Т. 7. №1. 2021
https://doi.org/10.33619/2414-2948/62

UDC 37.022 https://doi.org/10.33619/2414-2948/62/49

# TEST DESIGNING PRINCIPLES AND RELATED PROBLEMS

**©Kertaeva Z.,** *Alisher Navoi Tashkent State University of the Uzbek Language and Literature Tashkent, Uzbekistan, zaurekertaeva@gmail.com*

# ПРИНЦИПЫ ПРОЕКТИРОВАНИЯ ТЕСТОВ И ПРОБЛЕМЫ, СВЯЗАННЫЕ С НИМИ

**©Кертаева З. С.,** *Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои, г. Ташкент, Узбекистан, zaurekertaeva@gmail.com*

*Abstract.* The article discusses the role of following assessment principles while designing tests. Problems and obstacles of test designing and administering at national higher education context are illustrated, being grouped into seven respective principles.

*Аннотация.* В статье обсуждается роль следования принципам оценивания во время проектирования тестов. Проблемы и препятствия в разработке и проведении тестов в национальном контексте высшего образования, которые сгруппированы в семь соответствующих принципов, проиллюстрированы.

*Keywords*: test designing principles, usefulness, validity, practicality, reliability, authenticity, washback, transparency, test-taking competence, achievement test, proficiency.

*Ключевые слова:* принципы проектирования тестов, полезность, достоверность, практичность, надежность, подлинность, обратное влияние, прозрачность, компетентность при сдаче теста, тест достижения, мастерство.

## *Introduction*

Many developing countries are having reforms in their education system, inclusively in assessment. Uzbekistan, being post-Soviet Union country, is also experimenting a great number of evaluation frameworks and tools without yet resulting in application of consistent assessment system. In foreign language learning and teaching, assessment literacy is under great attention and investigation by many scholars of language teaching methodology, being claimed as an important tool and concept in forming and developing effective and meaningful assessment framework.

Before we go down to provide theoretical base for EFL assessment and justify the role of assessment literacy, the term itself should be defined. According to Rogier, generally, teachers who are literate at assessment are knowledgeable about the principles, practices and key concepts of testing and decisions surrounding their usage. [1] A link of assessment to learning and teaching process, which is achieved by matching it to instructional objectives is another essential element of assessment literacy.

The problems arisen due to absence or lack of assessment-literate professionals can be the following (in experience of higher education system). Misconception of students about the real purpose of assessment. Teachers often feel unprepared to test designing or taking; and they do not know how to set purposes for assessing. The function of assessment is often limited with summative objectives. Instructional objectives of the course do not match the form or content of testing.

Т. 7. №1. 2021
https://doi.org/10.33619/2414-2948/62

Very few teachers are able and willing to interpret the results of tests and use them for further development of the course or teaching skills. Students are not competent enough at test taking. Absence or limit of using more practical and authentic methods of assessing because of higher institutions` conservative control and requirements. Students may be reluctant to accept and receive alternative ways of assessment.

Below each of the problems will be discussed thoroughly grouping them into different principles of test designing. Rogier classifies them into seven key concepts – usefulness, reliability, validity, practicality, authenticity, washback and transparency [1]. Being aware of all these principles can assist a teacher to become assessment-literate and prevent related problems.

## *Main Part*

*Usefulness*

With reference to Bachman and Palmer [3], usefulness is the most essential consideration when selecting or designing a test. This criterion for test designing is closely connected with the purpose which means all language tests are supposed to be developed with a specific purpose and should be congruent with teaching aims and content. International tests of English can be prime examples where purpose is specific and oriented to particular audience

.

| *Cambridge exams* | *The purpose and audience* |
|---|---|
| IELTS | Has an academic and general-training version. The academic version is for those who want to study in English-speaking universities; the general version focuses on basic "survival skills in broad social and workplace contexts" [IELTS 2013] |
| Level based tests | Is an acidic test. For those who intend to test their proficiency in English and apply for related educational institutions or occupations. E.g CPE is for those who want to work as EFL teacher.(examenglish.com) |
| Business exams | Intended for those who learn English as a specific (business) purpose. E.g BEC is for students who are studying business. [examenglish.com] |

If we explain this feature of test developing in teaching context, many teachers do not consider the purpose and audience of the test before administering it. Instead they tend to use pre-made test without reviewing its suitability to the age and background of language learners. For example, pre-intermediate level for children/teenagers and adults differ and the content of the test should respectively be specific for age groups which many teachers fail to consider. Another problem can be about the purpose of testing. In particular, if we would like to test reading skills of a learner who want to migrate to an English-speaking country, giving them to read a fictional story and asking him/her to retelling it would be out of use as a person who is going to live in a foreign country would need the skills such as understanding road signs, announcements, menus at cafes etc.

