Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 15 (1). 2025. 171-198. Published online: 13.01.2025
https://doi.org/10.14746/ssllt.25141
http://pressto.amu.edu.pl/index.php/ssllt

Trends in language assessment and testing: A bibliometric study

Xian Zhang ✉
University of North Texas, Denton, United States
https://orcid.org/0000-0001-8472-5380
xian.zhang@unt.edu

Abstract

The current bibliometric study employed citation analysis and keyword analysis to perform a review of language assessment and testing. Based on citation counts and keywords, this study identified the recent trends/changes and the most influential regions, institutions, scholars, and publications in the field. In addition, the intellectual structures of the field reviewed by the network maps of the most influential documents and scholars showed how these eminent documents and authors were related to each other. It was found that the field experienced significant changes with the emergence of new scholars, research themes, and topics. This study is also a tribute to hundreds of scholarly documents in the field, which keep the field moving forward.

Keywords: bibliometrics; language testing; language assessment; citation analysis; co-citation analysis; keyword analysis

1. Introduction

A systematic review offers useful information of a field that may include an introduction to the key documents and concepts, prominent scholars, and the most recent trends of development in the field. However, surveying a discipline or a field

to provide a systematic review usually requires sufficient knowledge in this respect. In addition, a review by experts may involve subjectivity. A bibliometric study, which surveys hundreds or thousands of publications in a discipline, can generate valuable quantitative data for a systematic review that may alleviate the issue of subjectivity and bias (de Bellis, 2009).

Based on the meta-data of publications (e.g., citations, dates, and places of publications), the bibliometric method applies statistical techniques to provide insights into a discipline (de Bellis, 2009). It can be used to evaluate the impact of entities at various levels, including geographical regions, research institutions, authors, and documents (Holden et al., 2005). Bibliometrics allows researchers to monitor the trends of a field and to plan their research (Chang et al., 2015).

Although bibliometrics has existed for decades, this technique has not been used to study language assessment and testing. Bibliometric studies, however, have appeared in major journals and books in applied linguistics, investigating a variety of subjects, such as discourse analysis (Swales, 1986), English for academic purposes (Hyland & Jiang, 2021a), English for specific purposes (Hyland & Jiang, 2021b), listening (Lei et al., 2023), vocabulary (Meara, 2012, 2023), task difficulty (Wang & Zhang, 2019), written interaction (Hyland & Jiang, 2023), and second language acquisition (Chen, 2023; Zhang, 2020). Bibliometrics was also applied to study the productivity of regions (e.g., Lei & Liao, 2017), academic journals (Lei & Liu, 2019b; Riazi et al., 2023; Xu et al., 2023), and applied linguistics as a discipline (de Bot, 2015). Swales (1986) was among the first to use bibliometrics to study discourse analysis in applied linguistics. White (2004), a leading scholar in library science, introduced bibliometric methods for applied linguistics and advocated for cross-field collaborations. More recently, de Bot (2015) published a monograph that applied citation analysis to the field of applied linguistics. He successfully identified not only the most influential scholars but also the most prominent topics in the field. Lei and Liu (2019a) implemented co-citation analysis and keyword analysis to capture key scholars and topics in applied linguistics.

Despite these existing applications of bibliometrics in various areas in applied linguistics, the intellectual structure of language assessment and testing has not been studied through the lens of bibliometrics. Language testing and assessment is an important branch in applied linguistics. McNamara (2004) laid out arguments that language assessment and testing is a core area in applied linguistics as assessment and testing not only lies "at the forefront" for defining and validating constructs in language ability/skills, but also plays significant "social and political roles" (p. 764) that have great impacts on the modern world, some of which include access to education, language policy, immigration, and more (e.g., Shohamy, 2001; Spolsky, 1981).

Language assessment and testing as a field has experienced many developments since the 1960s, when large-scale language tests were being developed and

industrialized, largely influenced by Lado's (1961) test of English and Carroll's (1961) integrative testing (Spolsky, 2017). Spolsky (1995, 2017) offered a comprehensive review of the history of the field and synthesized some notable developments that were associated with the reliability and validity of language tests, integrative testing, test scales, social contexts, and social impacts. The reliability and validity of language tests have been closely investigated for years from the perspective of psychometrics. Integrative testing, heavily influenced by Carroll (1961), advocates assessment in a holistic and integrative manner as opposed to language tests with only discrete point elements. Test scales, or verbal descriptions that define the specific language ability at a set of proficiency levels, have attracted significant research interest as some of the highly influential test scales were developed, such as, for example, the *Common European Framework* (Council of Europe, 2001). The social aspect of language assessment and testing concerns some key aspects such as contexts of test use, ethics, and fairness. The social aspect of testing and assessment also led to the development of alternative assessments that link assessments to a situated learning context, such as self-assessment, classroom assessment, and task-based assessment (Farhady, 2018). Recently, advancement in technology has made computer-assisted language assessment (CALA) more popular, partially due to its efficiency and cost-effectiveness (e.g., Chapelle & Douglas, 2006). However, CALA also brings new challenges to language assessment and testing, such as test validity (Bachman, 2000), test consequences (Chapelle & Douglas, 2006), and test fairness (Chen et al., 2011).

These developments identified by language assessment experts offer an overview of some key topics in the field of language assessment and testing. These topics, such as reliability and validity, are some of the key components that make up the intellectual structure of the field. While identification of the key topics is critical to understand this structure, it is not enough to depict the full picture. This is because the key topics are not isolated. In fact, many of them are interconnected, and more or less related. For example, validity is closely linked to CALA and test fairness, as discussed above. In addition, each key topic has different significance or impact. The impact of these key components can change over time as the field evolves: Some of the topics become more critical, while others may fade. Finally, the full picture of the intellectual structure is far from comprehensive without recognizing the key scholars and experts who actually do the heavy lifting to produce academic documents that move the field forward. The key components of language assessment and testing can be identified via keyword analysis. Based on frequency counts of content words, keyword analysis can pinpoint prominent topics as highly frequent content words represent the key concepts and knowledge within a discipline (Callon et al., 1983). Moreover, the frequency counts of keywords across time can be used to

detect the major changes and trends in language assessment and testing, which provides valuable information for scholars and researchers in the field to plan future research.

To measure the impacts of scholars, academic documents, and topics, a commonly used method is citation analysis, which uses citation information to perform quantitative analysis solely based on data. One prominent application is the impact factor (IF), first introduced by Garfield (1955), and later used to compile the famous social citation index (SCI) and the social science citation index (SSCI). Slightly different from citation analysis, which is largely based on citation counts, co-citation analysis (Small, 1973) uses the frequency of co-occurrence of citations and references to organize scholarly works (or scholars) into network maps that show how the components of the intellectual structure of a field are related to each other. By combining these bibliometric techniques, this study aims to gain a comprehensive understanding of the field of language assessment and testing.

2. Purpose

Depicting the intellectual structure of the field allows us to capture how the field values the work of scholars in assessment and testing through systematic and objective analysis. It also allows us to understand developments in language assessment and testing over time through content analysis, which helps researchers, both in the field and from other fields, design and plan their future activities in assessment and testing that are crucial for moving the field forward.

The current bibliometric study was aimed at conducting an objective review of the field between 2008 and 2019 so as to identify the most recent trends/changes as well as the most influential scholars and academic publications in language assessment and testing. In order to capture the changes and trends, we compared the most influential regions/institutions/authors/publications and the key topics between two time periods: 2008~2013 and 2014~2019 (2008 is the year when bibliometric information, such as citation counts and impact factors, became available for the major journals in language assessment and testing). Three research questions were formulated:

RQ1: Which publications, authors, institutions, and regions have the most impact according to citation counts in language assessment and testing?
RQ2: What are the major themes and topics in language assessment and testing?
RQ3: What are the changes and trends over the last 12 years in language assessment and testing?

