



An Extensive Knowledge Mapping Review of Measurement and Validity in Language Assessment and SLA Research

Vahid Aryadoust^{1*}, Azrifah Zakaria¹, Mei Hui Lim² and Chaomei Chen^{3,4}

¹ National Institute of Education, Nanyang Technological University, Singapore, Singapore, ² Nanyang Technological University, Singapore, Singapore, ³ College of Computing and Informatics, Drexel University, Philadelphia, PA, United States, ⁴ Department of Information Science, Yonsei University, Seoul, South Korea

OPEN ACCESS

Edited by:

Thomas Eckes,
Ruhr University Bochum, Germany

Reviewed by:

John Read,
The University of Auckland,
New Zealand
Stefanie A. Wind,
University of Alabama, United States

***Correspondence:**

Vahid Aryadoust
vahid.aryadoust@nie.edu.sg

Specialty section:

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

Received: 04 February 2020

Accepted: 14 July 2020

Published: 04 September 2020

Citation:

Aryadoust V, Zakaria A, Lim MH and Chen C (2020) An Extensive Knowledge Mapping Review of Measurement and Validity in Language Assessment and SLA Research. *Front. Psychol.* 11:1941.
doi: 10.3389/fpsyg.2020.01941

This study set out to investigate intellectual domains as well as the use of measurement and validation methods in language assessment research and second language acquisition (SLA) published in English in peer-reviewed journals. Using Scopus, we created two datasets: (i) a dataset of core journals consisting of 1,561 articles published in four language assessment journals, and (ii) a dataset of general journals consisting of 3,175 articles on language assessment published in the top journals of SLA and applied linguistics. We applied document co-citation analysis to detect thematically distinct research clusters. Next, we coded citing papers in each cluster based on an analytical framework for measurement and validation. We found that the focus of the core journals was more exclusively on reading and listening comprehension assessment (primary), facets of speaking and writing performance such as raters and validation (secondary), as well as feedback, corpus linguistics, and washback (tertiary). By contrast, the primary focus of assessment research in the general journals was on vocabulary, oral proficiency, essay writing, grammar, and reading. The secondary focus was on affective schemata, awareness, memory, language proficiency, explicit vs. implicit language knowledge, language or semantic awareness, and semantic complexity. With the exception of language proficiency, this second area of focus was absent in the core journals. It was further found that the majority of citing publications in the two datasets did not carry out inference-based validation on their instruments before using them. More research is needed to determine what motivates authors to select and investigate a topic, how thoroughly they cite past research, and what internal (within a field) and external (between fields) factors lead to the sustainability of a Research Topic in language assessment.

Keywords: document co-citation analysis, language assessment, measurement, review, Scientometrics, validity, visualization, Second language acquisition

INTRODUCTION

Although the practice of language testing and/or assessment can be traced back in history to ancient eras in China (Spolsky, 1990), many language assessment scholars recognize the pioneering book of Lado (1961) and the book chapter of Carroll (1961), as the beginning of the modern language testing/assessment field (Davies, 2008, 2014). The field was routinely referred to as language testing,

at least from the 1950s until the 1990s. In contemporary usage, it is possible to make a distinction between testing and assessment, in terms of the formality and stakes involved in the procedures, the use of quantitative vs. qualitative approaches in design and implementation and other aspects¹. Nonetheless, in the present study, testing, and assessment are used interchangeably. Despite the general recognition of 1961 as the beginning of the field of language testing, there had been many language testing studies published before 1961, particularly in the field of reading (e.g., Langsam, 1941; Davis, 1944; Hall and Robinson, 1945; see also Rosenshine, 2017; Aryadoust, 2020 for reviews). By definition, these studies qualify as language testing research and practice since they meet several criteria that Priscilla Allen, Alan Davies, Carol Chapelle and Geoff Brindley, and F. Y. Edgeworth set forth in their delineations of language testing, most notably the practice of evaluating language ability/proficiency, the psychometric activity of developing language tests, and/or decision making about test takers based on test results Fulcher (n.d.).

In order to build a fair portrayal of a discipline, researchers often review the research outputs that have been generated over the years to understand its past and present trends (Goswami and Agrawal, 2019). For language assessment, several scholars have surveyed the literature and divided its development into distinct periods (Spolsky, 1977, 1995; Weir, 1990; Davies, 2014), while characterizing its historical events (Spolsky, 2017). Alternatively, some provided valuable personal reflections on the published literature (Davies, 1982; Skehan, 1988; Bachman, 2000; Alderson and Banerjee, 2001, 2002). Examples of personal reflections on specific parts of language assessment history also include Spolsky's (1990) paper on the "prehistory" of oral examinations and Weir et al.'s (2013) historical review of Cambridge assessments.

These narrative reviews offer several advantages such as the provision of "experts' intuitive, experiential, and explicit perspectives on focused topics" (Pae, 2015, p. 417). On the other hand, narrative reviews are qualitative in nature and do not use databases or vigorous frameworks and methodologies (Jones, 2004; Petticrew and Roberts, 2006). This contrasts with quantitative reviews, which have specific research questions or hypotheses and rely on the quantitative evaluation and analysis of data (Collins and Fauser, 2005). An example of such an approach is Scientometrics which is "the quantitative methods of the research on the development of science as an informational process" (Nalimov and Mulcjenko, 1971, p. 2). This approach comprises several main themes including "ways of measuring research quality and impact, understanding the processes of citations, mapping scientific fields and the use of indicators in research policy and management" (Mingers and Leydesdorff, 2015, p. 1). This wide scope makes Scientometrics a specialized and "extensively institutionalized area of inquiry" (De Bellis, 2014, p. 24). Thus, it is appropriate for analyzing the entire areas of research across various research fields (Mostafa, 2020).

¹We are grateful to one of the reviewers for suggesting this note.

Present Study

The present study had two main aims. First, we adopted Scientometrics to identify the intellectual structure of language assessment research published in English peer-reviewed journals. Although Scientometrics and similar approaches such as Bibliometric have been adopted in applied linguistics to investigate the knowledge structure across several research domains (Arik and Arik, 2017; Lei and Liu, 2019), there is currently no study that has investigated the intellectual structure of research in language assessment. Here, intellectual structure refers to a set of research clusters that represents specialized knowledge groups and research themes, as well as the growth of the research field over time (Goswami and Agrawal, 2019). To identify an intellectual structure, a representative dataset of the published literature is firstly generated and specialized software is subsequently applied to mine and extract the hidden structures in the data (Chen, 2016). The measures generated are then used to portray the structure and dynamics of the field "objectively," where the dataset represents the research field in question (Goswami and Agrawal, 2019). Second, we aim to examine the content of emerged research clusters, using two field-specific frameworks to determine how each cluster can be mapped onto commonly adopted methodologies in the field: validity argument (Chapelle, 1998; Bachman, 2005; Kane, 2006; Chapelle et al., 2008; Bachman and Palmer, 2010) and measurement frameworks (Norris and Ortega, 2003). The two research aims are discussed in detail next.

First Aim

To achieve the first aim of the study, we adopted a Scientometric technique known as document co-citation analysis (DCA) (Chen, 2006, 2010) to investigate the intellectual structure for the field of language assessment as well as assessment-based research in second language acquisition (SLA). Co-citation refers to the frequency with which two or more publications are referenced in another publication (Chen, 2003, 2016). When a group of publications cites the same papers and books, this means that they are not only thematically related but they also take reference from the same pool of papers (Chen, 2003). Moreover, co-citations can be also generalized to authors and journals by identifying the frequency with which they have been written by the same authors or cited using the same journal resource (Chen, 2004, 2006; Chen and Song, 2017). Of note, co-citation analysis is similar to factor analysis that is extensively used for data reduction and pattern recognition in surveys and tests. In the latter, items are categorized into separate clusters called factors based on their correlation patterns. Factor loadings indicate the correlation of the item in question with other items that are categorized as a factor (Field, 2018). Some items have high loadings on latent variables, whereas others have low loading coefficients. The items with low loading coefficients do not make a significant contribution to the measurement of the ability or skill under assessment and can be removed from the instrument without affecting the amount of variance explained by the test items (Field, 2018). Similarly, co-citation analysis categorizes publications as discrete research clusters based on the publications that are co-cited in each cluster. When two

publications co-cite a source or reference, this suggests that they may be related. If these publications share (co-cite) at least 50% of their references, it is plausible that there is a significant thematic link between them. Identifying the publications that co-cite the same sources facilitates the identification of the related research clusters via their pool of references. The publications that are clustered together (like factors in factor analysis) may be then inspected for their thematic relationships, either automatically through text-mining methods or manually by experts who read the content of the clustered publications. Furthermore, there may be influential publications in each cluster that have received large numbers of co-citations from other publications, and this is termed as “citation bursts.” Reviewing the content of the citation bursts can further help researchers characterize the cluster in terms of its focus and scope (Chen, 2017).

Second Aim

To achieve the second aim of the study, we developed a framework to describe measurement and validation practices across the emerged clusters. Despite the assumption that testing and assessment practices are specific to the language assessment field, SLA researchers have employed certain assessment techniques to investigate research questions pertinent to SLA (Norris and Ortega, 2003). Nevertheless, there seems to be methodological and conceptual gaps in assessment between the language testing field and SLA, which several publications attempted to bridge (Upshur, 1971; Bachman, 1990; see chapters in Bachman and Cohen, 1998). Bachman (1990, p. 2) asserted that “language testing both serves and is served by research in language acquisition and language teaching. Language tests, for example, are frequently used as criterion measures of language abilities in second language acquisition research.” He extended the uses and contributions of language assessment to teaching and learning practices, stressing that language tests are used for a variety of purposes like assessing progress and achievement, diagnosing learners’ strengths and weaknesses, and as tools for SLA research. He stressed that insights from SLA can reciprocally assist language assessment experts to develop more useful assessments. For example, insights from SLA research on learners’ characteristics and personality can help language testing experts to develop measurement instruments to investigate the effect of learner characteristics on assessment performance. Therefore, in Bachman (1990) view, the relationship between SLA and language assessment is not exclusively unidirectional or exclusive to validity and reliability matters. Despite this, doubts have been voiced regarding the measurement of constructs in SLA (Bachman and Cohen, 1998) and the validity of the instruments used in SLA (Chapelle, 1998). For example, Norris and Ortega (2003) critiqued SLA research on the grounds that measurement is not often conducted with sufficient rigor.

Measurement is defined as the process of (i) construct representation, (ii) construct operationalization, (iii) data collection via “behavior elicitation” (Norris and Ortega, 2003, p. 720), (iv) data analysis to generate evidence, and (v) the employment of that evidence to draw theory-based conclusions (Messick, 1989, 1996). To establish whether measurement instruments function properly, it is essential to investigate

their reliability and, where applicable and plausible, validate interpretations and uses of their results (scores) (Messick, 1996; Kane, 2006). Reliability refers to the evidence that the measurement is precise or has low error of measurement (Field, 2018) and its output is reproducible across occasions, raters, and test forms (Green and Salkind, 2014; Grabowski and Oh, 2018). In addition, since the publication of Cronbach and Meehl (1955) paper, validation has been primarily treated as the process of developing arguments to justify the meaning and utility of test scores or assessment results. Messick (1989) emphasized that validation should encompass evidentiary and consequential bases of score interpretation and meaning and Kane (2006) proposed a progressive plan for collecting various sorts of evidence to buttress inferences drawn from the data and rebut counter-evidence (if any). Like the theory of measurement, Messick (1989) and Kane (2006) frameworks have had a lasting impact on language assessment (Bachman, 2005; Chapelle et al., 2008; Bachman and Palmer, 2010; Aryadoust, 2013).

We note that, in addition to the argument-based validation framework, there are several validation frameworks such as Weir (2005b) socio-cognitive framework or Borsboom and Mellenbergh (2007) test validity framework which have been adopted in some previous research. However, Borsboom and Mellenbergh (2007) work is less well-known in language assessment and SLA and has a heavy focus on psychometrics. In addition, certain components of Weir (2005a) framework such as cognitive validity are relatively under-researched in language assessment and SLA and coding the studies for these components would not generate as useful information. Therefore, the choice of argument-based validation framework seems to be more plausible for this study, although we do recognize the limitations of the approach (see *Conclusion*).

Bachman (2005) stressed that, before using an assessment for decision-making purposes, a validity argument should be fully fledged in terms of evidence supporting test developers’ claims. On the other hand, empirical validation studies have demonstrated that collecting such evidence to establish an all-encompassing validity argument is an arduous and logically complex task (Chapelle et al., 2008; Aryadoust, 2013; Fan and Yan, 2020). We are, hence, keen to determine the extent to which language assessment and SLA studies involving measurement and assessment have fulfilled the requirements of validation in the research clusters that are identified through DCA.

METHODOLOGY

Overview

This study investigated the intellectual structure in the language assessment field. It examines the literature over the period 1918–2019 to identify the network structure of influential research domains involved in the evolution of language assessment. The year 1918 is the lower limit as it is the earliest year of coverage by Scopus. The study adopted a co-citation method that comprises document co-citation analysis (DCA) (Small and Sweeney, 1985; Chen, 2004, 2006, 2010, 2016; Chen et al., 2008, 2010). The study also adopted CiteSpace Version 5.6.R3 (Chen, 2016), a computational tool used to identify highly cited publications

and authors that acted as pivotal points of transition within and among research clusters (Chen, 2004).

Data Source and Descriptive Statistics

Scopus was employed as our main database, with selective searches carried out to create the datasets of the study. We identified several publications that defined language assessment as the practice of assessing first, second or other languages (Hornberger and Shohamy, 2008), including the assessment of what is known to be language “skills and elements” or a combination of them. Despite the defined scope, the bulk of the publications concerns SLA (as will be seen later). We treated the journals that proclaimed their focus to be exclusively language assessment as the “core journals” of the field, while using a keyword search to identify the focus of language assessment publications in applied linguistics/SLA journals. Accordingly, two datasets were created (see Appendix for the search code).

