ALANYAZIN
Eğitim Bilimleri Eleştirel İnceleme Dergisi
CRES Journal Critical Reviews in Educational Sciences

The purpose and principles of testing in ELT

Koray Kaçar[1]

Abstract: Language is learned for different reasons. As a result, the aim of the testing differs. Testees sit for an exam for various reasons, which causes preparing different tests to utilize them. Applying a test with considering its purpose increases the validity. Excluding validity, it is impossible to say that the test assesses the performance. Consequently, all the basic principles are linked to each other. On the other hand, language tests are designed for a purpose. Each test type has different aims. Similar to the testing principles, each exam type is connected to one another in some aspect. In this context, the purposes of foreign language exams are explained in detail. Then, the generally accepted principles are identified elaborately. Finally, essential opinions are advanced so as to produce well-prepared and appropriate test. Correspondingly, the review mainly tries to shed light on the purpose and principles of testing in order to produce appropriate exams containing the fundamental principles. Furthermore, it is aimed to give a short and clear guidance to the EFL teachers to create better exams..

Keywords: *ELT, testing, assessment, principles of testing, purposes of testing*.

ELT'de ölçmenin amaç ve ilkeleri

Özet: Dil farklı amaçlar için öğrenilir. Bunun sonucunda da ölçmenin amacı değişir. Öğrenenler farklı amaçlar için hazırlanan farklı sınavlara girer. Amacını düşünmeden bir sınav hazırlamak sınavın geçerliliğini düşürür. Geçerliliğin ihmal edilmesi onun istenilen performansı ölçmesini engeller. Dolayısıyla, temel test ilkeleri birbirleri ile ilişkilidir. Öte yandan, dil sınavları farklı amaçlar için oluşturulur. Her sınav türünün farklı amaçları vardır. Test ilkelerinde olduğu gibi sınavların amaçları da bazı açılardan birbirleriyle bağlantılıdır. Bu bağlamda, ilk olarak yabancı dil sınavlarının amaçları detaylı bir şekilde açıklanmıştır. Daha sonra, genel olarak kabul edilen ölçme ilkeleri ayrıntılı bir şekilde tanımlanmıştır. Sonuç olarak iyi hazırlanmış ve amacına uygun sınavların hazırlanması için bazı fikirler verilmiştir. Tüm bunlar bağlamında, bu derlemenin maksadı uygun testlerin hazırlanabilmesi için sınavların amacı ve ilkelerini açığa kavuşturmaktır. Ayrıca, İngilizce öğretmenlerinin daha iyi sınavlar hazırlayabilmeleri için kısa ve net bir rehberlik etmek hedeflenmiştir

Anahtar Kelimeler: *ELT, ölçme, değerlendirme, ölçmenin ilkeleri, ölçmenin amaçları..*

Başvuru/Submitted
22 Şub / Feb 2023
Kabul/Accepted
23 May / May 2023
Yayın/Published
31 May/May 2023

DOI:
10.59320/alanyazin.1255021

İnceleme Makalesi
Review Article

Alanyazın Eğitim Bilimleri Eleştirel İnceleme Dergisi
CRES Journal
Critical Reviews in Educational Sciences
2023, 4/1, 33-41

Kaçar, K.. (2023). The purpose and principles of testing in ELT, *Alanyazın*, *4*(1), 33-41.

Öncü Okul Yöneticileri Derneği
e-ISSN: 2718-0808

[1] MEB, koray_kacar22@hotmail.com ORCID:0000-0002-0258-0016

Introduction

In language assessment, tests are designed for different purposes. For a successful assessment, the goal should be decided accurately beforehand. Besides the purpose, the principles of testing should be taken into consideration. These principles are the innermost parts of a language test (Jumaniyozova, 2021). In other words, a well-decided purpose of an exam and adhering to the main principles of testing are the basic conditions for an accurate assessment. Furthermore, the results of testing are the main data to determine the quality of the education system and to advise the policy makers to update the system (Wurman, 2022).

