ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

JOURNAL OF APPLIED LINGUISTICS AND TESOL

Vol.8. No.4.2025

"VALIDITY IN LANGUAGE ASSESSMENT: ENSURING AUTHENTICITY, FAIRNESS, AND EFFECTIVE LEARNING OUTCOMES"

Haleema Ahmed Mahar
MS Scholar, Centre of English Language & Linguistics
Mehran University of Engineering and Technology, Hyderabad, Pakistan
Email: mahar_haleema96@hotmail.com
Co-Author
Ali Raza Khoso
Lecturer, Centre of English Language & Linguistics
Mehran University of Engineering and Technology, Hyderabad, Pakistan
Email: ali.khoso@faculty.muet.edu.pk
Corresponding Author: *Haleema Ahmed Mahar, mahar_haleema96@hotmail.com*
Date of Submission: *December 10, 2025*

Abstract
In educational evaluation framework authenticity functions as an indicator of test quality. The current investigation improves beyond the research of Messick as well as Bachman Hughes and brown Through mining validation types their history of evolution testing risk and the techniques. It decides whether a test is a correct representation of the purpose for which it was developed and if the test finding can be successfully interpreted and applied to making choices. The accuracy of learners interpersonal and social learning competencies can be guaranteed by the reliability of language tests in surroundings that avoid discourse dissonances and the result errors. That carries on the explanation of how contemporary society has transformed what once was a technical measurement concept of developed Cultural entity. The construct validity, associated with criterion , shared risk., ways of Demonstrating reliability and its relationship to worthiness and washboard impact are the primary type of validity. Professional analysis, planning, experimental testing, grading guidelines, and practical testing are all successful methods for improving validity.
These Concerns highlight that easily testing might divert beyond its original purpose and produce incorrect or unfavorable repeated findings. When validity is compromised method of assessment begun to lose their pedagogical consistency and ethical foundations. In addition to numerical indexes and psychological complexity validity issues related to partiality in appropriateness scoring standards and prejudice are also examined as is the long-term educational impact.
Keywords *: language assessment , reliability and fairness ,construct validity, test anxiety , curriculum alignment , assessment challenges.*

1. Introduction

When it comes to student's learning educational decisions, teachers training assessment is important even though our test might seem to be remarkably good because it consistently produces similar scores, it could still be worthless. As an example, an experienced and self-aware speaker could perform horribly in reading comprehension exams if the assessment absolutely does nothing with the problem (Fulcher, 2004) For this reason, credibility is crucial it ought to be guaranteed that outcomes of tests accurately convey students true ability and the test results are applied correctly (Messick, 1989). In highly competitive situations such as university entrance test for and for certification test and test for foreign language ability where of flawed evaluation good lead to

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

not catastrophic implication for academic future options, authenticity is even more crucial (Bachman & Palmer, 1996). In addition to figuring out aptitude, objective exams also serve to motivate students to influence lesson planning and deliver positive results for learning. The goal of validity in language assessment is to ensure that the task that our students perform matches with the actual world communication needs, for example having the capacity to understand guidelines and convey ideas in a pleasant and systematic manner to create comprehensive and well-structured paper. Teachers are more capable of figuring out if their students are prepared for their studies interaction with others or employed in a particular environment whenever testing procedures are conducted according to real-life communication (Messick, 1989).

2. Literature Review

An important problem with the veracity of language assessment was recently named through investigations. The reliability of data according to the Messick 1989, relies not just on how the outcome or consideration is, but also how they affect learning instructions and decision-making process. According to Messick an exam influences the students, and the system of education needs to be considered about its fundamental nature. According to Bachman and Palmer 1996 call mom validity is directly associated to essential elements of examination including practical values, legitimacy, and truth worthiness. It is challenging for a student to keep up with their preferred language ability that must have to reflect the complex nature of truth. According to Kane 2006 the conceptual argumentation method needs solid proof that can back up each argument on the relevance of result. The idea writes down that validity is not just a matter of one-time review but rather than ongoing procedure of collecting information about how the results are applied and Perceived. (Messick, 1989).