3. Method

3.1. Bibliometric data

The current study retrieved bibliometric data from Web of Science (WoS). WoS is arguably one of the best-known and most trusted databases for bibliometrics with over 7,000 official subscribers, twice as many as the closest competitor Scopus (Roemer & Borchardt, 2015). A larger subscription base is important as it represents the reputation of the database among funding agents, research institutes, and scholars, which is associated with funding, grant, promotion, hiring, etc. (de Bellis, 2009). Moreover, although Scopus covers more journals, the overall quality of journals is lower according to citation impacts (de Groote & Raszewski, 2012), which would also affect the quality of citations because citations from different sources may not have equal scholarly values (de Bellis, 2009). Google Scholar is another bibliometric database. However, it offers only very limited types of bibliometric information (e.g., citation counts). When computing citations, Google Scholar does not distinguish between academic and non-academic sources (e.g., blog posts, news). This limits its academic usefulness. Since WoS is superior in terms of accessibility and data quality, it was selected to be the database for the current bibliometric study.

Two international journals specialized in language assessment and testing, *Language Testing* and *Language Assessment Quarterly*, were chosen for analysis. These two leading journals are the backbone of language assessment and testing. Publications in the two journals provide first-hand knowledge about the developmental trends in the field. *Assessing Writing* was excluded from the analysis due to the relatively focused research scope of the journal. Different from *Language Testing* and *Language Assessment Quarterly* which publish research on language assessment of all four language skills (listening, speaking, reading, and writing), *Assessing Writing* focuses on writing in general education. Since the current bibliometric analysis focused on different issues, such as analysis of impact (authors, regions, key documents, etc.) and keyword analysis, including *Assessing Writing* would have decreased the balance among the four language skills, thus having a significant impact on all types of analyses by, for example, favoring the impacts of scholars/documents specialized in writing assessment.

This study focused exclusively on international journals due to their more rigorous peer-review processes, which helps ensure the quality of publications. Additionally, international journals offer greater accessibility and visibility to a global readership as compared to regional journals (Benson et al., 2009). We included only full-length articles published between 2008 and 2019, excluding other publication types such as book reviews and blogs. A total of 501 full-length

articles have been published by the two journals between 2008 and 2019. Several types of bibliometric information were analyzed: titles, authors, affiliations, abstracts, citation counts, author-supplied keywords, and references.

3.2. Data cleaning

Before analyzing the data, it was necessary to perform data cleaning to remove coding variations. This is because different names may be used by different formats of references that can denote the same authors. For example, the references extracted from the two major journals used various formats of name, such as, for example, "Alderson Charles," "Alderson, C," "Charles Alderson," "Alderson J. Charles," "Alderson CJ," and so on. All these names were recoded as "Alderson C." In addition to author name, keywords also need cleaning as the same concept can be represented by different keywords, for example, *second language vocabulary* and *L2 vocabulary*. These keywords were coded as *L2 vocabulary*. To give another example, *test use, use of test*, and *use of tests* were all coded into *test use*. To recode the author names, Microsoft Excel was used to sort the author list by the last names of authors. This name list was then manually checked and recoded. The recoded list was loaded into VOSviewer via the author-thesaurus function to compute citation counts for authors. For the keyword list, each keyword was manually examined, grouped, and recoded one by one.

3.3. Data analysis

Data analysis included citation analysis, co-citation analysis, and keyword analysis. Citation analysis pinpoints the most-cited articles published in the two journals in each of the two time periods (2008~2013, 2014~2019). This part of the analysis also showed which research institutions and regions were most active in each time period.

Regarding co-citation analysis, VOSviewer (Waltman et al., 2010; see Waltman & Van Eck, 2017 for an introduction to the software) was used to visualize the core structure of the field by constructing the intellectual maps of documents (and authors), which were made up of the most influential scholarly documents (and authors) in the field. To build these maps, two types of information were needed: citation counts and co-citation patterns extracted from the references of the 501 full-length articles. The citation count of a document (or an author) equaled the number of times a document (or an author) was cited by 501 full-length articles, which is an important indicator of impact. Co-citation patterns of documents (authors), or how often documents (authors) were cited together, showed

the relationship between documents (authors). The full-length articles published in the first period (2008~2013) cited more than 7,500 unique sources. The full-length articles published in the second period (2014~2019) cited over 9,600 unique documents. The intellectual network maps at the two time periods produced in the co-citation analysis consisted of only the 50 most-cited documents (or authors) as nodes. The reason to include only a limited number of nodes was that too many nodes in a network map would have made the map difficult to interpret. The smart local moving algorithm (Waltman & Van Eck, 2013) was employed to generate the network maps, which grouped the 50 most cited documents (or authors) into clusters based on the associational strength between documents (or authors). These clusters revealed the major themes in the field at each time period.

The keyword analysis examined the trends in the field via frequency changes of key topics between the two time periods. The first step for the keyword analysis was to identify key topics in the form of keywords. The common method for keyword identification is through author-supplied keywords (e.g., Chiu & Ho, 2007; Courtial, 1994; Law & Whittaker, 1992), which could be retrieved from WoS. Although the author(s) of a document know their publication well, they may not always list out all the keywords. Therefore, some key topics may be overlooked if a keyword analysis only uses author-supplied keywords. To offer a more comprehensive list of keywords in the field, the current keyword analysis also extracted words and phrases (in the form of n-grams) from the abstracts of the full-length articles using Antconc based on the frequency of the words and phrases (Anthony, 2018). N-grams are multiple-word units, for example, a bigram is made up of two words (e.g., *content validity*) and a trigram contains three words (e.g., *paired speaking tests*). The current study only analyzed n-grams that constituted stand-alone concepts that could be regarded as topics. In other words, most of the n-grams that contain function words (e.g., pronouns and modal words) were excluded, such as, for example, *students can, our data, of great interest, belong to* (some exceptions were *can-do statements*). Moreover, n-grams too broad to be considered as useful topics were excluded. Some examples were *language learner(s)*, *test taker(s)*, *test score(s)*, *test result(s)*, *language assessment(s), L2 acquisition, target language, school students, strengths and weaknesses*, and *data analysis*. N-grams that functioned as transitions were not included, for example, *as a result*, and *studies show that*. Finally, since the decision to include/exclude n-grams involved subjectivity, we tried to be more inclusive when determining which n-grams were included in the list in order to compile a more comprehensive keyword list for the analysis, which would leave more room for readers to interpret the results. It should be emphasized that keywords derived from abstracts are complementary to the author-supplied keywords.

After the keyword list had been assembled, frequency counts of the keywords were computed for both time periods by searching the keywords in each abstract. Many keywords appeared multiple times in one abstract. To avoid bias, keywords that

appeared multiple times in an abstract were counted as one occurrence. Log-likelihood (LL) tests, using the keywords' frequency information across both time periods, were employed to examine which keywords were unusually frequent during a specific period. A significant LL value of a topic indicated that a significant change of interest towards the topic had taken place. The LL tests used the following formulas to calculate the LL values of the keywords (Rayson & Garside, 2000):

(a) $$Ei = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

(b) $$-2 \ln \lambda = 2 \sum_i O_i \ln(\frac{O_i}{E_i})$$

Figure 1 Formulas to run the log-likelihood tests (*Oi* = raw frequency of the target word; *Ni* = the size of a corpus; *Ei* = the expected frequency value of the target word)

The criterion to treat words/phrases as keywords uses LL = 3.84 as the threshold (Rayson & Garside, 2000), which has been widely accepted as the adequate criterion in social sciences (Wilson, 2013). For the current study, words/phrases with LL > 3.84 and an increase in frequency from the first period to the second period suggested that these topics had received more interest over time. Words/phrases with an absolute LL value < 3.84 suggested that these topics had remained stable. Words/phrases with LL < - 3.84 and a decrease in frequency from the first period to the second period indicated that these topics had received less interest.