Although teachers of English department at Tashkent State Uzbek Language and Literature University try to relate English lessons to students` major (translation), in terms of test designing, we cannot claim teachers always take the students` future needs and interests into consideration. All this derives from lack of materials and assessment literacy levels and competence of local teachers.

*Reliability*

Reliability refers to the consistency of test scores. This means a student taking the same test for many times should show the same or similar results. Brown defines reliable tests as being unambiguous to test taker; being consistent in its conditions across two or more administrations; giving clear directions for scoring and evaluating; having uniform rubrics for scoring and evaluating; and lending itself to consistent application of those rubrics by the scorer [2]. Moreover,

Т. 7. №1. 2021
https://doi.org/10.33619/2414-2948/62

he classifies principle of reliability into four types: student – related reliability (involves student factors of illness, physiology, anxiety, test-wiseness, test taking efficiency); rater reliability (involves factors of human error, subjectivity, bias; test administration reliability (involves factors of test tasking conditions and quality of test materials); test reliability (involves factors of equal difficulty of test, reliable and unambiguous distracters).

To understand the case of unreliability, a few examples should be illustrated. Although most factors causing student-related unreliability are their temporary illness, fatigue or anxiety; the main problem in our context is test- wiseness of our students. Many students are not competent to understand the instructions and follow them. From my personal experience as a university teacher and writing instructor, during the final exams majority of the students tend to ask extra questions or need extra help despite the fact all instructions are clearly illustrated in the exam paper and orally explained during the exam and before the exam in classes. For example, even though it is written "Write in pens only", they ask again "Can we write in pencils"; or despite the note "Copy your answers to an answer box", they ask "Do we have to copy the answers to an answer sheet?". Another instance is that most of the test takers fail to manage their time, they do not manage to copy from their draft and ask for extra time violating the exam rules. Moreover, if the task says to write answers as letters, they write the words as an answer, which causes hurdle and discomfort during the test checking process. This all could be related to the fact that students have not been taught to carefully read and follow instructions or they are not aware of culture of test taking, which shows their test taking incompetence. As a researcher and a higher education professional, I consider the issues mentioned above need to be investigated and the ways to improve the situation should be suggested and tried out.

Our local problems concerned with rater-reliability usually happens when test designers take ready-made speaking test rubric from the Internet, which often do not coincide with test takers` level and the purpose of exam. Besides, speaking test examiner-teachers are not pre-instructed about how to use those rubrics, they fail to follow the rubric or teachers may just refuse to mark students` speaking abilities according to assessment criteria, instead they may give scoring relying on biased "good or bad speaking" perceptions. The second case is usually the result of great number of test takers for one speaking examiner-teacher, therefore preventing teachers from giving objective, unbiased and reliable scoring.

At university exams test administration reliability cannot be achieved, especially when students have to seat very close to each other in the exam room, allowing them to copy from each other or causing discomfort to take a test.

*Test reliability*, from my own observations, occur when teachers design open ended questions without providing a specific assessment criterion for students to make them aware of what is expected to be tested. The criteria may be about whether students are supposed to give brief or extended answers; if they are checked for grammar range and accuracy or for task achievement only; or whether they have to justify their answers with relevant examples or they do not play a role in marking. This again may lead to additional questions from students or confusions for teachers during the exams and result in very subjective and unreliable test results.