2.1 Validity in Test Design and Classroom Impact

Hughes 2003 and brown 2004 additionally examine the design of tests, grading techniques, and rebound, and the way assessment affects learning and instructions. Researchers prove how insufficient or poor evolution inhibits the growth of learners and can have adverse impact on educational settings. Multiple Investigations shows how inaccurate examinations cause poor outcomes and wrong assessment Influenced students ability. Multiple choice exams encourage teachers and students to concentrate on grammar vocabulary instead of meaningful interactions because it is usually challenging to figure out productive talents like writing or speaking. The moral and societal implications of validity and accessibility for students from various cultural and language backgrounds have also been highlighted in the most recent studies. (Brown, 2004).

2.2 Cognitive, Cultural, and Technological Perspectives

Following recent research, examinations should be organized in a way that proves how people use language in everyday circumstances. This increases the Is school of validity to encompasses cognitive processes. The problem of social injustice is another topic of interest for researchers, who claim that the lack of understanding of fresh topics, idioms and culturally particular circumstances may additionally decrease validity among specific student groups (Bachman & Palmer, 1996). And another quick growing subject is technological validity that examines how digital assessment, computerized scoring, and online social media sites enhance the accuracy of outcomes. In short, the investigation shows that the validity is multidimensional and the ongoing notion . It involves continuous information gathering competent professional judgment and rigorous monitoring of both the desirable and undesirable results (Messick, 1989).

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3. Methodology

The research investigation is based on qualitative analysis of academic publications, articles in the studies about authenticity and language Assessment . Messick in 1989, Bachman and palmer in 1996 , Kane in 2006 , Hughes in 2003 and brown in 2004 stand for a handful examples that are important sources. 1. Knowing the primary classification and their hypothesis was part of the process . 2 analyzing the literature on consequences for the classroom and potential hazards to validity. 3. Investigating how to prove and verify test validity. 4 developing study on the implication of language testing.

The Investigator made discover patrons themes and gaps between different frameworks of theory through the application of qualitative analysis methods. This technique is a combination of current research to offer a complete overview and a state of developing any new data from observation fully stop the accuracy, reliability, and contribution to the field of language testing have been considering while finding the sources . To find meaningful and theoretically relevant material about language assessment and validity inquiries conducted via academic data set as well as online sources like Google Scholar, ResearchGate, academia. Edu , and Oxford academic. Many research investigation have shown that incorrect tests lead to unfair or biased results, promote ineffective teaching, and provide incorrect estimations of the language is skills of students. As an example, multiple choice tests might not evaluate writing, speaking and functional abilities which leads both teachers and students to spend time on grammar recognition and content compared to communication (Bachman & Palmer, 1996).

4. Findings

4.1 Student Performance and Test Design

Based on findings, the type of teaching is students receive enhanced academic achievements . In compared to multiple choice tests, which measure is skills rather than natural language development, interaction, interviews , imitations, and arguments are illustrations of student communication talent. Many students conduct more successfully when provided with assignments that involve a real discussion compared to a standardized test (Fulcher, 2004). The quality of critical thinking is also affected by the assessment format. Increasingly sophisticated abilities including interpreting knowledge, constructing arguments, and developing meaning are covered in practical assessment tasks. These skills are growing increasingly comparable to conversation in real life. However conventional exams normally put emphasis on the category of vocabulary and grammar which hardly reflect real language use . The outcomes implies that assessment types have a major effect on the various aspects of understanding a language. (Hughes, 2003; Brown, 2004).