4. Results

In the first part of this section, the results of the citation counts will identify the most cited articles published in the two journals, as well as the top institutions and regions that generated the most citations. In the second part, the results of the co-citation analysis will be presented to show the intellectual structure of the field, composed of the most-cited researchers and the most influential scholarly publications. Finally, the keyword analysis will be presented to identify the key topics and trends in the field.

4.1. Citation analysis

4.1.1. Most-cited publications in the two journals

Table 1 presents the raw citation counts and normalized citation counts of the 20 most-cited articles published in the two journals during each time frame. Citation

counts are affected by time (citation counts tend to increase as a function of time). To control for the time effect, citation counts of the journal articles were normalized based on annual citation counts. For example, the citation count of an article published in 2008 would be normalized using the citation count of all articles published in 2008. This normalized process made citation counts of papers published in different years comparable (Waltman & van Eck, 2017). The top 20 most cited articles suggest that writing and speaking dominated both time periods. Vocabulary assessments were also of great interest to researchers.

Table 1 The most cited articles in the journals at each time period (ordered by normalized citation)

	Documents	Raw citation	Norm. citation
	2008-2013		
1	Fulcher (2012). Assessment literacy for the language classroom.	55	4.21
2	Kane (2012). Validating score interpretations and uses.	54	4.13
3	Wigglesworth & Storch (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy.	97	3.77
4	Bosker et al. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs.	53	3.73
5	Winke & Myford (2013). Raters' L2 background as a potential source of bias in rating oral performance.	51	3.59
6	Eckes (2008). Rater types in writing performance assessments: A classification approach to rater variability.	92	3.53
7	Isaacs & Thomson (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions.	49	3.45
8	Hulstijn (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment.	52	3.28
9	Beglar (2010). A Rasch-based validation of the vocabulary size test.	67	3.25
10	Fulcher & Kemp (2011). Effective rating scale development for speaking tests: Performance decision trees.	47	2.96
11	Cheng (2008). The key to success: English language testing in China.	77	2.95
12	Bernstein et al. (2010). Validating automated speaking tests.	58	2.81
13	Butler & Lee (2010). The effects of self-assessment among young learners of English.	53	2.57
14	Scarino (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning.	36	2.54
15	Matsuno (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms.	63	2.45
16	Knoch (2009). Diagnostic assessment of writing: A comparison of two rating scales.	62	2.41
17	Hill & McNamara (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment.	31	2.37
18	Cho & Bridgeman (2012). Relationship of TOEFL iBT (r) scores to academic performance: Some evidence from American universities.	31	2.37
19	Carey et al. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews?	37	2.33
20	Choi (2008). The impact of EFL testing on EFL education in Korea.	60	2.30
	2014-2019		
1	Wind & Peterson (2018). A systematic review of methods for evaluating rating quality in language assessment.	13	6.57
2	Khabbazbashi (2017). Topic and background knowledge effects on performance in speaking assessment.	10	5.66
3	Segbers & Schroeder (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size.	9	5.10
4	Vogt & Tsagari (2014). Assessment literacy of foreign language teachers: Findings of a European study.	30	4.39
5	McCray & Brunfaut (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking.	8	4.04
6	Roever & Kasper (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking.	8	4.04
7	Trace et al. (2017). Measuring the impact of rater negotiation in writing performance assessment.	7	3.96
8	Davis (2016). The influence of training and experience on rater performance in scoring spoken language.	15	3.71
9	Bouwer et al. (2015). Effect of genre on the generalizability of writing scores.	28	3.62
10	Ginther & Yan (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles.	7	3.54
11	Chapelle et al. (2015). Validity arguments for diagnostic assessment using automated writing evaluation.	25	3.23
12	Zumbo et al. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding.	25	3.23
13	Poehner et al. (2015). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation.	24	3.10
14	Han (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach.	11	2.72
15	Youn (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods.	21	2.71
16	Nitta & Nakatsuhara (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance.	18	2.63
17	Lam (2015). Language assessment training in Hong Kong: Implications for language assessment literacy.	20	2.58
18	Zhang et al. (2014). Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: A multi-sample SEM approach.	16	2.34

19	Préfontaine et al. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language?	9	2.23
20	Gu (2014). At the interface between language testing and second language acquisition: Language ability and context of learning.	15	2.19
	Kuiken (2014). Rating written performance: What do raters do and why?	15	2.19
	Park (2014). Corpora and language assessment: The state of the art.	15	2.19

Note. See Table 1 in the supplementary material[1] for the full references

4.1.2. Most influential institutions and regions

The articles published in the two journals during the two time periods came from 340 institutions. Table 2 in the supplementary material presents the top 20 most productive institutions. Besides ETS, some major research institutions in the field were University of Melbourne, Lancaster University, Georgia State University, University of Toronto, Michigan State University, and University of Hawaii, which had remained on the top 10 list in both time periods. It is worth noting that the productivity of an institution or a region is not completely independent of the number of researchers in the institution or the region as more researchers may likely generate more documents.

Table 3 in the supplementary material lists the top 20 most productive regions that contributed research articles to the field. The most productive regions in the 2008~2013 period included the United States, Australia, England, Japan, and China. The United States continued to remain at the top of the list in the 2014-2019 period, with the highest citation counts much bigger than other regions in both time periods. Canada, England, China, and Australia were ranked in the top 5 alongside the United States during the 2014-2019 period. Both England and China improved their rankings compared to the previous time frame, while Canada entered the top 5 in the second period.

4.2. Co-citation analysis and network mapping

4.2.1. The most influential documents

Table 4 in the supplementary material summarizes the 50 most cited articles among the 7500+ unique references in the 238 documents published in the first period. The raw citation counts of these references were all larger than 9. Among the 9600+ unique references cited in the 265 articles of the second time period, all top 50 references were cited at least 9 times.

The smart local moving algorithm (Waltman & Van Eck, 2013) assigned the top 50 references in the first period to 4 clusters and the top 50 references in the second period to 5 clusters. Each cluster is represented by a group of circular nodes and lines

[1] Supplementary material can be accessed at: http://ssllt.amu.edu.pl/download/docs/SSLLT%2025141%20Zhang%20supplementary.pdf

in the same color (more grouping information is given in Table 4 and Table 5 in the supplementary material). Each node represents one document. The sizes of the nodes reflect the citation counts of the documents since: larger nodes have more citations. Documents that were frequently cited together would be located closer to each other.

