English Language Teaching and Linguistics Studies
ISSN 2640-9836 (Print) ISSN 2640-9844 (Online)
Vol. 8, No. 3, 2026
www.scholink.org/ojs/index.php/eltls

Original Paper

A Bibliometric Analysis on Validity of IELTS and TOEFL in Language Testing

Yujie Ji*

School of Foreign Languages, Southeast University, Jiangsu, Nanjing, 211189, China

* Corresponding Author

Received: April 12, 2026 Accepted: June 09, 2026 Online Published: June 24, 2026

doi:10.22158/eltls.v8n3p268 URL: http://dx.doi.org/10.22158/eltls.v8n3p268

Abstract

This study conducts a bibliometric analysis of research on the validity of the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) from 2009 to 2025. Using Biblioshiny, 70 articles from the Web of Science Core Collection were analyzed to explore publication trends, key journals, influential authors, hot topics, and emerging research directions. Results indicate a growing interest in validity research, with LANGUAGE TESTING as the dominant journal. Key themes include construct validity, predictive validity, and automated assessment. The findings highlight the need for more diverse research contexts and advanced validation methodologies, providing insights for test developers, educators, and researchers.

Keywords

IELTS, TOEFL, Validity, Bibliometric analysis, Language testing

1. Introduction

In an era defined by unprecedented global mobility and interconnectedness, English has solidified its position as the lingua franca of higher education, international business, and cross-cultural communication (Gagen & Faez, 2024). This linguistic dominance has amplified the significance of standardized English proficiency tests, with the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) emerging as the most widely recognized and utilized assessments worldwide. Administered to millions of test-takers annually, these high-stakes evaluations serve as critical gatekeepers, influencing decisions related to university admissions, professional licensure, and immigration in English-speaking contexts. Central to their legitimacy and utility is the concept of validity, a multifaceted construct that refers to the degree to which a test measures what it purports to measure and the appropriateness of inferences drawn from its scores.

Validity, according to Messick (1989), is not a static attribute of a test but an evolving accumulation of evidence that supports the intended interpretation and use of results. For IELTS and TOEFL, which claim to assess "academic English proficiency," validity evidence must span multiple dimensions: content validity, construct validity, criterion-related validity, and consequential validity (Bachman & Palmer, 2010). Messick's conceptualization of validity has been fundamental for subsequent discussion of validity in educational measurement and professional language testing community (Carlsen & Rocca, 2021). Together, these dimensions form a validity argument that justifies the tests' role in high-stakes decision-making (Aryadoust, 2023).

Over the past few decades, research on the validity of IELTS and TOEFL has expanded significantly, reflecting growing scholarly and practical scrutiny of their impact on global education systems (Gagen & Faez, 2024). Early studies predominantly focused on content and criterion-related validity, investigating whether test tasks authentically mirrored academic contexts and whether scores could reliably predict academic success. More recent scholarship has broadened to encompass construct validity, leveraging advances in psychometrics and natural language processing to unpack the underlying language skills measured by test performance. Consequential validity has also gained prominence, with scholars examining how these tests shape classroom practices, perpetuate or mitigate inequities in score interpretation, and influence educational disparities (Hashemi & Daneshfar, 2018).

Despite this growing body of research, critical gaps remain in our understanding of the validity landscape. Existing studies are often fragmented, focusing on single dimensions of validity or specific test components (e.g., writing tasks) without synthesizing findings across contexts, methodologies, or time periods (Uysal, 2009). This fragmentation impedes the development of a comprehensive validity argument for IELTS and TOEFL, leaving educators, test developers, and policymakers, without a clear grasp of cumulative evidence or emerging trends. For instance, while numerous studies have explored the predictive validity of TOEFL scores for undergraduate GPA, few have systematically compared these findings across regions or academic disciplines (Llosa & Malone, 2019), or examined how validity coefficients may have shifted following test revisions. Similarly, research on IELTS's construct validity has primarily centered on academic contexts (Dang, C. N., & Dang, 2023), with limited attention to its General Training module, despite its widespread use in immigration and professional settings.