Purposes of Testing

Tests are designed and administered for different purposes in education. In addition, their purposes are the means to categorize tests (Brown & Abeywickrama, 2010). Generally; achievement tests, proficiency tests, diagnostic tests, placement tests, and aptitude tests are considered for testing purposes (Gonzalez, 1996; Brown, 2003; Brown & Abeywickrama, 2010; Ivonava, 2011). As well as these types, Rahman and Gautam (2012) mention prognostics tests as a purpose of testing.

Achievement tests, or progress tests, are designed and used to evaluate test-takers's language and skill progress based on the syllabus (Harmer, 2007). Achievement tests are restricted to specific objectives (Brown & Abeywickrama, 2010). According to these objectives, it can be mentioned that two kinds of achievement tests are available: final achievement and progress achievement (Hughes, 2003). The former ones occur at the end of a course and are often summative (Hughes, 2003; Brown & Abeywickrama; 2010). On the other hand, the latter ones are formative and are often designed to monitor the progress of students during the course (Harmer, 2007; Demirezen, 2013). Achievement tests can be internal and external; the internal tests are prepared and scored by the lecturers but the external ones are written and checked by a tester out of the institute (Riddel, 2003). That is to say, progress achievement tests are planned to measure the progress learners have made whereas final ones are given whether the objectives that set at the beginning are reached. Also, the learning problems can be detected earlier due to the formative tests, which enables teachers to correct them conveniently.

Proficiency tests are not confined to any specific course or curriculum and measure general language competence (Gonzalez, 1996). Proficiency tests measure general language ability and desire to prepare testees for a particular communicative role (Desheng & Varghese, 2013). As Rahman and Gautam (2012) suggest proficiency tests are designed to reveal how much of a language a person know.

Placement tests are intended to assign testees to a program by their knowledge and abilities. These tests are planned to get information about test-takers to place them in the most appropriate program (Ivanova, 2001). They are generally for new students (Harmer, 2007). As Brown (2007) emphasises since these tests consist of items from the curriculum, they have content validity. Ideal placement tests should be neither easy nor difficult and should serve a diagnostic purpose (Brown, 2007; Brown & Abeywickrama, 2010). Placement tests give the opportunity to classify the academically similar students together.

Diagnostic tests are developed to diagnose a particular aspect of language (Brown, 2007) which test-takers need to build (Brown & Abeywickrama, 2010). They aim to determine the testees' weaknesses and strengths in this aspect (Bachman & Palmer, 1996; Hughes, 2003). They enable to improvement teaching quality by detecting defects (Gonzalez, 1996). However, Yin and Sims (2006) claim that diagnostic tests contrast with placement and proficiency tests that brand test takers' level. Yet, diagnostic tests give meaningful feedback and serve as a guide for appropriate planning. They also shed light upon the difference between diagnostic and achievement test in that achievement tests happen during or at the end of a course and are covered with the content of the system, but diagnostic tests happen independently of a system and not limited to the course content. This type of tests are generally used to identify the specific techniques to improve the deficiencies.

Aptitude tests are invented to measure a test taker's capacity and ability to learn a foreign language and success (Brown, 2007). Rysiewicz (2008) differentiates ability and aptitude as abilities are available at present but aptitudes characterise the potential for achievement. Based on many types of research about language aptitudes, Rysiewicz (2008) summarises the facts about language aptitude as independent from affective and cognitive factors, independent of academic background and steady. In other words, tests designed based on these principles provide data about test-takers preferred styles and potential abilities. Aptitude tests have limitations and flaws therefore they are rarely used as a measurement tool in education now (Brown, 2007).

All these tests have their own functions that are different from each other. Considering the differences enables the testers to design affective exams. The types of question items, the testing methods, the number of question items and time can change depending on the test type. Moreover, learners' expectations and preparation for the language tests may differ. Hence, a suitable test for a specific purpose helps testees display improved performance and get high scores.

