

A BIBLIOMETRIC ANALYSIS OF RESEARCH ON LANGUAGE SKILL RATING SCALES

Ying Zhou^a and Weijie Hu^{b*}

^aMA candidate, College of Foreign Languages, Zhejiang Normal University; No. 688 Yingbin Avenue, Jinhua City, Zhejiang 321000, China

^bProfessor, College of Foreign Languages, Zhejiang Normal University; No. 688 Yingbin Avenue, Jinhua City, Zhejiang 321000, China

<https://doi.org/10.59009/ijlllc.2025.0226>

Received Date:10 Sept 2026/Published Date:21 Oct 2026

ABSTRACT

Rating scales are a core tool connecting test-takers' language performance and score interpretation in language assessment, and their psychometric properties directly affect test reliability and validity. However, existing reviews mostly adopt traditional narrative methods and lack systematic bibliometric analysis of international research on language skill rating scales. Using the Web of Science Core Collection, this study screened 99 core publications (2008–2026) and applied CiteSpace to analyze publication trends, institutional and national collaboration networks, keyword co-occurrence and clustering, and burst detection. Findings: publication trends show fluctuating growth with a 2023 peak followed by a sharp decline, reflecting a paradigm shift from classical scale validation to technological integration and cross-cultural adaptation; the institutional collaboration network is highly fragmented, with North America and Western Europe leading; the field forms six core clusters—language testing, generalizability theory, rating criteria, many-facet Rasch measurement, automated scoring, and localization; hotspot evolution has gone through three stages—scale development and reliability, psychometric validation and rating processes, and construct validity and comprehensibility. This study provides a knowledge map for rating scale development, validation, and cross-context adaptation.

Keywords: Language Skill Rating Scales, Citespace, Visualization Analysis, Research, Hotspots, Frontier Evolution.

1. INTRODUCTION

In second language assessment, rating scales are the core tool connecting test-takers' language performance and score interpretation. Luoma (2004) pointed out that rating scales are a direct manifestation of test developers' understanding of the test construct, and constitute one of the core elements in defining the test construct. Tang and Ma (2024) also argued that rating scales can not only diagnose dynamic assessment in a timely manner, but also organically connect diagnostic assessment with teaching intervention.

Scholars have reviewed rating scales in the field of language assessment from different perspectives. Min et al. (2020) conducted a systematic review of 70 empirical studies in the field of language assessment in China, listing "rating and rating scales" as one of six major themes. Knoch (2009) compared two writing rating scales and found that scales with detailed level descriptors could significantly improve rater reliability. In the field of speaking assessment, Li et al. (2019) found that when developing speaking scales and training raters,

one should not focus only on grammatical accuracy, but should also pay special attention to productive features such as fluency and pronunciation, as well as vocabulary knowledge. Iwashita et al. (2008) found that measures of grammatical accuracy and complexity, vocabulary, pronunciation, and fluency all helped distinguish overall performance levels, but vocabulary and fluency had the most significant impact.

However, previous research still has clear research gaps. Min et al. (2020) empirical research on language assessment in China focused on the Chinese context and adopted traditional narrative methods, and did not conduct a bibliometric analysis of rating scale research at the international level. Aryadoust et al. (2020) used bibliometric methods to conduct a knowledge mapping analysis of measurement and validity research in language assessment and second language acquisition; however, that study covered the macro panorama of language assessment, and rating scales, as a core subfield, has not yet received dedicated bibliometric examination regarding their research hotspots, knowledge base, and frontier trends.

Based on the above gaps, this study takes the Web of Science Core Collection as the data source and uses CiteSpace to conduct a bibliometric analysis of international language skill rating scales, attempting to address the following questions: First, from 2008 to 2026, what phased characteristics do the publication trends in this field show? Second, what are the major research countries and institutions in this field, and what are their collaborative relationships? Third, what research themes have formed in this field? Fourth, from 2008 to 2026, what phased characteristics and evolutionary trajectories do the hotspot evolution in this field show, and what are the underlying academic endogenous dynamics and external driving mechanisms?

The academic value and practical significance of this study are mainly reflected in the following aspects: First, this study breaks through the methodological limitations of existing reviews. Existing studies mostly adopt traditional narrative reviews. This study uses CiteSpace bibliometric analysis, achieving a breakthrough in analytical methods compared with existing reviews. Second, existing reviews mostly adopt traditional narrative methods, focusing on specific skill domains or specific geographic regions (e.g., Min et al., 2020), and lack a macro knowledge mapping review of rating scales as a core subfield. This study uses CiteSpace to conduct keyword co-occurrence, cluster analysis, and burst detection on 99 core publications, which helps reveal the thematic structure, knowledge base, and frontier evolutionary path of this field, providing a systematic theoretical reference for subsequent research. Third, the findings of this study have reference value for the development, validation, and cross-context adaptation of rating scales. By identifying core research themes and emerging frontiers, such as automated scoring, multimodal assessment, and CEFR localization, this study can provide a knowledge map for rating scale developers, testing organizations, and teachers, helping them grasp the mainstream methods and emerging trends in this field, and thus make more evidence-based decisions in scale design and rating scheme formulation.

2. RESEARCH DESIGN

This research design includes explaining the data source and screening method, as well as the research tools.