Bibliometric analysis offers a powerful tool to address these gaps by systematically mapping the intellectual structure of a field, identifying key themes, influential scholars, and evolutionary patterns (Ellegaard, & Wallin, 2015). By quantifying and visualizing publication trends, citation networks, and keyword co-occurrences, bibliometrics can reveal hidden connections between studies, highlight underrepresented areas, and guide future research agendas. In the context of language testing, such analysis can illuminate how validity research has evolved in response to shifts in global education, technological advancements (e.g., automated scoring), and theoretical developments in language assessment theory. Specifically, it addresses the following research questions:

1. What are the temporal trends, key publication venues, and influential authors in research on the validity of IELTS and TOEFL?
2. What are the dominant themes and emerging trends in this research field?

2. Methods

This study used the tool called Biblioshiny to conduct the analysis on studies related to validity of IELTS and TOEFL, aiming to find out the development dynamics and research trends in past 15 years (2009-2025).

2.1 Data Collection

This study employed the bibliometric approach, which is the quantitative analysis of literature based on relevant information such as authors and citation numbers, making it clearer to conduct systematic literature review. The methods for data collection are described as follows. The Web of Science (WoS) Core Collection was chosen as the primary data source. WoS is widely regarded as a high-quality database indexing peer-reviewed journals across disciplines, ensuring the inclusion of rigorous scholarship in language testing (Jiménez, Maz, & Bracho, 2013). It also provides comprehensive metadata (e.g., authors, keywords, citations) necessary for bibliometric analysis. The search strategy applied in this study was TS= (“IELTS” OR “TOEFL” AND “validity”). These articles were selected from SSCI indexed journals. The time span was set from January 1st, 2009 to May 31st 2025. The inclusion criteria are as follows: 1). the document type is article; 2). the publication language is English; 3). keywords include validity. The exclusion criteria are as follows: 1). repeated publication of literature; 2). studies focusing solely on test-taking strategies without discussing validity; 3). articles comparing IELTS/TOEFL to other tests (e.g., PTE) without centering validity. After reading the titles and abstracts of the literature carefully, some irrelevant articles which are not suitable for this topic were removed. Lastly, a total of 70 articles were collected for the study.

2.2 Data Processing

Following data collection, all retrieved records were exported in bibtex format, including bibliographic information such as titles, authors, keywords, and abstracts. The dataset was subsequently imported into Biblioshiny for bibliometric analysis. Annual publication outputs were calculated to examine the publication trends over the selected study period. Journal information was analyzed to identify the principal publication venues within the field. Citation data were used to determine the most influential publications based on citation counts.

To explore the intellectual structure of the research domain, keyword co-occurrence networks were constructed using binary counting, with a minimum occurrence threshold of 10 to ensure the inclusion of representative and meaningful themes. Research hotspots were identified through the analysis of high-frequency keywords, keyword clustering, and highly cited publications. In addition, thematic maps were employed to examine the evolution of research themes and to reveal emerging trends and future directions in the field.

3. Results and Discussion

This study has reviewed the relevant literature on validity of IELTS and TOEFL. This section presents the results of the bibliometric analysis, organized by research question, with critical discussion of trends and implications.

3.1 Publication Trend

The volume of literature output in a field serves as a critical barometer of its maturity, trajectory, and relevance to contemporary needs (Lei & Liu, 2019). For research on the validity of IELTS and TOEFL, the annual and cumulative publication data presented in Figure 1 reveal a dynamic landscape of evolving scholarly engagement. Over the years, the cumulative number of articles shows a generally upward trend, indicating a growing body of research in the field. The annual publication output, however, fluctuates, with some years having relatively higher numbers of new publications and others having lower ones. From 2009 to around 2015, the cumulative growth is relatively slow, but after that, it accelerates, suggesting an increase in research activity and interest in the field. The annual publication line, with its variations, reflects the yearly dynamics of research output, possibly influenced by factors such as emerging trends, funding opportunities, or changes in research focus within the field.