- (i) A core journals dataset consisting of 1,561 articles published in *Language Testing*, *Assessing Writing*, *Language Assessment Quarterly*, and *Language Testing in Asia*, which were indexed in Scopus. These journals focus specifically on publishing language assessment research and were, accordingly, labeled as core journals. The dataset also included all the publications (books, papers etc.) that were cited in the *References* of these articles.
- (ii) A general journals dataset consisting of 3,175 articles on language assessment published in the top 20 journals of applied linguistics/SLA. The dataset also included all the publications cited in these articles. This list of journals was identified based on their ranking in the “Scimago Journal and Country Rank (SJR)” database and their relevance to the current study. The journals consisted of *Applied Psycholinguistics*, *System*, *Language Learning*, *Modern Language Journal*, *TESOL Quarterly*, *Studies in Second Language Acquisition*, *English Language Teaching*, *RELC Journal*, *Applied Linguistics*, *Journal of Second Language Writing*, *English for Specific Purposes*, *Language Awareness*, *Language Learning and Technology*, *Recall*, *Annual Review of Applied Linguistics*, and *Applied Linguistics Review*. There was no overlap between *i* and *ii*. To create *ii*, the Scopus search engine was set to search for generic keywords consisting of “test,” “assess,” “evaluate,” “rate,” and “measure” in the titles, keywords, or abstracts of publication². These search words were chosen from the list of high-frequency words that were extracted by Scopus from the core journal dataset *(i)*. Next, we reviewed the coverage of 1,405 out of 3,175 articles³, as determined by CiteSpace analysis, that contributed to the networks in this dataset to ascertain if they addressed a topic in language assessment. The publications were

²We did not include methodological journals such as ‘*Journal of Educational Measurement*’ in the search, as the majority of the papers in those journals include the search keywords, even though they are not relevant to language assessment.

³In DCA, some publications may not have a clear link with the rest of the publications in the dataset. These were not listed among the contributory publications to the major clusters that were visualized by CiteSpace in the present study.

found to either have an exclusive focus on assessment or used assessment methods (e.g., test development, reliability analysis, or validation) as one of the components in the study.

Supplemental Table 1 presents the total number of articles published by the top 20 journals, countries/regions, and academic institutes. The top three journal publishers were *Language Testing*, *System*, and *Language Learning*, with a total of 690, 389, and 361 papers published between 1980 and 2019—note that there were language testing/assessment studies published earlier in other journals. In general, the journals published more than 100 papers, with the exceptions of *Language Learning Journal*, *ReCall*, *Language Awareness*, *Journal of Second Language Writing*, *Language Learning and Technology*, and *English for Specific Purposes*. The total number of papers published by the top five journals (2,087) accounted for more than 50% of the papers published by all journals.

The top five countries/regions producing the greatest number of articles were the *United States (US)*, the *United Kingdom*, *Canada*, *Iran*, and *Japan*, with 1,644, 448, 334, 241, and 233 articles, respectively. Eleven of the top 20 countries/regions, listed in **Supplemental Table 1**, published more than 100 articles. The top three academic institutes publishing articles were the *Educational Testing Service (n = 99)*, the *University of Melbourne (n = 92)*, and *Michigan State University (n = 68)*. In line with the top producing country, just over half of these institutions were located in the US.

First Aim: Document Co-Citation Analysis (DCA)

The document co-citation (DCA) technique was used to measure the frequency of earlier literature co-cited together in later literature. DCA was used to establish the strength of the relationship between the co-cited articles, identify ‘popular’ publications with high citations (bursts) in language assessment, and identify research clusters comprising publications related via co-citations⁴. DCA was conducted twice—once for each dataset obtained from Scopus, as previously discussed. We further investigated the duration of burstiness (the period of time in which a publication continued to be influential) and burst strength (the quantified magnitude of influence).

Visualization and Automatic Labeling of Clusters

The generation of a timeline view on CiteSpace allowed for clusters of publications to be visualized on discrete horizontal axes. Clusters were arranged in a vertical manner descending in size, with the largest cluster at the top. Colored lines representing co-citation links were added in the time period of the corresponding color. Publications that had a citation burst and/or were highly cited were represented with red tree rings or appear larger than the surrounding nodes.

⁴CiteSpace, by default, shows the largest connected component. If a cluster does not appear in the largest connected component, this means it must appear in the second-largest connected component or other smaller components. The present study was limited to clusters within the largest connected component, which is a widely adopted strategy in network analysis.

The identified clusters were automatically labeled. In CiteSpace, three term ranking algorithms can be used to label clusters: latent semantic indexing (LSI), log-likelihood ratio (LLR), or mutual information (MI). The ranking algorithms use different methods to identify the cluster themes. LSI uses document matrices but is “underdeveloped” (Chen, 2014, p.79). Both LLR and MI identify cluster themes by indexing noun phrases in the abstracts of citing articles (Chen et al., 2010), with different ways of computing the relative importance of said noun phrases. We chose the labels selected by LLR (rather than MI) as they represent unique aspects of the cluster (Chen et al., 2010) and are more precise at identifying cluster themes (Aryadoust and Ang, 2019).

While separate clusters represent discrete research themes, some clusters may consist of sub-themes. For example, our previous research indicated that certain clusters are characterized by publications that present general guidelines on the application of quantitative methods alongside publications focused on a special topic, e.g., language-related topics (Aryadoust and Ang, 2019; Aryadoust, 2020). In such cases, subthemes and their relationships should be identified (Aryadoust, 2020).

Temporal and Structural Measures of the Networks

To evaluate the quality of the DCA network, temporal and structural measures of networks were computed. Temporal measures were computed using citation burstness and sigma (Σ). Citation burstness shows how favorably an article was regarded in the scientific community. If a publication receives no sudden increase of citations, its burstness tends to be close or equal to zero. On the other hand, there is no upper boundary for burstness. The sigma value of a node in CiteSpace merges the citation burstness and betweenness centrality, demonstrating both the temporal and structural significance of a citation. Sigma could also be indicative of novelty, detecting publications that presented novel ideas in their respective field (Chen et al., 2010). That is, the higher the sigma value, the higher the likelihood that the publication includes novel ideas.

Structural measures comprised the average silhouette score, betweenness centrality, and the modularity (Q) index. The average silhouette score ranges between -1 and 1 and measures the quality of the clustering configuration (Chen, 2019). This score defines how well a cited reference matches with the cluster in which it has been placed (vs. other clusters), depending on its connections with neighboring nodes (Rousseeuw, 1987). A high mean silhouette score suggests a large number of citers leading to the formation of a cluster, and is therefore reflective of high reliability of clustering; by contrast, a low silhouette score illustrates low homogeneity of clusters (Chen, 2019).

The modularity (Q) index ranges between -1 and 1 and determines the overall intelligibility of a network by decomposing it into several components (Chen et al., 2010; Chen, 2019). A low Q score hints at a network cluster without clear boundaries, while a high Q score is telling of a well-structured network (Newman, 2006).

The betweenness centrality metric ranges between 0 and 1 and assesses the degree to which a node is in the middle of a link that connects to other nodes within the network (Brandes,

2001). Moreover, a high betweenness centrality indicates that a publication may contain groundbreaking ideas; if a node is the only connection between two large but otherwise unrelated clusters, this is evidence that the author scores are high on betweenness centrality (Chen et al., 2010).

However, it must be noted that these measures are not absolute scales where a higher value automatically indicates increased importance. Rather, they show tendencies and directions for the analyst to pursue. In practice, one should also consider the diversity of the citing articles (Chen et al., 2010). For example, a higher silhouette value generated from a single citing article is not necessarily indicative of greater importance than a relatively lower value from multiple distinct citing articles. Likewise, the significance of the modularity index and the betweenness centrality metric is subject to interpretation, dependent on further analyses, including of citing articles.

Second Aim: The Analytical Framework

In DCA, clusters reflect what *citing* papers have in common in terms of how they cite references together (Chen, 2006). Therefore, we designed an analytical framework to examine the citing publications in the clusters (Table 1). In addition, we took into account the bursts (cited publications) per cluster in deciding what features would characterize each cluster. The framework was informed by a number of publications in language assessment research such as Aryadoust (2013), Bachman (1990), Bachman and Cohen (1998), Bachman and Palmer (2010), Chapelle et al. (2008), Eckes (2011), Messick (1989), Messick (1996), Kane (2006), Norris and Ortega (2003), and Xi (2010a). In Table 1, “component” is a generic term to refer to the inferences that are drawn from the data and are supported by warrants (specific evidence that buttress the claims or conclusions of the data analysis) (Kane, 2006; Chapelle et al., 2008; Bachman and Palmer, 2010). In addition, it also refers to the facets of measurement articulated by Messick (1989, 1996) and Norris and Ortega (2003) in their investigation of measurement and construct definition in assessment and SLA. It should be noted that the validity components in this framework, i.e., generalization, explanation, extrapolation, and utilization, are descriptive (rather than evaluative) and intended to record whether or not particular studies reported evidence for them. Thus, the lack of reporting of these components does not necessarily indicate that this evidence was not presented when it should have been, unless it is stated otherwise.

Using this framework, we coded the publications independently and compared their codes. Only few discrepancies were identified which were subsequently resolved by the first author.

RESULTS

DCA of the Core and General Journals Networks

Supplemental Table 2 presents the top publications in the core and general journals datasets with the strongest citation bursts sustained for at least 2 years. (Due to space constraints, only the top few publications have been presented). Overall, the

TABLE 1 | The analytical framework to address the second aim of the study.

Component	Definition	Relevant procedures and/or warrants	References
Domain specification	The definition of the target use domain (TLU) domain and the components of the representation of the construct in question (construct representation)	Generating a theoretical framework to explain (i) the cognitive processes of the latent trait under investigation (competency-based approach) and/or (ii) the characteristics of the tasks that represent the TLU domain (task-based approach)	Messick, 1989; Norris and Ortega, 2003; Chapelle et al., 2008
Construct operationalization	The realization of the construct or translating the construct definition into actual assessment instruments	(i) Using one or more task formats such as open-ended questions or discrete-point/selected response methods like multiple choice questions, and (ii) experts' evaluation of the tasks	Messick, 1989; Norris and Ortega, 2003
Evaluation (scoring)	Eliciting the intended behavior from the test taker and using a scale to translate the test performance to a score, mark, or grade	(i) Developing or adapting a scale to grade or provide feedback on students' performance. This can be conducted by human raters or machines (e.g., automated writing evaluators), (ii) establishing the reliability of the scale using reliability analysis (e.g., internal consistency or rater reliability)	Norris and Ortega, 2003; Kane, 2006; Chapelle et al., 2008; Bachman and Palmer, 2010; Xi, 2010a; Grabowski and Oh, 2018
Generalization	Establishing whether the observed scores represent a "universe score" and are not exclusive to the test form, rater, or test item formats in the assessment	Generalizability theory analysis or many-facet Rasch measurement to investigate the sources of variance and error in data as well as the erratic marking patterns on	Kane, 2006; Eckes, 2011; Aryadoust, 2013; Grabowski and Lin, 2019; Sawaki and Xi, 2019
Explanation (analogous to traditional construct validation)	Establishing whether the test engages the target construct or whether the test takers' performance can primarily be explained by the target construct	Latent variable analysis such as exploratory or confirmatory factor analysis or Rasch measurement	Chapelle et al., 2008
Extrapolation (analogous to traditional criterion evidence of validity)	Establishing whether the test scores can be extrapolated to or predict test takers' performance in the TLU domain	Correlation analysis, regression analysis, or structural equation modeling (SEM) to examine the relationships between test results and future performance of the test takers in the TLU domain	Kane, 2006; Bachman and Palmer, 2010
Utilization (analogous to traditional washback research or consequential validity)	Establishing whether the test results are used appropriately and whether their use has any positive impact on the individual, educational system, and society	Investigation of washback through collecting evidence from classrooms, work places, or test takers, using questionnaires or interviews and analysis methods such as SEM or regression analysis.	Bailey, 1999; Bachman and Palmer, 2010;

publications had a low betweenness centrality index ranging from 0.01 to 0.39. Bachman (1990; centrality = 0.35) and Canale and Swain (1980; centrality = 0.39) had the highest betweenness centrality index among the core and general journals datasets, respectively. Of these, Bachman (1990) and Skehan (1998) appeared on both core and general journals lists. The books identified in the analysis were not included directly in the datasets; they appeared in the results since they were co-cited by a significant number of citing papers (i.e., they came from the *References* section of the citing papers).

The top five most influential publications in the core journals were Bachman and Palmer (1996; duration of burst = 6, strength = 17.39, centrality = 0.11, sigma = 6.4), Bachman and Palmer (2010; duration of burst = 4, strength = 14.93, centrality = 0.02, sigma = 1.25), Bachman (1990; duration of burst = 5, strength = 11.77, centrality = 0.35, sigma = 32.79), Fulcher (2003; duration of burst = 5, strength = 11.54, centrality = 0.01, sigma = 1.10), and Council of Europe (2001; duration of burst = 3, strength = 11.17, centrality = 0.01, sigma = 1.11).

In addition, four publications in the general journals dataset had a burst strength higher than 11: Skehan (1988; duration of burst = 9, strength = 13.42, centrality = 0.05, sigma = 1.85), Bachman and Palmer (1996; duration of burst = 7, strength = 12.15, centrality = 0.05, sigma = 1.81), Norris and Ortega (2009; duration of burst = 7, strength = 13.75, centrality = 0.01, sigma = 1.08), and Nation (1990; duration of burst = 6, strength = 11.00, centrality = 0.05, sigma = 1.67).