Principles of Testing

A test must have essential measurement principles to fulfil its aim. A finite number of principles can be listed that serve as guidelines for a good test. Rajhy (2014) explains reliability, validity, practicality, discrimination, and authenticity as primary principles of a good test. Rahman and Gautam (2012) also list discrimination, reliability, validity, scalability, economy and administrability as the requirements of a test. Harmer (2007) mentions the two principles, reliability and validity, as the most important principles. Apart from these principles, Hughes (2003) attaches importance to beneficial backwash. Yet, Bachman and Palmer (1996) use the term quality instead of principle and state the qualities as reliability, construct validity, authenticity, interactiveness, impact, and practicality. They also formulised the usefulness of a test as follows:

Usefulness = Reliability + Construct Validity + Authenticity + Interactiveness + Impact +Practicality

Five principles in this formulation (practicality, reliability, validity, washback, authenticity) are universally recognised criteria (Paradowski, 2002). Brown (2007) also lists these abovementioned five qualities as five basic principles for a practical test. Brown and Abeywickrama (2010) describe these principles as five chief criteria for testing a test. The importance of these principles is that "they present a synthesis of what various assessment specialists cite as priorities for the design of language assessment'' (p.446).

Reliability is defined as the consistency of test scores (Bachman & Palmer, 1996; Brown, 2003). Namely, reliability is getting the same results from the same test on different occasions and days (Harmer, 2007). As Bachman (1990) states, reliability is the concurrence between similar measures of the same trait. He also defines reliability as "the agreement between two efforts to measure the same trait through maximally similar methods" (p. 240). External test factors, such as health, lack of interest, de-motivation, and negative situations in the testing environment should be decreased to increase reliability. The more minimised these various factors which are not related to language ability, the more maximised reliability is created. In other words, the less these elements affect test scores, the more excellent achieved the reliability of language scores (Bachman, 1990).

A reliable test is consistent and dependable. It is independent of situation, time, and administration. It provides clear guidance for scoring (Brown & Abeywickrama, 2010). As a reliable test measures consistently, testees get nearly the identical scores on the same day or the next day (Hughes, 2003). It should also include fewer errors (Akıncı, 2010).

Moreover, a test is reliable if different teachers or experts grade it and the results are similar when the scores are compared. The test is free from rater bias unless there is a difference, and it is called an objective test. Thus, objective tests increase reliability. In contrast, subjective tests need personal opinions and judgement to determine the correct answer. This situation decreases reliability of the test (Brown, 2003).

There are three aspects of reliability: stability, equivalence, and internal consistency (Bachman, 1990; Rajhy, 2014). Stability, also called the test-retest reliability, occurs when the same or similar results are obtained by testing the same group after a while (Rajhy, 2014). In this method, the test is administered twice to the same group and then the results of both tests are compared. The correlation of the results demonstrates how stable they are over time (Bachman, 1990).

Another feature is equivalence. Equivalence called parallel forms reliability, is the agreement of two or more tests administered nearly simultaneously. One way of achieving equivalence is counterbalanced design. In this design, half of the testees take one form first, the others take the other, and vice versa.

	First	Second
Half 1	Form 1	Form 2
Half 2	Form 2	Form 1

Figure 1. An example of counterbalanced test design (Bachman, 1990: 183)

Finally, the last feature is internal consistency which is interested in the reasons for errors within the test and scoring procedures (Bachman, 1990). Rajhy (2014) suggests internal consistency as concerns the degree to which items in a test or instrument measure the same thing. To estimate the reliability of the item in a trial, a number of formulae have been developed. They can be examined in two categories:

- Spilt-Half Reliability Estimates: In the method, the test is divided into two halves and then decided how these halves are consistent with each other. The formulae are the Spearmen-Brown Split Half Estimate Formula and the Guttman Split-Half Estimate Formula
- Reliability Estimates Based on Item Variances: Kuder- Richardson Reliability Coefficients, Coefficients Alpha (Bachman, 1990; Garson, 2009).

Furthermore, in estimating the scoring procedures, two reliability types are significant. They are intra-rater reliability and inter-rater reliability.

- Intra-rater Reliability: In the case of a single rater, what is important is the consistency of the results scorer obtained at different times and occasions.
- Inter-rater Reliability: When there is more than one rater, the results taken from different raters are examined and the consistency between their results is decided. (Bachman, 1990).

Hughes (2003, p.46-50) suggests the ways to increase reliability:

- Having sufficient examples of behavior
- Excluding items not distinguishing weak and strong testees
- Do not allow testees too much freedom
- Writing clear items and instructions
- Using familiar format and techniques
- Providing suitable conditions for a test
- Using objective items
- Having a detailed scoring key
- Identifying candidates by number
- Employing multiple and independent scoring

Similarly, Henning (2012) lists the reasons for diminishing reliability.

