INTERNATIONAL INNOVATION AND RESEARCHES
International scientific journal Volume: 2 Issue: 3

THE CHALLENGES OF DESIGNING A FAIR LANGUAGE PROFICIENCY TEST

Umaraliyeva Maftuna Kamoliddin qizi
Student of the Faculty of tourism, Chirchik state pedagogical university
E-mail: mirzahakimovmaftuna@gmail.com

Scientific supervisor: Usmonova Gulsevar Abdulaziz qizi
Teacher at Chirchik state pedagogical university
E-mail: gulsevardesigner@gmail.com

Annotatsiya. *Testning validligi va xatolarni hisobga olishdan tashqari, tilni bilish bo'yicha testning amaliyoti ham muhim ahamiyatga ega. Test dizayni vaqt va resurslar cheklangan sharoitda amalga oshirilishi kerak bo'lib, uni o'tkazish va baholash oson bo'lishi zarur. Bundan tashqari, testning real dunyo tilidan foydalanishni aks ettirishi zarur, shunda o'quvchilar til ko'nikmalarini amaliy holatlarda samarali qo'llay olishadi. Adolatli testni ishlab chiqishdagi qiyinchiliklar, shuningdek, doimiy fikr-mulohazalar va baholashni talab qiladi, shunda test materiallari takomillashtirilib, uning ahamiyati vaqt o'tishi bilan saqlanadi.*

Kalit so'zlar: *amaliyot, testni o'tkazish, baholash, real dunyo tilidan foydalanish, doimiy fikr-mulohazalar, testni baholash, testni takomillashtirish, testning Ahamiyati, o'rganish natijalari, adolatli test dizayni.*

Аннотация. *Помимо учета валидности и предвзятости, важную роль играет также практичность теста на знание языка. Дизайн теста должен быть выполнимым в условиях ограниченных времени и ресурсов, а также должен быть легким для проведения и оценки. Более того, важно, чтобы тест отражал использование языка в реальных ситуациях, чтобы участники могли эффективно применять свои языковые навыки в практических условиях. Разработка справедливого теста также включает необходимость постоянной обратной связи и оценки для совершенствования тестовых заданий и обеспечения их актуальности с течением времени.*

Ключевые слова: *практика, проведение теста, оценка, использование языка реального мира, постоянная обратная связь, оценка теста, улучшение теста, важность теста, результаты обучения, справедливый дизайн теста.*

Annotation. *In addition to validity and bias considerations, the practicality of a language proficiency test also plays a significant role. The test design must be feasible within time and resource constraints, and should be easy to administer and score. Furthermore, it is essential that the test reflects real-world language use to ensure that test-takers can apply their language skills effectively in practical situations. The challenge of designing a fair test also involves the need for continuous feedback and evaluation to refine the test items and ensure its relevance over time.*

Key words: *practicality, test administration, scoring, real-world language use, continuous feedback, test evaluation, test refinement, test relevance, learning outcomes, fair test design.*

INTRODUCTION

Language proficiency tests are essential tools used to assess an individual's ability to understand, speak, read, and write in a second language. These tests play a critical role in various contexts, including academic admissions, professional qualifications, and immigration processes. However, designing a fair language proficiency test is far from straightforward. The process involves a delicate balance between accurately measuring a candidate's language skills and ensuring the test is free from biases that could disadvantage certain groups of individuals.

The challenges of designing such a test are multifaceted and require careful consideration of numerous factors, including cultural neutrality, the inclusion of diverse language skills, test format, and the avoidance of discrimination. Additionally, external factors such as anxiety and the test-taking environment can influence test results, complicating the task of creating an equitable assessment. In this article, we will explore the key challenges involved in designing a fair language proficiency test, examining the obstacles faced by test designers and the strategies that can help overcome them to ensure that language assessments are both accurate and just.

The five principles—practicality, reliability, validity, authenticity, and washback—serve as valuable guidelines for assessing an existing evaluation method or creating a new one. Whether it is quizzes, tests, final exams, or standardized proficiency exams, these principles can be applied to critically examine the effectiveness of any assessment procedure.