*Validity*

Validity, being considered to be the most complex, yet the most important principle of test designing, is "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment" [5, p. 226]. Another expert in validity, Samuel Messick defines it as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of

Т. 7. №1. 2021
https://doi.org/10.33619/2414-2948/62

inferences and actions based on test scores or other modes of assessment" [5, p 11]. From this we can infer that a valid test assesses what is intended to assess; provides useful, meaningful information about test- taker`s ability; and is supported by a theoretical rationale or argument. It is important to note that, according to Messick complete validity cannot be achieved or there is no absolute frame of measuring validity, yet it can be provided to some extent (the greater the extent, the more valid it is) [4]. Brown defines validity through four different forms of evidence: content (related to objectives and their sampling); construct (referring to the theory underlying the target); criterion (related to concrete criteria that it should reach to); consequential (correlating high with another measure already validated and capable of anticipating some later measure) and face (related to test`s overall appearance: whether test takers see it as a fair, unbiased and objective test) [2].

The greatest challenges in effective assessment at the university context are about content validity. Teachers at the university often fail to achieve content validity due to many external and internal factors. External factors can be related to the fact that requirements of authorities do not always correlate with course objectives. From my experience at Tashkent State University of the Uzbek Language and Literature (TSUUL), the syllabus for Fall Term in writing course mainly intends to develop students` paragraph writing skills and competence; however, as authority required it to design multiple choice tests for final exam, the test could not measure what it intended to measure. For, multiple choice tests could, at maximum, check students` awareness on paragraph development and grammar knowledge rather than their competence of developing effective paragraphs. Another case could be about specific listening sub-skills. During the particular period, students were trained to listen for gist, they were listening to recordings, answering comprehension based questions and discussing them generally. However, for a mid-term test, they did listening where they had to fill in a form listening to specific details. Here the test form, which tests students` ability to extract particular detail, could not measure achievement of the course content (which prepares students for general listening comprehension).

Criterion related validity is in most cases overlooked by test designers and university professionals. For example, students who have the certificate of English at B1 level verified by state testing centre do not show similar results in university exams which are also designed according to B1 level requirements. Alternatively, due to the fact tests are not designed effectively, it may not truly represent students` general performance on the course or skill — it may be too high or too low from actual case. Especially, in current national assessment system, where final exam grade is accepted as an overall course grade, invalid tests are risky giving false results and turning an excellent student into a student with poor results or vice versa. All the abovementioned were cases of how concurrent criterion related validity, which is a principle of showing students` current true performance, is not achieved. The predictive criterion related validity, a principle of measuring a test-taker`s likelihood of future success, is also often underestimated, as many teachers do not analyze or compare the results of previous tests with new ones to see whether predicted success of a student has been confirmed.

With reference to Brown construct- related evidence/validity is a model which refers to the theory underlying the target (proficiency, fluency, accuracy) [2]. For example, from my own observations, there was a case when students were asked to make presentations for at the end of speaking course and were just evaluated according to general speaking skills. However, the criteria such as body language, interaction with audience, public speaking skills, persuasiveness, novelty; which were essential features of real-life presentations, could have been taken into consideration. That means assessment did not measure students` particular competence but general language proficiency and led to mismatch of form and assessment criterion.

Т. 7. №1. 2021
https://doi.org/10.33619/2414-2948/62

As to Brown, “Consequential validity encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its effect on the preparation of test-takers, and the(intended and unintended) social consequences of a test`s interpretation and use” [2, p. 34] This validity principle is similar to washback effect, which will be discussed further later.

Face validity is achieved when students consider the test as objective, to the point and helpful in developing skills [3]. Similarly to what Bachman claims “it is purely factor of the `eye of the beholder`”, it is about an opinion of test-taker [3]. Therefore, it is highly unlikely to be empirically measured or theoretically justified under the category of validity [2]. Although the following statement is not justified with objective survey results and analysis; when I ask students` opinion about language tests in their majors, majority of them tend to have negative impressions concerned with test-paper quality and format, its user-friendliness, unrehearsed tasks. Besides, usually they have fixed misconceptions that teachers are interested in students` failure, which affects their performance in test to great extent.

To avoid face invalidity, Brown suggests test designers to make tests well-constructed with expected format and familiar tasks; with clear instructions; at a level of difficulty with a reasonable challenge; and manageable to allotted time [2].