2.1 Data Source

The literature samples in this study were all collected from Web of Science. Combining this with the academic development trajectory of bibliometric analysis of international language skill rating scale research, this study limited the search time range to January 1, 2008 to December 31, 2026. The search topic revolved around "language skill rating scales and their psychometric analysis." Since terms such as "rating scale" and "oral" are highly generic in the

medical and clinical fields, the initial search results contained a large number of medical publications unrelated to language testing. To ensure the disciplinary purity of the dataset, this study set exclusion conditions in the search formula, excluding only terms that are highly specific to the medical and clinical fields and almost impossible to appear in language testing rating scale research. Therefore, the specific search formula is:

TS=("speaking" OR "oral" OR "writing" OR "listening" OR "L2" OR "second language") AND TS=("rating scale" OR "rating rubric" OR "scoring rubric" OR "assessment criteria" OR "rating criteria" OR "scoring scale" OR "rating system") AND TS=(validity OR reliability OR psychometric* OR Rasch OR "many-facet" OR MFRM OR "generalizability") NOT TS=(dental OR dentistry OR "oral health" OR "oral cancer" OR sleep OR pain OR sensor OR algorithm OR nursing OR pharmacotherapy OR "clinical trial"), selected from the WOS Core Collection, thereby ensuring the academic quality and authority of the initial sample. The initial search yielded 630 publications.

2.2 Data Screening and Inclusion

To ensure disciplinary purity and the academic nature of literature types, this study conducted three rounds of screening. First, the WOS category refinement function was used to remove categories completely unrelated to language testing, education, and psychology, leaving 162 publications. Second, conference abstracts, book reviews, letters to the editor, and other informal academic literature were removed, retaining only research articles and reviews. Finally, through reading titles and abstracts one by one, medical scale studies unrelated to bibliometric analysis of language skill rating scales, purely applied studies, and duplicate records were excluded. Through the above screening, 99 valid publications were ultimately included, with an inclusion rate of 15.7%.

2.3 Research Tools

"Bibliometrics is a quantitative analysis method based on mathematics and statistics, which can describe, evaluate, and predict the research status and future development trends of a field" (Zhu and Meng, 2013). CiteSpace is a visualization analysis tool based on co-citation analysis theory and pathfinder network algorithms, etc., which conducts bibliometric analysis of literature in a specific field, draws scientific knowledge maps, and reveals the research frontiers and knowledge base of a field (Chen, 2009). This study takes CiteSpace 6.4.R1 visualization analysis software as the core, supplemented by Microsoft Excel data processing software, to conduct knowledge map drawing and statistical analysis of the collected 99 publications.

3. RESEARCH RESULTS

3.1 Publication Trends

This study first used Microsoft Excel data processing software to draw the annual publication volume trend chart of international language skill rating scale research, as shown in Figure 1. The publication volume in this field shows significant phased fluctuation characteristics. From 2008 to 2023, it fluctuated and grew, reaching a peak. During this period, the publication volume oscillated sharply between 2 and 14 publications, with multiple local small peaks in 2009, 2013, 2016, 2018, and 2020, finally reaching the historical highest peak of 14 publications in 2023. After 2023, the publication volume declined sharply, dropping to 3 publications in 2025.

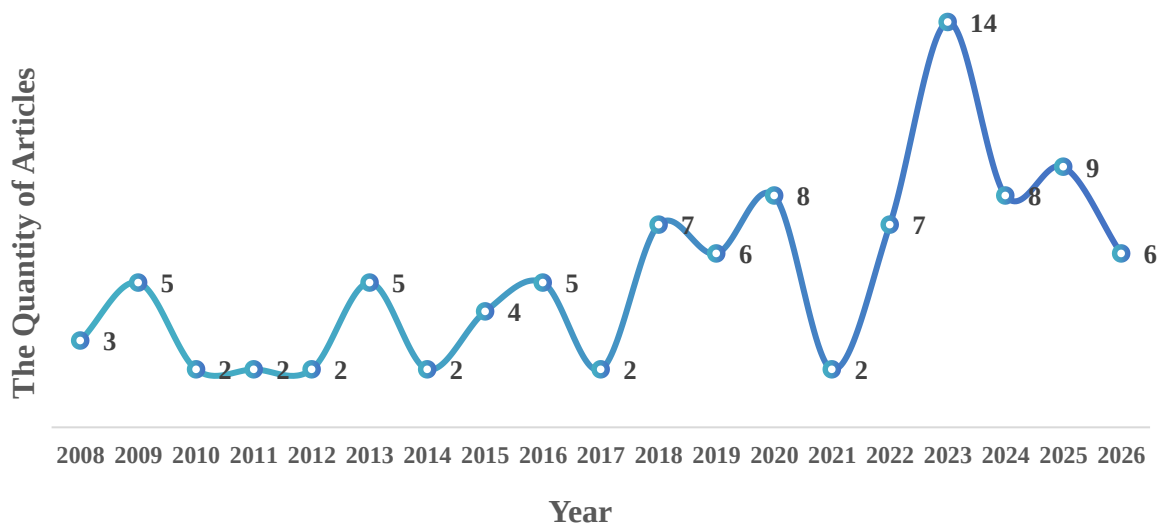


Figure 1 Publication Trend Chart

3.2 Major Research Institutions

This study used CiteSpace to draw the major research institution map, setting the time from January 1, 2008 to December 31, 2026, with a time slice of 1, selecting "Institution" as the node type, setting the g-index screening criterion ($k=25$), and keeping other parameters at default settings, obtaining the statistical results of research institution publication

