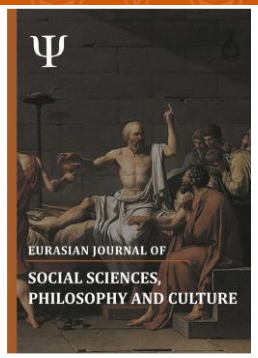
EURASIAN JOURNAL OF SOCIAL SCIENCES, PHILOSOPHY AND CULTURE
Innovative Academy Research Support Center
IF = 9.3 www.in-academy.uz/index.php/ejsspc



A CASE STUDY APPROACH: EVALUATING THE VALIDITY AND RELIABILITY OF A B2 READING COMPREHENSION TEST

Nilufarkhon Shokirova
Instructor
Uzbekistan State World Languages University
Tashkent, Uzbekistan
Email: nilufarshokirovam@gmail.com
https://orcid.org/0009-0004-0789-7452
https://doi.org/10.5281/zenodo.20744652

ARTICLE INFO

Received: 02nd June 2026
Accepted: 08th June 2026
Online: 09th June 2026

KEYWORDS

Language assessment, reading comprehension, validity, reliability, authenticity, practicality, language testing.

ABSTRACT

Assessment plays a significant role in language education as it provides information about learners' proficiency and assists in formulating teaching strategies. The research work examines a reading comprehension test administered to a 16-year-old grade 10 student at school, who is studying English as a foreign language. The study analyzes validity and reliability of the chosen assessment tool. A qualitative case-study approach was applied to analyze a B2-level reading comprehension test adapted from CEFR examination materials. The participant completed the reading comprehension test, and the assessment items were evaluated against fundamental principles of language testing proposed by Brown (2010), Bachman and Palmer (1996), and Hughes (2003). The findings demonstrated that while the assessment test showed high scoring reliability and practicality through its objective multiple-choice format, several items lacked construct validity as these items measured vocabulary knowledge rather than reading comprehension. The assessment test was reformulated utilizing True/False/Not Given items and provided concise instructions to complete to eliminate the shortcomings. The research work accentuates the significance of developing assessment tasks that can accurately measure intended constructs and provide reliable evidence of learners' reading abilities. The findings contribute to a better understanding of effective language assessment practices and test development in English language teaching contexts.

КЕЙС-СТАДИ ПОДХОД: ОЦЕНКА ВАЛИДНОСТИ И НАДЁЖНОСТИ ТЕСТА НА ПОНИМАНИЕ ЧТЕНИЯ УРОВНЯ B2

Нилуфархон Шокирова
Преподователь
Узбекский государственный университет мировых языков

EURASIAN JOURNAL OF SOCIAL SCIENCES, PHILOSOPHY AND CULTURE
Innovative Academy Research Support Center
IF = 9.3 www.in-academy.uz/index.php/ejsspc

Ташкент, Узбекистан
Электронная почта: nilufarshokirovam@gmail.com
ORCID: https://orcid.org/0009-0004-0789-7452
https://doi.org/10.5281/zenodo.20744652

ARTICLE INFO

Received: 02nd June 2026
Accepted: 08th June 2026
Online: 09th June 2026

KEYWORDS

Языковое оценивание, понимание прочитанного, валидность, надёжность, аутентичность, практичность, языковое тестирование.

ABSTRACT

Оценивание играет значительную роль в языковом образовании, поскольку оно предоставляет информацию об уровне владения языком учащимися и помогает формировать стратегии обучения. Данное исследование рассматривает тест на понимание прочитанного, проведённый среди 16-летнего ученика 10 класса школы, изучающего английский язык как иностранный. В работе анализируются валидность и надёжность выбранного инструмента оценивания. Был применён качественный кейс-стади подход для анализа теста уровня B2 на понимание прочитанного, адаптированного из материалов экзаменов CEFR. Участник выполнил тест на понимание прочитанного, а задания были оценены с точки зрения основных принципов языкового тестирования, предложенных Brown (2010), Bachman и Palmer (1996), а также Hughes (2003). Результаты показали, что, хотя тест демонстрировал высокую надёжность оценивания и практичность благодаря объективному формату множественного выбора, некоторые задания не обладали конструктной валидностью, поскольку они измеряли знание лексики, а не понимание прочитанного. Тест был переработан с использованием заданий формата True/False/Not Given и дополнен краткими и чёткими инструкциями для устранения выявленных недостатков. Исследование подчёркивает важность разработки оценочных заданий, которые точно измеряют предполагаемые конструктные характеристики и предоставляют надёжные данные о навыках чтения учащихся. Полученные результаты способствуют лучшему пониманию эффективной практики языкового оценивания и разработки тестов в контексте преподавания английского языка.