*Practicality*

In accordance with Rogier, practicality is about how teacher friendly the test is including the cost of test developing, time for administration, convenience of marking and presence of appropriately trained markers. [6]. Brown, along with those, adds the factor of not exceeding available material resources [2]. The former scholar also claims that the amount of time and effort we spend for test designing and marking should be relevant to its worth in overall course mark. He thinks it would not be sensible to spend two hours for each student`s paper if the test is only for continuous formative assessment and has very little point distribution in overall grading. In fact, some teachers decide to test just the day before or that day and make up some open ended questions on covered topics. They forget to specify word and time limit, and assessment criteria. This causes teachers to wait for unlimited and unpredicted time until last student finishes, and spend their whole weekend to check too long or too short written answers which steal their valuable time or prevent from getting expecting answers. A careful planning, clear instructions and assessment criteria would aid teachers to design teacher-friendly tests.

*Authenticity*

The principle of authenticity describes the suitability of the form and content of the tests to real - world situations [6]. Bachman and Palmer also give a similar definition to the concept of authenticity “the degree of correspondence of the characteristics of a given language test task to the features of a target language task” [7, p. 23]. As stated by Brown, authenticity is achieved with representation of natural language; contextualization (not isolation) of items; inclusion of meaningful, relevant and interesting topics; provision of some thematic organization to items; and offering tasks that reflect real – world demands.[2]

Many scholars including Bachman & Palmer [3], Lewkowicz [8] and Brown [2] considered this principle difficult to define, measure and reflect in assessment yet did not underestimate its importance. Another expert, Chun [7] believes many test types fail to stimulate real-world tasks. However, I hold the opinion that the level of authenticity in assessment has increased with the introduction of communicative approach. As one of the main objectives of this approach is to enable students to use language skills learnt in the classroom in a real- world communication and interaction, it encourages using authentic materials, tasks and methods in both teaching and evaluation. Replacement of story retelling with oral presentations as a speaking assessment or

Т. 7. №1. 2021
https://doi.org/10.33619/2414-2948/62

introduction of alternative communicative skills — based language tests instead of multiple choice grammar and vocabulary tests serve as justification to the viewpoint put forward above. It is important to note that available time and resources do not always allow EFL teachers to think of or apply authentic ways of assessment, specifying it to the future needs and purposes of students.

*Washback*

Washback effect is the influence of testing on teaching and learning. It is also the potential impact that the form and content of a test may have on learners` conception of what is being assessed and what it involves. Therefore, test designers, deliverers and raters have a particular responsibility, considering testing process may have a substantial impact, either positive or negative. In particular, Messick [4] emphasized the role of impression the students would have after the test claiming that they play a part in promotion and inhibition of learning both in beneficial and harmful way, while Spratt [6] considered teachers should become agents of beneficial washback in their language classrooms. Overall, as concluded by Brown [2], a test with beneficial washback has positive influence on what and how teachers teach; has positive impact on what and how learners learn; creates adequate preparation opportunities for students; provides constructive feedback to boost their language progress; is more formative rather than summative; allows learners to reach maximum results; and encourages a number of basic principles of language learning such as intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, strategic investment.

Another important idea maintained by Brown [2] is that informal performance assessment has more tendency to give positive washback rather than official assessment as the latter consists of only letter grade or overall numerical score. To be more specific, when teachers assess students` a written work, they can give a thorough feedback which includes comments on both achievements and problems of the student. Consequently, even the grade is not an expected result, consideration of their strengths by their teacher will motivate them to work on their problems rather being focused on weaknesses or low grade.

The duty of teachers is to create classroom tests with positive washback which can praise what students have successfully acquired and where the learning gaps can serve as insight into further development. However, this is not always feasible for several reasons. One of them can be that teachers are usually overloaded with unaffordable number of teaching hours or their additional duties which leave no time to reflect and work on their students` test results. Unawareness of teachers about this principle and its importance can also hinder achievement of positive washback. Assessment`s becoming part of more formal grading and administrative procedure rather than language learning and teaching is another obstacle to apply this principle into tests; as especially after final tests, teachers, being busy with scoring and reporting about it to administration, are not able or willing to feedback students` results.

