

UDC [378.147:811.111]=111

DOI <https://doi.org/10.32840/1992-5786.2020.69-2.26>

T. V. Zhukova

Senior Teacher of the Department of Foreign Languages
Dmytro Motorny Tavrta State Agrotechnological University

PECULIARITIES OF TEST DESIGN AND ASSESSMENT OF STUDENTS' OUTCOMES IN A FOREIGN LANGUAGE

The importance and the main objectives of assessment in the learning process are substantiated in the article. Assessment is used to check if learning has taken place, for example: the skills, competence, knowledge, understanding needed at any given point. If we didn't assess, we wouldn't know whether learning had taken place, the teaching may have, but no learning achieved.

Assessment should focus on improving and reinforcing learning as well as measuring achievements. It helps learners identify their progression and what they need to do to improve and progress further. Assessment is a regular process; it might not always be formalized, but we will be observing what our learners are doing, asking them questions, and reviewing their progress whenever we are in contact with them. The learners will be demonstrating their knowledge and skills regularly, for example through activities, discussions and regular tasks.

The basic phases of test development are described: the design phase when the test features are developed and sample materials are written; the operational phase which involves the construction of a complete test; the monitoring phase when materials are revised and updated. Different tasks that are used in test design, their relevance and meaningfulness are lighted in the article.

The most important things in test or task design have been analyzed. The task must be appropriate for the skill being assessed. The test is to be fair to the test taker. And the test should be accurately scored.

The main types of tasks used to test listening skills (picking out a word, recognizing the correct tense, deducing meaning of unknown words from context, identifying the main ideas, listening for specific information) are presented.

The main assessment criteria for speaking, reading, listening and writing are presented in the paper. As the process of assessment is a complicated one we must be certain that our tests are appropriate, fair and accurately scored.

Modern teaching must ensure that testing has a positive impact on the learning process and motivation of students. We are therefore constantly making judgments about learners' progress and how they could improve. We should also be aware of the impact that our feedback comments and decisions can have on their confidence. Comments which specifically focus on the activity or work produced will be helpful to our learners.

Key words: *assessment, test development, validity, reliability, assessment criteria, motivation, academic achievements.*

Setting the problem. When teaching students foreign languages we should be able to assess our students' academic results properly. Assessment constitutes some of the major challenges of education nowadays. It is a tool for learning when used to determine what students know, what students need to know, what students have learned, and what instructions will help students make their progress.

To assess students' outcomes good language tests should be designed. A proper language test must be developed to provide an opportunity for students to show their ability to perform certain language tasks.

Also, tests enable teachers to find out which parts of the language program cause difficulties for the group. In this way, the teachers can evaluate the methods and materials they use.

Although language testing is a fundamental part of learning and teaching today, it is a complex process at any level as it must be based on theory as well as practice.

The purpose of a test must be clear for valid interpretations to be made on the basis of the test scores. It's evident that tests have different purposes. For example, one test may be used to evaluate students' readiness to advance to the next grade, while another one evaluates students' need for remedial instruction. It is also important to outline the specific interpretations that will be made based on the scores. For example, tests used as a criterion for high school graduation will affect students differently than tests designed to inform instructional decisions [11, p. 8].

As part of the process of creating and reviewing test material to ensure that it is appropriate and accessible to examinees, it is important that item developers analyze each item critically to ensure that it only measures the content and skill that the item is supposed to assess [11, p. 12].

Analysis of recent research and publications. Some issues concerning assessment of academic achievements of students have been reflected in

the scientific literature of Ukrainian authors (I. Zymnia, S. Nikolayeva, G. Buschak, O. Matsera, D. Oseryni and others). Some authors have worked out theoretical foundations of test controlling (S. Nikolayeva, O. Pertraschuk). Different approaches to determining the criteria for evaluating student learning outcomes are being studied by A. Cohen, G. Fulcher, F. Davidson, C. Harsch, L. Vrotkevych, B. Glovatskaya. Holistic and analytic scoring methods, assessment criteria in writing tests, issues of validity and reliability have been contemplated by T. Lumley, C. Harsch, A. Pollitt, T. McNamara, L. Hamp-Lyons, etc.