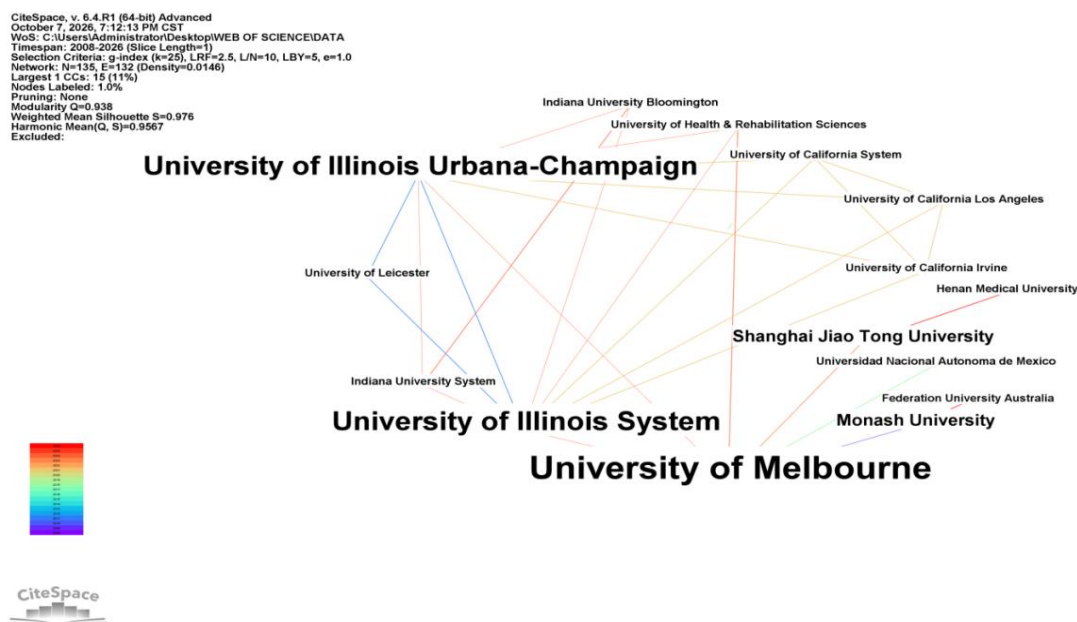


Figure 2 Institutional Collaboration Network Map

volume (see Figure 2). As shown in Figure 2, the institutional collaboration network has 135 nodes and 132 links, with a network density of 0.0146, showing an overall highly fragmented collaboration characteristic. The graph modularity value $Q=0.938$, and the weighted mean

silhouette value $S=0.976$, indicating that the clustering structure is significant and the homogeneity is extremely high.

In terms of node size, University of Illinois Urbana-Champaign is the largest node in the graph, leading in publication volume; University of Illinois System, University of Melbourne, Shanghai Jiao Tong University, and other nodes are also relatively prominent, constituting the core research forces in this field. In terms of collaboration relationships, there is a close link between Indiana University System and Indiana University Bloomington, and there are also collaborative relationships between University of California System and University of California Los Angeles as well as University of California Irvine. However, overall, the vast majority of institutions are isolated nodes, and large-scale cross-institutional and cross-national collaboration networks have not yet formed.

3.3 Major Countries

This study used CiteSpace software to explore collaboration relationships at the national level, setting the time from January 1, 2008 to December 31, 2026, with a time slice of 1, selecting

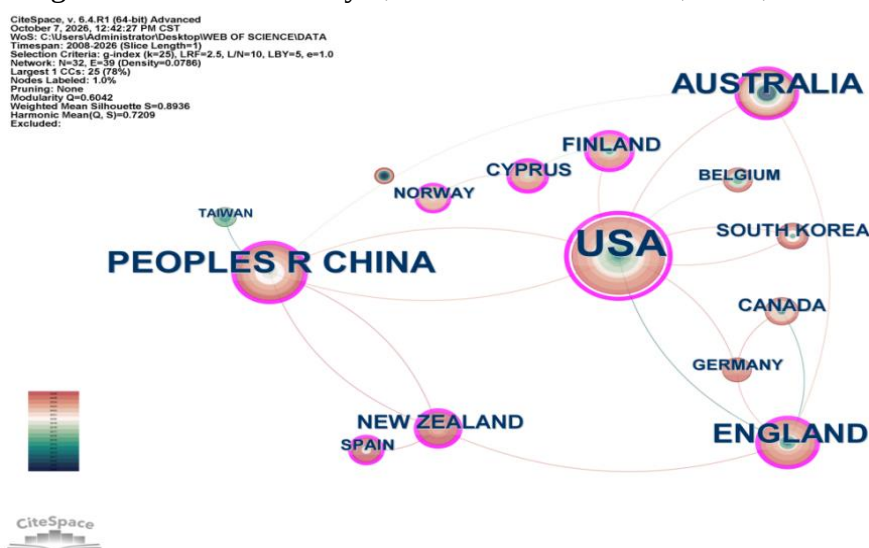


Figure 3 National Collaboration Network Map

"Country" as the node type, setting the g-index screening criterion ($k=25$), and keeping other parameters at default settings, obtaining the national collaboration network map (see Figure 3). In terms of nodes, the United States has the largest node, leading by a wide margin in publication volume, followed by mainland China, which has become the second largest research force in this field. Australia, the United Kingdom, New Zealand, Finland, and other nodes are also relatively prominent, constituting the major research forces in this field. It is worth noting that the ring layers of the US node contain a complete sequence from early cool colors to recent warm colors, while the warm tones of the mainland China node are more obvious, indicating that the United States started early in this field and has remained active for a long time, while mainland China has seen rapid growth in publication volume in recent years. Although mainland China's publication volume has ranked among the top, the breadth of its links still has room for improvement; it currently mainly collaborates with the United States, the United Kingdom, Australia, and other countries, and has relatively few collaborations with continental European countries. Overall, this field shows a pattern of "North America and Western Europe leading, Oceania and Asia actively participating, with a global collaboration network initially formed but of low density." In the

future, cross-national collaboration can be further strengthened to promote the validation of rating scales in different languages and testing contexts.