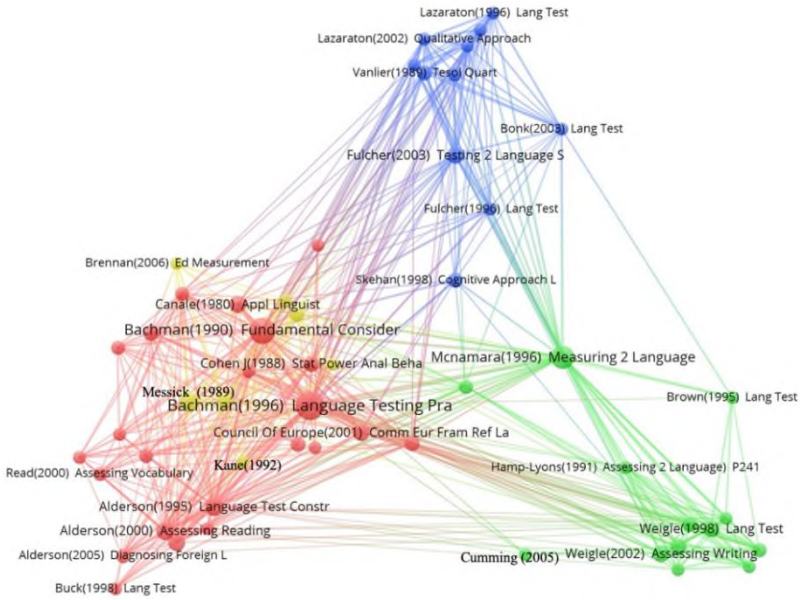

Figure 2 Network map of the most cited references (2008-2013)

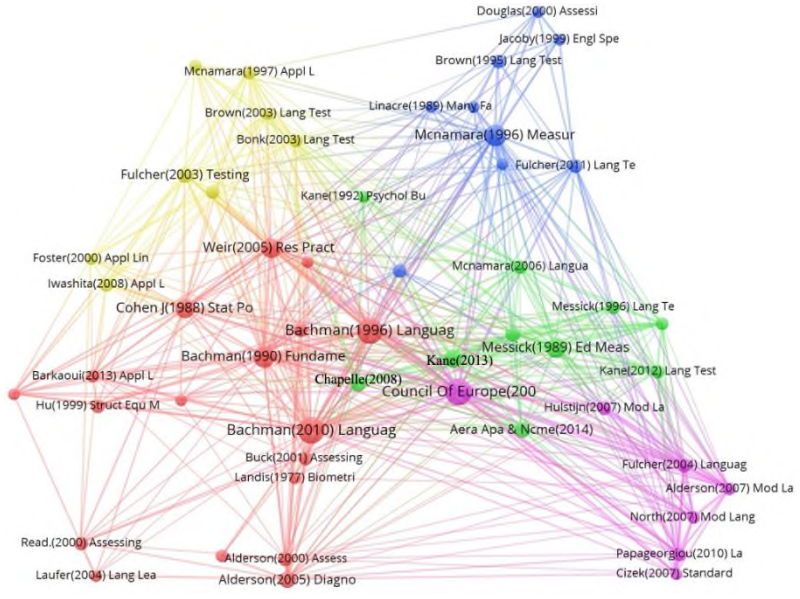

Figure 3 Network map of the most cited references (2014-2019)

4.2.2. The most influential authors

This section highlighted the prominent authors in the field according to their citation counts. The top 50 most cited authors in the two periods are given in Table 2. The intellectual network maps of the most influential authors are given in Figure 4 and Figure 5. These network maps show how the most influential authors are connected. For example, the works by L. F. Bachman and J. C. Alderson are often cited together, and they are positioned at the center of the largest cluster in both time periods. Some vocabulary specialists (e.g., J. Read, N. Schmitt) have published studies about vocabulary acquisition and vocabulary knowledge assessment. They form their own cluster quite far away from all other clusters.

Table 2 The most influential scholars in the two time periods

	2008-2013				2014-2019		
	Author	Raw citation	Norm. citation		Author	Raw citation	Norm. citation
1	Bachman, LF	220	92.4	1	Bachman, LF	166	62.6
2	Alderson, JC	137	57.6	2	Alderson, JC	152	57.4
3	McNamara, TF	116	48.7	3	McNamara, TF	93	35.1
4	Shohamy, E	83	34.9	4	Fulcher, G	92	34.7
5	Linacre, JM	75	31.5	5	Kane, MT	74	27.9
6	Brown, AL	69	29	6	Chapelle, CA	71	26.8
7	Davies, A	66	27.7	7	Weir, CJ	69	26
8	Cumming, A	65	27.3	8	Brown, AL	67	25.3
8	Messick, S	65	27.3	9	Linacre, JM	64	24.2
10	Fulcher, G	64	26.9	9	Nation, ISP	64	24.2
11	Brown, JD	60	25.2	11	Biber, D	61	23
12	Weigle, SC	59	24.8	12	Brown, JD	53	20
13	Weir, CJ	54	22.7	12	Elder, C	53	20
14	Lumley, T	52	21.9	12	Messick, S	53	20
15	Buck, G	51	21.4	12	Purpura, JE	53	20
16	Elder, C	48	20.2	12	Taylor, L	53	20
17	Kane, MT	47	19.8	17	Cumming, A	52	19.6
18	Chapelle, CA	44	18.5	17	Douglas, D	52	19.6
18	Meara, PM	44	18.5	17	Weigle, SC	52	19.6
20	Kunnan, AJ	40	16.8	20	Shohamy, E	51	19.2
21	Lazaraton, A	38	16	21	Xi, XM	49	18.5
22	Hamp-Lyons, L	35	14.7	22	Eckes, T	47	17.7
22	Read, J	35	14.7	22	Laufer, B	47	17.7
24	Skehan, P	34	14.3	24	Davies, A	46	17.4
25	Chalhoub-Deville, M	32	13.5	25	Bailey, AL	44	16.6
25	Cheng, LY	32	13.5	25	Cheng, LY	44	16.6
27	Eckes, T	31	13	25	Lee, YW	44	16.6
28	Davidson, F	30	12.6	28	Read, J	42	15.8
29	Brennan, RL	29	12.2	29	Kunnan, AJ	39	14.7
30	Cohen, J	28	11.8	30	Lumley, T	38	14.3
30	Nation, ISP	28	11.8	31	Barkaoui, K	37	14
30	Taylor, L	28	11.8	31	Hulstijn, JH	37	14
30	Xi, XM	28	11.8	31	North, B	37	14
34	Leung, C	27	11.3	34	Knoch, U	34	12.8
35	Brindley, G	26	10.9	35	Cohen, J	33	12.5
35	Canale, M	26	10.9	35	Crossley, SA	33	12.5
35	Douglas, D	26	10.9	37	Cizek, GJ	31	11.7
35	Young, RF	26	10.9	38	Brennan, RL	30	11.3
39	Bentler, PM	25	10.5	38	Plakans, L	30	11.3
39	Cronbach, LJ	25	10.5	40	Sawaki, Y	29	10.9
39	Lee, YW	25	10.5	41	Chalhoub-Deville, M	28	10.6
42	Cohen, AD	23	9.7	41	Jang, EE	28	10.6

42	Hambleton, RK	23	9.7	41	Papageorgiou, S	28	10.6
42	Laufer, B	23	9.7	44	O'sullivan, B	27	10.2
42	Spolsky, B	23	9.7	45	Ellis, R	26	9.8
42	Swain, M	23	9.7	45	Ockey, GJ	26	9.8
47	Abedi, J	22	9.2	47	Buck, G	25	9.4
47	Hughes, A	22	9.2	47	Iwashita, N	25	9.4
47	Iwashita, N	22	9.2	47	Poehner, ME	25	9.4
47	Knoch, U	22	9.2	47	Schmitt, N	25	9.4
47	North, B	22	9.2				
47	Oller, JW	22	9.2				
47	Qian, DD	22	9.2				