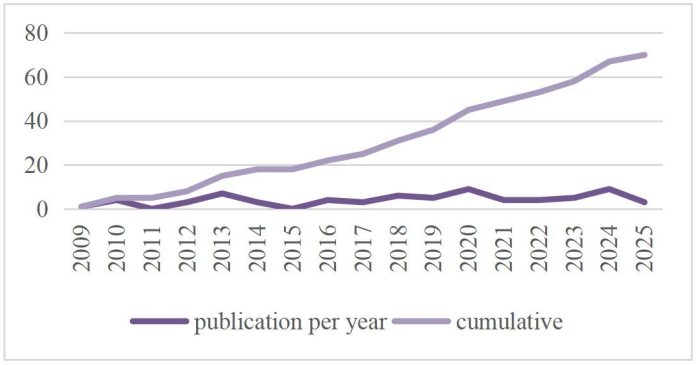

Figure 1. The Annual Output of Articles in This Field

3.2 Analysis of Journals, Countries and Authors

Table 1 illustrates the top 5 most productive journals in the relevant research field, ranked by the number of publications. As depicted, “LANGUAGE TESTING” emerges as the leading journal, with a substantial output of 29 publications, indicating its prominent role in disseminating research within this domain. “LANGUAGE ASSESSMENT QUARTERLY” follows in the second position, contributing 9 publications, while “ASSESSING WRITING” secures the third rank with 8 publications. “LANGUAGE TESTING IN ASIA” and “SYSTEM” occupy the fourth and fifth ranks, with 5 and 4 publications respectively. This distribution highlights the concentration of research efforts in a relatively small set of specialized journals, which likely serve as key platforms for scholars to share findings related to language testing and assessment. The dominance of “LANGUAGE TESTING” suggests its centrality in shaping and disseminating knowledge within this research area, while the

presence of region-specific and other specialized journals reflects the multidisciplinary and international nature of the field.

The dominance of these five journals indicates a high degree of specialization in the field. While this concentration fosters cumulative knowledge building through focused dialogue, it also raises concerns about disciplinary insularity. Research on validity in less specialized journals, such as those focused on higher education or applied linguistics, may be underrepresented, limiting cross-disciplinary insights into how language assessment interacts with broader educational outcomes.

Table 1. Top 5 Most Productive Journals

Rank	Journal	Publications
1	LANGUAGE TESTING	29
2	LANGUAGE ASSESSMENT QUARTERLY	9
3	ASSESSING WRITING	8
4	LANGUAGE TESTING IN ASIA	5
5	SYSTEM	4

Figure 2 illustrates the geographic distribution of publications, with the United States, United Kingdom, Australia, Canada, and China emerging as the most productive nations. This pattern reflects the global reach of IELTS and TOEFL, as well as the concentration of research capacity in countries with large international student populations and well-established assessment research institutions. The United States leads in output, consistent with its role as a hub for language testing innovation (home to ETS, developer of TOEFL) and its large cohort of international students. UK and Australian contributions are similarly unsurprising, given their status as major destinations for international education and as the home of IELTS (co-owned by the British Council and IDP Education). China's notable presence in the data reflects its position as the largest source of international test-takers, driving demand for research on how IELTS/TOEFL scores predict academic success for Chinese students.

Regional disparities, however, are evident. African and South American countries are minimally represented, despite growing interest in IELTS and TOEFL in these regions. This gap may stem from limited access to research infrastructure, language barriers in publishing, or a focus on local rather than international assessments. Addressing this imbalance is critical, as validity is context-dependent; findings from Western contexts may not generalize to test-takers from diverse linguistic or educational backgrounds.