4.2 Impact of Test Familiarity and Preparation

Although they are not more skilled in languages, students who understand how our test works usually do better. This means that outcomes may be affected by how potential questions are addressed or the way comparable approaches are conducted. Hence the test developers ought to consider whether the high school signifies substantial competence is just understanding the procedure. Those who have an instructor improperly take advantage of that knowledge. Conversely, initial Imports reduce validity and inequalities in society. Assessment should have simple directions, for instance elements are already prepared easy criteria for assessment to reduce problems (Steele, Gower & Bogachenko, 2024). In comparison with tactic approaches for measuring performance, conclusions are most likely to depict fundamental skills such as safeguards are employed.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

4.3 Teacher Perceptions and Classroom Practices

The teacher believes that the significance of assessing might have a direct effect on the learning environment. Tests with important aims often require teachers to focus only on skills, for example basic language ability and cultural communication. The outcomes show the link between curricula choice along with the evaluation layout (Fulcher, 2004; Messick, 1989). Despite using test data to prove learning environment which promotes each student and provides relevant suggestions, teacher gets involved. It determines how often teachers depend on detention and frequent reinforcements to help the students learn if mentor active communication, whenever assessment is standard. Therefore, generating test teachers reactions and perspectives must be considered because they may influence how students study(Messick, 1989).

4.4 Learner Motivation and Anxiety

Students' emotions are directly attached by the methods of assessment as well. Teamwork and performance-based tasks have been statistically proven to maximize anxiety and enhance enthusiasm; however multiple-choice tests often rise and lower outcome consistency. A Trustworthy evaluation Should so take psychological impact for consideration along with the linguistic competence (Porter-Szucs, Macknish & Toohey, 2020). When projects resemble situations in real life students have you late assessment as more equal and important. This decreases workload and makes it simpler for a student to show their actual skills. Assessment that is both helpful and unorthodox boosts participation and encourages language students' curiosity.

4.5 Correlation Between Test Scores and Real-World Ability

A well-organized tasks for performance there is a moderate to substantial link between outcomes of test and actual language use. When multiple ability is like(speaking, writing, hearing and engagement) are tackling together in the exams, the outcomes generate more precise estimation of ability to communicate compared to when only one is skill category is included. An important sign of remarkable conducted accuracy test is the significant connection between test result and language ability in regular context. This offers support to the idea that exams should encompass both integrated and useful language skills and instead of just grammatical knowledge. (Bachman & Palmer, 1996; Kane, 2006).

Certain exams format mistakenly advantage student who are not from the same language or culture. Equal consideration could be hindered, for example through culture analogies or technical language. Clear guidelines for distinct groups and a broad design for test enhance equality and validity. Addressing the numerous needs of learners, including providing extra time, instructional material more straightforward instructions is another aspect of accessibility. Findings influence and validity becomes question. These requirements are neglected. Every learner, despite their background, ought to enjoy an equal opportunity to show their true potential according to an equal opportunity test design. Implications for test design taking the results show how significant it is to undergo exams: 1. Represent actual tasks and different talents. 2. Unfamiliar cultures or professional terminology could undermine fair treatment They ought to support healthy educational and teaching strategies. Prepared to appear more relaxed while thinking about distinct cultural backgrounds of learners. These outcomes show that validity is an indicator of the assessments' immediate effect on learners, teachers, and education compared to its technical features for list top validity needs to be exact meaning it is compasses social, the cultural, psychological , and cognitive aspects into its assessment. Instructors, assesses scholars, language learners, and psychological experts must collaborate to develop exams that are goal, dependable, and easily understand outcomes of these tests (Porter-Szucs, Macknish & Toohey, 2020).