Too difficult or too easy questions

Insufficient number of items

- Lack of reliance measures
- Trick questions
- Obvious cues
- Convergence clues
- Option number
- Cheating
- Problems with instructions
- Subjectivity of scoring
- Lack of piloting
- Problems with administration

The term validity refers to whether or not the test measures what it claims to measure. According to the CEFR, a test has validity to the extent that what it is assessed can be showed (CoE, 2001). Brown (2007) also defines it as "the degree to which the test actually measures what it is intended to measure" (p. 448). Harmer (2007) clarifies validity by explaining that the test is valid so long as there is validity in the mode of grading. Moreover, Brown (2007, p. 22), Brown and Abeywickrama (2010, p.30) produce a long and explicit definition of validity by stating that "a valid test measures what it aims, does not measure irrelevant variables, offers clear information about performance, is supported by theoretical justification". Moreover, validity is the use of test scores and the ways of interpreting them; therefore, it is closely related to the aim of the test (Alderson, Clapham & Wall, 1995).

There is a close link between validity and reliability. Bachman (1990, p.227) attaches importance to validity since it is the most important feature of a test, while reliability is a necessary condition of validity. Likewise, Hughes (2003) sees validity as a prior requirement of a good test. He also states that everything to make a test reliable must be done to create a valid test. "If a test is not reliable, it cannot be valid" (p. 34). Like Huges (2003), Chapelle (1999) argues that reliability is the precondition for validity. Similar to his definition about reliability, Bachman (1990) defines validity as "the consistency of two attempts to measure the same traits through different methods. In addition, Cohen (2001) says that before being valid, a test must be reliable.

On the other hand, Paradowsky (2002, p.40) compares validity and reliability in that "Practicality and reliability are particularly significant in norm-referenced placement and proficiency tests, whereas in criterion-referenced testing the most prominent role is given to validity".

What's more, Hughes (2003, pp. 33-34) proposes the following items to increase validity:

- Writing explicit specifications for the test
- Using direct testing as long as possible
- Increasing reliability
- Defining clearly the scoring of responses relating them to what is being tested

Likewise, Hughes (2003), Henning (2012) note validity concerns on a test as follows:

- Mixed content
- Wrong medium
- Common knowledge
- Unsuitable syllabus
- Unnecessary words
- Content matching

Since validity is a broad term, validity is divided into several different types. In their studies, researchers examined different types and numbers of validity. To include all types of validity, four types are examined here: face, content, criterion, and construct. In addition, there is consequential validity. According to Bachman (1990), it refers to the term impact. Furthermore Brown (2007), Hughes (2003), Brown and Abeywickrama (2010) use the term washback instead of it. Thus, consequential validity is admitted as washback and defined below.

The first term, face validity refers to whether a test seems to be suitable and appropriate as to what is measured. Hughes (2003, p.33) defines it as ''if it seems as if it measures what it is supposed to measure''. As an example of face validity in a grammar test, vocabulary knowledge should not be tested; hence, the vocabulary ought not to be challenging. Similarly, in a vocabulary test grammar should be simple.

Content validity, the second term, is an attempt to demonstrate that the content of a test is a specimen from the domain to be tested (Fulcher & Davidson, 2007). Demirezen (2013) defines content validity as a system in which the relationship between test items and the purpose of the test is created. Each question item is supposed to match to the indicated content area. Namely, a test has content validity when it includes the correct samples of appropriate structure.

Correspondingly, criterion validity is how much a test predicts data, which is important (Rajhy, 2014). According to Hughes (2003), criterion validity describes the extent to which results from a test provide a dependable assessment of test takers' abilities. The information gathered from a test shows the relation between test scores and some criteria which are signs of the tested ability (Bachman, 1990). Bachman (1990, p.248), also, states "the criterion may be a level of ability". There are two types of criterion: predictive and concurrent. Predictive validity is using test results for future criteria whereas concurrent validation is using the scores at the same time the test is conducted (Fulcher & Davidson, 2007). Demirezen (2013) states that the difference between these two terms as concurrent validity which allows classifying examinees correctly. "It uses a statistical method using correlation, rather than a logical method" (p. 169). If the correlation is strong, the concurrent validity is high. In comparison, predictive validity discusses the examinees' future specialities. Hence, this is useful for selection and admissions of the schools.

