for errors of grammar, spelling or vocabulary depending on the purpose of the test. The problem in scoring free writing tests is manifold. Some examinees write longer passages than others, some avoid all complex structures and sophisticated vocabulary for fear of mistakes, while others try more creative use of language and thus generate comparatively more errors. Subjectivity may be minimized if we use a rating schedule which operationally distinguishes between superior and inferior performance:

Sample rating schedule for use with précis, essays and compositions
Mechanics content

Area	Weight	Area	Weight
Spelling	1	Organization	1
Grammar use	1	Relevance to the topic	1
punctuation	1	Creativity/interest	1
orthography	1	Range and sophistication of syntax	
paragraphing	1	Richness of vocabulary	1
Total	5	Total	5

This schedule permits maximum 10 points with equal weighting of mechanics and content with one point awarded for satisfactory in each component area. There is still a problem with such a scale, t. i. the selection of component areas and weights is somewhat arbitrary but from the teaching practice proved to the most decisive ones. Thus, one teacher may prefer to give more weight for grammar use or organization and so on. Still it allows us to analyze each of the components measured.

Another variety of writing schedule might specify the kind of writing skills (behavior) that warrants specific marks.

A simple behavior-specific rating schedule for use in foreign language writing evaluation.

Behavior	Rating
1. Writing is indistinguishable from that of a competent native speaker	5
2. Writing is grammatically correct but employs nonnative usages	4
3. Writing contains infrequent errors of grammar, lexis, spelling or punctuation	3
4. Writing contains numerous errors of grammar, lexis, spelling and punctuation	2
5. Writing is incomprehensible. Orthography is illegible	1

Tests of oral communications. It is equally desirable to have rating schedules and multiple independent raters who are experts. It is equally preferable to judge performance in more than one topic or situational context. It is advisable to record answers if possible. A sample rating schedule shows the examinee' s performance on a scale of 1–5, depending on how nearly it approximates a native speaker's performance:

Name of examinee	Fluency (1–5)	Accuracy of pronunciation (1–5)	Grammar accuracy (1–5)	Expressive content (1–5)	Total (the average)

And one more important remark. Usually a test can't be exhaustive but must be selective in content. A reasonable approach here will be to develop elaborate specifications for items in variety of domains. A large number of items would be written in each domain so as to cover the principal aspects or parts of the program and allow the examinee to complete them in prescribed time limits.

This material is far from being exhaustive as the area of application is really broad. It may require additional specifications for each domain. But it gives the principal directions for the estimation of the results of efforts both the teachers and their students have made, helps to be aware of purposes and objectify the marks students receive.

References

Lyle F. Bachman. Fundamental Considerations in Language Testing. Oxford Univ. Press, 1990. 408 p.

Robert L. Ebel. Measuring Educational Achievements, Prentice Hall International, 1995. 481 p.

Tests and Measurements in High Education, Chicago. III. The Univ. of Chicago Press, 1976. 237 p.

Материал поступил в редколлегию 28.01.2012

Е. В. Лихачева

ЭФФЕКТИВНЫЕ МЕТОДЫ ОЦЕНКИ ЯЗЫКОВОЙ КОМПЕТЕНЦИИ

Статья представляет собой исследование внутренней структуры языковой способности в отношении уровня владения языком и методов оценки. Обучение и тестирование теоретически и практически взаимосвязаны. Теоретически на них оказывают влияние принципы лингвистики, психологии и других областей науки, связанных с языком. С практической стороны обучение и тестирование дополняют друг друга, поскольку нацелены на повышение уровня эффективности обучения. Основная цель данного исследования – рассмотреть факторы, влияющие на испытуемых в языковых тестах, чтобы избежать ошибок при окончательной оценке их квалификации.

Ключевые слова: тестирование, тест на определение уровня, языковая компетенция.