Visualization of the DCA Network for the Core Journals Dataset

Figure 1 depicts the cluster view of the DCA network of the core journals. Each cluster consists of nodes, which represent publications, and their links which are represented by lines and show co-citation connections. The labels per clusters are representative of the headings assigned to the citing articles within the cluster. The color of a link denotes the earliest time slice in which the connection was made, with warm colors like red representing the most recent burst and cold colors like blue

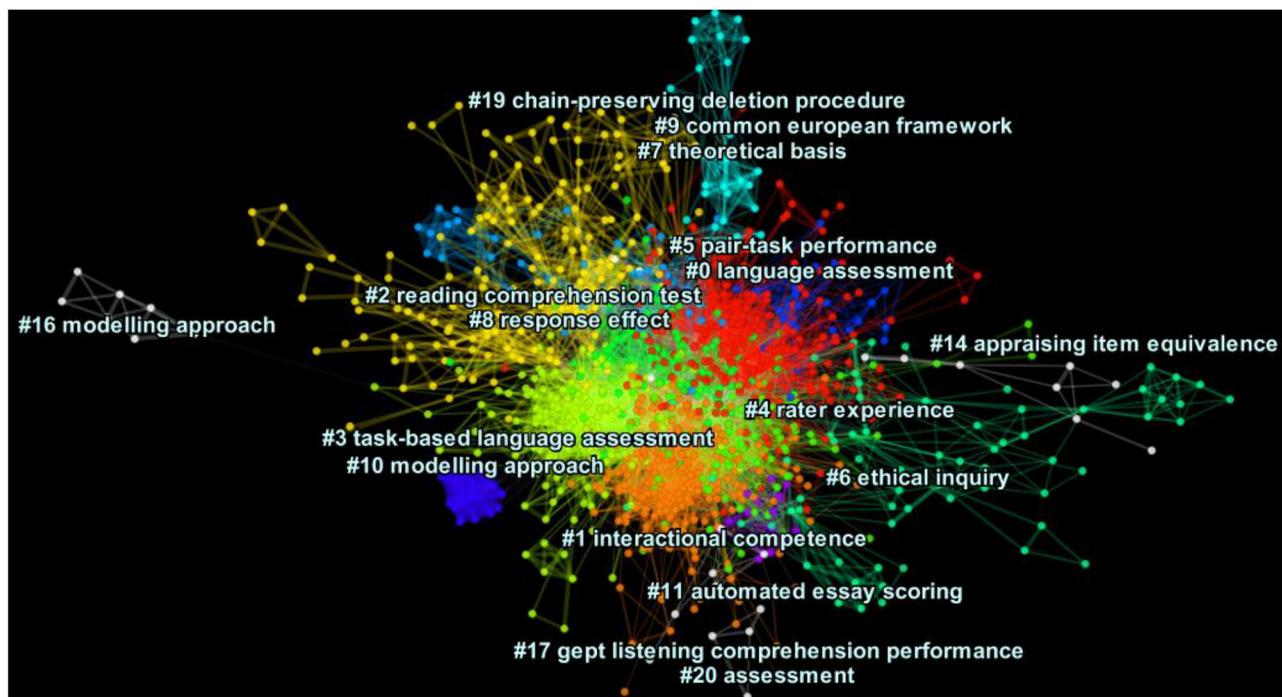


FIGURE 1 | The cluster view of network in the core journals dataset (modularity $Q = 0.541$, average silhouette score = 0.71), generated using CiteSpace, Version 5.6.R3.

representing older clusters. As we can see from the denseness of the nodes in **Figure 1**, there were six largest clusters experiencing citation bursts: #0 or language assessment (size=224; silhouette value = 0.538; Mean year of publication = 1995), #1 or interactional competence (size= 221; silhouette value = 0.544; Mean year of publication = 2005), #2 or reading comprehension test (size= 171; silhouette value = 0.838; Mean year of publication = 1981), #3 or task-based language assessment (size= 161; silhouette value = 0.753; Mean year of publication = 1994), #4 or rater experience (size=108; silhouette value = 0.752; Mean year of publication = 1999), and #5 or pair task performance (size = 78; silhouette value = 0.839; Mean year of publication = 1993). Note that the numbers assigned to the clusters in this figure (from 0 to 20) are based on the cluster size, so #0 is the largest, followed by #1, etc. Smaller clusters with too few connections are not presented in cluster views. This DCA network had a modularity Q metric of 0.541, indicating a fairly well-structured network. The average silhouette index was 0.71, suggesting medium homogeneity of the structures (See **Supplemental Table 3** for further information). It should be noted that after examining the content of each cluster, we made some revisions to the automatically generated labels to enhance their consistency and precision (see *Discussion*).

Visualization of the DCA Network for the General Journals

Figure 2 depicts a cluster view of the major clusters in the general journals dataset visualized along multiple horizontal lines

(modularity $Q = 0.6493$, average silhouette score = 0.787). The clusters are color-coded, with their nodes (publications) and links being represented by dots and straight lines, respectively. Among the clusters visually represented, there were nine major clusters in the network, as presented in **Supplemental Table 4**. The largest cluster is #2 (incidental vocabulary learning); the oldest cluster is #0 (foreign language aptitude), whereas the most recent one is #4 (syntactic complexity). As presented in the **Supplemental Table 4**, although the dataset represented co-citation patterns in the general journals, we noted that there were multiple cited publications in this dataset that were published in the core journals. It should be noted that only major clusters are labeled and displayed in **Figures 1, 2** and therefore the running order of the clusters are different across the two.

Second Aim: Measurement and Validity in the Core Journal Clusters

Next, we applied the analytical framework of the study in **Table 1** to examine the measurement and validation practices in each main cluster.

Domain Specification in Core Journals

For the core dataset, **Table 2** presents the domains and constructs specified in the six major clusters. (Please note that the labels under the “The construct or domain specified” column were inductively assigned by the authors based on the examination of papers in each cluster). Overall, there were fewer constructs/domains in the core dataset ($n = 15$) as compared

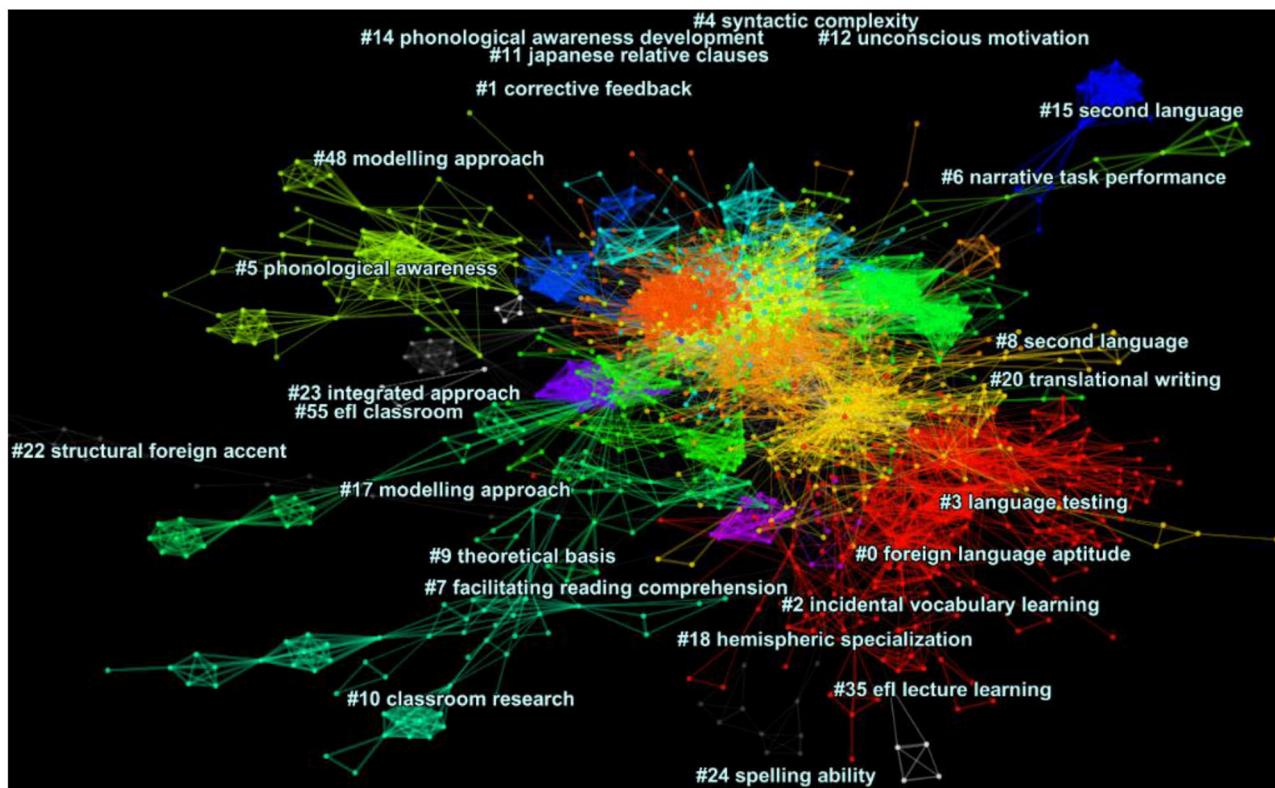


FIGURE 2 | The cluster view of network in the general journals dataset (modularity $Q = 0.6493$, average silhouette score = 0.787), generated using CiteSpace, Version 5.6.R3.

to the 26 in the general journals dataset below. The top four most frequently occurring constructs or domains in the core dataset were speaking/oral/communicative skills, writing and/or essays, reading, and raters/ratings. The most frequently occurring construct, Speaking/oral/communicative skills, appeared in every cluster, which is indicative of one of the major foci of the core journals. A series of χ^2 tests showed that all categories of constructs or domains were significantly different from each other in terms of the distribution of the skills and elements ($p < 0.05$). Specifically, Clusters #0 and #2 were primarily characterized by the dominance of comprehension (reading and listening) assessment research while Clusters #1, #4, and #5 had a heavier focus on performance assessment (writing and oral production/interactional competence), thus suggesting two possible streams of research weaving the clusters together. The assessment of language elements such as vocabulary and grammar was significantly less researched across all the clusters.

Other Components in Core Journals

Table 3 presents the other components of the analytical framework in the core journals consisting of construct operationalization, evaluation, generalization, explanation, extrapolation, and utilization. The domains and constructs were operationalized using (i) a discrete-point and selected response format comprising 61 assessments that used cloze, Likert scales,

and multiple-choice items, and (ii) production response format comprising 61 essays and writing assessments, and 59 oral production and interview. Specifically, the two most frequently occurring methods of construct operationalization were through cloze/ Likert/ multiple choice and essays and writing assessments in the major clusters of the core journals dataset.

In addition, reliability coefficients were reported in slightly more than half of the publications (56.7%), whereas generalizability was underreported in all the clusters with a mere 7.1% of the studies presenting evidence of generalizability. Likewise, only 7.5% presented criterion-based evidence of validity; 10.8% of the studies reported or investigated evidence supporting construct validity or the explanation inference; and 5% (12/240) of the studies addressed the utilization inference of the language assessments investigated. Among the clusters, Cluster #5 and #0 had the highest respective ratios of 4/19 (21%) and 6/59 (10%) studies investigating the utilization inference.

Measurement and Validity in the General Journal Clusters

Domain Specification in General Journals

Table 4 presents the domains and constructs specified in the major clusters in the general journals dataset. Of the 26 constructs/domains specified in the nine clusters, the top five constructs/domains in the clusters were grammar, speaking/

TABLE 2 | Domain specification in major clusters in the core journals.

Cluster #	The construct or domain specified	# of papers
Cluster 0		
	Reading	18
	Listening	8
	Speaking/ oral/ communicative ability	8
	Writing	5
	Overall language proficiency	7
Cluster 1		
	Reading	8
	Writing	29
	Speaking/ oral/ communicative ability	16
	Interactional competence	6
	Corpus linguistics	3
	Overall language proficiency	9
	Feedback	3
Cluster 2		
	Reading	6
	Listening	2
	Speaking/ oral/ communicative ability	3
Cluster 3		
	Reading	3
	Vocabulary	7
	Speaking/ oral/ communicative	5
	Overall language proficiency	2
Cluster 4		
	Vocabulary	3
	Writing/ essays	15
	Raters/ ratings	18
	Speaking/ oral/ communicative ability	8
Cluster 5		
	Speaking/ oral/ communicative ability	13
	Washback	2

oral interactions, reading, vocabulary, and writing (ranked by frequency of occurrence in the clusters). Grammar appeared in every cluster except Cluster 8 which was distinct from other clusters as papers in this cluster did not examine linguistic constructs but the affective aspects of language learning, with a relatively low number of publications ($n = 13$). Looking at the number of papers for each respective domain in each cluster, we can observe that some clusters were characterized by certain domains. By frequency of occurrence, papers in Cluster 0 was mostly concerned with language comprehension (reading and listening), whereas Cluster 1 was characterized by feedback on written and oral production; Cluster 2 by vocabulary; and Cluster 4 by writing, with syntactic complexity being secondary in importance. A series of χ^2 tests showed that 20 of the 26 categories of construct or domains occurred with significantly unequal probabilities, i.e., fluency, speaking, oral ability/proficiency, language proficiency/competence, feedback, collocations, semantic awareness, syntactic complexity, task complexity, phonological awareness, explicit/ implicit

TABLE 3 | Measurement methods and evidence of validity in major clusters in the core journals.

Construct operationalization				
Cluster ID	Cluster ID	Cloze/ Likert/ multiple choice	Essays and writing	Oral/interview
1	1	10	32	21
4	4	17	17	9
0	0	20	5	13
5	5	4	0	11
Reliability				
Cluster ID	Cluster ID	Reported reliability	Did not report reliability	Total
1	1	49	36	85
0	0	30	29	59
Generalization				
Cluster ID	Cluster ID	Reported generalizability evidence	Did not report generalizability evidence	Total
1	1	6	79	85
0	0	1	58	59
Criterion Evidence of Validity				
Cluster ID	Yes	No		Total
1	5	80		85
0	5	54		59
4	1	29		30
3	2	24		26
2	5	16		21
5	0	19		19
Utilization				
Cluster ID	Yes	No		Total
1	1	82		85
0	6	50		59
4	0	27		30
3	0	24		26
2	1	20		21
5	4	14		19
Explanation				
Cluster ID	Yes	No		Total
1	10	75		85
0	8	51		59
4	3	27		30
3	0	26		26
2	3	18		21
5	2	17		19

TABLE 4 | Domain specification in major clusters in the general journals.

Cluster #	The construct or domain specified	# of papers
Cluster 0		
	Reading	12
	Listening	10
	Speaking	6
	Writing	4
	Grammar	5
	Vocabulary	5
	Oral ability	1
	Oral proficiency	1
	Language proficiency	3
	Language competence	1
Cluster 1		
	Reading	1
	Listening	1
	Speaking/ Oral/ Interaction	15
	Writing	3
	Grammar	6
	Vocabulary	1
	Memory	4
	Feedback*	15
Cluster 2		
	Reading	9
	Listening	9
	Speaking/ Oral/ Interaction	1
	Writing	5
	Grammar	1
	Vocabulary	43
	Collocations	5
	Semantic awareness	2
Cluster 3		
	Reading	2
	Listening	1
	Speaking/ Oral/ Interaction	5
	Writing	3
	Grammar	2
	Vocabulary	3
Cluster 4		
	Speaking/ Oral/ Interaction	5
	Writing	21
	Grammar	3
	Vocabulary	1
	Fluency	5
	Syntactic complexity	7
	Task complexity	2
Cluster 5		
	Reading	2
	Speaking/ Oral/ Interaction	2
	Grammar	1
	Vocabulary	3
	Phonological awareness	3

(Continued)

TABLE 4 | Continued

Cluster #	The construct or domain specified	# of papers
Cluster 6		
	Reading	1
	Speaking/ Oral/ Interaction	1
	Grammar	1
	Fluency	2
	Explicit/ implicit knowledge	3
	Listening comprehension	2
Cluster 8		
	Anxiety	4
	Attitudes	3
	Motivation	6
Cluster 11		
	Grammar	2
	Relative clauses	3
	Language awareness	2

*Papers on feedback were double-counted in other categories. This consisted of 10 papers on speaking/oral/interaction, 1 paper on grammar, 1 on explicit feedback, 1 on the use of classifiers and the perfective -le in Chinese, and 2 papers on writing.

knowledge, comprehension, anxiety, attitudes, motivation, relative clauses, and language awareness ($p < 0.005$).