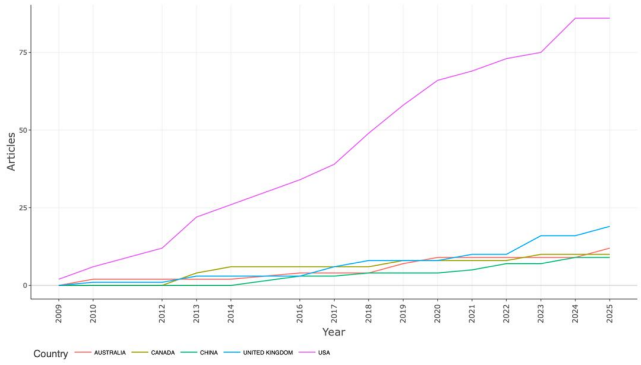

Figure 2. Countries' Production over Time

Table 2 and Figure 3 highlight the most productive and cited authors in the field, offering insights into intellectual leadership and collaborative networks. Plakans Lia stands out with 4 articles, focusing on writing assessment validity, while Cho Yeonsuk and Aryadoust Vahid each contributed 3 articles, specializing in construct validity and statistical validation methods, respectively. Fractionalized frequency scores, which account for co-authorship, further refine these rankings: Kyle Kristopher leads with 1.58, reflecting his collaborative work on natural language processing in TOEFL validation. Bridgeman Brent contributes 3 articles with a fractionalized score of 1.03, focusing on predictive validity, aligning with ETS's institutional focus on linking test scores to academic outcomes. Citation analysis (Figure 3) reveals that Weigle Sara Cushing and Llosa Lorena are among the most locally cited authors, with their work on writing assessment validity and second language acquisition forming foundational references. This overlap between productivity and citation impact indicates that these authors not only publish frequently but also shape the theoretical and methodological direction of the field. Notably, many top authors are affiliated with institutions like ETS, the University of Michigan, or the University of Melbourne, organizations with strong ties to test development or international education. This affiliation underscores the interplay between academic research and practical assessment design, as scholars collaborate with test developers to address real-world validity challenges.

Table 2. Top 5 Most Relevant Authors

Authors	Articles	Articles fractionalized frequency
PLAKANS LIA	4	1.67
ARYADOUST VAHID	3	1.33
BRIDGEMAN BRENT	3	1.03
CHO YEONSUK	3	1.33
KYLE KRISTOPHER	3	1.58

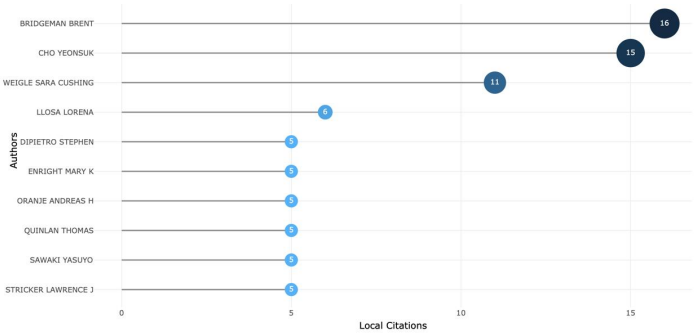

Figure 3. Most Locally Cited Authors

3.3 Most Highly Cited Articles and Highly Cited References

Citation patterns reveal which studies have most influenced the field, highlighting seminal contributions to validity theory and methodology. Figure 4 shows the most globally cited documents. Plakans (2013) leads with 114 citations, examining construct validity in TOEFL iBT integrated writing tasks, highlighting the importance of integrated skills assessment. Weigle (2013) and Cho (2012) also feature prominently, focusing on writing assessment validity and TOEFL's predictive power, respectively. These studies set benchmarks for validity research, emphasizing empirical rigor and alignment with contemporary assessment frameworks (e.g., the Standards for Educational and Psychological Testing). Figure 5 highlights articles frequently cited within the dataset, indicating their role in shaping the immediate research community. Enright (2010) leads, with its analysis of TOEFL iBT's predictive validity for graduate students becoming a touchstone for subsequent studies on academic success. Bax (2013) is also prominent, critiquing the "washback" effect of IELTS/TOEFL on language teaching, how test preparation may narrow curricula, raising questions about consequential validity (Bax, 2013). Together, these citation patterns reveal a field grounded in both theoretical rigor (e.g., validity frameworks) and practical application (e.g., AES, predictive validity). They also highlight the enduring influence of studies that bridge theory and practice, such as those linking test scores to real-world outcomes.