3.4 Keyword Analysis

3.4.1 Keyword Cluster Analysis

This study used CiteSpace software to analyze keywords in this field, setting the time from January 1, 2008 to December 31, 2026, with a time slice of 1, selecting "Keyword" as the node type, setting the g-index screening criterion ($k=25$), and

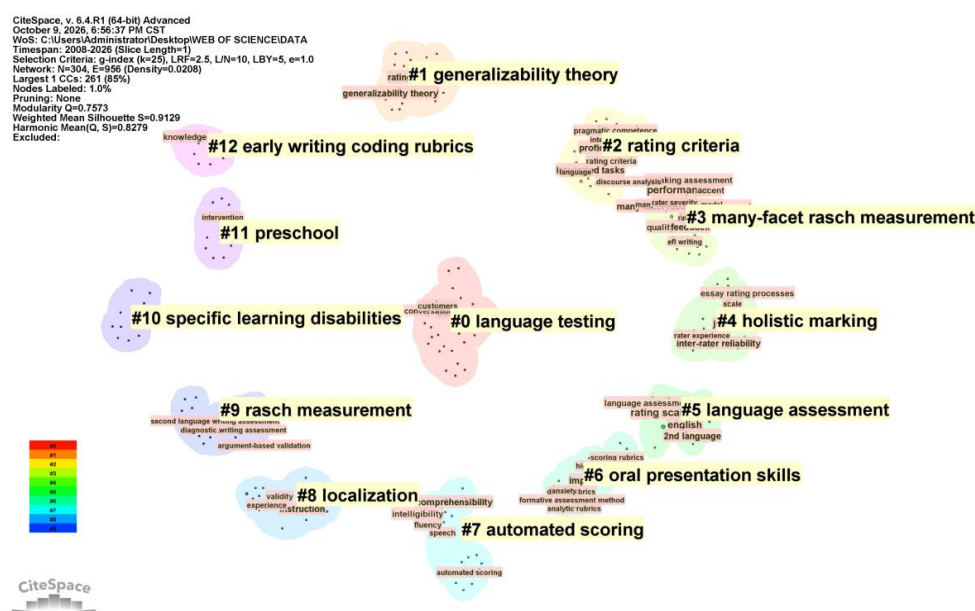


Figure 4 Keyword Cluster Map

keeping other parameters at default settings, obtaining the keyword cluster map (see Figure 4). The keyword cluster map identified 6 core research clusters in this field. The modularity value of the cluster map $Q=0.7573>0.3$, indicating a significant clustering structure, and the mean silhouette value $S=0.9129>0.7$, indicating high clustering efficiency and credibility (Chen Yue et al., 2015).

From the perspective of cluster structure, these six clusters clearly outline the research landscape of this field. Cluster 0 language testing is the largest cluster in the graph, constituting the disciplinary foundation framework of this field and guiding the overall direction of rating scale research. Cluster 1 generalizability theory and Cluster 3 many-facet Rasch measurement represent two core psychometric analysis methods in this field; the former is mainly used to analyze the source structure of rating errors, while the latter is widely used to validate rater severity, bias, and rating scale validity. Cluster 2 rating criteria focuses on the formulation of rating scales and the development of descriptors, directly responding to the core question of "how to define proficiency levels in speaking, writing, and listening." From the perspective of research frontiers, Cluster 7 automated scoring and Cluster 8 localization are particularly noteworthy. The former reflects the impact of natural language processing and artificial intelligence technology on traditional rating models, with researchers beginning to explore the consistency between machine scoring and human scoring; the latter reflects the adaptation needs of rating scales in different languages and cultural contexts, further driven by the global promotion of international frameworks such as CEFR. Together, they indicate that this field is

evolving from "scale structure validation" toward "technological integration" and "cross-cultural validity extension."

3. 4. 2 Hotspot Evolution

This study, based on the keyword timezone map and burst detection function of CiteSpace 6.4.R1, conducted a temporal analysis of international language skill rating scales from 2008

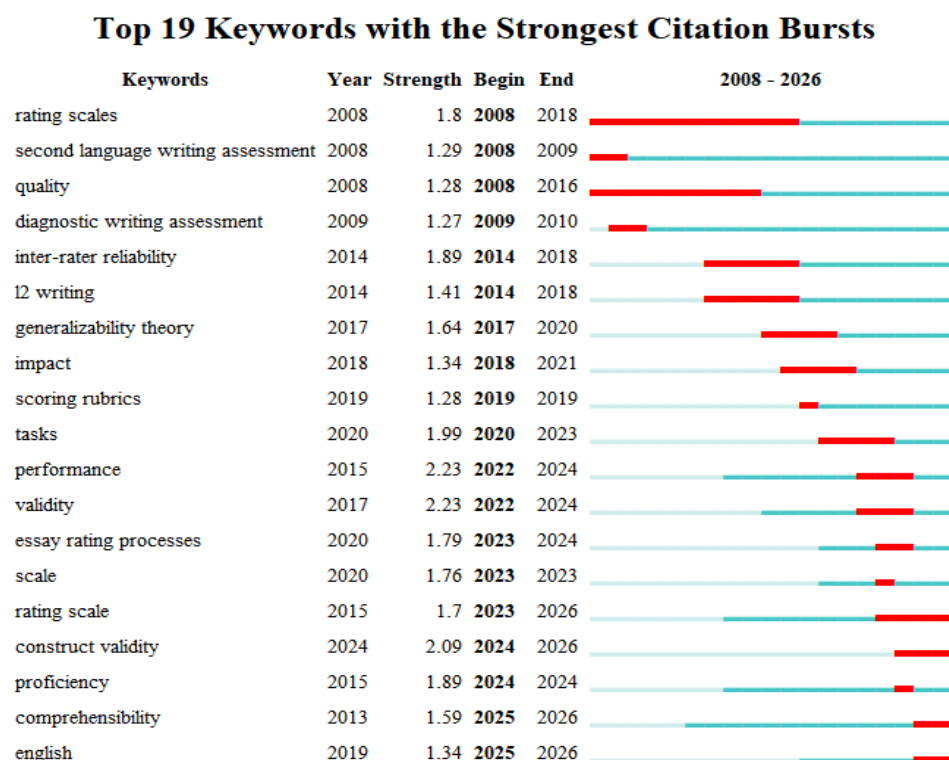


Figure 5 Keyword Burst Chart

to 2026, identifying a total of 19 keywords with burst strength (see Figure 5), clearly presenting the evolutionary path of research frontiers in this field.