*Transparency*

As to Rogier, transparency is about the availability of information to students. In particular, students beforehand should be aware of what is going to be tested, what they are expected to learn, how they are going to be assessed and graded [6]. Importantly, if students cannot do well in a test, the reason for this should be lack of their preparation and knowledge, not misunderstandings of directions or format of the test. For instance, IELTS has a number of official publications which present detailed explanation of assessment criteria, objectives, and audience of the test. In general, transparency facilitates objectivity and reliability of assessment by making students a part of it and allowing them to understand the link between course and assessment objectives, which every effective teacher is supposed to provide.

Т. 7. №1. 2021
https://doi.org/10.33619/2414-2948/62

*Conclusion*

To sum up, improving literacy in the field of language assessment is of great importance as it creates a bond between teaching and learning, teacher and student, course objectives and final achievement, learners needs and society`s interests. Without being aware of the principles of effective assessment, EFL teachers will not be able to

meet the needs of students and course objectives through assessment (usefulness)

achieve consistent/representative measurement (reliability);

measure what is supposed to measure (validity);

make the test designing, administering, scoring and interpreting easy (practicality)

prepare students to real – world experiences through tests (authenticity)

make students feel satisfied with achieved results and further motivate them to work on existing weaknesses (washback)

make their students involved in assessment process and ensure their understanding of course aims (transparency).

In short, only when these concepts are acquired by teachers, do they become assessment-literate professionals and bring a positive change to language learning and teaching.

*References:*

1. Rogler, D. (2014). Assessment Literacy: Building a Base for Better Teaching and Learning. In *English Teaching Forum* (Vol. 52, No. 3, pp. 2-13). US Department of State. Bureau of Educational and Cultural Affairs, Office of English Language Programs, SA-5, 2200 C Street NW 4th Floor, Washington, DC 20037.

2. Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (Vol. 10). White Plains, NY: Pearson Education.

3. Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

4. Messick, S. (1996). Validity and washback in language testing. *Language testing*, *13*(3), 241-256. https://doi.org/10.1177/026553229601300302

5. Gronlund, N. E. (1998). *Assessment of student achievement*. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.

6. Spratt. M. Washback and the Classroom. // Language Teaching Research. 2005. V 19, p13-20

7. Chun, C. W. (2006). Commentary: An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly: An International Journal*, *3*(3), 295-306. https://doi.org/10.1207/s15434311laq0303_4

8. Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language testing*, *17*(1), 43-64.

*Список литературы:*

1. Rogler D. Assessment Literacy: Building a Base for Better Teaching and Learning //English Teaching Forum. US Department of State. Bureau of Educational and Cultural Affairs, Office of English Language Programs, SA-5, 2200 C Street NW 4th Floor, Washington, DC 20037, 2014. V. 52. №3. P. 2-13.

2. Brown H. D., Abeywickrama P. Language assessment: Principles and classroom practices. White Plains, NY : Pearson Education, 2010. V. 10.

3. Bachman L. F., Palmer A. S. Language testing in practice: Designing and developing useful language tests. Oxford University Press, 1996. V. 1.

4. Messick S. Validity and washback in language testing // Language testing. 1996. V. 13. №3. P. 241-256. https://doi.org/10.1177/026553229601300302

5. Gronlund N. E. Assessment of student achievement. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele, 1998.

6. Spratt. M. Washback and the Classroom. // Language Teaching Research. 2005. V 19, p13-20

7. Chun C. W. Commentary: An analysis of a language test for employment: The authenticity of the PhonePass test // Language Assessment Quarterly: An International Journal. 2006. V. 3. №3. P. 295-306. https://doi.org/10.1207/s15434311laq0303_4

8. Lewkowicz J. A. Authenticity in language testing: some outstanding questions // Language testing. 2000. V. 17. №1. P. 43-64.

*Работа поступила*
*в редакцию 07.12.2020 г.*

*Принята к публикации*
*12.12.2020 г.*

Ссылка для цитирования:

Kertaeva Z. Test Designing Principles and Related Problems // Бюллетень науки и практики. 2021. Т. 7. №1. С. 426-433. https://doi.org/10.33619/2414-2948/62/49

Cite as (APA):

Kertaeva, Z. (2021). Test Designing Principles and Related Problems. *Bulletin of Science and Practice, 7*(1), 426-433. (in Russian). https://doi.org/10.33619/2414-2948/62/49