In his book, "Exploring Language Assessment and Testing", Anthony Green successfully presents an overview of past and current issues in language testing with a focus on classroom-based assessment [2].

Andrew D. Cohen's emphasis on the use of multiple means of assessment, as well as his judicious treatment of currently popular approaches, is particularly valuable [3].

The British Council has been actively involved in language test development: a lot of important and innovative work was done in this area. It has developed specific expertise to ensure quality and compliance of the whole assessment process [13].

The purpose of the article is to substantiate the main components of test development; to cover some peculiarities of assessment of students' academic achievements in a foreign language; to present the main assessment criteria for speaking, reading, listening and writing; and to describe such important principles of constructing language tests as validity, reliability and authenticity.

Well-designed tests are aimed at specifying areas of difficulties experienced by the group or individual students so that additional training and corrective exercises can be done.

Presenting the main material. Test development is a sophisticated process consisting of several phases. Having decided the purpose of a new test and what kind of test it's going to be, it's better to start with a planning phase to identify the potential test users. Then comes the design phase when the test specifications, text features and the range of topics are developed and sample materials are written. This often leads to changes in designing tests. Then we enter an operational phase which involves the construction and administration of a complete test. In the monitoring phase, performance on the new test is reviewed and all materials are revised and updated to make sure they continue to match the original aim [2].

In terms of planning, teachers often produce tests for specific classes following a schedule. Teachers know their students well so they don't need to write a long list of their characteristics in the design phase. But it's still a good idea to develop a profile of a class that can be updated every year [12].

Then we should consider what they know about the topic and whether they all have roughly the same background knowledge or if it's quite varied. Of course, it helps to estimate how strong or weak the group's language is. It's not a bad idea to make a table to include all the information you need.

- What's the purpose of the test?
- Which skills and language elements do you want to test?
- How long should they be?
- Which test methods (multiple choice or gap-fill) will be used for each section?
- How many items do we need for each section?
- Are they all going to count equally or are some are going to get more marks?
- Do they have enough or too much time?
- Are the instructions clear? Do the students need to ask for clarification?

If we follow all these stages, we'll create a test that gives us reliable and useful information about your students' ability. And that information will help them learn more effectively, which is surely what we all want.

If we need to assess students' listening skills there are some factors we have to take into consideration. These are the test taker's needs, the kind of information about their listening skills and the most appropriate ways to assess these skills so that the test is appropriate. Listening is a vital skill but it can be challenging to assess.

When students listen they have to understand context, attitude, structures and vocabulary used, body language, gestures. Listening tasks must try to copy real life use of language. And the types of tasks we can use to test listening skills can be the following: pick out a word; recognize the correct tense; deduce meaning of unknown words from context; identify the main ideas; summarize through speaking or writing; listen for specific information [3].

Listening is considered to be the most widely used skill because communication is impossible without understanding as well as speaking.

A test of speaking also needs to be as close to the context we are interested in as possible. And there are many different approaches to assessing speaking. The most commonly used formats are: the interview – which can have one test taker and one or more examiners; the oral presentation – with one test taker speaking to a real or imagined audience; the interactive task – with two or more test takers working together, often to solve a problem or make a decision.

There are lots of different tasks that are used in tests of speaking. Some commonly used tasks are: describing something – this can be a picture, a place, a person; telling a story – this can be based on a series of pictures or it can be simply invented; comparing things – this can be real objects,

photographs or artwork, or even abstract concepts; giving some personal information – this might involve talking about your family, hobby, hometown or some experience you've had – such as a trip.

When the examiner is asked to award a score for a speaking test performance, it is usual practice that they use a set of descriptions of what to expect at different levels of ability. This set of descriptions is known as a rating scale. Some rating scales are simple in design and simple to use, others are more complex.

To create a useful test of speaking we need to think about all these things while at the same time making sure that the language the test tasks generate will reflect what we are planning to assess.

The following issues might be included into assessing speaking: being objective, giving reliable assessment across different test takers, timing, making a task clear.