Other Components in General Journals

Table 5 presents the breakdown of construct operationalization and the presentation of evidence of validity in the papers in the major clusters of the general journals data set. Given the domain characteristics (writing) of Cluster 4, discussed above, it is not surprising that the constructs are operationalized mainly through writing/essay in 59.6% of the papers in the cluster. As with the core journals dataset, the evaluation of reliability in the papers is fairly split, with 54.63% of the publications reporting reliability. The vast majority of papers did not provide any generalizability evidence (98.83%). Likewise, the majority of papers did not investigate construct validity (extrapolation) (95.03%) nor did they provide criterion evidence of validity (93.27%). Finally, only 24 of the publications reported or investigated the utilization inference.

DISCUSSION

This study set out to investigate intellectual domains as well as the use of measurement and validation methods in language assessment research. We created two datasets covering the core and general journals, and employed DCA to detect research clusters. Next, we coded citing papers in each cluster based on an analytical framework for measurement and validation (Norris and Ortega, 2003; Kane, 2006; Bachman and Palmer, 2010). In this section, we will discuss bursts and citing publications per cluster to determine the features that possibly characterize each main clusters. Next, we will discuss the measurement and validation practices in the citing papers in the two datasets.

TABLE 5 | Measurement practices and evidence of validity in major clusters in the general journals.

Cluster ID	Cloze/Likert/ multiple choice	Essay/writing	Oral/interview	Total
Construct operationalization				
2	29	13	6	48
1	3	16	21	40
3	10	7	12	29
0	20	8	8	36
4	3	28	16	47
6	6	2	6	14
8	5	0	1	6
5	2	0	6	8
11	3	4	4	11
Cluster ID	Reported reliability	Did not report reliability	Non-English	Total
Reliability				
2	44	40	0	84
1	34	32	0	66
3	21	20	0	41
0	25	13	0	38
4	27	22	0	49
6	16	8	0	24
8	5	6	1	12
5	12	3	0	15
11	3	9	1	13
Cluster ID	Reported generalizability evidence	Did not report generalizability evidence	Non-English	Total
Generalization				
2	1	83	0	84
1	0	66	0	66
3	1	40	0	41
0	0	38	0	38
4	0	49	0	49
6	0	24	0	24
8	0	11	1	12
5	0	15	0	15
11	0	12	1	13
Cluster ID	Yes	No	non-English	Total
Criterion Evidence of Validity				
2	3	81	0	84
1	4	62	0	66
3	5	36	0	41
0	6	32	0	38
4	1	48	0	49
6	0	24	0	24
8	0	11	1	12
5	2	13	0	15
11	0	12	1	13

(Continued)

TABLE 5 | Continued

Cluster ID	Yes	No	Non-English	Total
Explanation				
2	2	82	0	84
1	4	62	0	66
3	4	37	0	41
0	6	32	0	38
4	1	48	0	49
6	0	24	0	24
8	0	12	0	12
5	0	15	0	15
11	0	13	0	13
Cluster ID	Yes	No	Claimed without evidence	Total
Utilization				
2	0	82	2	84
1	0	63	3	66
3	0	29	12	41
0	1	30	7	38
4	0	49	0	49
6	0	24	0	24
8	0	11	0	12
5	0	15	0	15
11	0	12	0	13

First Aim: Characterizing the Detected Clusters

Core Journals

Bursts (impactful cited publications) in the influential clusters in the core journals dataset are presented in **Table 6**. The review presented in the following sections is organized according to the content and relevance of these publications. We will further provide a broad overview of these publications. It should be noted that while narrative literature reviews customarily have specific foci, what we aim to do is to leverage the potentiality of clustering and highlight the linked concepts that might have resulted in the emergence of each cluster. Each cluster will be characterized by virtue of the content of the citing and cited publications. Due to space constraints, we provide a detailed review commentary on two of the largest clusters in the Core Journals dataset, and a general overview of the rest of the major clusters (see the Appendices for further information per cluster).

Cluster 0: Language assessment (and comprehension)

As demonstrated in **Table 7**, bursts in this cluster can roughly be divided into two major groups: (i) generic textbooks or publications that present frameworks for the development of language assessments in general (e.g., Bachman, 1990; Alderson et al., 1995; Bachman and Palmer, 1996, 2010; McNamara, 1996; Shohamy, 2001; Alderson, 2005), or of specific aspects in the

TABLE 6 | Selected cited publications (Bursts) in the core journals.

References	Burst strength	Frequency	Centrality	Sigma	Cluster ID
Bachman and Palmer (1996)	17.39	63	0.11	6.4	0
Alderson et al. (1995)	10.65	28	0.02	1.19	0
Bachman (1990)	9.58	67	0.16	4.13	0
Alderson (2000)	8.55	26	0.01	1.07	0
Bachman and Palmer (2010)	7.97	18	0.01	1.06	0
Shohamy (2001)	7.84	22	0.01	1.1	0
Alderson (2005)	7.7	22	0.02	1.13	0
McNamara (1996)	7.22	22	0.02	1.14	0
Buck (2001)	6.86	18	0	1.02	0
Bond and Fox (2007)	6.55	12	0	1.02	0
Bachman (2005)	5.99	32	0.03	1.17	0
Read (2000)	5.64	13	0	1.01	0
Taylor (2009)	5.33	10	0	1.02	0
Alderson and Hamp-Lyons (1996)	4.7	12	0.01	1.05	0
Douglas (2000)	4.47	8	0	1.01	0
Fulcher (2004)	4.16	11	0.01	1.03	0
Canale and Swain (1980)	4.13	49	0.22	2.29	0
Brennan (2001)	4.06	10	0	1.01	0
Alderson and Lukmani (1989)	3.75	15	0.02	1.07	0
Kobayashi (2002)	3.68	7	0	1.02	0
Davison (2007)	3.64	6	0	1.01	0
Brindley (2001)	3.62	6	0	1.01	0
Fulcher (2003)	11.55	27	0.01	1.1	1
Council of Europe (2001)	11.17	23	0.01	1.11	1
American Educational Research Association (2014)	9.17	19	0.01	1.05	1
Weigle (2002)	9.05	60	0.05	1.6	1
Knoch (2009)	7.77	21	0.01	1.08	1
Kane (2006)	7.3	30	0.03	1.24	1
Weir (2005a)	6.82	16	0.01	1.04	1
Luoma (2004)	6.74	14	0	1.02	1
Guo et al. (2013)	6.29	13	0	1.01	1
Messick (1989)	6.17	81	0.12	2.03	1
Cohen (1988)	5.99	19	0.01	1.07	1
Fulcher et al. (2011)	5.8	10	0	1.02	1
Kane (2013)	5.54	15	0.01	1.04	1
Chapelle et al. (2008)	5.1	12	0	1.02	1
Cumming (2013)	4.81	10	0	1.02	1
Biber and Gray (2013)	4.67	11	0	1.01	1
Iwashita et al. (2008)	4.44	17	0.01	1.05	1
Gebril (2009)	4.33	15	0	1.02	1
Flower and Hayes (1981)	4.32	8	0	1.01	1
McNamara et al. (2014)	4.32	8	0	1.01	1
May (2011)	4.26	10	0	1.01	1
Deane (2013)	4.07	14	0.01	1.03	1
Jacobs (1981)	3.98	7	0	1.02	1
Fulcher (1996)	3.81	15	0.01	1.03	1
Ortega (2003)	3.78	7	0	1	1
Plakans (2008)	3.69	11	0	1.02	1
Knoch (2011)	3.69	10	0.01	1.03	1
Wright and Stone (1979)	8.1	17	0.05	1.48	2

(Continued)

TABLE 6 | Continued

References	Burst strength	Frequency	Centrality	Sigma	Cluster ID
Henning (1987)	6.09	13	0.02	1.14	2
Oller (1979)	5.29	9	0.04	1.25	2
Rasch (1960)	5.25	8	0.01	1.05	2
Hambleton and Swaminathan (1985)	4.91	8	0.01	1.06	2
Hughes (1989)	4.55	7	0.01	1.05	2
McNamara (1990)	4.21	8	0.01	1.03	2
Chen and Henning (1985)	4.02	8	0.03	1.14	2
Skehan (1998)	7.9	16	0.01	1.1	3
Messick (1989)	7.18	12	0.01	1.05	3
Brindley (1998)	5.52	12	0.04	1.22	3
Clapham (1996)	4.8	8	0.01	1.03	3
Messick (1994)	4.58	12	0.03	1.12	3
Brown and Hudson (1998)	3.89	6	0.01	1.02	3
Bachman (1990)	3.73	6	0	1	3
Alderson and Wall (1993)	3.61	19	0.01	1.05	3
Cumming et al. (2002)	8.48	26	0.01	1.1	4
Lumley (2002)	7.94	43	0.04	1.32	4
Cumming (1990)	6.72	28	0.01	1.09	4
Eckes (2008)	6.05	24	0.01	1.06	4
Lumley and McNamara (1995)	5.27	26	0.01	1.07	4
Weigle (1998)	4.54	36	0.03	1.14	4
Weigle (1994)	4.49	17	0.01	1.04	4
Brown (1995)	4.26	22	0.04	1.17	4
Lim (2011)	4.06	7	0	1	4
Barkaoui (2010)	3.83	9	0	1	4
(Hamp-Lyons, 1991)	3.81	13	0.01	1.04	4
Brown (2003)	6.65	28	0.02	1.15	5
van Lier (1989)	4.81	13	0.02	1.08	5
Lazaraton (1996)	4.59	14	0.01	1.05	5
Messick (1996)	4.15	33	0.03	1.14	5
Chalhoub-Deville (2003)	3.95	17	0.01	1.04	5
Shohamy (1988)	3.88	6	0.01	1.03	5

development of language assessments (Alderson, 2000; Read, 2000; Brennan, 2001; Buck, 2001; Kobayashi, 2002; Bachman, 2005) and psychometric measurement (McNamara, 1996; Bond and Fox, 2007), and (ii) publications that describe the contexts and implementations of tests (Alderson and Hamp-Lyons, 1996; Fulcher, 2004; Davison, 2007; Taylor, 2009). The citing publications in this cluster, on the other hand, consist of papers that chiefly investigate the assessment of comprehension skills (The labels under *Focus area 1* and *Focus area 2* in Tables 7, 8 and Supplemental Tables 5 through 11 were inductively assigned by the authors based on the examination of papers).

Among the bursts in the first group, a few publications prove to be the pillars of the field: Alderson et al. (1995), Bachman (1990), and Bachman and Palmer (1996, 2010). This can be seen from the burst strength of these publications (Table 6) as well as from the citing publications. The articles that cite the publications in Cluster 0 span from reviews or editorials that provide an overview of the field of language assessment

to looking at aspects of language assessment. Reviews of the field of language assessment (e.g., Harsch, 2014; McNamara, 2014) consistently mention the works of Bachman. Bachman's influence is such that his publications merited mention even when reviewing specific areas in the field as in Phakiti and Roever (2011) on regional issues in Australia and New Zealand, Xi (2010b) on scoring and feedback, and Lee and Sawaki (2009) on cognitive diagnostic assessment. Bachman and Palmer (1996, 2010) have wide appeal and are referenced with respect to a wide range of topics like reading (Carr, 2006; Zhang et al., 2014), listening (Papageorgiou et al., 2012), and pragmalinguistics (Roever, 2006) in Cluster 0. Bachman and Palmer (1996) and Bachman (1990) are also frequent sources for definitions, e.g., of which are too numerous to recount exhaustively. Two examples are that of reliability in Winke (2011) and of practicality in Roever (2006), which show the influence of these two texts in explicating core concepts of language assessment.

TABLE 7 | Major citing and cited publications in clusters 0 in the core journals.

Cluster	References	Citing	Cited (bursts)	Focus area 1	Focus area 2
0	(Bachman and Palmer, 1996)		X	Test usefulness	Test development
0	(Alderson et al., 1995)		X	Test specification	Test development
0	(Bachman, 1990)		X	Test development	Test methods facets
0	(Alderson, 2000)		X	Test development (reading)	-
0	(Bachman and Palmer, 2010)		X	Validation	Test development
0	(Shohamy, 2001)		X	Tests and policy-making	Democratic assessment
0	(Alderson, 2005)		X	Test development (diagnostic assessment)	The DIALANG assessment system
0	(McNamara, 1996)		X	Test development	Psychometric measurement
0	(Buck, 2001)		X	Test development (listening)	Theories of listening
0	(Bond and Fox, 2007)		X	Rasch measurement	-
0	Bachman (2005)		X	Validation	-
0	(Read, 2000)		X	Test development (Vocabulary)	Theories of vocabulary acquisition and assessment
0	(Taylor, 2009)		X	Language assessment literacy	Test wiseness
0	(Alderson and Hamp-Lyons, 1996)		X	Washback	The TOEFL
0	(Douglas, 2000)	X	X	Assessment of language for specific purposes	-
0	(Fulcher, 2004)		X	The Common European Framework of Reference	Language assessment (political dimensions)
0	(Canale and Swain, 1980)		X	Communicative competence framework	-
0	Brennan (2001)		X	Generalizability theory	-
0	(Kobayashi, 2002)		X	Test method effect	-
0	(Davison, 2007)		X	Hong Kong Examinations and Assessment Authority (HKEAA) School Based Assessment	Perceptions toward school-based assessments
0	(Harsch, 2014)	X		Review of General Language Proficiency	-
0	(McNamara, 2014)	X		Review of Communicative Language Testing (Editorial)	CEF
0	(Phakiti and Roever, 2011)	X		Review of Language Assessment in Australia and New Zealand (Editorial)	-
0	(Xi, 2010b)	X		Review of Automated scoring and feedback systems (Editorial)	-
0	(Lee and Sawaki, 2009)	X		Review of cognitive diagnostic assessment	-
0	(Carr, 2006)	X		Reading comprehension	Test task characteristics
0	(Zhang et al., 2014)	X		Reading comprehension	-
0	(Papageorgiou et al., 2012)	X		Listening comprehension	Test task characteristics (Dialogic vs. monologic assessment)
0	(Roever, 2006)	X		Pragmalinguistics	Validity
0	(Winke, 2011)	X		U.S. Naturalization Test	Reliability
0	Gao and Rogers (2011)	X		Reading comprehension	Test task characteristics
0	(Green and Weir, 2010)	X		Reading comprehension (textual features)	Validity
0	(Jang, 2009a)	X		Reading comprehension	Cognitive diagnostic assessment
0	(Jang, 2009b)	X		Reading comprehension	Cognitive diagnostic assessment
0	(Sawaki et al., 2009)	X		Reading and listening comprehension	Cognitive diagnostic assessment
0	(Harding et al., 2015)	X		Reading and listening comprehension	Diagnostic assessment
0	(Eckes and Grotjahn, 2006)	X		(German) General Language Proficiency (reading, listening, writing, speaking)	Validity

TABLE 8 | Major citing and cited publications in clusters 1 in the core journals.