Finally, "construct validation is the process of building a case that test scores support a particular interpretation of ability, and it thus subsumes content relevance and criterion relatedness" (Bachman, 1990, p. 290). The deductions from the test results should comply with the underlying construct that is being measured. The aforementioned construct means any underlying trait, which is hypothesized in a language theory (Hughes, 2003).

Practicality means to conduct a test without too much effort. A practical test is economical, easy to administer and appropriate for time (Brown, 2007). "Practicality is the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities" (Bachman & Palmer, 1996, p.36). It is important for testers and testees to know the test design and format in order to create practicality. Unless they are familiar with the format and technique, it will be time-consuming not only in the process of the test but also in the scoring phase. As Brown and Abeywickrama (2010) discuss, a practical test does not cost too much; it is finished within convenient time limits, and has clear instructions, appropriately utilizes the sources. Bachman and Palmer (1996, p. 36) formulate practicality as follows:

$$\text{Practically} = \frac{\text{Available Resource}}{\text{Required Resource}}$$

Figure 2. Practicality

Based on the information above, practicality can be easily available, affordable and familiar, and administered within a given time.

For a test, authenticity means being natural as much as possible. An authentic test contextualizes items (Brown & Abeywickrama, 2010). It includes cohesive and coherent texts, real-world paragraphs, and tasks. (Brown, 2003). An authentic test also includes real world samples. In this kind of test, a discourse is produced so that the items are not independent from each other but rather "provides some thematic organization to items, such as through a storyline or episode'' (Brown & Abeywickrama, 2010, p. 37).

The notion "of authenticity" emerged in the 1970s when the communicative approach got on the stage and the interest increased for "real-life" situations in both teaching and testing (Lewkowicz, 2000). Since then, authenticity has gained importance in communicative language teaching, in which communication and real use of language are important. When it comes to authentic assessment, is learner-centred and continuous assessment is required (Finch, 2002). Besides, it provides enthusiasm in teaching and learning, teacher commitment and public assistance (Archbald & Newmann, 1988).

Washback is the effect of the test on testees. "The effect of testing on teaching and learning is known as backwash or washback, and can be harmful or beneficial". (Hughes, 2003, p.1). Washback is a type of impact, which relates to the effects of high-stakes tests on classroom practices – particularly teaching and learning (Alderson & Wall, 1993; Bailey, 1999). Bachman and Palmer (1996) also discuss washback as the impact on society, educational systems, and individuals. They state that the test impact performs at two levels: the micro level (i.e., the effect of the test on individual students and teachers) and the macro level (the impact on society and its educational systems).

The advantage of this effect for learners is to provide learners a chance to be prepared adequately (Brown & Abeywickrama, 2010). Moreover, it gives learners feedback about their development. Brown (2007, p.451-452) says that washback empowers basic principles of language acquisition, such as intrinsic motivation, autonomy, self-confidence, and language ego.

Alderson and Wall (1993, pp.120-121) propose 15 possible hypotheses on washback:

- A test affects teaching.
- A test affects learning.
- A test affects what teachers teach.
- A test affects how teachers teach.
- A test affects what learners learn.
- A test affects how learners learn.
- A test affects the speed and procession of teaching.
- A test affects the speed and procession of learning.
- A test affects the quality and quantity of teaching.
- A test affects the quality and quantity of learning.
- A test affects the method of teaching.
- A test having important consequences has washback.
- A test having no important consequences has no washback.
- A test has washback on all learners and teachers.
- A test has washback on some learners and teachers.