Note. Normalized citation is in citation counts per hundred documents

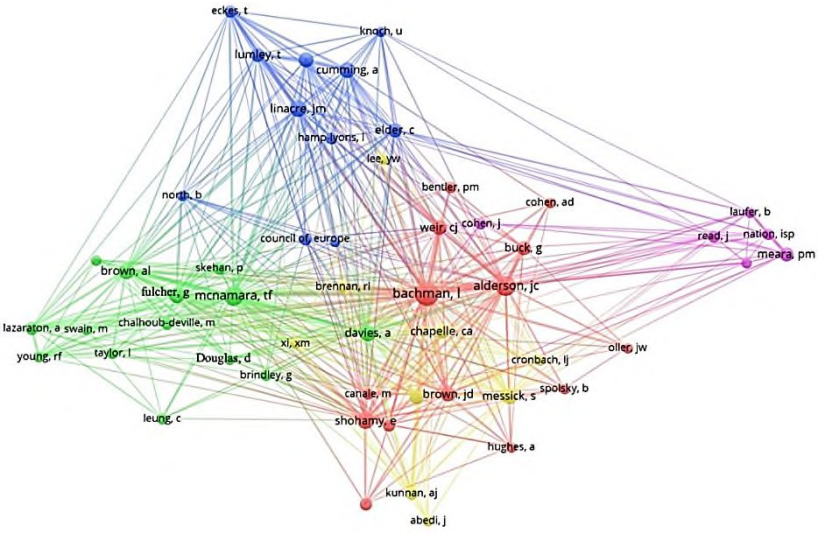

Figure 4 Author network map (2008-2013)

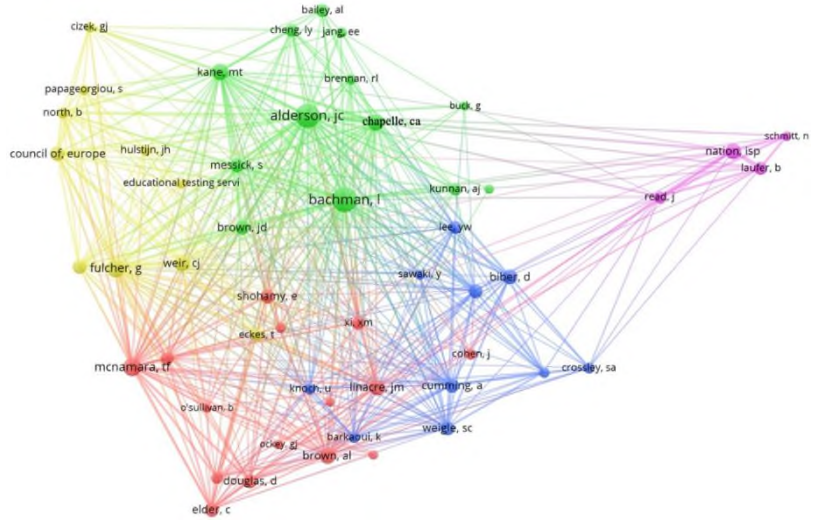

Figure 5 Author network map (2014-2019)

4.3. Keyword analysis

The keyword analysis identified 5,823 keywords. A total of 422 keywords remained following the data cleaning/recoding. This final set comprised 121 author-supplied keywords and 301 keywords extracted from abstracts. In cases where a keyword appeared in both categories, it was classified as an author-supplied keyword. Tables 6, 7, and 8 in the supplementary material give these keywords. The author-supplied keywords are marked by "AU" and the keywords-from-abstracts are marked by "AB." The frequency counts of keywords in both time periods with their LL values are summarized in the three tables. There were 27 keywords with absolute LL values larger than 3.84, suggesting that these topics experienced significant changes, for example, corpus, integrated speaking tasks, and *Common European Framework of Reference for Languages* (CEFR, Council of Europe, 2001).

5. Discussion

The current study documented a number of trends/changes that were supported by objective quantitative data, namely, citation information and citation patterns between documents and authors. The three types of analyses employed in this study revealed notable patterns and trends.

5.1. The most influential documents

The two most influential books in the field were authored by L. F. Bachman and A. S. Palmer (Bachman & Palmer, 1996, 2010) on a number of major subjects in language assessment and testing, such as test design, test development, and test use. L. F. Bachman also wrote one of the most cited monographs in the field that covers fundamental topics such as use of language test, communicative competence, reliability, validity, and so on (Bachman, 1990). A number of the most cited documents were books or book chapters (e.g., Alderson, 2005; Fulcher, 2003; McNamara, 1996; Messick, 1989) on important subjects such as diagnostic assessment, speaking assessment, Rasch models, and validity frameworks.

As shown in Figure 2, the smart local moving algorithm (Waltman & Van Eck, 2013) assigned the top 50 references in the first period to four clusters. Cluster 1 in red includes 23 documents that cover fundamental issues and key topics such as statistics, assessment for teachers, washback, communicative competence, and so on. In addition to L. F. Bachman's prominent books, the influential book series edited by J. C. Alderson and L. F. Bachman (e.g., Buck, 2001; Purpura, 2004;

Read, 2000) also appear in the cluster. These 23 documents were frequently cited together and constitute the largest cluster in the field. Cluster 4 in yellow, concerning the central assessment topic of validity that features the work of S. Messick and M. Kane (Messick, 1989; Kane, 1992, 2006), embeds itself within Cluster 1, demonstrating the close relationship between validity and other key topics in the field. Cluster 2 in green is associated with important topics regarding raters/rating issues and writing assessment. The rater/rating issues, such as rater backgrounds, are closely related to writing and speaking assessment, given that written tests are usually graded by humans. Cluster 3 mainly concerns issues in speaking assessment, featuring topics related to interaction or dialog (Brown, 2003) and the works by G. Fulcher, A. Lazaraton and others (e.g., Chalhoub-Deville, 2003; Fulcher, 2003; Lazaraton, 2002).

The network analysis generated five clusters for the second period. Connections between these clusters and their components are given in Figure 3. The largest cluster is Cluster 1 in red located at the lower left of the map. Some of the key publications include L. F. Bachman's books and the book series edited by J. C. Alderson and L. F. Bachman, which can also be found in Cluster 1 of the map at the first time period. Cluster 2, in green, includes key publications associated with test validity by S. Messick, M. Kane, and others. Cluster 3, in blue, is associated with raters and assessment in language for specific purposes (LSP), which does not exist in the map of the first time period. Cluster 4, in yellow, is mainly about assessing speaking, which can also be found in the first time period. Cluster 5, in purple, is a new cluster concerning the CEFR (Council of Europe, 2001). Since the CEFR is one of the most rapidly growing keywords in the last decade, it will be further discussed below.

5.2. The most influential authors

L. F. Bachman, J. C. Alderson, and T. McNamara occupied the top three spots in both time periods, demonstrating their prominent status in the field. In addition to these three authors, thirty-three highly influential scholars stayed on the list in both time periods, including E. Shohamy, J. M. Linacre, A. L. Brown, A. Davies, A. Cumming, S. Messick, G. Fulcher, J. D. Brown, S. C. Weigle, C. J. Weir, T. Lumley, G. Buck, C. Elder, M. T. Kane, C. A. Chapelle, A. J. Kunnan, J. Read, M. Chalhoub-Deville, L. Y. Cheng, T. Eckes, R.L. Brennan, J. Cohen, ISP Nation, L. Taylor, X. M. Xi, D. Douglas, Y. W. Lee, B. Laufer, N. Iwashita, U. Knoch, and B. North. A number of scholars received a boost in citations in the second period, for example, M. T. Kane, G. Fulcher, and C. Chapelle. In addition, sixteen scholars made it to the list in the second period, including A. Bailey, K. Barkaoui, D. Biber, G. Cizek, S. Crossley, R. Ellis, J. Hulstijn, E. Jang, B. O'Sullivan, G. Ockey, S. Papageorgiou, L. Plakans, M. Poehner, J. Purpura, Y. Sawaki, and N. Schmitt.