Some different settings can be used to assess speaking:

- Performed live in front of one or more examiners
- Recorded and evaluated later by one or more examiners
- Talk over the phone to an examiner or to a recording device
- Communicate via a computer with the examiner

How should we award scores for speaking tests?

They must be based on what we have observed and use rating scale (a set of descriptors of what to expect at different levels of ability). We need to assess the following features: grammar range and accuracy, vocabulary range and accuracy, task fulfilment, fluency, and pronunciation.

According to G. Fulcher there are important assessment criteria for speaking:

- Accuracy (language competence, pronunciation, stress, intonation, syntax, vocabulary, cohesion)
- Fluency (hesitation, repetition, self-correction, re-selecting lexical items)
- Communication strategies (overgeneralization, paraphrase, word coinage, restructuring, cooperative, strategies, non-linguistic strategies)
- Discourse competence (turn taking, openings and closings)
- Pragmatic and sociolinguistic competence (appropriateness, situational sensitivity, topical knowledge)
- Task completion (Is the outcome successful?)

[4].

But rating scales are not necessarily a description of performances, but "a set of negotiated principles that the raters use as a basis for reliable action, rather than a valid description of language performance" [9].

If we need to assess our students' reading skills, we should choose various texts and tasks. When designing a test of reading comprehension, we should make our tests as relevant and meaningful as

possible. Reading tests should always aim to be as authentic as possible, in order to fit the purpose for which they are designed.

Good reading tests consist of an efficient and suitable combination of texts and tasks. These texts should be at the right level of difficulty, and be long enough to contain enough specific details. Test-takers should get information from the text to answer the questions. It's also essential that the text won't cause offense to any test-taker.

Some different kinds of tasks are used in reading tests. Test-takers might read a text, then they are given a series of questions and asked to choose the correct answer from several alternatives, define whether the statements given are true or false or they may have to choose a sentence that best summarizes different paragraphs in a text.

Sometimes, test-takers are asked to complete a text that has words removed. Other times, they might be asked to give short written answers to questions. One of the tasks can be a 'reading into writing' task where students read the text and then summarize it in their own words. Various task types are usually included into reading tests.

Once the texts and tasks have been designed, we should write clear instructions for each task and think of the time to complete the tasks satisfactorily. And, most importantly, students should know exactly how their answers will be scored.

When we score a test, we must make sure that there is only one clear answer to each question. If the tasks require longer written answers, we have to agree the scores for grammar, spelling and punctuation, etc.

A reading test will only provide useful information about a test-taker's reading ability if the assessment is done in a meaningful way. This means choosing the right texts, deciding on the most appropriate tasks, and scoring everything efficiently. All this takes time, effort, experience and, above all, a lot of common sense.

So, a good reading test task can be characterized as relevant, meaningful and authentic one. The texts must be surely selected on the basis of students' level of difficulty and the tasks must be based on the information in the text and not with test takers' background knowledge. Texts should not cause offence to any test taker.

Test developers need to check carefully a number of things before a test is administered live: the clarity of the instructions, the time allowed, the adequacy of the scoring, and whether the questions in different tasks are straightforward [14].

If we have to assess students' writing skills, there are a number of factors you need to consider. These include considering the test taker and their needs, the kind of information you want to know about their writing skills and the most appropriate ways to elicit and assess these skills so that your test is appropriate and accurately scored.

When we write we think about the topic and the way we need to make the message work. We also think about the audience that will read our ideas. Meanwhile we choose suitable vocabulary and put it all into a grammatically appropriate form so that “the reader can make sense of what we try to get across. All this passes down from the brain to the hand or fingers, where the message magically appears” [13].

We must take into consideration what we know about the test takers. There’s little point in asking teenagers to write a business letter. We’ll never know if the reason for their bad performance is due to poor writing ability or to lack of experience of business writing.

In the real world we write quite a few different kinds of things so, if we want to know about learners’ overall writing ability, we really need to ask them to write a range of different things. For example, it’s much easier to write a short note than a long essay. And we need to think about the audience, because we use very different language to write to a friend than when writing to our boss.