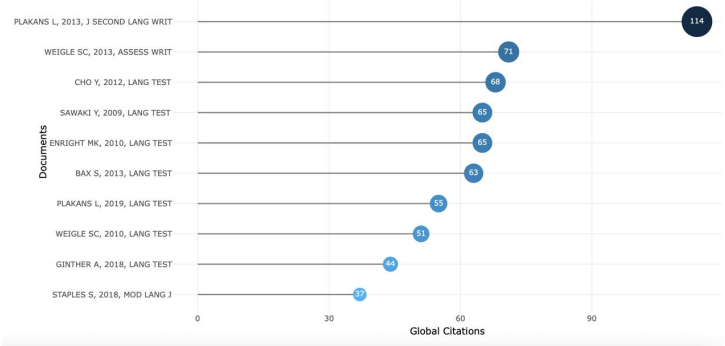

Figure 4. Most Global Cited Documents

Cited References	Citations
BACHMAN L. F., 1996, LANGUAGE TESTING IN PRACTICE: DESIGNING AND DEVELOPING USEFUL LANGUAGE TESTS	17
CHAPELLE CA, 2008, ESL APPL LING PROF, P1	11
CHAPELLE C, 2008, ESL APPL LING PROF, PIX	10
MESSICK S., 1989, EDUCATIONAL MEASUREMENT, V3RD, P13, DOI DOI 10.7203/RELIEVE.22.1.8248	10
CHO Y, 2012, LANG TEST, V29, P421, DOI 10.1177/0265532211430368	9
WEIGLE SC, 2010, LANG TEST, V27, P335, DOI 10.1177/0265532210364406	9
CUMMING A., 2005, ASSESSING WRITING, V10, P5, DOI DOI 10.1016/J.ASW.2005.02.001	7
KANE MT, 1992, PSYCHOL BULL, V112, P527, DOI 10.1037/0033-2909.112.3.527	7
BACHMAN L. F., 1990, FUNDAMENTAL CONSIDERATIONS IN LANGUAGE TESTING	6
BACHMAN L. F., 2010, LANGUAGE ASSESSMENT IN PRACTICE: DEVELOPING LANGUAGE ASSESSMENTS	6

Figure 5. Most Local Cited References

3.4 Keyword Co-Occurrence Analysis

Keyword co-occurrence and thematic mapping identify core research themes and emerging trends, offering a snapshot of the field's intellectual structure (Lei & Liu, 2019). Figure 6 offers a concise visual summary of the key themes in research regarding the validity of IELTS and TOEFL assessments. At the center, "validity" stands out prominently, highlighting its status as the core concept under investigation, researchers are deeply concerned with determining whether these tests measure what they claim to, across different language skills and contexts. Surrounding it, terms like "writing assessment," "listening assessment," and "speaking assessment" reflect a focus on evaluating specific language abilities, while "toefl ibt," "ielts," and "toefl" anchor the research firmly to these two major English-proficiency tests. Concepts such as "construct validity" and "predictive validity" delve into the nuanced dimensions of validity, with the former examining if tests measure theoretical language-proficiency constructs (e.g., academic English skills) and the latter exploring whether scores can forecast outcomes like academic success. Additionally, "automated essay scoring" points to technological advancements in assessment methods, and "international students" and "academic success" underscore the real-world stakes, as these tests often influence decisions about students' access to education and their performance within academic settings. Overall, the word cloud maps out a research landscape centered on rigorously evaluating IELTS and TOEFL, blending theoretical inquiries into validity with practical concerns about test design, skill measurement, and real-world impact.