INTRODUCTION

Assessment is a fundamental component of the teaching and learning process. Effective assessment tools provide significant information about learners' language proficiency, strengths,

and weaknesses, enabling teachers to make informed instructional decisions. Efficient assessment not only evaluates students' achievement but also contributes to the improvement of curriculum design and planning, material selection, and classroom instruction. Reading comprehension plays a pivotal role among four language skills since it enables language to access information, develop critical thinking skills, and support language acquisition. Assessing reading comprehension accurately is crucial for understanding learners' abilities and ensuring that instructional objectives are achieved. However, developing a valid and reliable reading assessment remains a challenging task. Test items should measure reading comprehension rather than unrelated linguistic knowledge such as vocabulary or grammar. Language testing scholars accentuates several qualities of effective assessments. The present study investigates a reading comprehension assessment administered to a Grade 10 learner studying English as a foreign language. Specifically, the study examines the strengths and weaknesses of the selected test by analyzing its validity and reliability. Based on the findings, modifications are proposed to improve the effectiveness of the assessment instrument. The study aims to contribute to the growing body of research on language assessment and provide practical recommendations for English language teachers and test developers.

LITERATURE REVIEW

Reliability is considered to be an essential characteristic of any test which refers to the accuracy and consistency of data gathered in a study. In other words, tests should produce the similar results on repeated trials. Brown (2010) stated the items need to be evenly difficult, distractors need to be well designed, and items need to be well distributed to make the test reliable.

Validity of the test refers whether test measures or does not measure what it intends to measure. It is often considered the most important characteristic of assessment because all interpretations of test scores depend on the validity of the instrument. Bachman and Palmer (1996) emphasize that validity concerns the meaningfulness and appropriateness of inferences drawn from test results. Several types of validity are commonly discussed in language assessment. Content validity refers to the degree to which test items adequately represent the target skill domain. Construct validity concerns whether the test accurately measures the theoretical construct it intends to assess.

METHODOLOGY

The subject that was selected for the research work was a 16-years old girl. She studies at an International school in the 10th grade. Instructors utilize "Solutions for Upper Intermediate students" by Oxford University Press in conducting the lessons for 10th grades. However, there has been decided by the school administration that 10th grade students should be taught in the IELTS program. The same class is taught by two teachers: native English teacher and local teacher. Native English teacher works on the reading and writing comprehension of the 10th graders and local teacher works on their listening, speaking, vocabulary and grammar skills. The

participant is Uzbek. She is a bilingual learner. She speaks fluently both English and Russian languages. Her native language is Uzbek. However, she prefers to speak Russian language. The subject is from researcher's own English class where she conducts English language. This learner has already become an independent learner and acquired autonomous learning skills. This English learner only needs to be coached, trained and provided with the right materials. She does not have to be explained every single detail since she can study on her own and learn if the right path is shown for her. She has been studying English language intensively for 5 years. This learner is planning to take an IELTS test in summer. For this reason, the researcher has recently checked the level of the learner. All the four skills: speaking, listening, writing and reading skills of the learner were checked. In addition, paper- based grammar and vocabulary tests were also taken from the subject.

FINDINGS

The table below demonstrates the results of the test:

Skill	Level	Correct answers
Listening	B2 (40/32)	40/32
Speaking	B2	-
Writing	B1+	-
Reading	B2 (40/30)	40/30
Vocabulary test	B2 (10/8)	10/8
Grammar test	B2	15/12

The level assessment was taken a month ago. Generally, the learner does possess good command of English. She speaks fluently and can discuss desperate topics in unfamiliar situations. However, she might encounter challenges in discussing some topic because of her lack of knowledge and vocabulary case on these fields. The learner wrote an opinion essay about unpaid community. The learner can utilize simple and complex sentences effectively. However, she does have some problems in connecting the ideas logically and most of the given ideas were not developed in her essay. The learner possesses range of vocabulary on different topics and utilizes it efficiently both in her speech and written production.

The level assessment test was appropriate and reliable in identifying the level of the learner since researcher made a great effort to check all the skills of the learner in order to foster teaching process.

This placement test was taken by the researcher in order to identify the level of the learner and diagnose her weaknesses. After analyzing the results of this test, researcher prepared appropriate lesson plans and selected materials thoroughly in order to meet the needs of the learner. To be more precisely, this assessment test truly altered the researcher's instructions in the class.

Researcher has taken a test in order to identify the reading comprehension of the learner. This test was taken from the past CEFR exam materials. Researcher specifically selected B2 level reading test for the participant. This reading test does

include three parts. The total number of the question items are 26. However, the researcher decided to take the first part of the reading test. The overall number of the question items in this part are seven. There is a given a magazine article about a young mother whose house was burgled. The instructions are given at the top of the reading passage for the learners. The type of the questions is selected response: multiple choice. Learners get two points for each correct answer.

The outlined below essential aspects of the test were analyzed thoroughly:

- Validity
- Reliability
- Content goals
- Standards
- Biases

In addition, the accessibility of the test both linguistically and culturally to the learner's profile was also taken into consideration.