These are five principles:

1. Practicality

Practicality refers to the logistical and administrative aspects of creating, administering, and scoring an assessment. This includes factors like costs, time required to develop and administer the test, ease of scoring, and how easily results can be interpreted and reported (Mousavi, 2009). A test is considered impractical if it does not meet these criteria. Here are the key attributes of practicality:

A practical test:

- fits within budgetary constraints;
- can be completed within an appropriate timeframe by the test-taker;
- has clear instructions for administration;
- makes efficient use of available human resources;
- stays within the limits of available materials;
- considers the time and effort required for both design and scoring.

For example, a language proficiency test that takes five hours to complete is impractical, as it exceeds the time available for its intended purpose. Similarly, a test that requires one-on-one proctoring for a large group of test-takers is impractical. A test that takes minutes for a student to complete but several hours for an examiner to grade is also impractical for most classroom settings. Additionally, if a test requires scoring only by computer and is administered in a location far from the nearest computer, it would be impractical. A test that depends on the subjective judgment of the scorer might also be impractical and unreliable due to the time it takes to grade.

Here is an example of practicality gone wrong: An administrator needed to place about 50 students into ability-based sections for a six-week summer course. After a quick search, the administrator found an old English Placement Test from the University of Michigan, which had 20 listening items based on an audio recording and 80 grammar, vocabulary, and reading comprehension items—all multiple-choice. The test booklet was ready, a proctor was assigned, and the plan was to score the test that same afternoon so students could start classes the next day. However, things went wrong when the wrong audio was played for the listening section, confusing the students. In response, the proctor quickly improvised and gave the students a dictation. While the rest of the test proceeded smoothly, scoring the dictation turned out to be a subjective process. The administrators struggled to complete the grading due to time constraints, and by the end of the day, they still had not finished determining student placements.

This situation highlights the importance of practicality. Although the listening section seemed practical, the administrator failed to check the materials beforehand, leading to issues with the test's reliability. Furthermore, the scoring process didn't fit the available time, which is a crucial factor in classroom-based testing.

2. Reliability

A reliable test is one that consistently produces dependable results. When the same test is administered to the same student—or to equivalent students—on different occasions, it should yield similar outcomes. The principle of reliability can be summarized as follows:

A reliable test:

- maintains consistent conditions across multiple administrations;
- provides clear instructions for scoring and evaluation;
- uses uniform scoring rubrics;
- ensures consistent application of rubrics by the scorer;
- includes unambiguous items/tasks for the test-taker.

Reliability issues can arise from several factors, including those related to the student, the scoring process, the test administration, and the test itself. Let's explore these factors:

Student-Related Reliability

Common issues related to student reliability include factors like illness, fatigue, anxiety, or emotional states, which can lead to test scores deviating from a student's true ability. Additionally, a student's test-taking strategies, such as using certain techniques for answering questions, can also impact reliability.

While these student-related factors may seem beyond a teacher's control, many teachers find ways to minimize their impact. For instance, offering test-taking strategies or creating a comfortable test environment can help reduce anxiety and improve reliability.

Rater Reliability

Human error, subjectivity, and biases can affect the scoring process. Inter-rater reliability occurs when multiple scorers give consistent scores for the same test. If scorers do not adhere to the same criteria or are inexperienced, inconsistent results can occur. Intra-rater reliability refers to inconsistencies in a single scorer's judgment over time. For example, a teacher may be more lenient or strict depending on fatigue or how they apply their grading criteria. To enhance intra-rater reliability, it's helpful for teachers to review a portion of the tests first and then recheck all tests before finalizing scores.

In subjects like writing, achieving rater reliability is particularly challenging due to the subjective nature of writing assessments. However, using detailed scoring rubrics can help improve both inter- and intra-rater reliability.

Test Administration Reliability

Unreliability can also arise from the conditions under which a test is administered. For instance, background noise or technical issues, like an audio player malfunctioning during a listening test, can negatively impact reliability. Other factors, such as variations in lighting, temperature, and seating arrangements, can also cause inconsistencies in test performance.

Test Reliability

Sometimes, the test itself may contribute to unreliability. In multiple-choice tests, for example, poorly designed items or distractors can affect the reliability of the test. In classroom assessments, the reliability of subjective tests, such as essays, can be impacted by rater bias. Objective tests, on the other hand, tend to be more reliable because they have fixed answers.