Figure 7 presents a tree-map visualization of keywords in the relevant research domain, offering a structured and intuitive overview of key themes. At a glance, "performance" (15 occurrences, 10% share) and "validity" (13 occurrences, 9% share) emerge as the most prominent themes, occupying the largest rectangles. This underscores their centrality in the research, likely reflecting a strong focus on evaluating how language tests perform in measuring intended constructs and the overall validity of assessments like IELTS and TOEFL. "Scores" (11 occurrences, 7% share) and "language" (8 occurrences, 5% share) follow closely, highlighting the significance of test results and the linguistic aspects under scrutiny. Other notable keywords include "students" (7 occurrences, 5% share), "proficiency" (7 occurrences, 5% share), and "tasks" (6 occurrences, 4% share), which point to research interests in understanding test - taker characteristics, language proficiency levels, and the

design and impact of test tasks. Smaller rectangles represent a diverse range of more specialized or less frequently discussed concepts, such as “anxiety,” “comprehension,” “fluency,” and “GPA.” These cover areas like test-taker affect, cognitive processes, language skills, and academic outcomes, illustrating the breadth of the research landscape. In essence, this tree-map not only reveals the dominant research foci but also maps out the interconnected web of sub-themes, providing a comprehensive snapshot of the keyword ecosystem in studies related to language testing and assessment. It helps researchers and readers quickly identify core areas of inquiry and discover potential links between different aspects of the research domain.

Figure 6. Word Cloud of Authors’ Keywords

Figure 7. Tree-Map of Keywords Plus

3.5 Current Status and the Emerging Trends

The thematic map in Figure 9 and trend topics in Figure 8 reveal evolving interests. “Cognitive validity” (linked to “eye-tracking” and “discourse analysis”) is a growing theme, as researchers move beyond score correlations to explore test-takers’ cognitive processes (e.g., how they comprehend reading passages in IELTS). This reflects a shift toward “process validity”, ensuring tests measure the intended cognitive skills, not just outcomes. “Predictive validity” remains dominant but has expanded to include nuanced variables, such as the impact of test preparation courses and discipline-specific performance

(e.g., STEM vs. humanities). “Automated speech scoring” is also rising, with studies like Xi (2012) paving the way for research on AI in speaking assessments. Notably, “international students” and “academic success” are increasingly linked, reflecting a focus on equity, ensuring IELTS/TOEFL do not disadvantage students from non-Western educational systems. This trend responds to critiques that current validity research underrepresents diverse test-taker populations.