Cluster	References	Citing	Cited (bursts)	Focus area 1	Focus area 2
1	(Fulcher, 2003)		X	Speaking	
1	(Council of Europe, 2001)		X	Assessment	
1	American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014		X	Assessment	Validation
1	(Weigle, 2002)		X	Writing	
1	(Knoch, 2009)		X	Rating scales	Writing
1	(Kane, 2006)		X	Validation	
1	(Weir, 2005a)		X	Validation	
1	(Luoma, 2004)		X	Speaking assessment	
1	(Guo et al., 2013)		X	Linguistic features and rating	Coh-Metrix
1	(Messick, 1989)		X	Validation	
1	(Fulcher et al., 2011)		X	Rating scales	Speaking
1	(Kane, 2013)		X	Validation	
1	(Chapelle et al., 2008)		X	Validation	
1	(Cumming, 2013)		X	Review of Integrated Writing Tasks	
1	(Iwashita et al., 2008)		X	Rating scales	Speaking
1	(Gebril, 2009)		X	Integrated Writing Tasks	
1	(Flower and Hayes, 1981)		X	Writing process	
1	(McNamara et al., 2014)		X	Coh-Metrix	Linguistic features
1	(May, 2011)		X	Rating scales	Speaking
1	(Deane, 2013)		X	Automated scoring	Writing
1	(Jacobs, 1981)		X		
1	(Fulcher, 1996)		X	Rating scales	Speaking
1	(Ortega, 2003)		X	Review of syntactic complexity	
1	(Plakans, 2008)		X	Integrated Writing Tasks	
1	(Knoch, 2011)		X	Rating scales	Writing
1	(Plakans et al., 2019)	X		Integrated writing tasks (reading-writing)	Process
1	(Plakans and Gebril, 2017)	X		Integrated (reading-listening-writing) tasks	The TOEFL iBT
1	(Banerjee et al., 2015)	X		Writing assessment	Rating scale
1	(Barkaoui and Knouzi, 2018)	X		Writing assessment	Mode effect
1	(Guo et al., 2013)	X	X	Writing assessment	Linguistic features
1	(Isbell, 2017)	X		Writing assessment	Rating
1	(Lallmamode et al., 2016)	X		Writing assessment	Validation of scoring rubric
1	(Lu, 2017)	X		Writing assessment	Syntactic Complexity
1	(Rakedzon and Baram-Tsabari, 2017)	X		Writing assessment	Scoring rubric
1	(Wilson et al., 2017)	X		Writing assessment	Automated scoring (using linguistic features measures)
1	(Zhao, 2017)	X		Writing assessment	Scoring rubric (Voice)
1	(Zheng and Yu, 2019)	X		Writing assessment	Review of writing assessment
1	(Lam, 2018)	X		Speaking assessment	Interactional competence
1	(van Batenburg et al., 2018)	X		Speaking assessment	Interactional competence
1	(Römer, 2017)	X		Speaking assessment	Lexicogrammar

Articles on the assessment of reading comprehension (e.g., Jang, 2009a,b; Sawaki et al., 2009; Green and Weir, 2010; Gao and Rogers, 2011; Harding et al., 2015) often reference Charles Alderson: Alderson (2000), Alderson (2005) and to a lesser extent, Alderson et al. (1995) and Alderson and Lukmani (1989). For example, Jang's (2009a,b) studies on reading comprehension investigated the validity of LanguEdge test materials and the notion of reading subskills using cognitive diagnosis assessment. Prior discussions on the various aspects of reading assessment—like subskills—in Alderson's various works feature strongly in such studies (see also Sawaki et al., 2009). An exception is Carr (2006) study on reading comprehension. While mentioning Alderson (2000), Bachman and Palmer (1996) task characteristics model undergirds Carr (2006) investigation on the relationship between test task characteristics and test taker performance.

Just like Alderson's works for reading, Buck (2001) seems to be the definitive textbook on assessing the listening component of language. For example, in influential citing papers such as Harding et al. (2015), Papageorgiou et al. (2012), as well as Sawaki et al. (2009), Buck's conceptualization of the subskills involved in listening is discussed.

Similarly, McNamara (1996) is a sourcebook on the development and validation of performance tests. McNamara (1996) introduced many-facet Rasch measurement (Linacre, 1994) as a useful method to capture the effect of external facets—most notably rater effects—on the measured performance of test takers. Relatedly, Bond and Fox (2007) guide readers through the general principles of the Rasch model and the various ways of applying it in their textbook. The importance of the Rasch model for test validation makes this accessible text oft-cited in studies concerned with test validity (e.g., Eckes and Grotjahn, 2006; Winke, 2011; Papageorgiou et al., 2012).

Another group of bursts in the cluster describe the then-current contexts of language assessment literacy (Taylor, 2009), frameworks (Fulcher, 2004), language tests after implementation (Alderson and Hamp-Lyons, 1996; Davison, 2007), and language for specific purposes (LSP, Douglas, 2000). In a call for the development of "assessment literacy" (Taylor, 2009) among applied linguists, Taylor described the state of the field of language assessment at that moment, looking at the types of practical knowledge needed and the scholarly work that offer them. This need for "assessment literacy" (Taylor, 2009) when implementing tests was already highlighted by Alderson and Hamp-Lyons (1996) some years before. Emphasizing the need to move beyond assumptions when hypothesizing about washback, Alderson and Hamp-Lyons (1996) observed and compared TOEFL and non-TOEFL classes taught by the same teachers in order to establish the presence of the oft-assumed washback effect of the TOEFL language tests. Davison (2007) takes a similar tack in looking at teachers' perception of the challenges in adapting to Hong Kong's shift to school-based assessment (SBA) of oral language skills. Although Davison (2007) and Alderson and Hamp-Lyons (1996) describe different tests, both sources highlight the importance of moving beyond theory and looking at implementation. That test development does not end at implementation is similarly highlighted by Fulcher (2004), who tackles the larger contexts surrounding the Common European

Framework (CEF) in his critical historical overview of the development of said framework. Finally, Doughty (2001) work on the assessment of LSP has become a major sourcebook in the field. Douglas's model of LSP ability drew inspiration from the communicative competence model of Canale and Swain (1980) and comprised language knowledge, strategic competence, and background knowledge.

Cluster 1: Rating (and Validation)

Moving from the global outlook on language assessment that largely characterizes Cluster 0, Cluster 1 narrows down on two related aspects of language testing: validation and rating. The unitary concept of validity (Messick, 1989), the socio-cognitive validity framework (Weir, 2005a), and the argument-based approach to validation (Kane, 2006, 2013) are the three main frameworks of validity featured in Cluster 1. The second major line of research in Cluster 1 is focused on improving rating scales. Fulcher (1996) proposed a data-driven approach to writing rating scales, coding transcripts from the ELTS oral examination to pinpoint "observed interruptions in fluency" (Fulcher, 1996, p. 216) present in candidates' speech. Using discriminant analysis, Fulcher (1996) linked linguistic descriptions to speaker performance, and at the same time, validating the rating scale produced. Iwashita et al. (2008) took a similar approach but expanded the range of measures beyond fluency with a more comprehensive set: grammatical accuracy and complexity, vocabulary, pronunciation, and fluency. Along the same idea, Fulcher et al. (2011) criticized the low richness of the descriptions generated from the measurement-driven approach and proposed Performance Decision Trees (PDTs), which are based on a non-linear scoring system that comprises yes/no decisions. In contrast, May (2011) took a different approach, using raters' perspectives to determine how raters would operationalize a rating scale and what features are salient to raters. Unlike the previous studies, however, the rating scale in May (2011) was for the paired speaking test. Mirroring the concerns about rating descriptors of speaking tasks, Knoch (2009) compared a new scale with more detailed, empirically developed descriptors with a pre-existing scale with less specific descriptors. Raters using the former scale reported higher rater reliability and better candidate discrimination. In a separate study, Knoch (2011) explained the features of diagnostic assessments of writing, stressing the uses and interpretations of rating scales.

With regards to the citing publications, papers describing the development of rating or scoring scales often cited the above publications, irrespective of what task the scale is for, resulting in the emergence of Cluster 1. For example, Banerjee et al. (2015) article focused the rating scale of writing assessment but discussed Fulcher (2003) and Fulcher et al. (2011). In addition, it is noted that rating scales are exclusively discussed with reference to the assessment of writing and speaking, with integrated tasks forming the nexus between these strands. Fulcher (2003) is the major publication of the speaking component of language assessment in this cluster, cited in studies focusing on speaking (Römer, 2017; van Batenburg et al., 2018) as well as meriting mention in studies on other topics like writing (Banerjee et al., 2015; Lallamamode et al., 2016). Akin to Fulcher

(2003) for speaking, Weigle (2002) is a reference text on the subject of writing. It is cited in studies with a range of topics like integrated tasks (Plakans, 2008; Gebril, 2009; Plakans and Gebril, 2017), rubrics (Banerjee et al., 2015), validation (Lallmamode et al., 2016) and linguistic features of writing (Guo et al., 2013; Lu, 2017). Other citing papers focusing on writing assessment were Isbell (2017), Zhao (2017), Lam (2018), and Zheng and Yu (2019).

Measures of linguistic features in rater-mediated assessments have a significant importance in the cluster. Ortega (2003) research synthesis quantified the effect size of syntactic complexity on assessed proficiency levels. More sophisticated ways of quantifying linguistic features have emerged since. A notable example is Coh-Metrix, a computational linguistic engine used to measure lexical sophistication, syntactic complexity, cohesion, and basic text information (Guo et al., 2013). McNamara et al. (2014) discussed the theoretical and practical implications of Coh-Metrix and provided an in-depth discussion of the textual features that Coh-Metrix measures. In a review article on syntactic complexity, Lu (2017) highlighted the increasing popularity of this tool. Coh-Metrix is used to operationalize and quantify linguistic and discourse features in writing, so as to predict scores (Banerjee et al., 2015; Wilson et al., 2017), test mode effect (Barkaoui and Knouzi, 2018).

Cluster 2: Test development (and dimensionality)

Cluster 2 is characterized by test development and dimensionality (see **Supplemental Table 5**). Publications in this cluster center around the development of tests (for teaching) (e.g., Oller, 1979; Henning, 1987; Hughes, 1989) and the implications of test scores, like Chen and Henning (1985), one of the initial works on bias. As well, a large part of the language test development process outlined in these publications include the interpretation and validation of test scores through item response theory (IRT) and Rasch models (Wright and Stone, 1979; Hambleton and Swaminathan, 1985; Henning, 1987). Rasch's (1960) pioneering monograph is the pillar upon which these publications stand. Citing articles are largely concerned with dimensionality (Lynch et al., 1988; McNamara, 1991) and validity (Lumley, 1993). From the publication dates, Cluster 2 seems reflective of prevailing concerns in the field specific to the 1980s and early 1990s.

Cluster 4: Rater Performance

As demonstrated in **Supplemental Table 6**, Cluster 4 concerns rating, which links it to Cluster 1. Chief concerns on variability in rating include raters' characteristics (Brown, 1995; Eckes, 2008), experience (Cumming, 1990; Lim, 2011) and biases (Lumley and McNamara, 1995) that affect rating performance, the effect of training (Weigle, 1994, 1998) and the processes by which the raters undergo while rating (Cumming et al., 2002; Lumley, 2002; Barkaoui, 2010). Citing articles largely mirror the same concerns (rater characteristics: Zhang and Elder, 2010; rater experience: Kim, 2015; rater training: Knoch et al., 2007; rating process: Wiseman, 2012; Winke and Lim, 2015), making this cluster a tightly focused one.

Cluster 5: Spoken Interaction

Cluster 5 looks at a specific aspect of assessing speaking: spoken interaction. Unlike Cluster 1 which also had a focus on assessing speaking, this cluster centers on a different group of bursts, thus its segregation: Brown (2003), Lazaraton (1996), Shohamy (1988), van Lier (1989) who explored the variation in the interactions between different candidates and testers during interviews. The social aspect of speaking calls into question validity and reliability in a strict sense, with implications for models of communicative ability, as Chalhoub-Deville (2003) highlighted. These developments in language assessment meant citing articles move beyond interviews to pair-tasks (O'Sullivan, 2002; Brooks, 2009; Davis, 2009), while maintaining similar concerns about reliability and validity (see **Supplemental Table 7** for further information).

Clusters in the General Journals Dataset

Table 9 demonstrates bursts in the influential clusters in the general journals dataset. The main clusters are discussed below.

Cluster 0: Test development (and dimensionality)

Cluster 0 in the General journals dataset overlapped in large part with Cluster 2 of the Core journals. Publications in Cluster 0 described the processes of test development (Oller, 1979; Wright and Stone, 1979; Henning, 1987; Hughes, 1989; Bachman, 1990). As with Cluster 2 (Core), there is a subfocus on IRT and Rasch models (Rasch, 1960; Wright and Stone, 1979; Hambleton and Swaminathan, 1985; Henning, 1987). Bachman (1990), Bachman and Palmer (1982), and Halliday and Hasan (1976) feature in this cluster but not in Cluster 2 (Core). There is a similar overlap in terms of the citing literature: 42% of the citing literature of the cluster overlaps with the citing literature of the Cluster 2 (Core), with little differences in central concerns of the articles (see **Supplemental Table 8** for further information).

Cluster 1: Language Acquisition (Implicit vs. explicit)

Cluster 1 of the General journals dataset is a rather large cluster, which reflects the vastness of research into SLA. Long (2007)'s book is one such attempt to elucidate on decades of theories and research. Other publications looked at specific theories like the output hypothesis (Swain, 1995), communicative competence (Swain, 1985) and the cognitive processes in language learning (Schmidt, 1994, 2001; Miyake and Friedman, 1998; Doughty, 2001). A recurrent theme in the theories of SLA is the dividing line between implicit and explicit language knowledge, as Ellis N. (2005) summarized. Research in the cluster similarly tackle the implicit and explicit divide in instruction (Ellis N., 2005; Erlam, 2005; Spada and Tomita, 2010). A subset of this is related to corrective feedback, where implicit feedback is often compared with explicit feedback (e.g., Ammar and Spada, 2006; Ellis et al., 2006). Along the same lines, Gutiérrez (2013) questions the validity of using grammaticality judgement tests to measure implicit and explicit knowledge (see **Supplemental Table 9** for further information).