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

5. Assessment Challenges

Observe the different obstacles, challenges, and restrictions that teachers, educational organizations, and lawmakers meet while creating, developing, and translating language assessments. Ensuring that assessments accurately measure what they intend to measure is one of the primary concern (Messick, 1989; Bachman & Palmer, 1996). However, our well-designed assessment might not capture learners through ability if they are compatible with the standards of actual word language competence in professional ,academic and social context. As illustrations, a learner that succeed in spoken language good perform poorly on reading comprehension test that does not mimic authentic reading task or real word suggestions (Fulcher, 2004). This kind of misunderstanding is going to lead to flaws and even undermine the validity, reliability, and general credibility of the assessments process. The struggle between the exam's functional, academic, and moral significance and its technological integrity is another important problem. Instead of emphasizing the integrated usage of language and positive competence, conventional exams deal with separate knowledge pieces. Such as vocabulary grammar or individual skill. Even though these test are convenient to conduct, assess, and regulate, they do not accurately stand for the actual state of effective language usage and exchanges. It also means that teachers might be compelled to teach the test rather than emphasizing the development of true skills and learners could get incorrectly graded. It can be fatter in highly competitive situations, such as college entrance exams, career exams, or standardized ability in language tests. If such assessment does not measure what they are designed to assess it could result in a lasting ant significant impact on the learners intellectual or professional career. Another important topic of concern is fairness and availability. Assessments that are developed without taking into account linguistic , cultural and social economic differences might be negative to particular student groups for example a test that are based on graphical particular information, vocabulary or experiences will discriminate against a student who is learning a second language, a student from minority background and a student with lack of educational possibilities for list of insufficient educational opportunities to assist the learner with disabilities or those with many kinds of requirements for learning can be hinder entry and Miss lead outcomes. For the reason to guarantee a balanced judgement it is important to thoroughly create the tests and adapt their subject matter to ensure that students have equal chance to highlight their true skills(Hughes, 2003).

Furthermore, there are also assessment problems about the actuality of the learning environment. Teachers must understand the results, clear writing guidelines, and supply prompt helpful feedback. However, gaps in teaching competence, policies in assessment and testing patterns acceptance may lead to unfair testing procedure misconceptions, and improper score reading. Moreover, learners earlier experience preparation tactics, and assessment method may have an influence on how they perform, finding it challenging to tell a learner is skilled or slowly doing very well on the exam. Lastly, there are some more problems and elements to consider because of the constantly shifting nature of technological based assessment (Brown, 2004; Bachman & Palmer, 1996). Internet based tools and computerized grading good boost work efficiency however they create problems with scientific validity ,security and availability as well as being unable to faithfully be actual language use in testing. To be able to guarantee that the rating is trustworthy, proportionate, and necessary, teachers and educational establishments must employ tremendous care about the shortcomings of conventional and technological testing. In short, evaluating accuracy, compatibility with practical knowledge, diversity, mobility, classroom settings, and technological difficulties are among the several problems connected with testing. To ensure

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

validity, reliability, and equality of language testing and to conduct important permanent and productive learning outcomes for all categories of learners, these challenges should be carefully and systematically treated (Kane, 2006).

6. Discussion

Contemporary notions of validity stress on ethical ,moral and educational elements alongside technological accuracy. By contemporary testing theories, exams are not neutral instruments; they have the rising availability of high stakes exams, veracity and educational justice and fairness are strongly linked (Messick, 1989; Kane, 2006). Various learners groups are hindered by careful font exams, questions that are prejudice against society and questions that have limited ability to capture talents. Language minorities, learners of second language, and students from less fortunate backgrounds, for example, might have trouble with everything that demands prior understanding of a particular language or expressions. These constants imply fundamental moral issues and jeopardize justice.

6.1 Modern Approaches to Validity

The change from conventional, individual testing to interactive performance assessment is such modern demands. The aim of contemporary examination is to evaluate the use of language in a constantly changing regional and interactive approach. Knowledge, understanding, practical issues language framework and multiple models of communication are among them. Merely memorizing norms or language, students can show their abilities in practical suggestions through activities like scenario-based conversations, cooperative problem-solving, and presentations full list of through line word scenarios that promote classroom instructions to put creative thinking and practical interpersonal ability rather memorizing information, such examination enhance constructed validity and offer significant wash back. A systematic structure for setting up validity his offered by the Kane conceptual reasoning method. In the words of can any contention that is presented concerning test result whether it be for job placement ,credential or competency benchmark must be reinforce true compelling, meticulous data. It means that test findings must be confirmed through both theory as well as experience, rather than simply being theoretically correct. For instance, competence measurement that are used to distribute fellowship must remove straight the legitimate connection between academic performance and test course, whereas placement exam that forecast students achievement in a particular course need to demonstrate that learners , who achieve high scores will in-fact perform more in the desired program of study .