The outlined below strengths of the selected test/assessment were observed:

1. High scoring reliability: The test consists of discrete 26 point dichotomous items which make the test objective
2. Authenticity: Bachman & Palmer (1996) stated that authenticity is "the degree of correspondence of the characteristics of a given language test task to the features of a target language task Selected response (multiple choice) items can make test reliable.
3. Brown (2010) sated that a fixed response format items increase test reliability
4. Test can be considered practical because it can be administrated easily and scoring is easy. Practicality can include "costs, amount of time, administer, and ease of scoring" (Brown, 2010, p.26)

The outlined below weaknesses were observed and analyzed:

1) Has poor construct in validity: it does not measure exactly what it proposes to measure

2) Test reliability: Brown (2010) stated the items need to be evenly difficult, distractors need to be well designed, and items need to be well distributed to make the test reliable

3) Authenticity:

- Decontextualized
- irrelevant topics/questions
- the tasks do not replicate real – world tasks.

DISCUSSION

The test was modified after analyzing its weaknesses thoroughly. First and foremost, True-false question items and Not given distractor were added to decrease the probability of guessing answer. Hughes (2016) mentions that True/False/Not selected-response test items are considered to be the variety of multiple-choice items. Nevertheless, when the test has one distractor, it might increase 50% of probability and students can guess the answer.

Then researcher decided to add clear instructions in order to provide learners with concise description of the test items. The instructions from the IELTS Cambridge were taken with the aim to specify the instructions for students (Jakeman and McDowell, 2014). Learners might accelerate solid understanding of the task which

eventually might enhance the reliability level of the test.

True	If the statement agrees
False	If the statement contradicts the information
Not Given	If there is no information on this

The current test transformed into computer-based testing. Computer-based testing might bring several benefits namely reducing the workload of a teacher and paper-printing. In addition, it gives a chance to get immediate feedback and scoring which accelerates practicality and reliability (Chapelle & Douglas, 2006).

The test items were modified in order to make them clear and enhance validity. All the question items were altered into True/False/Not given question types in order to check the reading comprehension of the learner since these items mostly check vocabulary case of the learner.

Test item 1. Original version: *How was Lisa feeling as she walked home from work?*

This item was altered because it checks the vocabulary case of the learners not the reading comprehension.

Modified version: Lisa was very tired as she walked home from work

Test item 2. Original version: What does "pick up" mean in line 5?

Modified version

Lisa did not take her three years old kid from her grandmother's home

Test item 3. Original version: What first led Lisa to think there was a burglar in her house?

Modified version: Lisa noticed that that the door was open while watching TV

Test item 4. Original version: Why did not Lisa wait in her neighbor's house until the police arrival?

Modified version

Lisa was so curious what was happening in her house that she did not want to wait for the police arrival

Test item 5. Original version. What does "Lisa saw red" mean?

Modified version: Lisa was frightened when she saw that the window was open

In conclusion, the current test had its own strengths and weaknesses. Weaknesses were excluded by altering the test items into True/False/Not given type of questions since it might increase the test validity.

CONCLUSION

This study examined the effectiveness of a B2-level reading comprehension assessment administered to a Grade 10 English language learner. The analysis focused on key principles of language assessment, particularly validity, reliability, authenticity, and practicality. The findings revealed that while the original test demonstrated strong reliability and practicality through its objective scoring procedures and ease of administration, several items lacked construct validity because they primarily assessed vocabulary knowledge rather than reading comprehension skills.

To address these shortcomings, the assessment was revised by transforming selected multiple-choice questions into True/False/Not Given items and by providing clearer instructions. These

modifications were intended to ensure closer alignment between the assessment tasks and the reading comprehension construct being measured. Furthermore, the transition to a computer-based format enhanced the practicality of the assessment by facilitating efficient administration, scoring, and feedback.

Overall, the study highlights the importance of carefully designing assessment tasks that accurately measure the intended language skills. Effective reading assessments should not only be reliable and practical but also demonstrate strong construct validity. The findings suggest that thoughtful revision of test items can significantly improve assessment quality and provide more meaningful evidence of learners' reading abilities. Future research may investigate the effectiveness of the revised assessment with a larger group of learners to further evaluate its validity and reliability in different educational contexts.

References:

1. Bachman, L.F., and Palmer, A. S. (1996) Language Testing in Practice: Designing and Developing Useful Language Tests. New York. Oxford University Press.
2. Brown, D.H. (2010). Language Assessment: Principles and Classroom Practices (2nd Edition). White Plains, NY
3. Cambridge University Press. (2009). War record of the Cambridge University Press (p.21). Cambridge (England)
4. Chappelle, C., & Douglas, D. (2006). Assessing language through computer technology. Cambridge, England: Cambridge University Press
5. Hughes, A. (2003). Testing for language teachers. Cambridge: Cambridge University Press.