Conclusion

In this review, the purposes and principles of testing are analyzed in detail. The purpose of a test varies depending on the testees' needs and the programme. Tomlinson (2005) regards language tests as an opportunity for learners to improve their skills and knowledge. Buck (2001) states that thanks to language tests, learners, teachers, and the administration have the chance to revise and develop the learning atmosphere. Similarly, Alabi and Babatunde (2001) see the test as a diagnostic tool to provide feedback about students' learning outcomes. It can be concluded that tests are highly used to determine a learner's knowledge and skills. In this sense they are useful tools to guide educators. However, all the tests mentioned above have different functions to use. Ignoring the function may lead to a negative washback on the learner. Therefore, teachers should be careful about the reason why they test their students.

Additionally, another important factor is to design the language test on the basis of the testing principles. Even though the reason for testing changes, the basic principles remain the same (Ali, Ahmad & Khan, 2019). Noticing these principles allows the teachers develop their assessment procedure (Tosuncuoglu, 2018).

To sum up, testing is not an easy process. It has different sub-scales and subsets. In order to create an exam that satisfies all parties, an educator should know the basic procedure that includes test techniques, assessment types, and basic principles. Teaching and assessment cannot be separated from each other.

Assessment has a main role in teaching thus everybody should comprehend the goal of the exam. The exam type should be well-decided before and the principles should be followed properly. Preparing and taking an exam regardless of its aim brings about demotivation and anxiety rather than motivation and self-esteem. In that event, testing creates reverse effect on teaching instead of supportive one. So, foreign language tests are devices that back up learning. Appropriate and well prepared exams let learners develop their competence in foreign languages and show better performance in the exams. After all, it should be noted that all these efforts are done to make foreign language teaching efficiently. All in all, the present study deals with only theoretical background of the purposes and principals of testing. Further studies and research can be conducted in their practical sides. Also, the history and possible future of the concepts can be analysed. Today, alternative assessment tools and technology in assessment are highly recommended. Hence, implementation of these fundamental principles to the new trends in testing can be investigated. Technology provides many opportunities to design the different test patterns. Applications can be designed and utilized in favour of every purposes of testing.

References

Akıncı, T. (2010). *Opinions of English teachers in state primary schools on the test they apply, the effect of SBS on their tests and the problems faced*. Unpublished Master's Thesis, Pamukkale University, Denizli,Turkey.

Alabi, A.O. and Babatunde, M.A. (2001*). Primary English curriculum and methods*. Oyo, Odumatt Press and Publishers.

Alderson, J. C. and Wall, D. (1993). Does wasback exist? *Applied Linguistics, 14*, 115-129. http://dx.doi.org/10.1093/applin/14.2.115

Alderson, J. C., Clapham, C., and Wall, D.(1995). *Language test construction and evaluation.* Cambridge University Press.

Ali, S. R., Ahmad, H., Khan, R. (2019). Testing in English Language Teaching and its Significance in EFL Contexts: A Theoretical Perspective. *Global Regional Review, 4*(2), 254-262. http://dx.doi.org/10.31703/grr.2019(IV-II).27

Archbald, D. A. and Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic achievement in the secondary schoo*l. National Association of Secondary School Principals. https://files.eric.ed.gov/fulltext/ED301587.pdf

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

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

Bailey, K.M. (1999). *Washback in language testing. Educational Testing Service*. https://www.ets.org/Media/Research/pdf/RM-99-04.pdf

Brown, S. (2003). Assessment that works at work. *The Newsletter for the Institute of Learning and Teaching in Higher Education 11*: 6–7.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. (3rd Edition). Pearson Education, Inc.

Brown, H.D. and Abeywickrama, P (2010). *Language assessment: Principles and classroom practices.* (2nd Edition). Pearson Education, Inc.

Buck, G. (2001). *Assessing listening.* Cambridge University Press. https://anekawarnapendidikan.files.wordpress.com/2014/04/assessing-listening-by-gary-buck.pdf

Chapelle, C. A.(1999). Validity in language assessment. *Annual Review of Applied Linguistics. 19*, 254-272. https://doi.org/10.1017/S0267190599190135

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge University Press.

Cohen, A. D. (2001) Second language assessment. In M. Celce- Murcia (Ed.), *Teaching English as a second or foreign language*. (3rd Edition)(pp. 515-535).Heilne &Heilne.