As shown in the figure, the early stage (approximately 2008–2014) core keywords include "rating scales," "second language writing assessment," "quality," "diagnostic writing assessment," "inter-rater reliability," and "L2 writing." Burst detection results show that "second language writing assessment," "diagnostic writing assessment," and "L2 writing" were the earliest keywords to burst. "Inter-rater reliability" was the keyword with the highest burst strength in this stage, with a burst strength of 1.89 and a burst period from 2014 to 2018. Next were "rating scales" and "quality," with burst strengths of 1.8 and 1.28 respectively, and burst periods from 2008 to 2018 and 2008 to 2016 respectively, lasting 10 years and 8 years respectively.

The middle stage (approximately 2017–2022) core keywords include "generalizability theory," "impact," "scoring rubrics," "tasks," "performance," and "validity." Among them, "performance" and "validity" had the highest burst strength, both 2.23, with the same burst period from 2022 to 2024. "Tasks" had a burst strength of 1.99, with a burst period from 2020 to 2023. "Generalizability theory" had a burst strength of 1.64, with a burst period from 2017 to 2020. "Impact" had a burst strength of 1.34, with a burst period from 2018 to 2021. In

addition, it is worth noting that "scoring rubrics" had a burst period of 2019–2019, belonging to a single-year burst.

The later stage (approximately 2023–2025) core keywords include "essay rating processes," "scale," "rating scale," "construct validity," "proficiency," "comprehensibility," and "English." Burst detection results show that "construct validity" was the most core burst keyword in this stage, with a burst strength of 2.09, and its burst period continued until 2026. "Proficiency" had a burst strength of 1.89, with a burst period of 2024–2024, belonging to a single-year burst. "Essay rating processes" had a burst strength of 1.79, with a burst period from 2023 to 2024. "Scale" and "rating scale" belong to the same concept as the earlier "rating scales." The burst periods of "comprehensibility" and "English" were both 2025–2026.

4. DISCUSSION

Based on the quantitative analysis results of international language skill rating scale research from 2008 to 2026, this section systematically responds to the three research questions raised in the introduction, in order to explain the internal logic and deep motivations of academic development in this field, and to fill the gap in current research regarding the insufficient explanation of the development and evolution patterns and the internal connections of the knowledge framework in this field. Based on 99 core publications on international language skill rating scale research from 2008 to 2026, this study used CiteSpace for bibliometric analysis, responding to the four research questions raised in the introduction. The following discussion unfolds from four aspects: publication trends, research landscape, thematic structure, and hotspot evolution.

4.1 Publication Trends and Field Development Stages

The emergence of the 2023 peak is closely related to the widespread application of many-facet Rasch measurement and generalizability theory in rating scale validity validation (Aryadoust et al., 2021), and is also driven by the global promotion of international frameworks such as CEFR. The sharp decline after 2023 may reflect two factors: first, research on classical scale validation issues has become saturated, and the space for innovation has narrowed; second, the field may be in a period of methodological transformation, with researchers beginning to turn to emerging directions such as multimodal assessment and automated scoring, causing the publication volume of original themes to temporarily decline. This "peak—decline" trend is not a sign of field decline, but a signal of paradigm transformation, indicating that the research focus is shifting from "scale structure validation" to "technological integration and cross-cultural validity extension."

4.2 Research Landscape and Academic Collaboration Networks

The national and institutional collaboration network analysis in this study jointly reveals a core characteristic: academic collaboration in this field shows a pattern of "broad macro participation, loose micro collaboration."

First, the "active participation" at the national level has not translated into "deep collaboration" at the institutional level. Although countries such as the United States, mainland China, Australia, and the United Kingdom lead in publication volume, the density of institutional collaboration networks is extremely low, indicating that the core research institutions of these countries are mostly isolated nodes. The links between countries mainly reflect "existence of collaboration" rather than "close synergy." This suggests that international collaboration in this field remains more at the macro level of exchange, and has not yet penetrated into substantive collaborative research between institutions.

Second, the distribution of core research forces highly coincides with national publication volume, but the "core-periphery" structure of the collaboration network has not yet formed. The United States, mainland China, Australia, and the United Kingdom are both the countries with the highest publication volume and the countries with the largest nodes in the institutional network. However, the links between the core institutions of these countries, such as the University of Illinois, the University of Melbourne, and Shanghai Jiao Tong University, are sparse, indicating that this field has not yet formed a stable collaboration network with a few core institutions as hubs. This is consistent with the findings of Min et al.'s (2020) review of empirical research on language assessment in China: research in this field is mostly small-scale empirical inquiry by single teams, and relatively few large-scale cross-institutional collaboration projects exist.

Third, the pattern of "North America and Western Europe leading, Oceania and Asia actively participating" is closely related to the strong context dependence of rating scale research. The development and validation of rating scales need to be deeply embedded in specific testing systems, language policies, and teaching traditions. The differences in testing systems and rating criteria among different countries are significant, which makes large-scale cross-institutional and cross-national collaboration face higher coordination costs. Deygers et al.'s (2015) research on the co-construction of CEFR scales also showed that even if raters can reliably use the scale, their understanding of descriptors remains inconsistent. These findings indicate that cross-context adaptation of scales itself faces the challenge of interpretation bias, and cross-national collaboration is therefore difficult to achieve at the operational level.