Some of the top 50 most cited authors were not language testing specialists, but their works were frequently cited by research articles in the field. A few examples were I. S. P. Nation/P. Meara/B. Laufer (vocabulary specialists), J. Cohen (psychologist and statistician), S. Crossley (corpus specialist), D. Biber (corpus specialist), and R. Ellis/P. Skehan/M. Swain (second language acquisition specialists who also published research related to language assessment and testing). J. Cohen's book (Cohen, 1988) on power analysis and effect size had great influences on statistical analysis in the field. The vocabulary specialist I. S. P. Nation authored some of the most important L2 vocabulary documents (e.g., Nation, 2001) that informed the design and development of many widely used vocabulary tests (e.g., the Vocabulary Levels Test and the Vocabulary Size Test). P. Skehan's (2009) language performance model based on language accuracy, complexity, and fluency also inspired many studies in the field.

5.3. Trends and changes in the field

The four skills in assessment and testing were found to be unbalanced. Between 2008 and 2013, more studies involved speaking (frequency count of 63 times, a percentage weighting of 33.2% among the four skills, computed as the frequency count as a percentage of the frequency count of all four skills), followed by writing (frequency count of 52, weighting of 28.3%) and reading (frequency count of 43, weighting of 23.4%). The least assessed skill was listening (frequency count of 28, weighting of 15.2%). Speaking continued to be the most assessed skill between 2014 and 2019, with a frequency count of 83. Its weighting increased to 38.2%. Reading, with a frequency count of 51 and a weighting of 23.5%, became the second most investigated skill in the second time period. Writing, however, dropped to 45 with a weighting of 20.7%. Listening increased slightly to 38 with a weighting of 17.5%, which made it remain the least studied skill in the field. Despite the increase of listening research, the field was losing balance with speaking becoming more dominant.

Validity was one of the most frequent keywords in the field as it appeared in 48 studies in the first period and in 60 studies in the second period. Its relatively small LL value and high frequency indices suggest that the topic remained crucial for language assessment and testing. In terms of different types of validity, several types of validity experienced a small growth (e.g., construct validity and content validity). Concurrent validity, however, decreased in frequency. Similar to validity, the keyword reliability was quite stable over the two time periods (22 at each time period). The stability of validity and reliability could be due to the fact that these two concepts could be viewed as unitary concept facets (Messick, 1989). While validity and reliability are essential for language tests, the distinction between the two is

not always clear (Bachman, 1990). The frequency of the keyword *inter-rater reliability* increased from 0 to 8, which generated an LL value of 9.90, suggesting a much higher use in the second period. A closer look at the keyword in the abstracts showed that it was mainly applied in writing/speaking or test validation.

The keyword analysis showed that the keyword CEFR received more interest in the second period (LL = 4.0). This trend was also confirmed by the co-citation analysis of the key documents of the field. Eight of the top cited documents were associated with the CEFR. The intellectual network map of the second period (see Figure 3) captured the emergence of a new cluster associated with the CEFR, suggesting that research drawing on this framework had indeed become a main theme in the field. Ever since its official appearance in 2001, the CEFRhad a significant influence on language testing, language policy, language teaching, and learning in Europe (e.g., Deygers et al., 2018; Little, 2007), which was well beyond its original purpose to inform language curriculums and examinations (Little, 2007). While the CEFR is not without criticism (e.g., Fulcher, 2004), the number of research studies associated with it was undeniably growing. The framework was often used as the alignment framework for various assessment methods, such as tests of English for academic purposes (Green, 2018), academic speaking (Berger, 2019), self-assessment (Denies & Janssen, 2016), and so on (also see the *LAQ* special issue concerning the issues for using CEFR in the higher education context).

One notable trend was that corpus-associated studies surged over time as the frequency of the keyword *corpus/corpora* increased from 2 to 21 (LL = 15.49), the biggest increase among all keywords. The keyword *learner corpus/corpora*, a sub-category in corpus linguistics, also increased from 0 to 4. This trend pushed the corpus specialist D. Biber to the most cited list in the second period. Language corpora have a lot to offer for language testing (e.g., Alderson, 1996; Cushing, 2017; Park, 2014; Xi, 2017). Despite the growing interest in using corpus/corpora for assessment and testing for more than two decades since Alderson (1996), the current bibliometric analysis found that corpus-informed studies only started to appear more frequently in top language testing journals in the past 5 years. Language corpora have now been used to conduct content analysis and validation for speaking (e.g., LaFlair & Staples, 2017), writing (e.g., Lu & Ai, 2015), vocabulary (Jarvis, 2017; Römer, 2017; Segbers & Schroeder, 2017), and grammar (e.g., Alderson & Kremmel, 2013; Pan & Qian, 2017). Corpora were also used to inform formative assessment (Can Daşkın & Hatipoğlu, 2019) and diagnostic assessment (Xie, 2019).

An emerging topic was associated with occupational tests. The keyword analysis found a significant LL value for the keyword *occupational English tests* (LL = 7.42). The 2016 special issue in *Language Testing* published a series of articles associated with language for specific-purposes (LSP) tests in the health profession, which involved several leading scholars in the field such as T. McNamara

and C. Elder. The increasing interest in language testing for LSP was also confirmed by the co-citation analysis of documents. Cluster 3 in the network map of most cited articles in the second period included three documents associated with LSP testing. In addition to the health profession, aviation English also received some interest, featuring the work of C. Elder and colleagues.

Some topics appeared for the first time in the two international journals. One example is dynamic assessment. This topic also made it to the keyword list in the second period. One of the most-cited articles was Poehner et al. (2015), which investigated a computerized version of dynamic assessment (see Table 1). M. Poehner also made it to the top cited list in the second period. All these findings suggest that the method, guided by sociocultural theory (Vygotsky, 1978), had become more influential. Given that sociocultural theory is getting more popular in the field of applied linguistics (de Bot, 2015; Lei & Liu, 2019a) and that it offers pedagogical implications to inform second language learning and teaching (e.g., Lantolf & Poehner, 2014; Lantolf & Thorne, 2006), it is foreseeable that dynamic assessment will continue to influence the field.

While a number of topics received more attention, some topics experienced a drop in interest. The most significant decrease was associated with the language testing context, which referred to whether the language being tested/assessed was the official language in a specific region (the second language context vs. the foreign language context). The frequency of *English as a second language* decreased from 25 to 8, the biggest drop among all keywords. The most frequently investigated context was *English as a foreign language* (EFL), which had frequency counts of 52 and 71 in the first and second time period, respectively. The small LL value of the keyword *EFL* (.74) suggested that the testing context remained relatively stable. The drop of interest in the ESL context might be associated with: (1) a growing body of research conducted in regions such as Europe and China where English was a foreign language, and (2) a growing interest in the CEFR. Besides the testing context, the keyword analysis also showed that assessment and testing in different levels of education (e.g., university, secondary school, primary school) remained quite stable as well.

There was a shifting interest in some important topics in the field. One of these topics was found to be washback. The frequency of the keyword *washback* dropped from 8 to 3, which generated an LL value of -3.20. The keyword *washback effect* had an LL value of -9.28 (its frequency count dropped from 6 to 0). This is surprising given the important status of the topic in language policy, teaching, and learning (e.g., Alderson & Wall, 1993; Cheng & Curtis, 2004; Messick, 1996; see the special issue on language testing on washback, edited by Alderson & Wall, 1996). One possible explanation is that washback has been viewed as one aspect of test consequences that can impact learners, teachers, tests, the education system, and the society. In addition to the topic of washback, some other topics that experienced a drop in interest included

automated scoring, *discourse analysis*, and *pronunciation*. Due to space limitation, it is not possible to get into full detail to evaluate why these topics experienced a drop. Further research is needed to explain these trends.