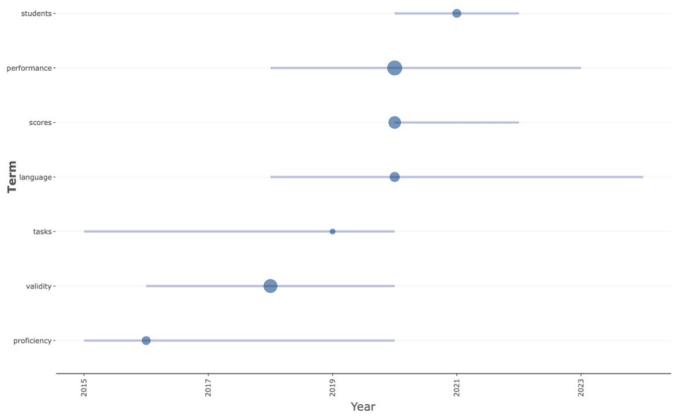

Figure 8. Trend Topics

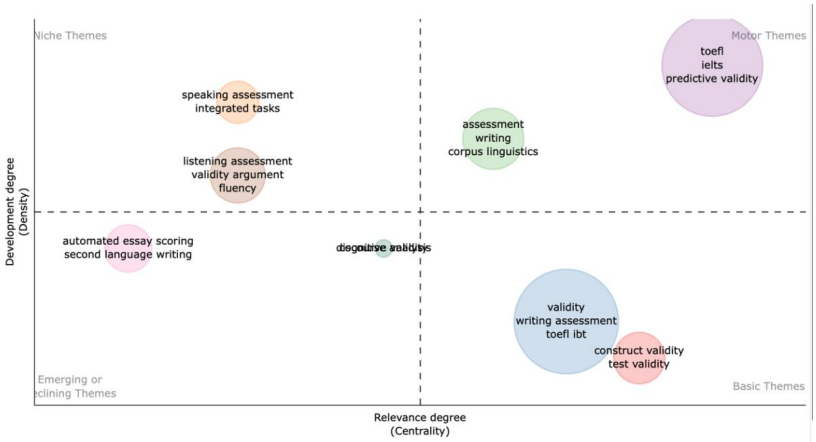

Figure 9. Thematic Map

4. Conclusion

In this study, the bibliometric analysis provides a comprehensive overview of the research landscape surrounding the validity of IELTS and TOEFL, spanning 15 years (2009-2025). By synthesizing 70 articles from the Web of Science Core Collection, the study identifies key trends, influential contributors, and dominant themes, offering critical insights into how scholars have interrogated the validity of these high-stakes language assessments. First, the publication trends reveal a field in steady growth, with accelerating research activity post-2015. This expansion reflects the increasing global demand for rigorous validation of English proficiency tests, driven by rising international student mobility and the need for evidence-based admissions practices. Second, the analysis of journals,

countries, and authors highlights a concentrated yet global network of scholarship. “LANGUAGE TESTING” emerges as the preeminent venue, underscoring its role in shaping discourse on validity in language testing. Geographically, productivity is dominated by English-speaking nations with large international student populations, such as the United States, the United Kingdom, and Australia, reflecting their institutional stakes in IELTS and TOEFL. Lastly, citation patterns and keyword analyses identify core themes that define the field. These themes reflect scholars’ efforts to determine whether IELTS and TOEFL measure theoretical constructs of academic English proficiency and predict real-world outcomes like academic success. Emerging trends, such as “cognitive validity” and “automated speech scoring,” point to a shift toward investigating the cognitive processes underlying test performance and the impact of technology on assessment validity, areas likely to gain prominence as AI-driven scoring becomes more prevalent. The findings collectively underscore the complexity of validity in language testing, as scholars grapple with theoretical frameworks (e.g., Messick’s unified validity) and practical concerns.

However, this study is not without limitations. Firstly, this study only gets data from one database, which is Web of Science core collection database. Though this database is highly reputable and widely used, it may not include all relevant publications, especially those from regional or less widely indexed journals. This could potentially lead to an underrepresentation of certain regions or emerging research areas. Secondly, the present study only uses one software, Biblioshiny to conduct the analysis. While Biblioshiny is a powerful tool, using multiple software tools or methods could provide a more robust and nuanced analysis. Future research could benefit from a more comprehensive list of search terms or the use of additional search strategies. In conclusion, while the bibliometric analysis provides valuable insights into the development and current state of validity, the limitations outlined above suggest areas for further research and refinement. Future studies should aim to address these limitations to provide a more comprehensive and nuanced understanding of validity.

References

Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. *Language Testing*, *40*(1), 8-14.

Bachman, L. F., & Palmer, A. S (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441-465.

Carlsen, C. H., & Rocca, L. (2021). Language Test Misuse. *Language Assessment Quarterly*, *18*(5), 477-491.

Dang, C. N., & Dang, T. N. Y. (2023). The Predictive Validity of the IELTS Test and Contribution of IELTS Preparation Courses to International Students’ Subsequent Academic Study: Insights from Vietnamese International Students in the UK. *RELC Journal*, *54*(1), 84-98.

Ellegaard, O., & Wallin, J.A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, *105*, 1809-1831.

Gagen, T., & Faez, F. (2024). The predictive validity of IELTS scores: A meta-analysis. *Higher Education Research and Development*, *43*(4), 873-888.

Hashemi, A., & Daneshfar, S. (2018). A Review of the IELTS Test: Focus on Validity, Reliability, and Washback. *A Review of the IELTS Test Indonesian Journal of English Language Teaching and Applied Linguistics* (Vol. 3, Issue 1).

Ihlenfeldt, S. D., & Rios, J. A. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, *40*(2), 276-299.

Jiménez, N., Maz, A., & Bracho, R. (2013). Bibliometric analysis of mathematics education journal in the SSCI. *International Journal of Research in Social Sciences*, *2*(3), 26-32.

Lei, L., & Liu, D. (2019). Research trends in Applied Linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, *40*(3), 540-561.

Llosa, L., & Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing*, *36*(2), 235-263.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan Publishing Company.

Uysal, H. H. (2009). A critical review of the IELTS writing test. *ELT Journal*, *64*(3), 314-320.