4.3 Thematic Structure and Core Research Clusters

From the results, the six clusters constitute two levels: the foundational level and the methodological level. Cluster 0 language testing and Cluster 2 rating criteria belong to the foundational level; the former is the disciplinary matrix on which this field depends, and the latter is the core research object directly faced by this field. Cluster 1 generalizability theory and Cluster 3 many-facet Rasch measurement belong to the methodological level; they provide core psychometric tools for the validity validation of rating scales. Cluster 7 automated scoring and Cluster 8 localization constitute the frontier level; the former is a technology-driven transformation of the assessment paradigm, and the latter is a context-driven need for cross-cultural adaptation. This three-level structure of "foundation—method—frontier" is precisely a microcosm of the academic evolution logic of this field.

The essence of rating scale research is a kind of measurement tool research. The core mission of a scale is to provide reliable and valid score interpretation, which determines that it must rely on psychometric methods to solve the problem of error sources in the rating process. The reason why generalizability theory and many-facet Rasch measurement can become two core methodological clusters is that they exactly cover two levels of rating error analysis. First, generalizability theory estimates the relative contribution of variance components such as raters, tasks, and test-takers from a macro level, answering "how should the rating scheme be designed"; many-facet Rasch measurement examines the functioning of specific raters, rating dimensions, and tasks from a micro level, answering "where problems occur in the rating process." Aryadoust et al.'s (2021) review found that many-facet Rasch measurement is the most frequently used Rasch model in the field of language assessment, and Facets software is the most commonly used analysis tool, which precisely explains why Cluster 3 occupies a prominent position in the graph.

The shift in frontier areas is jointly driven by two structural forces. The technological force comes from the impact of natural language processing and artificial intelligence; the emergence

of automated scoring technology has recontextualized the traditional problem of "human rater consistency." (Attali et al., 2013) found that more refined rating scales can improve reliability, but raters referencing automated scoring results can improve consistency while also increasing the risk of similarity to machine scoring. Gjorevski et al. (2025) found that ChatGPT and human ratings are highly consistent in criterion-based writing assessment, and Lee and Hwang (2026) further pointed out that prompt design affects ChatGPT's scoring behavior, but cannot fully control domain bias. These studies indicate that automated scoring has not made rating scales obsolete; instead, it has reactivated the issue of rating scale validity validation. The contextual force comes from the global promotion of international frameworks such as CEFR. When the same set of standards is used in different languages and different testing systems, the issue of "standard comparability" becomes prominent. Deygers et al. (2018) found that two tests based on the same CEFR speaking standards had significantly different interpretations of the B2 level, suggesting that "using the same standard" does not equal "comparable rating results."

4.4 Hotspot Evolution and Frontier Trends

This study found that keyword bursts in this field present three stages, driven jointly by three forces: methodological maturity, technological impact, and contextual updating.

In the early stage, rater consistency was the core issue. The earliest burst keywords "second language writing assessment," "diagnostic writing assessment," and "L2 writing" indicate that research in this stage focused on writing assessment. The high burst of inter-rater reliability indicates that research was still remained at the basic question of "whether different raters give consistent scores." The limitation of this stage is that consistency is only one aspect of rating quality; it cannot answer "whether high consistency means accurate rating." Next were "rating scales" and "quality," with burst strengths of 1.8 and 1.28 respectively, and burst periods from 2008 to 2018 and 2008 to 2016 respectively, lasting 10 years and 8 years respectively, indicating that the development and validation of scales themselves and rating reliability were long-term issues.

The middle stage shifted to systematic error source analysis, with psychometric validation and rating processes as the core. Burst detection results show that the burst strength in this stage was overall higher than that in the first stage, indicating that research hotspots in this stage were more concentrated. Among them, "performance" and "validity" had the highest burst strength, indicating that in this stage, test-taker performance assessment and validity validation were highly intertwined; researchers no longer examined the validity of the scale in isolation, but took test-takers' actual performance in tasks as the core source of evidence for validity validation. The burst of tasks indicates that researchers realized that the same scale used on different task types may produce different rating results. The methodological deepening in this stage expanded the research perspective from "the scale itself" to the complete chain of "task—performance—rating." The burst of "generalizability theory" marks the field's shift from simple reliability analysis to deep analysis that decomposes raters, tasks, test-takers, and their interactions into different variance components. "Impact" indicates that this stage introduced the dimension of social consequences of assessment. In addition, it is worth noting that the burst period of "scoring rubrics" belongs to a single-year burst, indicating that this issue received concentrated attention in the short term and then quickly faded, without forming a sustained research trend.

The later stage returned to the fundamental issue of construct validity. Construct validity became the strongest burst keyword, indicating that the frontier issue of current research centers on answering whether the rating scale truly measures the language ability construct it

claims to measure. This shift is not accidental. Chapelle (2012) pointed out that validity validation in the field of language assessment has shifted from "testing whether the scale is valid" to "constructing and evaluating a validity argument," that is, systematically asking whether the inferential chain of score interpretation is complete and credible. When automated scoring and cross-context adaptation bring new rating contexts, the original validity argument chain of the rating scale must inevitably be re-examined. The burst of construct validity is precisely a mapping of this theoretical framework at the bibliometric level. Corresponding to "construct validity" is the burst keyword "proficiency," whose issue—namely, whether the levels divided by the rating scale truly reflect differences in test-takers' language ability levels—and the burst of "construct validity" both attempt to explore whether the rating scale can accurately measure and distinguish different levels of language ability. The burst of essay rating processes indicates that the research perspective has shifted from rating results to rating processes, with researchers beginning to focus on raters' cognitive processing mechanisms. Baker (2012) pointed out that rating scales are a concrete operationalization of test developers' language ability construct. "Scale" and "rating scale" belong to the same concept as the earlier "rating scales," indicating that rating scales, as a core object, have once again received concentrated attention in current research.