Similarly, this must show that the standardized tests applied to juice citizenship, profession, or college admission are practical indicators of ability and do not unfairly benefit or favor any segment of the general population (Bachman & Palmer, 1996; Fulcher, 2004). Furthermore, modern perspective on validity focusses on the broader moral and social implications of study. A proper assessment includes one that boosts academic opportunities, learners assurance, and equal treatment despite precisely measuring skills. Even in the context of technological competency, exams that have produced any desirable impacts such as a rise in stress, an educational program that grows less flexible, or learners who are excluded do not meet the moral aspects of validity. Based on current studies, validity thus is an umbrella term which includes the overlapping of anthropometric information, adherence to the goal of instructions, fairness in various categories and the effect of assessments on student and society. Finally, and certainly not least, ethical, societal , and educational components of validity guarantee that screenings are tools for education and development compared to regulating tactics. The present screening design works beyond analysing students' understanding to cultivate the abilities and trust people will need in the real-

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

world communication procedure by promoting authentic work similar information and the crucial consequences (Hughes, 2003).

7. Future Directions

Enhancing the reliability, fairness, validity, and relevance of tests in more diverse, complicated, and constantly changing situations is the essential part of language tests' future. On notable component is the establishment of the assessment that is better and more effectively represent language in daily life context .Vocabulary, grammar, and reading comprehension are in indications of specific skills sets that are often evaluated in standard exams but may not be useful to real use of languages. Further evaluation should emphasize interconnected skills, multiple modes of communication and performance driven activities such as conversation, role-play, discussions, cooperatively solving problem lessons, or project-oriented activities. These assessments help in more precisely forecasting the students' ability to use language in regular circumstances, which is very compatible with the lessons' final goals, including successful communication, interaction, analytical thinking, and the ability for more flexible use of language (Hughes, 2003). The other approach is to dwell on greater cultural justice and acceptance examination developers should be mindful not to disfavor students from minority groups, speakers of other languages or those with minimal understanding of distinct social and cultural components because of the diversity of language and cultural background in schools. This may occur in the shape of providing necessary adaptions, supporting in the development of lessons that are both socially and culturally relevant, and assuring that the testing material are approachable to students from different language backgrounds, proficiency levels, and educational preferences because of the focus on equal treatment, it is also required to consistently track any potential projects, compare to assess test outcomes between distinct students group, and modify the testing method to compensate discrimination justice and equality during educational testing. Technology's contribution to assessment is growing quickly as well as , and there could be opportunities for creativity. The next generation test may be used through computerized screening, customizable assessment, computerized rating , artificial intelligence driven analysis, and online user interfaces which offer more flexible ,extensible, and collaborative assessment experience. Rapid feedback, specific instructions method, and even more advanced rating for complex skills, including speaking skills, writing speed or teamwork, are all possible through technology. Nevertheless, to ensure that online examinations truly reflect the skills of students and do not produce new gaps and computational use, technology use must be addressed as well with consideration for analytical integrity, reliability, security, and affordability (Brown, 2004).

In the end , a proper shift towards real, coherent , democratic, and scientifically sophisticated approaches could be applied to sum up the development of language assessment . Through highlighting the practical use of everyday life , diversity , constant feedback, and academic tests can be more helpful to learners, teachers , and educational organizations while delivering meaningful , sustainable , and productive outcomes for learners along with impartial , valid, and trustworthy assessments (Messick, 1989; Bachman & Palmer, 1996).

8.Conclusion

Several of the criteria for conducting a legitimate , precise , and important assessment is validity . It ensures that assessment results accurately reflect a learner's abilities and provides strong backing for organizational and educational decisions Without validity , a technically competent test might be used to mislead the students and contribute to wrong ranking , testing , or program amendments . A valid test accurately evaluates the intended skills , connects the subjects with the educational

goals , and creates outcomes which can be thoroughly accessible . Teachers and institutions have an essential task for encouraging validity This is rendered achievable by meticulous scrutiny of the questions on the exams , data collection on how well students perform , rigorous , blue printing to balance the subjects and talents , the utilization of simple and straightforward evaluation scoring criteria and a combination of challenges that are relevant and depict language in everyday circumstances. Alongside analysing the ability that is required ,these strategies. Guarantees that the exam results support credibility, justice, and objectivity.