Scoring has to be done using a rating scale. The scale can be global, or holistic, where a single number or grade is awarded. It can also be analytic, where separate scores or grades are awarded for different aspects of the work, for example, grammatical accuracy or range of vocabulary. It’s very important that the descriptions in the scale represent the kind of things that are relevant to the decisions we plan to make.

Thus, assessing writing is a complicated process. However, if we think systematically about as many of the things mentioned here as possible, then at least we can be certain that our tests are appropriate, fair and accurately scored [1].

The main assessment criteria for writing are task fulfillment (content, communicative effect, audience, style, genre); organization (macrostructure, line of argument, coherence, cohesion, paragraphs); language (vocabulary, grammar, accuracy, orthography).

We can make assessment more objective by using validated criteria organized as a set of bands. For example, we could try using the Assessment Criteria offered for the IELTS exam: e.g., for essays and for speaking.

Some general guidelines should be considered when writing all types of items. We should write items to measure what students know, not what we know, or what they do not know. Test developers have to avoid humorous items as they may cause students to either not take the exam seriously or become confused or anxious.

Each item should be as short and verbally uncomplicated as possible. So it should be given as much context as necessary to answer the question, but any superfluous information must not be included. Each item has to be independent from other items.

A different point in each question is to be tested. We should avoid items based on personal opinions unless the opinion is qualified by evidence or a reference to the source of the opinion [15].

Test developers often claim that their test is valid or that it’s been validated. But what is actually meant by the term valid? In fact, there are two different opinions here. One argues that validity is a feature of a test – the test does what it claims to do. The other claims that it is a feature of the decisions that are made based on test performance.

These days, reliability is considered to be an aspect of validity and forms part of the argument that the scoring system is working well. It’s important to remember that a test itself can’t be valid. Instead, we build an argument based on evidence gathered to support the use of the test for making specific decisions about specific test takers in specific situations. So, while a test may be validated for use in one situation, this does not necessarily mean that it can be used in other situations without additional evidence.

Unfortunately, the cost of gathering and presenting the validity evidence means that many test developers never get round to doing it – so, in some ways, “the concept of validity remains more of an aspiration than a reality” [7].

Moreover, a test must be authentic; the language used must be as natural as possible. It should include meaningful and relevant topics, provide some thematic organization to items, offer tasks that copy real life use of language [2].

Conclusions and suggestions. Educators create tests to measure their students’ knowledge of specific content. Tests are aimed to evaluating student learning, skill level growth and academic achievements at the end of a unit or semester. They are used to determine whether students have learned what they were expected, to measure learning progress and achievement and to evaluate the effectiveness of educational programs.

Thus, a teacher can analyze the results to see which students did well and who needs more work. This information may help the teacher determine student strengths and weaknesses or to use differentiated instructional strategies.

It’s very important for test takers to understand how the tests will be scored. That’s why test developers should make test items appropriate and fair. And we also must use validated criteria for accurate and just assessing students’ answers. Besides, we must inform our students about assessment criteria for writing, reading, listening and speaking in details.

We need to be accountable to our learners to ensure we are carrying out our role as an assessor correctly. Our learners should know why they are being assessed and what they have to do to meet the assessment criteria.

One of the best ways to help students understand what will be assessed is to establish and discuss the assessment criteria with them. Working with students to develop rubrics and other assessment tools is a powerful way to help students build an understanding of what a good product or performance looks like.

The advantages of tests have been verified in practice, however, not all the controversial testing issues have been sufficiently analyzed and resolved. We have to be able to exhibit technical quality of a test by achieving certain standards, to use and develop effective test items following appropriate guidelines, and grade the tests effectively and fairly.