TABLE 9 | Selected cited publications (Bursts) in the general journals dataset.

References	Burst strength	Frequency	Centrality	Sigma	Cluster ID
Bachman (1990)	11.13	37	0.11	3.06	0
Oller (1979)	8.36	15	0.06	1.61	0
Henning (1987)	7.86	13	0.01	1.1	0
Wright and Stone (1979)	7.7	13	0.02	1.15	0
Halliday and Hasan (1976)	7.01	15	0.05	1.41	0
Hughes (1989)	5.7	9	0	1.03	0
Rasch (1960)	5.22	8	0.01	1.05	0
Chen and Henning (1985)	5.2	9	0.02	1.13	0
Bachman and Palmer (1982)	5.19	8	0.02	1.08	0
Hambleton and Swaminathan (1985)	4.78	8	0	1.01	0
Cohen (1988)	10.67	63	0.04	1.45	1
Swain (1995)	10.61	56	0.03	1.43	1
Ellis N. (2005)	10.3	56	0.03	1.33	1
Spada and Tomita (2010)	8.7	25	0.01	1.06	1
Pica (1994)	8.3	18	0.01	1.1	1
Lyster and Saito (2010)	8	20	0	1.03	1
Lyster and Ranta (1997)	7.48	38	0.02	1.18	1
Schmidt (1994)	7.2	18	0.01	1.08	1
Swain (1985)	7.08	42	0.03	1.2	1
Long (2007)	6.73	13	0	1.01	1
Goo (2012)	6.72	13	0	1.02	1
Harrington and Sawyer (1992)	6.61	19	0.01	1.04	1
Daneman and Carpenter (1980)	6.26	26	0.05	1.34	1
Ammar and Spada (2006)	6.03	28	0.01	1.04	1
Li (2010)	5.99	27	0	1.03	1
Doughty (2001)	5.96	14	0	1.01	1
(Ellis et al., 2006)	5.93	27	0.01	1.05	1
Schmidt (2001)	5.76	78	0.08	1.58	1
Ellis N. (2005)	5.69	11	0	1.02	1
Rebuschat (2013)	5.57	12	0	1	1
Sheen (2004)	5.41	15	0	1.01	1
(Ellis et al., 2001)	5.38	18	0.01	1.05	1
Gutiérrez (2013)	5.24	10	0	1.02	1
Lyster (1998)	5.24	10	0	1.01	1
Lyster (2004)	5.09	25	0.01	1.04	1
Long (1991)	5	15	0.02	1.09	1
Miyake and Friedman (1998)	4.8	13	0	1.01	1
Erlam (2005)	4.7	8	0	1	1
Mackey and Goo (2007)	4.66	8	0	1.01	1
Nation (1990)	11	33	0.05	1.67	2
Nation (2001)	8.95	67	0.03	1.36	2
Laufer and Hulstijn (2001)	7.1	23	0	1.03	2
Read (2000)	6.88	31	0.01	1.05	2
Nation (2006)	6.82	31	0.01	1.07	2
Read (2000)	6.74	18	0.01	1.06	2
Schmitt (2010)	6.68	20	0	1.01	2
(Godfroid et al., 2013)	6.5	14	0	1.02	2
Plonsky and Oswald (2014)	6.25	11	0	1.01	2
Laufer (1992)	6.12	16	0	1.03	2
Coxhead (2000)	6.02	31	0.04	1.24	2

(Continued)

TABLE 9 | Continued

References	Burst strength	Frequency	Centrality	Sigma	Cluster ID
Laufer and Ravenhorst-Kalovski (2010)	5.77	11	0	1.01	2
Nation (2013)	5.68	10	0	1	2
Waring and Takaki (2003)	5.58	14	0	1.01	2
Wray (2002)	5.56	13	0	1.01	2
Hulstijn (2003)	5.31	13	0	1.01	2
O'Malley and Chamot (1990)	5.16	11	0.01	1.05	2
Barr et al. (2013)	5.12	9	0	1.02	2
Boers et al. (2006)	5.05	11	0	1.01	2
Schmidt (2001)	4.72	9	0	1	2
Schmitt et al. (2001)	4.65	8	0	1	2
Canale and Swain (1980)	10.36	57	0.39	31.21	3
Alderson and Wall (1993)	6.15	11	0	1.03	3
Bachman and Palmer (1996)	4.82	27	0.02	1.1	3
Norris and Ortega (2009)	11.72	35	0.01	1.08	4
Norris and Ortega (2000)	9.81	48	0.03	1.37	4
Ellis (2003)	9.76	37	0.01	1.09	4
Skehan (1998)	8.59	65	0.08	1.91	4
Foster et al. (2000)	8.24	28	0.03	1.27	4
Skehan (2009)	8.02	24	0.01	1.07	4
Wolfe-Quintero et al. (1998)	7.01	21	0	1.02	4
Housen and Kuiken (2009)	6.65	13	0	1.02	4
Biber (1999)	6.38	16	0	1.03	4
Chandler (2003)	6.25	19	0.01	1.07	4
Levett (1989)	6.2	12	0	1.02	4
Ellis (2009)	6.01	13	0	1.01	4
Vygotsky (1978)	5.68	10	0	1	4
Bates et al. (2015)	5.68	10	0	1	4
Larsen-Freeman (2006)	5.66	10	0	1	4
Ellis (2008)	5.65	20	0.01	1.03	4
Biber et al. (2011)	5.58	14	0	1.02	4
Kormos and Dénes (2004)	5.29	9	0	1	4
Ortega (2003)	5.18	13	0	1.02	4
Plonsky (2013)	4.78	12	0	1.02	4
Swain (2000)	4.74	12	0	1.01	4
Robinson (2005)	4.64	10	0	1	4
Dörnyei (2007)	4.64	10	0	1	4

Cluster 2: Vocabulary Learning

Cluster 2 comprises of vocabulary learning research. General textbooks on theoretical aspects of vocabulary (Nation, 1990, 2001, 2013; O'Malley and Chamot, 1990; Schmitt, 2010) and Schmitt (2008) review provide a deeper understanding of the crucial role of vocabulary in language learning, and in particular in incidental learning (Laufer and Hulstijn, 2001; Hulstijn, 2003; Godfroid et al., 2013). Efforts to find more efficient ways of learning vocabulary have led to the adoption of quantitative methods in research into vocabulary acquisition. Laufer (1992), Laufer and Ravenhorst-Kalovski (2010) and Nation (2006) sought the lexical threshold—the minimum number of words a learner needs for reading comprehension while the quantification of lexis allows for empirically-based vocabulary wordlists

(Coxhead, 2000) and tests like the Vocabulary Levels Test (Schmitt et al., 2001). The use of formulaic sequences (Wray, 2002; Boers et al., 2006) is another off-shoot of this aspect of vocabulary learning. Read's (2000) text on assessing vocabulary remains a key piece of work, as it is in Cluster 0 of the Core journals. Finally, with the move toward quantitative methods, publications on relevant research methods such as effect size (Plonsky and Oswald, 2014) and linear mixed-effects models (Barr et al., 2013) gain importance in this cluster (see **Supplemental Table 10** for further information).

Cluster 4: Measures of Language Complexity

Cluster 4 represent research on language complexity and its various measures. A dominant approach to measuring linguistic

ability in this cluster is the measurement practices of complexity, accuracy, and fluency (CAF). In their review, Housen and Kuiken (2009) traced the historical developments and summarized the theoretical underpinnings and practical operationalization of the constructs, forming an important piece of work for research using CAF. Research in this cluster largely looked at the effect of methods of language teaching on one or more of the elements of CAF: for example, the effect of corrective feedback on accuracy and fluency (Chandler, 2003) and corrective feedback and the effect of planning on all three aspects in oral production (Ellis, 2009). Another line of research was to look at developments in complexity, accuracy, and/or fluency in students' language production (Ortega, 2003; Larsen-Freeman, 2006).

The CAF is not without its flaws, which are pointed out by Skehan (2009) and Norris and Ortega (2009). Norris and Ortega (2009) suggested that syntactic complexity should be measured multidimensionally and Biber et al. (2011), using corpus methods, suggested a new approach to syntactic complexity. As with Biber et al. (2011), another theme emerging from this cluster was the application of quantitative methods in language learning and teaching research (Bates et al., 2015). Methodological issues (Foster et al., 2000; Dörnyei, 2007; Plonsky, 2013) form another sub-cluster, as researchers attempt to come up with more precise ways of defining and measuring these constructs (see *Supplemental Table 11* for further information).

Second Aim: Measurement and Validation in the Core and General Journals

The second aim of the study was to investigate measurement and validation practices in the published assessment research in the main clusters of the core and general journals. **Figures 3–5** present visual comparisons in measurement and validation practices between the two datasets. Given the differing numbers in the two data sets, numbers presented in the histograms have been normalized for comparability (frequency of publications reporting the feature divided by the total number of papers). As demonstrated in **Figure 3**, studies in the general journals dataset covered a wider range of domain specifications, providing more coverage of more fine-grained domain specifications as compared to the core journals dataset. On the other hand, the four "basic" language skills—reading, writing, listening and speaking (listed here as Oral Production) were well-represented in both the general and core journals dataset, unsurprisingly. Cumulatively, reading, writing/essays, oral production dominate both the general journals and core journals datasets, with listening comparatively less so in both datasets. Of considerable interest is the predominance of vocabulary in the general journals dataset, far outstripping the four basic skills in the dataset.

In addition, as **Figure 4** shows, the numbers of studies in both the core journals and general journals datasets that operationalized the constructs using Cloze/Likert/MCQ, Writing and Oral Production was fairly evenly matched. Writing is used most in the Core journals while Oral Production is used most in the General Journals. Finally, **Figure 5** shows the importance placed on reliability by authors, in both datasets. In comparison, other measurement practices are scarcely given

mention. Generalization and utilization had extremely poor showing in the general journals, in comparison to core journals, as the disparity between the four bars in **Figure 5** shows.

Limitations and Future Directions

The present study is not without limitations. As the focus of the study was to identify research clusters and bursts and the measurement and validation practices in language assessment research. However, the reasons why certain authors were co-cited by a large number of authors were not investigated. Merton (1968, 1988) and Small (2004) proposed two reasons for bursts in citations based on the sociology of science whereby the Matthew effect and the halo effect constitute possible contributors to the burstness of publications. First, Merton (1968, 1988) proposed that eminent authors often receive comparatively more credit from other authors than less known authors—Merton (1968, 1988) called this the Matthew Effect. This results in a widening lacuna between unknown and well-known authors (Merton, 1968, 1988) and in many cases the unfortunate invisibility of equally superior research published by unknown authors (Small, 2004). This is because citations function like "expert referral" and once they gain momentum, they "will increase the inequality of citations by focusing attention on a smaller number of selected sources, and widening the gap between symbolically rich and poor" (Small, 2004, p. 74). One way that this can be measured in future research is using power laws or similar mathematical functions to capture the trends in the data (Brzezinski, 2015). For example, a power law would fit a dataset of cited and citing publications wherein a large portion of the observed outcomes (citations) result from a small number of cited publications (Albarrán and Ruiz-Castillo, 2011). Albarrán et al. (2011, p. 395) provided compelling evidence from an impressively large dataset to support this phenomenon, concluding that "scientists make references that a few years later will translate into a highly skewed citation distribution crowned in many cases by a power law."

In addition, the eminence of scholars or the reputation of journals where the work is published can make a significant contribution to their burstness—this is called the halo effect (Small, 2004). In a recent paper, Zhang and Poucke (2017) showed that journal impact factor has a significant impact on the citations that a paper received. Another study by Antoniou et al. (2015, p. 286) identified "study design, studies reporting design in the title, long articles, and studies with high number of references" as predictors of higher citation rates. To this list, we might add seniority and eminence of authors and the type of publication (textbooks vs. paper), as well as "negative citation, self-citation, and misattribution" (Small, 2004, p. 76). Future research should investigate whether these variables have a role in citation patterns and clusters that emerged in the present study.

While self-citation was not filtered out and may present a limitation of this study, self-citation can be legitimate and necessary to the continuity of the development of a line of research. In CiteSpace, to qualify as a citing article, the citations of the article must exceed a selection threshold, either by g-index, top N most cited per time slice, or other selection modes. Although this process does not prevent the selection of a self-cited reference, the selection is justifiable to a great extent. If

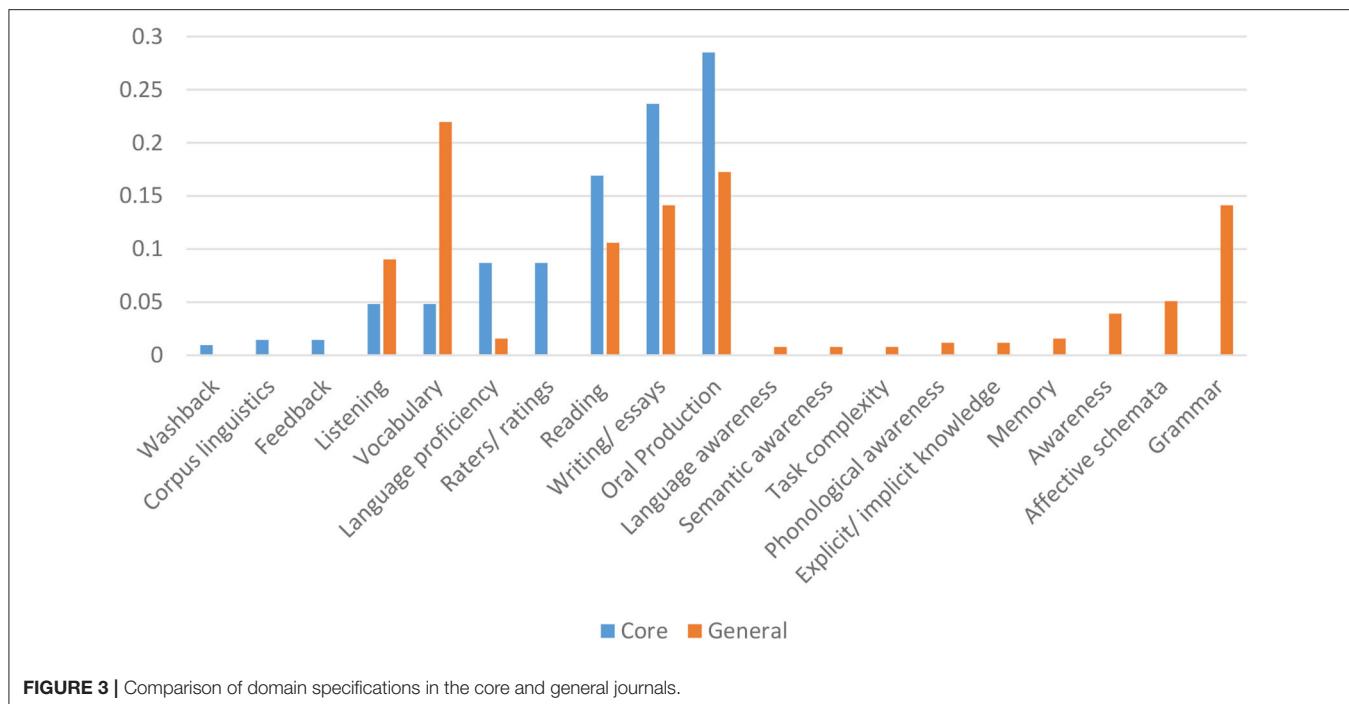


FIGURE 3 | Comparison of domain specifications in the core and general journals.