Unreliability can also result from ambiguously worded test items or from tests that are too lengthy, which can lead to student fatigue. Time-limited tests may disadvantage students who perform poorly under pressure, even if they understand the material. Therefore, test characteristics, including how they interact with student factors, play a significant role in determining test reliability.

3. Validity

Validity is the most intricate and arguably the most crucial criterion for an effective test. It refers to "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in relation to the purpose of the assessment" (Gronlund, 1998, p. 226). In more technical terms, Samuel Messick (1989), a well-known expert on validity, defined it as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other forms of assessment".

From these definitions, we can deduce the following key attributes of validity:

A valid test:

- accurately measures what it is intended to measure;
- avoids measuring irrelevant or "contaminating" factors;
- relies heavily on empirical evidence (performance-based);
- involves performance that reflects the test's intended criteria;
- provides meaningful and useful insights into test-taker's abilities is backed by a theoretical rationale.

For example, a valid reading test should measure reading ability, not factors like visual acuity or prior knowledge of a subject. Similarly, a writing test that merely counts the number of words written in 15 minutes would be easy to administer and reliable in terms of scoring but would not be valid because it would overlook crucial aspects like clarity, structure, and rhetorical effectiveness.

How is the validity of a test established? According to various experts (Broadfoot, 2005; Chapelle & Voss, 2013; Kane, 2016; McNamara, 2006; Weir, 2005), there is no absolute measure of validity, but several types of evidence can support it. Moreover, as Messick (1989) pointed out, validity is a matter of degree—it is not an all-or-nothing concept. Sometimes, we may assess how closely the test reflects the performance expected from the course or unit being evaluated. In other cases, we focus on how well the test measures whether students have achieved certain goals or competencies. Another common form of evidence is the statistical correlation between the test and other related measures. Additionally, concerns about validity might include the consequences of the test beyond simply measuring the

criteria or even how the test-taker perceives its validity. The following sections will explore four types of evidence used to establish validity.

4. Authenticity

Authenticity is another key principle in language testing, though it is challenging to define, particularly in the context of test design and evaluation. Bachman and Palmer (1996) described authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task", and they outlined a process for identifying relevant real-world tasks and transforming them into valid test items. However, authenticity does not lend itself easily to empirical measurement or clear operationalization (Lewkowicz, 2000). Determining whether a task or language sample is "real-world" often involves subjective judgment, making it a concept that has generated considerable attention from language-testing experts (Bachman & Palmer, 1996; Fulcher & Davidson, 2007). According to Chun (2006), many test formats fail to replicate real-world tasks.

When authenticity is claimed in a test task, it means that the task is likely to occur in real-world situations. Many test items fail to simulate real-world tasks as they may be artificial or overly focused on specific grammatical forms or vocabulary. For example, a series of unrelated test items lacks authenticity, and reading comprehension passages in proficiency tests may not resemble real-world reading materials.

An authentic test may include the following features:

- language that is as natural as possible;
- contextualized items rather than isolated ones;
- topics that are relevant, meaningful, and engaging;
- a thematic organization of tasks, such as through a storyline or scenario;
- tasks that mirror real-world activities.

In recent years, there has been significant progress in increasing the authenticity of test tasks. A few decades ago, disconnected and artificial items were considered standard in testing. However, the landscape has changed. It was once believed that large-scale tests couldn't incorporate productive skills like speaking and writing within budget constraints, but now many tests include these components. Reading passages are now drawn from real-world materials that test-takers are likely to encounter. Listening sections often feature natural language with hesitations, background noise, and interruptions. Additionally, many tests now feature "episodic" items, where the questions are organized into meaningful units like paragraphs or stories.

As you design tests in your classroom, we encourage you to embrace the challenge of authenticity. In the following chapters of this book, we will explore various types of tasks where the principle of authenticity plays a central role.

5. Washback

Washback refers to the impact of testing on teaching and learning, a component of consequential validity (Hughes, 2003). Messick (1996) emphasized that washback can have both positive and negative effects on learning. Alderson and Wall (1993) developed a washback hypothesis to describe how tests influence both teaching and learning, and Cheng, Watanabe, and Curtis (2004) explored this further. Spratt (2005) also encouraged teachers to foster beneficial washback in their classrooms.