In terms of statistical methods, Rasch models were undoubtedly among the most widely used methods in the field as 43 studies employed the technique between 2008 and 2019. The Rasch method, named after George Rasch, models the probability of a correct response as a function of the respondent's abilities and item difficulty (e.g., Wright, 1977) that have been commonly used for item analysis in language tests. Correlation analysis was also very common as it had a total frequency count of 49. Correlation analysis has been frequently used in test alignment, exploratory construct validity, trait discrimination, and so on. Other methods included sophisticated multivariate statistical procedures that are regularly used by the testing community, e.g., factor analysis (Freq. = 19), multiple regression (Freq. = 11), structural equation models (Freq. = 10), multilevel/hierarchical models (Freq. = 6), and quantile regression (Freq. = 2). All these methods remained quite stable over time, except confirmatory factor analysis, which dropped from 9 to 3 (LL = -4.14).

6. Limitations

One limitation of the current study is the scant bibliometric data in the field. The WoS database was chosen to retrieve data for the current study because WoS is arguably the most reliable database to track the impact of research (de Bellis, 2009). However, the current study only used data from *Language Testing* and *Language Assessment Quarterly* between 2008 and 2019 that were available in WoS. Scopus maintained bibliometric data of *Language Testing* since 1994. However, the data of *Language Assessment Quarterly* before 2008 were not available in Scopus (the first issue of *Language Assessment Quarterly* was published in 2004). Due to the compatibility issue (data in the two databases are not compatible) and the inclusion criteria, the current study could only evaluate the post-2008 impact of publications and scholars. Future research may perform a pre/post 2008 bibliometric analysis based solely on data from *Language Testing* to evaluate the impact of publications and scholars at the specific time frame.

In addition, as pointed out above, our citation analysis was largely quantitative and descriptive, focusing mainly on the number of citations. The qualitative aspects of citations, (e.g., how and why a document was cited) were not analyzed. Although the nature of the current study was not qualitative in nature, future bibliometric studies are needed to take into account the qualitative aspects of citations to depict a more comprehensive picture of the intellectual structure of language assessment and testing.

7. Implications and future research

One significant finding of the current study is related to the unbalanced research among the four skills. It was found that listening was the least studied skill, a phenomenon that has also been found in areas such as second language acquisition (Zhang, 2020). While all four skills are necessary when learning a second language, listening plays a fundamental role in language development as listening is central to learning and communication (Brown, 2000). Thus, assessing listening abilities plays a crucial role in language learning and teaching. Despite its importance, Buck (2001) suggested that listening assessment is the least understood in the field of assessment, possibly due to the complexity of the listening processes and the difficulty to measure the construct of listening abilities (Batty, 2015; Buck, 2001; Wagner, 2021). Thus, there is a great need for more future research to foster listening assessment.

Another finding is associated with the application of corpora and computational tools for assessment. Although using corpora is not new in language assessment, developed computational tools such as automatic syntactic complexity analyzers (Lu, 2010, 2011) and Coh-Metrix (Graesser et al., 2004) have made it easier to generate linguistic indices such as lexical sophistication, cohesion, and syntactic complexity (e.g., Crossley et al., 2011) for a large number of texts including multi-million-word corpora within a short period of time. These tools have significant implications for language teaching and assessment. Language educators and test administrators can use these tools to obtain objective linguistic indices that are difficult to calculate manually. These indices offer new insights into the language abilities of learners that serve various purposes, such as tracking language development and performing group comparisons between learners (e.g., Lu & Ai, 2015).

8. Conclusion

The current study used the bibliometric method to offer an overview of language assessment and testing in the last decade. It identified the key documents, authors, research institutions, and topics. In addition, this study also evaluated the changes and trends in the field by comparing the intellectual maps and keywords during two time periods, which can be useful for researchers, experienced or new to the field of language assessment and testing, to plan the topics and scopes of their future research agenda. The changes and trends indicated that the field had continuously developed between 2008 and 2019. Hundreds of researchers in the field have contributed thousands of publications that kept pushing the field forward. Due to space limitations, the current study could only

list a few dozen authors and publications. This is not, however, meant to ignore other publications and authors in the field, without which the field would not have become what it is today.

References

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *Using corpora for language research: Studies in the honor of Geoffrey Leech* (pp. 248-259). Longman. https://doi.org/10.1075/ijcl.3.1.10bar

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* Continuum.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535-556. https://doi.org/10.1177/0265532213489568

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129. https://doi.org/10.1093/applin/14.2.115

Alderson, J. C., & Wall, D. (1996) (Eds.). *Language Testing*, *13*(3) [Special issue].

Anthony, L. (2018). *AntConc* (Version 3.5.6) [Computer Software]. Waseda University. http://www.laurenceanthony.net/software

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1-42. https://doi.org/10.1191/026553200675041464

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Batty, A. O. (2015). A comparison of video-and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing, 32*(1), 3-20. https://doi.org/10.1177/0265532214531254

Benson, P., Chik, A., Gao, X., Huang, J., & Wang, W. (2009). Qualitative research in language teaching and learning journals, 1997-2006. *Modern Language Journal, 93*(1), 79-90. https://doi.org/10.1111/j.1540-4781.2009.00829.x

Berger, A. (2019). Specifying progression in academic speaking: A keyword analysis of CEFR-based proficiency descriptors. *Language Assessment Quarterly*, *17*(1), 85-99. https://doi.org/10.1080/15434303.2019.1689981

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*, 1-25.

Brown, H. D. (2000). *Principles of language learning and teaching* (Vol. 4). Longman. https://doi.org/10.1191/0265532203lt242oa

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information, 22*(2), 191-235. https://doi.org/10.1177/053901883022002003

Can Daşkın, N., & Hatipoğlu, Ç. (2019). Reference to a past learning event as a practice of informal formative assessment in L2 classroom interaction. *Language Testing*, *36*(4), 527-551. https://doi.org/10.1177/0265532219857066

Carroll, J. B. (1961). Fundamental considerations in testing of English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 313-321). McGraw-Hill.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369-383. https://doi.org/10.1191/0265532203lt264oa

Chang, Y. W., Huang, M. H., & Lin, C. W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics, 105*, 2071-2087. https://doi.org/10.1007/s11192-015-1762-8

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing, 16*(1), 49-71. https://doi.org/10.1016/j.asw.2010.11.001

Chen, M. L. (2023). SLA as an interdiscipline: A bibliometric study. *Studies in Second Language Learning and Teaching*, *13*(4), 843-882. https://doi.org/10.14746/ssllt.40218

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng & Y. Watanabe (Eds.), *Washback in language testing* (pp. 25-40). Routledge. https://doi.org/10.4324/9781410609731-9

Chiu, W. T., & Ho, Y. S. (2007). Bibliometric analysis of tsunami research. *Scientometrics, 73*(1), 3-17. https://doi.org/10.1007/s11192-005-1523-1

Cohen, J. (1988). *Statistical power analysis for the social sciences.* Erlbaum.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Courtial, J. (1994). A coword analysis of scientometrics. *Scientometrics, 31*(3), 251-260. https://doi.org/10.1007/bf02016875

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561-580. https://doi.org/10.1177/0265532210378031

Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing, 34*(4), 441-449. https://doi.org/10.1177/0265532217713044

de Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow Press. https://doi.org/10.12775/tsb.2012.009

de Bot, K. (2015). *A history of applied linguistics: From 1980 to the present*. Routledge.

de Groote, S. L., & Raszewski, R. (2012). Coverage of Google Scholar, Scopus, and Web of Science: A case study of the h-index in nursing. *Nursing Outlook, 60*(6), 391-400. https://doi.org/10.1016/j.outlook.2012.04.007

Denies, K., & Janssen, R. (2016). Country and gender differences in the functioning of CEFR-based can-do statements as a tool for self-assessing English proficiency. *Language Assessment Quarterly*, *13*(3), 251-276. https://doi.org/10.1080/15434303.2016.1212055

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3-15. https://doi.org/10.1080/15434303.2016.1261350