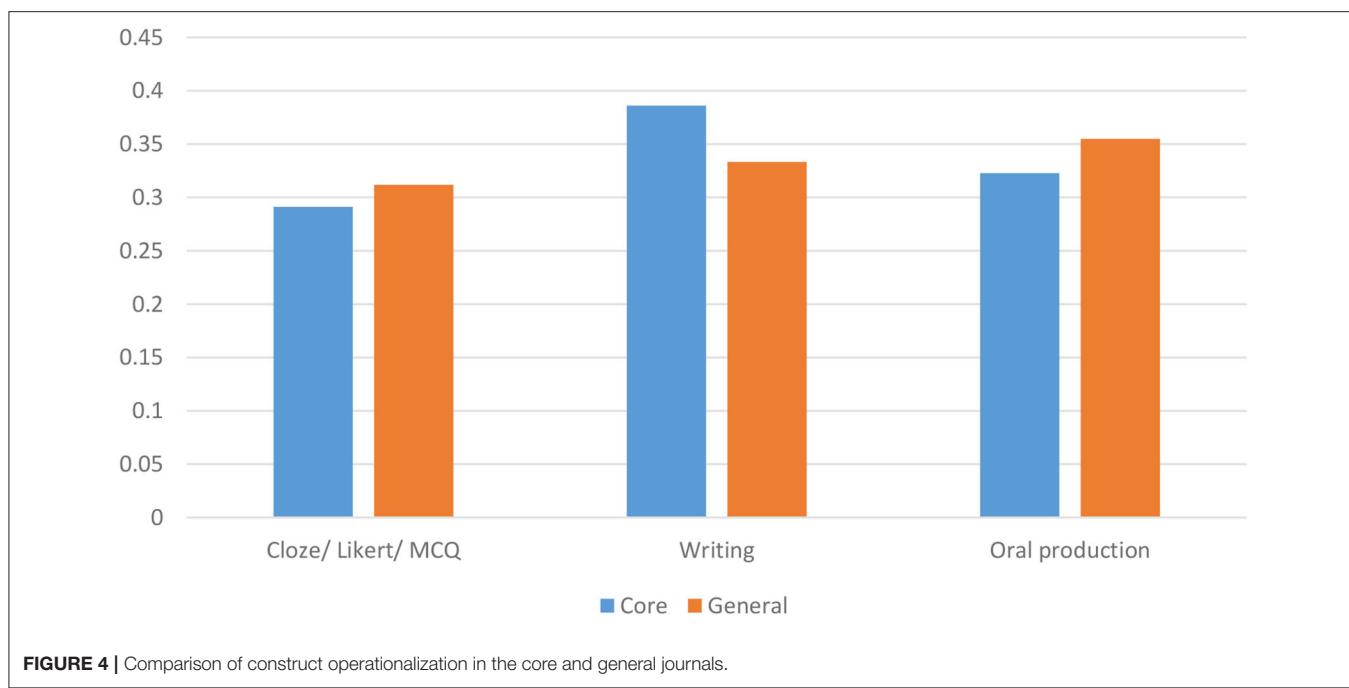


FIGURE 4 | Comparison of construct operationalization in the core and general journals.

a highly cited reference involves some or even all self-citations, then it behooves the analyst to establish the role of the reference in the literature. They should verify whether the high citations are due to inflated citations or if indeed, there is intellectual merit that justifies self-citation.

Another limitation of the study is that we did not include methodological journals such as “Journal of Educational Measurement” in the search, as indicated earlier. This was

because we adopted a keyword search strategy in this study and the majority of the papers in methodological journals include the search keywords we used such as measurement and assessment, even though many of them are not relevant to language assessment. This would affect the quality and content of the clusters. We suggest future research can explore the relationship between language assessment and methodological journals through, for example, the dual-map overlay method

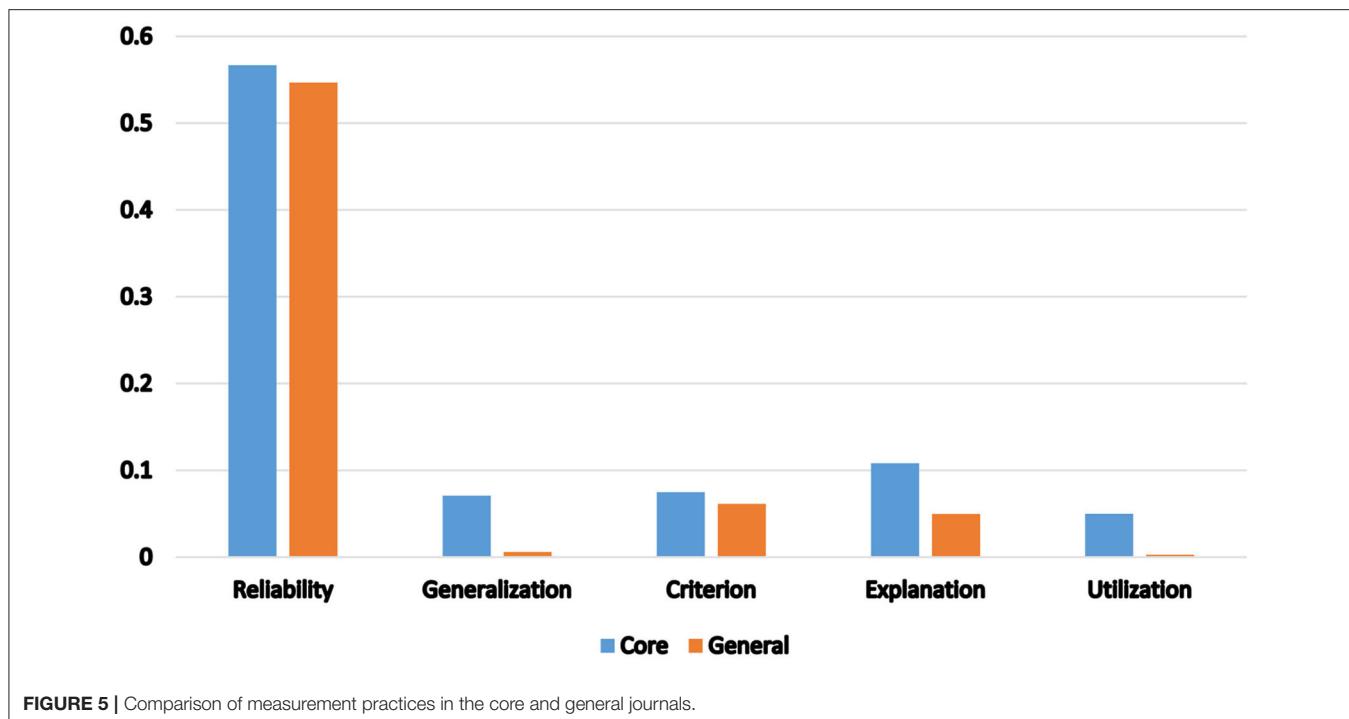


FIGURE 5 | Comparison of measurement practices in the core and general journals.

which is available in CiteSpace. Similarly, technical reports and book chapters were not included in the datasets, as the former are not indexed in Scopus and coverage of Scopus of the latter is not as wide as its coverage of journal articles.

Finally, it should be noted that for a recent publication to become a burst, it will take at least 1 year as our present and past analyses show (Aryadoust and Ang, 2019). Therefore, the dynamics of the field under investigation can change in a few years, as new bursts and research clusters emerge and drag the direction of research to a different direction.

CONCLUSION

The first aim of the study was to identify the main intellectual domains in language assessment research published in the core and general journals. We found that the primary focus of general journals was on vocabulary, oral proficiency, essay writing, grammar, and reading. The secondary focus was on affective schemata, awareness, memory, language proficiency, explicit vs. implicit language knowledge, language or semantic awareness, semantic complexity. By contrast, with the exception of language proficiency, this second area of focus was absent in the core general journals. The focus of the core journals was more exclusively on reading and listening comprehension assessment (primary theme), facets of speaking and writing performance such as raters and (psychometric) validation (secondary theme), as well as feedback, corpus linguistics, and washback (tertiary theme). From this, it may be said the main preoccupation of researchers in SLA and language assessment was the assessment of reading, writing, and oral production,

whereas assessment in SLA research additionally centered around vocabulary and grammar constructs. There were a number of areas that were underrepresented including affective schemata, awareness, memory, language proficiency, explicit vs. implicit language knowledge, language or semantic awareness, semantic complexity, feedback, corpus linguistics, and washback. These areas should be investigated with more rigor in future research.

In both datasets, several textbooks, editorials and review articles feature prominently in and/or across the clusters. The heavy presence of certain publications (like Bachman's) can be attributable to the importance of the scholar to the field. However, certain types of publications, like review articles, do tend to disproportionately get cited more often (Bennet et al., 2019) although precisely why this is the case is yet to be determined. Aksnes et al. (2019) cautions on overreliance on bibliometric analysis ring true here as well. Thus, we have provided additional analyses on the statistics to complete the picture behind the numbers, inasmuch that is possible.

The second aim of the study was to describe measurement and validation practices in the two datasets. Collectively, the data and comparisons presented demonstrated strong evidence that the majority of citing papers did not carry out inference-based validation that was spelled out by Bachman and Palmer (2010), Kane (2006), or Messick (1989) in both core and general journals. In language assessment, Bachman (2005) and Bachman and Palmer (2010) stressed that an all-encompassing validation program is “important and useful” before an assessment can be put to any use (Bachman, 2005, p. 30, emphasis in original). However, the feasibility and heavy demands of a strong validity program remain an open question (see Haertel, 1999). Particularly, it seems impracticable to validate both the

interpretations and uses of a language test/assessment before using the test for research purposes. The solution is Kane (2006) less demanding approach which holds that test instruments should be validated for the claims made. Accordingly, it would not be expected that researchers provide any “validity” evidence containing all the validity inferences explicated above for every instrument. Some useful guidelines include the report of reliability (internal consistency and rater consistency), item difficulty and discrimination range, person ability range, as well as evidence that the test measures the purported constructs. In sum, in our view, the lack of reporting of evidence for the above-mentioned components in the majority of studies was because these were not applicable to the objectives and design of the studies and their assessment tools.

The preponderance of the use of open-ended (essay/oral performance), which engage more communicative skills as compared to discrete point/selected response testing (like MCQ or Cloze), shows a tendency toward communicative testing approaches in both datasets. As format effects have been found on L1 reading and L2 listening, and L2 listening under certain conditions (see In’ami and Koizumi, 2009), the popularity of the relatively more difficult open-ended questions have implications for language test developers that cannot be ignored. Given the effect of format on scores impacts the reliability of tests in making discriminations on language ability, and consequently, fairness, the popularity of one type of format in language testing should be re-evaluated, or at the very least, examined more closely.

Finally, the sustainability of the intellectual domains identified in this study depends on the needs of the language assessment community and other factors such as “influence” of the papers published in each cluster. If a topic is an established intellectual domain with influential authors (high burstness and betweenness centrality), it stands a higher chance of thriving and proliferating. However, the fate of intellectual domains that have not attracted the attention of authors with high bursts and betweenness centrality could be bleak—even though these clusters may discuss significant areas of inquiry. There is currently no profound understanding of the forces that shape the scope and direction of language assessment research. Significantly more research is needed to determine what motivates authors to select and

investigate a topic, how thoroughly they cite past research, and what internal (within a field) and external (between fields) factors lead to the sustainability of a Research Topic.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The datasets can be reproduced from Scopus using the search formula provided in the Appendix.

AUTHOR CONTRIBUTIONS

VA conceptualized the study, downloaded the data, conducted data analysis, contributed to writing the paper, and led the team. AZ and ML helped with the data analysis and coding, and contributed to writing the paper. CC contributed conceptually to data generation and analysis and suggested revisions. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a research grant from Paragon Testing Enterprises, Canada, and partly by the National Institute of Education (NIE) of Nanyang Technological University (NTU), Singapore (Grand ID: RI 1/18 VSA). The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of NIE and NTU.

ACKNOWLEDGMENTS

We wish to thank Chee Shyan Ng and Rochelle Teo for their contribution to earlier versions of this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01941/full#supplementary-material>

REFERENCES

Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019). Citations, citation indicators, and research quality: an overview of basic concepts and theories. *Sage Open* 9, 1–17. doi: 10.1177/2158244019829575

Albarrán, P., Crespo, J. A., Ortúñoz, I., and Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics* 88, 385–397. doi: 10.1007/s11192-011-0407-9

Albarrán, P., and Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *J. Am. Soc. Inform. Sci. Technol.* 62, 40–49. doi: 10.1002/asi.21448

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511732935

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. London: A&C Black.

Alderson, J. C., and Banerjee, J. (2001). State of the art review: language testing and assessment Part 1. *Lang. Teach.* 34, 213–236. doi: 10.1017/S0261444800014464

Alderson, J. C., and Banerjee, J. (2002). State of the art review: language testing and assessment (part two). *Language Teach.* 35, 79–113. doi: 10.1017/S0261444802001751

Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., and Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Lang. Testing* 13, 280–297. doi: 10.1177/026553229601300304

Alderson, J. C., and Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. *Read Foreign Lang.* 5, 253–270.