Washback can have the following positive effects:

- it positively influences teaching methods;
- it positively influences how students learn;
- it provides students with an opportunity to adequately prepare;
- it offers feedback that supports language development;
- it is more formative than summative in nature;
- it creates optimal conditions for student performance.

In large-scale assessments, washback often refers to how tests shape student preparation, such as through “cram” courses or “teaching to the test,” which can have both positive and negative impacts. While standardized tests can encourage students to focus on achieving specific scores rather than improving language abilities, some test-preparation courses report that students gain competence in certain areas.

In classroom-based assessments, washback can be positive, such as when preparing and reviewing for a test helps students learn. Feedback from teachers can also serve as valuable diagnostic tools, highlighting students' strengths and weaknesses. The impact of washback is not limited to the test itself but also extends to preparation for the assessment. Informal assessments tend to have more built-in washback because teachers provide interactive feedback. Formal tests, on the other hand, may lack beneficial washback unless they offer more than a simple grade or score.

Teachers can enhance washback by turning classroom tests into learning tools. Incorrect answers can provide insight into areas that need further attention, while correct answers should be praised to encourage students. Teachers can also offer strategic advice to help students succeed, fostering intrinsic motivation, autonomy, self-confidence, and other key principles of language acquisition.

To improve washback, teachers should provide detailed feedback rather than just a grade or score. Comments that highlight strengths, offer constructive criticism,

and provide strategies for improvement make tests more motivating and help students gain a sense of accomplishment. Even specifying numerical scores for specific test sections can serve a diagnostic purpose, showing areas where students need more practice.

Formative tests are naturally more likely to provide washback, offering learners feedback on their progress. Summative tests, however, can also offer beneficial washback, particularly when teachers make the effort to provide feedback that supports further learning. For example, some teachers give final exams before the last class and return them during the final session, using the time to address areas of confusion.

Washback also involves ensuring that students have the opportunity to discuss their feedback with teachers. A classroom atmosphere where students can engage in dialogue with teachers about their performance promotes continuous learning. This allows students to clarify any uncertainties, receive further guidance, and set new learning goals.

Here are some examples of scholarly quotes and references related to the challenges of designing a fair language proficiency test[1]:

According to, Fulcher: "Fairness in language testing requires the development of clear, transparent test criteria and rubrics, as well as careful consideration of how different student populations might be affected by the test. The challenge lies in balancing the need for accurate assessment with the avoidance of unfair advantages or disadvantages caused by external factors such as socio-economic status or educational background"[2].

According to, Weir: "The principle of validity in language testing must be seen in light of fairness. A language proficiency test can only be considered valid if it measures what it intends to measure and if the results are not unfairly skewed by the test design itself. The challenge is in identifying and eliminating biases that could distort the accuracy and fairness of the test outcomes"[3].

CONCLUSION

In conclusion, designing a fair language proficiency test is a complex and multifaceted challenge that requires careful attention to various factors such as practicality, reliability, test validity, authenticity, and the washback effect. Ensuring

[1] Brown, H. Douglas, and Priyanvada Abeywickrama. Language Assessment: Principles and Classroom Practices. Pearson Education, 2010.

[2] Fulcher, G. "The Ethical and Practical Challenges of Designing Fair Language Tests." Language Testing, vol. 27, no. 2, 2010, pp. 175-190.

[3] Weir, C. J. "Language Testing and Validity: The Challenges of Ensuring Fairness." System, vol. 33, no. 2, 2005, pp. 211-221.

fairness means developing assessments that accurately measure language ability while minimizing the impact of external factors such as cultural, social, and educational differences. Test designers must also be mindful of the consequences of their tests on both teaching and learning, aiming to support rather than hinder the development of language skills. Ultimately, a fair language proficiency test should not only provide reliable and valid results but also contribute to a positive learning experience that benefits all test-takers, regardless of their background or circumstances.

References:

1. Brown, H. Douglas, and Priyanvada Abeywickrama. Language Assessment: Principles and Classroom Practices. Pearson Education, 2010.
2. Fulcher, G. "The Ethical and Practical Challenges of Designing Fair Language Tests." Language Testing, vol. 27, no. 2, 2010, pp. 175-190.
3. Weir, C. J. "Language Testing and Validity: The Challenges of Ensuring Fairness." System, vol. 33, no. 2, 2005, pp. 211-221.