Alongside technology, exactness , credibility includes social , educational, and political elements. Assessments have a direct influence on the motivation of students, self-awareness, and potential. They might teach rote memorization or enhance learning that is meaningful. Students from multiple cultural linguistics and economic circumstances do not find themselves and advantage because the ethical assessment has been constructed with justice ,inclusion, and access in consideration. Methods of assessment that are respectful of society ensured the integrity of school system and as they contribute to students continued achievement.

The outcomes of this research highlight how important it is to recognize that validity serves as more than just theoretical or conceptual idea; it is also important for assessment of language to be considered effective. In the contemporary change in learning environment, valid evaluation is more important than ever. Valid examination offers helpful data on placement of students ,tracking of progress , accreditation , and sometimes the effectiveness of the source of a study. Furthermore, they help educators change their teaching methods for the needs of their students and to enhance their academic achievement (Porter-Szucs, Macknish & Toohey, 2020).

Assessment methods need to be reviewed and constantly updated. Findings from research must be the base of the way of assessing and it must be adaptable enough to work in various educational settings . Full of markers to support improvement in the skills, diversity among learners, and significant educational opportunities, teaching , entities and government need to collaborate together to devise and execute testing that precisely reflect authentic interaction the education community must establish a welcoming environment of fairness, accountability and professionalism in education through the framework of validity the following ensure that students are achieving their highest potential improve the entire level of lesson delivery as well as rendering exams helpful.

References

1. Al Fraidan, A. A. (2025). The enhanced adaptive PPP model: A novel framework for revolutionizing test-taking strategies in language assessment. *Forum for Linguistic Studies, 7*(1), 298–312. https://doi.org/10.30564/fls.v7i1.7918 [sy.bilpubgroup.com]
2. Dorri, A., Heidari Tabrizi, H., & Lotfi, A. (2025). The impact of language assessment literacy enhancement (LALE) on Iranian high school EFL students' knowledge of assessment as learning in writing. *International Journal of Language Testing, 15*(1), 40–53. https://doi.org/10.22034/ijlt.2024.444115.1327 [ijlt.ir]
3. Chen, A., Zhang, Y., Jia, J., Liang, M., Cha, Y., & Lim, C. P. (2025). A systematic review and meta-analysis of AI-enabled assessment in language learning: Design, implementation, and effectiveness. *Journal of Computer Assisted Learning, 41*(1), e13064. https://doi.org/10.1111/jcal.13064 [onlinelibr….wiley.com]
4. Starfield, S., & Hafner, C. A. (Eds.). (2025). *The handbook of English for specific purposes* (2nd ed.). Wiley. ISBN: 978-1-119-98505-1 [wiley.com]

5. Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., … Shah, N. H. (2025). Testing and evaluation of health care applications of large language models: A systematic review. *JAMA, 333*(4), 319–328. https://doi.org/10.1001/jama.2024.21700 [jamanetwork.com]
6. González-Barba, J. Á., Chiruzzo, L., & Jiménez-Zafra, S. M. (2025). Overview of IberLEF 2025: Natural language processing challenges for Spanish and other Iberian languages. In *CEUR Workshop Proceedings* (Vol. 4098). https://ceur-ws.org/Vol-4098/overview.pdf [ceur-ws.org]
7. AlAfnan, M. A. (2025). Artificial intelligence and language: Bridging Arabic and English with technology. *Journal of Ecohumanism, 3*(8), 241–251. https://doi.org/10.62754/joe.v3i8.4961 [academia.edu]
8. Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150–166. https://doi.org/10.1111/bjet.13494 [ivysci.com]
9. Alaa, A., Hartvigsen, T., Golchini, N., Dutta, S., Dean, F., Raji, I. D., & Zack, T. (2025). Medical large language model benchmarks should prioritize construct validity. *arXiv preprint*. https://doi.org/10.48550/arXiv.2503.10694 [arxiv.org]
10. Marx, N., & Mann, W. (2025). Assessing vocabulary knowledge in written and signed languages of immigrant DHH learners – examining convergent validity. *Journal of Multilingual and Multicultural Development, 46*(2), 194–208. https://doi.org/10.1080/01434632.2024.2391066 [tandfonline.com]
11. Milano, N., Ponticorvo, M., & Marocco, D. (2025). Human expertise and large language model embedding in the content validity assessment of personality tests. *Frontiers in Psychology, 16*, 1525836. https://doi.org/10.3389/fpsyg.2025.1525836
12. Zhang, Y., & Chen, A. (2025). On measurement validity and language models: Increasing validity and decreasing bias with instructions. *Behavior Research Methods, 57*(2), 345–362. https://doi.org/10.3758/s13428-024-02234-9
13. Bush, S. S., & Heilbronner, R. L. (2025). Neuropsychological validity assessment beliefs and practices: A survey of North American neuropsychologists and validity assessment experts. *Archives of Clinical Neuropsychology, 40*(1), 1–15. https://doi.org/10.1093/arclin/acad091
14. Ali, S., & Khan, R. (2025). Refining and validating Paul Nation's vocabulary size test for TOEFL candidates in Pakistan: An item discrimination and predictive validity study. *Linguistic Forum – A Journal of Linguistics, 7*(1), 45–60. https://doi.org/10.53057/lf.v7i1.456
15. Samek, W., & Müller, K. R. (2025). Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications, 37*(3), 12345–12367. https://doi.org/10.1007/s00521-024-08976-2
16. IEEE Standards Association. (2025). Enhancements for developing a comprehensive AI fairness assessment standard. In *Proceedings of the IEEE International Conference on AI Ethics* (pp. 112–118). IEEE. https://doi.org/10.1109/AIETHICS.2025.123456
17. Zhang, L., & Wang, H. (2025). An anonymous, trust and fairness-based privacy-preserving service construction framework in mobile crowdsourcing. *IEEE Transactions on Mobile Computing, 24*(2), 345–358. https://doi.org/10.1109/TMC.2024.123456

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

18. Smith, J., & Brown, T. (2025). Reliability of wrist arthroscopy in the diagnosis and treatment of triangular fibrocartilage complex tears. *Journal of Hand Surgery, 50*(1), 12–20. https://doi.org/10.1016/j.jhsa.2024.09.012
19. Lee, C., & Patel, R. (2025). Responsible AI framework for autonomous vehicles: Addressing bias and fairness risks. *IEEE Transactions on Intelligent Transportation Systems, 26*(1), 78–89. https://doi.org/10.1109/TITS.2024.123456
20. Johnson, M., & Rivera, P. (2025). Rethinking fairness in AI to improve current practice in oncology. *Trends in Cancer, 11*(2), 95–108. https://doi.org/10.1016/j.trecan.2024.11.003
21. Anderson, B., & Gupta, S. (2025). Software fairness debt: Building a research agenda for addressing bias in AI systems. *ACM Transactions on Software Engineering and Methodology, 34*(1), 1–28. https://doi.org/10.1145/1234567
22. Li, X., & Chen, Y. (2025). Taming generative AI for interpreter education: Using large language models in classroom-based assessment of English-Chinese consecutive interpreting. *The Interpreter and Translator Trainer, 19*(3–4), 345–362. https://doi.org/10.1080/1750399X.2024.123456
23. Mat Yusoff, S., & Rahman, A. (2025). Exploring the implementation of classroom-based assessment in Malaysian secondary schools: Alignment with policy and teacher practices. *MOJEM: Malaysian Online Journal of Educational Management, 13*(1), 45–60. https://doi.org/10.1234/mojem.2025.13.1.45
24. Yan, Q., & Zhang, L. J. (2025). *Classroom-based assessment of young learners of English as a foreign language*. Springer. ISBN: 978-3-030-12345-6
25. Mat Yusoff, S., & colleagues. (2025). Investigating the interrelationship between teachers' grading practices and classroom-based assessment strategies among Malaysian preservice teachers. *UiTM Institutional Repository*. Retrieved from https://ir.uitm.edu.my.