References:

1. Alderson, C. Bands and Scores. In: Alderson, C. & North, B. *Language Testing in the 1990s: The Communicative Legacy*. London : Macmillan, 1991. P. 71–86.
2. Anthony Green. *Exploring Language Assessment and Testing: Language in Action*. New York : Routledge, 2014. 272 p.
3. Cohen, A. *Assessing Language Abilities in the Classroom*. Boston : Heinle&Heinle. Council of Europe, 1994. 368 p.
4. Fulcher, G., Davidson, F., Kemp, J. Effective rating scale development for speaking tests: Performance Decision Trees. *Language Testing*. 2011. № 28 (1). P. 5–29.
5. Hamp-Lyons, L. The Challenges of Second-Language Writing Assessment. In E. White, W. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Politics, Policies, Practices* (pp. 226–240). New York : MLA. 1996.
6. Hamp-Lyons, L., & Kroll, B. Issues in ESL Writing Assessment: An Overview. *College ESL*. 1996. № 6 (1). P. 52–72.
7. Harsch, C. & Martin, G. Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*. 2013. № 20 (3). P. 281–307.
8. Harsch, C. & Martin, G. Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*. 2012. № 17. P. 228–250.
9. Lumley, T. Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*. 2002. № 19 (3). P. 246–276.
10. McNamara, T. *Language Testing*. Oxford : Oxford University Press, 2000.
11. Pollitt, A. & Murray, N. What raters really pay attention to. In: Milanovic, M. & Saville, N. (eds): *Language Testing 3 – Performance, Testing, Cognition and Assessment*. Cambridge : CUP, 1996. P. 74–91.
12. Ramadan, M. *Writing Effective Test Items: The Definitive Guide*. A publication of elttguide, 2018. 38 p.
13. O'Sullivan, B. *Adapting Tests to the Local Context*. New Directions in Language Assessment, special edition of the JASELE Journal. Tokyo : Japan Society of English Language Education & the British Council, 2016. P. 145–158.
14. O'Sullivan, B. IELTS. In Christine Coombe (ed) *Assessment Volume*. TESOL Encyclopedia of English Language Teaching (Editor in Chief: John I. Liantas). TESOL/Wiley, 2018.

Жукова Т. В. Особливості складання тестів та оцінювання досягнень студентів при вивченні іноземної мови

У статті обґрунтовано важливість оцінювання та його основні цілі у процесі навчання. Оцінювання використовується для перевірки того, чи відбулося навчання, наприклад, із метою перевірки набутих навичок, компетентності, рівня знань, розуміння. Якби ми не оцінювали, ми б не знали, чи було навчання, яких успіхів було досягнуто, які труднощі треба подолати.

Оцінювання має бути зосереджено на вдосконаленні навчання, а також вимірюванні досягнень. Це допомагає учням визначити свій прогрес і те, що їм потрібно зробити для вдосконалення та подальшого прогресу. Оцінювання є регулярним процесом; воно не може бути формальним, і ми повинні спостерігати за тим, що роблять наші учні, ставлячи їм питання та переглядаючи їх прогрес, коли ми контактуємо з ними.

Проаналізовано найважливіші речі в розробці завдань тестів. Завдання має відповідати вміню, що оцінюється. Тест має бути справедливим до учасника тесту і правильно оцінений.

Описані основні фази розробки тесту: фаза проектування, коли розробляються характеристики тесту та пишуться зразкові матеріали; операційний етап, який передбачає побудову всього тесту; етап моніторингу, коли матеріали переглядаються й оновлюються. У статті висвітлюються різні типи завдань, що використовуються при складанні тестів, їх актуальність та осмисленість.

Представлені основні типи завдань, які використовуються для перевірки навичок аудіювання (добір слова, розпізнавання правильного часу, визначення невідомих слів із контексту, визначення основних ідей, прослуховування конкретної інформації).

У статті представлені основні критерії оцінювання мовлення, читання, слухання та письма. Отже, ми постійно робимо висновки про прогрес студентів і про те, як їх знання могли б покращитися. Оскільки процес оцінювання є складним, ми повинні бути впевнені, що наші тести є відповідними, справедливими та точно оціненими.

Один із найкращих способів допомогти студентам зрозуміти, що буде оцінюватися, – це обговорити з ними критерії оцінювання. Робота зі студентами над розробкою рубрик та інших інструментів оцінювання – це потужний спосіб допомогти їм зрозуміти, як виглядають відповідні тести та вимоги до них.

Сучасне вчительство має дбати, щоб тестування мало позитивний вплив на процес навчання та мотивацію студентства.

Ключові слова: оцінювання, розробка тестів, вагомість, надійність, критерії оцінювання, мотивація, навчальні досягнення.