The burst periods of "comprehensibility" and "English" were both 2025–2026. Their synchronous burst indicates that the current frontier focus is shifting toward validity validation of language skill comprehensibility in specific language contexts. Comprehensibility is an important focus of validity validation, while English, as the dominant language of international academic communication, remains the most concentrated application scenario for rating scale research. Isaacs and Thomson (2013) found that raters' judgments of L2 speech comprehensibility are affected by rating scale length and rater experience, and raters have more difficulty distinguishing adjacent levels under a 9-point scale condition. This indicates that comprehensibility is not a simple binary judgment of "understood/not understood," but a complex construct jointly regulated by scale design and rater cognition.

Overall, this field is continuously deepening along the path of "methodological deepening—perspective expansion—issue regression." From early attention to "whether ratings are consistent," to mid-term inquiry into "where errors come from," and then to later regression to "whether the scale measures what it should measure," this path indicates that the field is moving from tool-level scale validation toward theoretical-level construct validity argumentation.

In the early stage, rater consistency was the core issue. The earliest burst keywords "second language writing assessment," "diagnostic writing assessment," and "L2 writing" indicate that research in this stage focused on writing assessment. The high burst of inter-rater reliability indicates that research was still remained at the basic question of "whether different raters give consistent scores." The limitation of this stage is that consistency is only one aspect of rating quality; it cannot answer "whether high consistency means accurate rating." Next were "rating scales" and "quality," with burst strengths of 1.8 and 1.28 respectively, and burst periods from 2008 to 2018 and 2008 to 2016 respectively, lasting 10 years and 8 years respectively, indicating that the development and validation of scales themselves and rating reliability were long-term issues.

The middle stage shifted to systematic error source analysis, with psychometric validation and rating processes as the core. Burst detection results show that the burst strength in this stage was overall higher than that in the first stage, indicating that research hotspots in this stage were more concentrated. Among them, "performance" and "validity" had the highest burst strength, indicating that in this stage, test-taker performance assessment and validity validation were highly intertwined; researchers no longer examined the validity of the scale in isolation,

but took test-takers' actual performance in tasks as the core source of evidence for validity validation. The burst of tasks indicates that researchers realized that the same scale used on different task types may produce different rating results. The methodological deepening in this stage expanded the research perspective from "the scale itself" to the complete chain of "task—performance—rating." The burst of "generalizability theory" marks the field's shift from simple reliability analysis to deep analysis that decomposes raters, tasks, test-takers, and their interactions into different variance components. "Impact" indicates that this stage introduced the dimension of social consequences of assessment. In addition, it is worth noting that the burst period of "scoring rubrics" belongs to a single-year burst, indicating that this issue received concentrated attention in the short term and then quickly faded, without forming a sustained research trend.

The later stage returned to the fundamental issue of construct validity. Construct validity became the strongest burst keyword, indicating that the frontier issue of current research centers on answering whether the rating scale truly measures the language ability construct it claims to measure. This shift is not accidental. Chapelle (2012) pointed out that validity validation in the field of language assessment has shifted from "testing whether the scale is valid" to "constructing and evaluating a validity argument," that is, systematically asking whether the inferential chain of score interpretation is complete and credible. When automated scoring and cross-context adaptation bring new rating contexts, the original validity argument chain of the rating scale must inevitably be re-examined. The burst of construct validity is precisely a mapping of this theoretical framework at the bibliometric level. Corresponding to "construct validity" is the burst keyword "proficiency," whose issue—namely, whether the levels divided by the rating scale truly reflect differences in test-takers' language ability levels—and the burst of "construct validity" both attempts to explore whether the rating scale can accurately measure and distinguish different levels of language ability. The burst of essay rating processes indicates that the research perspective has shifted from rating results to rating processes, with researchers beginning to focus on raters' cognitive processing mechanisms. Baker (2012) pointed out that rating scales are a concrete operationalization of test developers' language ability construct. "Scale" and "rating scale" belong to the same concept as the earlier "rating scales," indicating that rating scales, as a core object, have once again received concentrated attention in current research.

The burst periods of "comprehensibility" and "English" were both 2025–2026. Their synchronous burst indicates that the current frontier focus is shifting toward validity validation of language skill comprehensibility in specific language contexts. Comprehensibility is an important focus of validity validation, while English, as the dominant language of international academic communication, remains the most concentrated application scenario for rating scale research. Isaacs and Thomson (2013) found that raters' judgments of L2 speech comprehensibility are affected by rating scale length and rater experience, and raters have more difficulty distinguishing adjacent levels under a 9-point scale condition. This indicates that comprehensibility is not a simple binary judgment of "understood/not understood," but a complex construct jointly regulated by scale design and rater cognition.

Overall, this field is continuously deepening along the path of "methodological deepening—perspective expansion—issue regression." From early attention to "whether ratings are consistent," to mid-term inquiry into "where errors come from," and then to later regression to "whether the scale measures what it should measure," this path indicates that the field is moving from tool-level scale validation toward theoretical-level construct validity argumentation.

5. CONCLUSION

This study takes the Web of Science Core Collection as the data source and uses CiteSpace to conduct a bibliometric analysis of 99 publications on international language skill rating scales from 2008 to 2026. It systematically examined the publication trends, research landscape, thematic structure, and frontier evolution of this field. The findings are as follows: First, the publication trends in this field show a phased characteristic of "fluctuating growth—peak then decline," and the research focus is shifting from "scale structure validation" to "technological integration and cross-cultural validity extension." Second, the research landscape of this field shows a characteristic of "high clustering, low density." At the national level, there is a pattern of "North America and Western Europe leading, Oceania and Asia actively participating." The strong context dependence of rating scale research, the small scale of research teams, and the high coordination cost of cross-context collaboration jointly constitute the deep motivations for this landscape. The thematic structure consists of six clusters, which can be summarized into three levels: foundation, method, and frontier. The three-level structure is not a linear replacement relationship, but a cyclical evolution in which "foundation provides the object, method provides the tool, and frontier raises new questions." Hotspot evolution can be summarized as a path "from consistency to construct validity," reflecting the continuous deepening of this field under the triple forces of methodological maturity, technological impact, and contextual updating.