Farhady, H. (2018). History of language testing and assessment. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-7). John Wiley & Sons. https://doi.org/10.1002/9781118784235.eelt0343

Fulcher, G. (2003). *Testing second language speaking*. Pearson Longman.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, *1*(4), 253-266. https://doi.org/10.1207/s15434311laq0104_4

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*(3159), 108-111.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193-202. https://doi.org/10.3758/bf03195564

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), 59-74. https://doi.org/10.1080/15434303.2017.1350685

Holden, G., Rosenberg, G., & Barker, K. (2005). Tracing thought through time and space: A selective review of bibliometrics in social work. In G. Holden, G. Rosenberg, & K. Barker (Eds.). *Bibliometrics in social work* (pp.1-34). Routledge. https://doi.org/10.1300/j010v41n03_01

Hyland, K., & Jiang, F. K. (2021a). A bibliometric study of EAP research: Who is doing what, where and when? *Journal of English for Academic Purposes*, *49*, Article 100929. https://doi.org/10.1016/j.jeap.2020.100929

Hyland, K., & Jiang, F. K. (2021b). Delivering relevance: The emergence of ESP as a discipline. *English for Specific Purposes*, *64*, 13-25. https://doi.org/10.1016/j.esp.2021.06.002

Hyland, K., & Jiang, F. (2023). Interaction in written texts: A bibliometric study of published research. *Studies in Second Language Learning and Teaching*, *13*(4), 903-924. https://doi.org/10.14746/ssllt.40220

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537-553. https://doi.org/10.1177/0265532217710632

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Eds.), *Educational measurement* (4th ed., pp. 17-64). American Council on Education/Praeger.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(4), 451-475. https://doi.org/10.1177/0265532217713951

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book.* McGraw-Hill.

Lantolf, J. P., & Poehner, M. E. (2014). *Sociocultural theory and the pedagogical imperative in L2 education: Vygotskian praxis and the research/practice divide*. Routledge. https://doi.org/10.4324/9780203813850

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press. https://doi.org/10.1093/applin/amm027

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics, 23*(3)*,* 417-461. https://doi.org/10.1007/BF02029807

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge University Press.

Lei, L., Deng, Y., & Liu, D. (2023). Research on the learning/teaching of L2 listening: A bibliometric review and its implications. *Studies in Second Language Learning and Teaching*, *13*(4), 781-810. https://doi.org/10.14746/ssllt.40216

Lei, L., & Liao, S. (2017). Publications in linguistics journals from Mainland China, Hong Kong, Taiwan, and Macau (2003-2012): A bibliometric analysis. *Journal of Quantitative Linguistics, 24*(1), 54-64. https://doi.org/10.1080/092 96174.2016.1260274

Lei, L., & Liu, D. (2019a). Research trends in Applied Linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, *40*(3), 540-561. https://doi.org/10.1093/applin/amy003

Lei, L., & Liu, D. (2019b). The research trends and contributions of System's publications over the past four decades (1973-2017): A bibliometric analysis. *System, 80*, 1-13. https://doi.org/10.1016/j.system.2018.10.003

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal*, *91*(4), 645-655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62. https://doi.org/10.5054/tq.2011.240859

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16-27. https://doi.org/10.1016/j.jslw.2015.06.003

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

McNamara, T. (2004). Language testing. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 763-783). Blackwell.

Meara, P. (2012). The bibliometrics of vocabulary acquisition: An exploratory study. *RELC Journal*, *43*(1), 7-22. https://doi.org/10.1177/0033688212439339

Meara, P. (2023). The Routledge handbook of vocabulary studies: A study in micro-bibliometrics. *Studies in Second Language Learning and Teaching*, *13*(4), 883-902. https://doi.org/10.14746/ssllt.40219

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). Macmillan/American Council on Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256. https://doi.org/10.1177/026553229601300302

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Pan, M., & Qian, D. D. (2017). Embedding corpora into the content validation of the grammar test of the National Matriculation English Test (NMET) in China. *Language Assessment Quarterly*, *14*(2), 120-139. https://doi.org/10.1080/15434303.2017.1303703

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly, 11*(1), 27-44. https://doi.org/10.1080/15434303.2013.872647

Poehner, M. E., Zhang, J., & Lu, X. (2015). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing*, *32*(3), 337-357. https://doi.org/10.1177/0265532214560390

Purpura, J. E. (2004) *Assessing grammar*. Cambridge University Press.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on comparing corpora* (pp. 1-6). Association for Computational Linguistics.

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Riazi, A. M., Ghanbar, H., Marefat, F., & Fazel, I. (2023). Review and analysis of empirical articles published in TESOL Quarterly over its lifespan. *Studies in*

Second Language Learning and Teaching, *13*(4), 811-841. https://doi.org/10.14746/ssllt.40217

Roemer, R. C., & Borchardt, R. (2015). *Meaningful metrics: A 21st-century librarian's guide to bibliometrics, altmetrics, and research impact*. American Library Association.

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, *34*(4), 477-492. https://doi.org/10.3115/1117729.1117730

Segbers, J., & Schroeder, S. (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing*, *34*(3), 297-320. https://doi.org/10.1177/0265532216641152

Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. Longman.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology*, *24*(4), 265-269. https://doi.org/10.1002/asi.4630240406

Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5-30). Peter Lang.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford University Press. https://doi.org/10.1177/026553229601300108

Spolsky, B. (2017). History of language testing. In E. Shohamy, I. G. Or, & S. May. (Eds.), *Language testing and assessment* (pp. 375-384). https://doi.org/10.1007/978-3-319-02261-1_32

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics, 7*(1), 39-56. https://doi.org/10.1093/applin/7.1.39

Wagner, E. (2021). Assessing listening. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 223-235). Routledge. https://doi.org/10.4324/9781003220756-18

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B, 86*, Article 471. https://doi.org/10.1140/epjb/e2013-40829-0

Waltman, L., & Van Eck, N. J. (2017). *VOSviewer manual.* Online retrieved from http://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.6.pdf.

Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4)*,* 629-635. https://doi.org/10.1016/j.joi.2010.07.002

Wang L., & Zhang L. J. (2019). Peter Skehan's influence in research on task difficulty in second language learners' acquisition of oral and written language. In E. Wen & W. Ahmadian (Eds.*), Researching L2 task performance and pedagogy: In honor of Peter Skehan* (pp. 183-198). John Benjamins. https://doi.org/10.1075/tblt.13.09wan

White, H. D. (2004). Citation analysis and discourse analysis revisited. *Applied Linguistics, 25*(1), 89-116. https://doi.org/10.1093/applin/25.1.89

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97-116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger & A. Koll-Stobbe (Eds.), *New approaches to the study of linguistic variability* (pp. 3-11). Peter Lang. https://doi.org/10.3726/978-3-653-03231-4

Vygotsky, L. S. (1978). *Mind in society*. Harvard University Press.

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing*, *34*(4), 565-577. https://doi.org/10.1177/0265532217720956

Xie, Q. (2019). Diagnosing linguistic problems in English academic writing of university students: An item bank approach. *Language Assessment Quarterly*. *17*(2), 183-203. https://doi.org/10.1080/15434303.2019.1691214

Xu, Y., Zhuang, J., Blair, B., Kim, A., Li, F., Thorson Hernández, R., & Plonsky, L. (2023). Modeling quality and prestige in applied linguistics journals: A bibliometric and synthetic analysis. *Studies in Second Language Learning and Teaching*, *13*(4), 755-779. https://doi.org/10.14746/ssllt.40215

Zhang, X. (2020). A bibliometric analysis of second language acquisition between 1997 and 2018. *Studies in Second Language Acquisition*, *42*(1), 199-222. https://doi.org/10.1017/s0272263119000573