Demirezen, M. (2013). Testing. In A Sarıçoban (Ed.), *Öğretmenlik alan bilgisi testi: inglizce öğretmenliği* (pp. 165- 194) Murat Yayınları.

Desheng, C. and Varghese, A. (2013). Testing and evaluation of language skills. *IOSR Journal of Research & Method in Education (IOSR-JRME), 1*(2), 31-33. https://www.iosrjournals.org/iosr-jrme/papers/Vol- 1%20Issue-2/F0123133.pdf

Finch, A. E. (2002). Authentic assessment: implications for EFL performance testing in Korea. *Secondary Education Research, 49,* 89-122. https://www.finchpark.com/arts/Authentic_Assessment_Implications.pdf

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Gardner, R.C. (1985). *Social psychology and second language learning, the roles of attitudes and motivation.* Edward Arnold Publishers Ltd.

Garson, G. D. (2009). Internal consistency reliability. *In Reliability analysis*. North Caroline State University College of Humanities and Social Sciences.http://faculty.chass.ncsu.edu/garson/PA765/reliab.htm.

Gonzalez, A. B. (1996). Teaching English as a foreign language: An overview and some methodological considerations. *RESLA, 11,* 17-49.

Harmer, J. (2007). *The practice of English language teaching*. (4th Edition). Pearson Education

Limited.

Henning, G. (2012). Twenty common testing mistakes for EFL teachers to avoid. *English Teaching Forum, 50* (3), 33-36. https://files.eric.ed.gov/fulltext/EJ997528.pdf

Hughes, A. (2003). *Testing for language teachers*. (2nd Edition). Cambridge: Cambridge University Press.

Ivanova, V. (2001). *Construction and evaluation of achievement tests in English. Institute of Mathematics and Informatics Bulgarian Academy of Sciences, Association for the Development of the Information Society* (pp. 276-285).http://sci-gems.math.bas.bg:8080/jspui/handle/10525/1554

Jumaniyozova, F.T. (2021). The Importance of Five Main Principles of Language Assessment in Designing the Language Tests in Uzbek Universities. *Current Research Journal of Philological Sciences*. 1 https://doi.org/10.37547/philological-crjps-02-06-01

Katz, A. (2013). Assessment in second language classrooms. In M. Celce- Murcia, D. M.Brinton & M. A. Snow (Eds.), *Teaching English as a second or foreign language*. (4th Edition) (pp. 320-340).Heinle ELT.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, 17* (1), 43-64. https://doi.org/10.1191/026553200669746135

Paradowski, M.B. (2002). *The features of good language tests*. The Teacher 1:40. http://files.eric.ed.gov/fulltext/ED503442.pdf

Rahman, M.M. and Gautam, A.M. (2012). Testing and evaluation: A Significant characteristic of language learning and teaching. *Language in India, 12*(1), 432-442. http://www.languageinindia.com/jan2012/motiurtestingevaluationfinal.pdf

Rajhy, H.A.A (2014). Five characteristics of a good language test. *National Journal of Extensive Education and Interdisciplinary, 2(*4), 61-66.

Riddell, D. (2003). *Teaching English as a second/foreign language*. The McGraw-Hill Companies, Inc.

Rysiewicz, J. (2008). Measuring foreign language learning aptitude. Polish adaptation of the modern language aptitude test by Carroll and Sapon. *Poznań Studies in Contemporary Linguistics 44*(4), 569–595. https://doi.org/10.2478/v10010-008-0027-6

Tomlinson, B. (2005). *Testing to learn: a personal view of language testing.* ELT Journal, Oxford University Press 59/1, 44

Tosuncuoglu, İ. (2018). Importance of Assessment in ELT. *Journal of Education and Training Studies. 6*(9),163-167. https://doi.org/10.11114/jets.v6i9.3443

Ze'ev Wurman (2022). Why Educational Testing is Necessary. Fraser Institute. https://www.fraserinstitute.org/sites/default/files/why-educational-testing-is-necessary.pdf

Yin, M. and Sims, J. (2006). Diagnostic language testing for Taiwanese University students: The online English assessment system (OEAS) project. *2006 International Conference on English Instruction and Assessment.* http://fllcccu.ccu.edu.tw/conference/2005conference_2/download/C44.pdf