The academic value of this study is mainly reflected in three aspects. First, this study breaks through the methodological limitations of existing reviews. Existing studies mostly adopt traditional narrative reviews; this study uses CiteSpace bibliometric analysis, achieving a breakthrough in analytical methods compared with existing reviews. Second, existing reviews mostly focus on specific skill domains or specific geographic regions (Min et al., 2020), lacking a macro knowledge mapping review of rating scales as a core subfield. This study uses CiteSpace to conduct keyword co-occurrence, cluster analysis, and burst detection on 99 core publications, revealing the thematic structure, knowledge base, and frontier evolutionary path of this field, and providing a systematic theoretical reference for subsequent research. Third, at the practical level, the findings of this study have reference value for the development, validation, and cross-context adaptation of rating scales. By identifying core research themes and emerging frontiers, such as automated scoring, multimodal assessment, and CEFR localization, this study can provide a knowledge map for rating scale developers, testing organizations, and teachers, helping them grasp the mainstream methods and emerging trends in this field, and thus make more evidence-based decisions in scale design and rating scheme formulation.

However, this study still has clear research boundaries. In terms of data sources, only the Web of Science Core Collection was used, which may miss important literature in other databases and affect the comprehensiveness of the sample. In terms of time range, the search time was limited to 2008–2026, which may miss earlier foundational research and provide an incomplete trace of the field's knowledge base. In terms of method, bibliometric analysis focuses on macro knowledge structure and has limited explanatory power for micro mechanisms; the thematic clusters and frontier trends revealed by the map still need to be further verified through close reading of texts. In addition, in the WOS category refinement and abstract screening stages, researchers' judgments may introduce subjective bias, affecting the representativeness of the final included literature.

Based on the above research conclusions and existing limitations, future research can be advanced in the following directions: First, conduct systematic close reading of core literature with high burst and high centrality to deeply understand their theoretical contributions and

methodological characteristics; second, use qualitative methods such as interviews and think-aloud protocols to explore the assessment beliefs and practices of scale developers, raters, and test-takers, establishing a closer connection between macro patterns and micro mechanisms; third, pay attention to emerging directions such as automated scoring and multimodal assessment, and explore the validity validation path of rating scales in human-machine collaborative assessment contexts; fourth, expand data sources and include multiple databases for cross-validation to improve the robustness of research conclusions.

Funding: This work is supported by the National Social Science Fund of China (NSSFC) Project "Construction of an Automated Speaking Assessment Model for International Chinese Language Education" (Grant No.: 25BYY143).

REFERENCES

- [1] Aryadoust V, Ng LY, Sayama H. A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Lang Test*. 2021;38(1):6-40.
- [2] Aryadoust V, Zakaria A, Lim MH, Chen C. An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. *Front Psychol*. 2020;11:1941.
- [3] Attali Y, Lewis W, Steier M. Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Lang Test*. 2013;30(1):125-141.
- [4] Baker BA. Individual differences in rater decision-making style. *Lang Assess Q*. 2012;9(3):225-248.
- [5] Chapelle CA. Validity argument for language assessment: The framework is simple.... *Lang Test*. 2012;29(1):19-27.
- [6] Deygers B, Van Gorp K. Determining the scoring validity of a co-constructed CEFR-based rating scale. *Lang Test*. 2015;32(4):521-541.
- [7] Deygers B, Van Gorp K, Demeester T. The B2 level and the dream of a common standard. *Lang Assess Q*. 2018;15(1):44-58.
- [8] Gjorevski A, Li M, Cox TL. Exploring the potential of ChatGPT for evaluating English essays in a criterion-based assessment. *TESOL Q*. 2025;59(S1):S251-S279.
- [9] Isaacs T, Thomson RI. Rater experience, rating scale length, and judgments of L2 pronunciation. *Lang Assess Q*. 2013;10(2):135-159.
- [10] Iwashita N, Brown A, McNamara T, O'Hagan S. Assessed levels of second language speaking proficiency: How distinct? *Appl Linguist*. 2008;29(1):24-49.
- [11] Knoch U. Diagnostic assessment of writing: A comparison of two rating scales. *Lang Test*. 2009;26(2):275-304.
- [12] Lee Y, Hwang M. ChatGPT for automated writing evaluation: Scoring and feedback across prompt conditions. *Lang Test*. 2026. Advance online publication.
- [13] Li S, Taguchi N, Xiao F. Variations in rating scale functioning in assessing speech act production in L2 Chinese. *Lang Assess Q*. 2019;16(3):271-293.
- [14] Luoma S. *Assessing speaking*. Cambridge University Press; 2004.
- [15] Min S, He L, Zhang J. Review of recent empirical research (2011-2018) on language assessment in China. *Lang Teach*. 2020;53(3):316-340.
- [16] Palmour L. Assessing speaking through multimodal oral presentations: The case of construct underrepresentation in EAP contexts. *Lang Test*. 2024;41(1):9-34.
- [17] Chen C. CiteSpace II: Identification and visualization of new trends and new dynamics in scientific literature. Chen Y, Hou J, et al, trans. *J Inf Sci*. 2009;(3):401-421.

-
- [18] Chen Y, Chen C, Liu Z, et al. Methodological functions of CiteSpace knowledge maps. *Stud Sci Sci*. 2015;33(2):242-253.
- [19] Tang Y, Ma X. An empirical study of rating scale intervention in college English writing teaching evaluation. *Foreign Lang Test Teach*. 2024;(3):39-49.
- [20] Zhu L, Meng X. A comparative study of bibliometrics and content analysis. *Libr Work Res*. 2013;(6):64-66.