Alderson, J. C., and Wall, D. (1993). Does washback exist? *Appl. Linguist.* 14, 115–129. doi: 10.1093/applin/14.2.115

American Educational Research Association (2014). *American Psychological Association, and National Council on Measurement in Education. Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Ammar, A., and Spada, N. (2006). One size fits all? Recasts, Prompts, and L2 Learning. *Stud. Second Lang. Acquis.* 28:543. doi: 10.1017/S0272263106060268

Antoniou, G. A., Antoniou, S. A., Georgakarakos, E. I., Sfyroeras, G. S., and Georgiadis, G. S. (2015). Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature. *Ann. Vasc. Surg.* 29, 286–292. doi: 10.1016/j.avsg.2014.09.017

Arik, B., and Arik, E. (2017). "Second language writing" publications in web of science: a bibliometric analysis. *Publications* 5:4. doi: 10.3390/publications5010004

Aryadoust, V. (2013). *Building a Validity Argument for a Listening Test of Academic proficiency*. Newcastle: Cambridge Scholars Publishing.

Aryadoust, V. (2020). A review of comprehension subskills: a scientometrics perspective. *System* 88, 102–180. doi: 10.1016/j.system.2019.102180

Aryadoust, V., and Ang, B. H. (2019). Exploring the frontiers of eye tracking research in language studies: a novel co-citation scientometric review. *Comput. Assist. Lang. Learn.* 1–36. doi: 10.1080/09588221.2019.1647251

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Lang. Testing* 17, 1–42. doi: 10.1177/026553220001700101

Bachman, L. F. (2005) Building and supporting a case for test use. *Lang. Assess. Quart.* 2, 1–34. doi: 10.1207/s15434311laq0201_1

Bachman, L. F., and Cohen, A. D. (1998). *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524711

Bachman, L. F., and Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quart.* 16:449. doi: 10.2307/3586464

Bachman, L. F., and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bachman, L. F., and Palmer, A. S. (2010). *Language assessment in practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford: Oxford University Press.

Bailey, K.M. (1999). *Washback in Language Testing. TOEFL Monograph Series MS-15, June 1999*. Educational Testing Service. Retrieved from: <https://www.ets.org/Media/pdf/RM-99-04.pdf>

Banerjee, J., Yan, X., Chapman, M., and Elliott, H. (2015). Keeping up with the times: revising and refreshing rating scale. *Assess. Writ. Int. J.* 26, 5–19. doi: 10.1016/j.asw.2015.07.001

Barkaoui, K. (2010). Variability in ESL essay rating processes: the role of the rating scale and rater experience. *Lang. Assess. Quart.* 7, 54–74. doi: 10.1080/15434300903464418

Barkaoui, K., and Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assess. Writ. Int. J.* 36, 19–31. doi: 10.1016/j.asw.2018.02.005

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Memory Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bennet, L., Eisner, D. A., and Gunn, A. J. (2019). Misleading with citation statistics? *J. Physiol.* 10:2593. doi: 10.1113/JP277847

Biber, D. (1999). *Longman grammar of spoken and written English*. London: Longman.

Biber, D., and Gray, B. (2013). Discourse Characteristics of Writing and Speaking Task Types on the "TOEFL iBT"® Test: A Lexico-Grammatical Analysis. "TOEFL iBT"® Research Report. TOEFL iBT-19. Research Report. RR-13-04. Princeton, NJ: ETS Research Report Series. doi: 10.1002/j.2333-8504.2013.tb02311.x

Biber, D., Gray, B., and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Q.* 45, 5–35. doi: 10.5054/tq.2011.244483

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., and Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test. *Lang. Teach. Res.* 10, 245–261. doi: 10.1191/1362168806lr195oa

Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Borsboom, D., and Mellenbergh, G. J. (2007). "Test validity in cognitive assessment," in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. P. Leighton, and M. J. Gierl (New York, NY: Cambridge University Press), 85–115. doi: 10.1017/CBO9780511611186.004

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer. doi: 10.1007/978-1-4757-3456-0

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: a review of the issues. *Lang. Test.* 15, 45–85. doi: 10.1177/026553229801500103

Brindley, G. (2001). Outcomes-based assessment in practice: some examples and emerging insights. *Lang. Test.* 18, 393–407. doi: 10.1177/026553220101800405

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: co-constructing a better performance. *Lang. Test.* 26, 341–366. doi: 10.1177/0265532209104666

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Lang. Test.* 12, 1–15. doi: 10.1177/026553229501200101

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Lang. Test.* 20, 1–25. doi: 10.1191/0265532203lt242oa

Brown, J. D., and Hudson, T. (1998). The alternatives in language assessment. *TESOL Q.* 32, 653–675. doi: 10.2307/3587999

Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. *Scientometrics* 103, 213–228. doi: 10.1007/s11192-014-1524-z

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511732959

Canale, M., and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Appl. Linguist.* 1, 1–47. doi: 10.1093/applin/1.1.1

Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Lang. Test.* 23, 269–289. doi: 10.1191/0265532206lt328oa

Carroll, J. B. (1961) "Fundamental considerations in testing for English language proficiency of foreign students," in *Testing Center for Applied Linguistics* (Washington, DC). Reprinted in Allen, H.B. & Campbell, R.N. (eds.). (1972) *Teaching English as a Second Language: A Book of Readings*. McGraw Hill.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Lang. Test.* 20, 369–383.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *J. Second Lang. Writing* 12, 267–296. doi: 10.1016/S1060-3743(03)00038-9

Chapelle, C. A. (1998). "Construct definition and validity inquiry in SLA research," in *Interfaces Between Second Language Acquisition and Language Testing Research*, eds L. F. Bachman and A. D. Cohen. (Cambridge: Cambridge University Press) 32–70. doi: 10.1017/CBO9781139524711.004

Chapelle, C. A., Enright, M. K., and Jamieson, J. M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.

Chen, C. (2004). Searching for intellectual turning points: progressive knowledge domain visualization. *Proc. Natl. Acad. Sci. U.S.A.* 101, 5303–5310. doi: 10.1073/pnas.0307513100

Chen, C. (2006). CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inform. Sci. Technol.* 57, 359–377. doi: 10.1002/asi.20317

Chen, C. (2010). "Measuring Structural Change in Networks Due to New Information," in *NATO IST-093/RWS-015 Workshop on Visualizing Networks: Coping with Change and Uncertainty*. Rome: Griffiss Institutes.

Chen, C. (2014). *The CiteSpace Manual*. Available online at: <http://cluster.ischool.drexel.edu/cchen/citespace/CiteSpaceManual.pdf>

Chen, C. (2016). *CiteSpace: A Practical Guide for Mapping Scientific Literature*. New York, NY: Nova Science Publishers.

Chen, C. (2017). Science mapping: a systematic review of the literature. *J. Data Inform. Sci.* 2, 1–40. doi: 10.1515/jdis-2017-0006

Chen, C. (2019). *How to Use CiteSpace*. Retrieved from <https://leanpub.com/howtousecitespace>

Chen, C., Ibekwe-SanJuan, F., and Hou, J. (2010). The structure and dynamics of co-citation clusters: a multiple-perspective co-citation analysis. *J. Am. Soc. Inform. Sci. Technol.* 61, 1386–1409. doi: 10.1002/asi.21309

Chen, C., Song, I. Y., Yuan, X., and Zhang, J. (2008). The thematic and citation landscape of data and knowledge engineering (1985–2007). *Data Knowl. Eng.* 67, 234–259. doi: 10.1016/j.ddata.2008.05.004

Chen, C., and Song, M. (2017). *Representing Scientific Knowledge: The Role of Uncertainty*. Princeton, NJ: Springer. doi: 10.1007/978-3-319-62543-0

Chen, Z., and Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Lang. Test.* 2:155. doi: 10.1177/0265532285002004

Chen, C. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. 1st Edn. Princeton, NJ: Springer. doi: 10.1007/978-1-4471-0051-5_1

Clapham, C. (1996). *The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension*. Cambridge: University of Cambridge Local Examinations Syndicate.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. L. New York, NY: Erlbaum Associates.

Collins, A. J., and Fauser, C.J.M. B. (2005). Balancing the strengths of systematic and narrative reviews. *Hum. Reprod. Update* 11, 103–104. doi: 10.1093/humupd/dmh058

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Press Syndicate of the University of Cambridge.

Coxhead, A. (2000). A new academic word list. *TESOL Quart.* 34, 213–238. doi: 10.2307/3587951

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Cumming, A. (1990). Expertise in evaluating second language compositions. *Lang. Test.* 7:31. doi: 10.1177/026553229000700104

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: promises and perils. *Lang. Assess. Quart.* 10, 1–8. doi: 10.1080/15434303.2011.622016

Cumming, A., Kantor, R., and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework. *Modern Lang. J.* 86, 67–96. doi: 10.1111/1540-4781.00137

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verb. Learn. Verb. Behav.* 19, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Davies, A. (1982). “Language testing parts 1 and 2,” in *Cambridge Surveys*, ed. V. Kinsella (Cambridge: Cambridge University Press), 127–159. (Originally published in *Language Teaching and Linguistics: Abstracts*, 1978).

Davies, A. (2008). Textbook trends in teaching language testing. *Lang. Test.* 25, 327–347. doi: 10.1177/0265532208090156

Davies, A. (2014). Remembering 1980. *Lang. Assess. Quart.* 11, 129–135. doi: 10.1080/15434303.2014.898642

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika* 9, 185–197. doi: 10.1007/BF02288722

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Lang. Test.* 26, 367–396. doi: 10.1177/0265532209104667

Davison, C. (2007). Views from the chalkface: english language school-based assessment in Hong Kong. *Lang. Assess. Quart.* 4, 37–68. doi: 10.1080/15434300701348359

De Bellis, N. (2014). “History and evolution of (biblio) metrics,” in *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, eds B. Cronin and C. Sugimoto (Cambridge, MA: MIT Press), 23–44.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assess. Writing* 18, 7–24. doi: 10.1016/j.asw.2012.10.002

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*. Oxford: Oxford University Press.

Doughty, C. (2001). “Cognitive underpinnings of focus on form,” in *Cognition and Second Language Instruction*, eds P. Robinson, M. H. Long, and J. C. Richards (Cambridge: Cambridge University Press) 206–257. doi: 10.1017/CBO9781139524780.010

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511732911

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Lang. Test.* 25, 155–185. doi: 10.1177/0265532207086780

Eckes, T. (2011). Introduction to many-facet Rasch measurement: analyzing and evaluating rater-mediated assessments. *Peter Lang.* 17, 113–116. doi: 10.1080/15366367.2018.1516094

Eckes, T., and Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Lang. Test.* 23, 290–325. doi: 10.1191/0265532206lt330oa

Ellis, N. (2005). At the interface: dynamic interactions of explicit and implicit language knowledge. *Stud. Second Lang. Acquis.* 27, 305–352. doi: 10.1017/S027226310505014X

Ellis, R. (2003). *Task-Based Language Learning and Teaching*. Oxford University Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: a psychometric study. *Stud. Second Lang. Acquis.* 27:141. doi: 10.1017/S0272263105050096

Ellis, R. (2008). *The Study of Second Language Acquisition 2nd Edn*. Cambridge: Oxford University Press.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Appl. Linguist.* 30, 474–509. doi: 10.1093/applin/amp042

Ellis, R., Basturkmen, H., and Loewen, S. (2001). Learner Uptake in Communicative ESL Lessons. *Lang. Learn. J. Res. Lang. Stud.* 51:281. doi: 10.1111/1467-9922.00156

Ellis, R., Loewen, S., and Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Stud. Second Lang. Acquisit.* 28, 339–368. doi: 10.1017/S0272263106060141

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Lang. Teach. Res.* 9, 147–171. doi: 10.1191/1362168805lr161oa

Fan, J., and Yan, X. (2020). Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Front. Psychol.* 11:330. doi: 10.3389/fpsyg.2020.00330

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics (5th Edn.)*. Cambridge: The Bookwatch.

Flower, L., and Hayes, J. R. (1981). A cognitive process theory of writing. *Coll. Compos. Commun.* 32:365. doi: 10.2307/356600

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Appl. Linguist.* 21, 354–375. doi: 10.1093/applin/21.3.354

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Lang. Test.* 13, 208–238. doi: 10.1177/026553229601300205

Fulcher, G. (2003). *Testing Second Language Speaking*. Cambridge: Pearson Education.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Lang. Assess. Quart.* 1, 253–266. doi: 10.1207/s15434311laq0104_4

Fulcher, G. (n.d.). *What Is Language Testing*. Language Testing Resources. Available online at: <http://languagetesting.info/whatis/lt.html>

Fulcher, G., Davidson, F., and Kemp, J. (2011). Effective rating scale development for speaking tests: performance decision trees. *Lang. Test.* 28, 5–29. doi: 10.1177/0265532209359514

Gao, L. and Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Lang. Test.* 28, 77–104. doi: 10.1177/0265532210364380

Gebril, A. (2009). Score generalizability of academic writing tasks: does one test method fit it all?. *Lang. Test.* 26, 507–531. doi: 10.1177/0265532209340188

Godfroid, A., Boers, F., and Housen, A. (2013). An eye for words: gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Stud. Second Lang. Acquisit.* 35, 483–517. doi: 10.1017/S0272263113000119

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Stud. Second Lang. Acquisit.* 34:445. doi: 10.1017/S0272263112000149

Goswami, A. K., and Agrawal, R. K. (2019). Building intellectual structure of knowledge sharing. *VINE J. Inform. Knowl. Manag. Syst.* 50, 136–162. doi: 10.1108/VJIKMS-03-2019-0036

Grabowski, K. C., and Oh, S. (2018). “Reliability analysis of instruments and data coding,” in *The Palgrave Handbook of Applied Linguistics Research Methodology*, eds A. Phakiti, P. De Costa, L. Plonsky, and S. Starfield (London: Palgrave Macmillan), 541–565. doi: 10.1057/978-1-37-59900-1_24

Grabowski, K. C., and Lin, R. (2019). “Multivariate generalizability theory in language assessment,” in *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*, eds V. Aryadoust and M. Raquel (New York, NY: Routledge), 54–80. doi: 10.4324/9781315187815-4

Green, A., Ünalı, A., and Weir, C. (2010). Empiricism versus connoisseurship: establishing the appropriacy of texts in tests of academic reading. *Lang. Test.* 27, 191–211. doi: 10.1177/0265532209349471

Green, S., and Salkind, N. (2014). *Using SPSS for Windows and Macintosh: Analyzing and understanding data*, 7th Edn. London: Person Education, Inc.

Guo, L., Crossley, S. A., and McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: a comparison study. *Assess. Writing* 18, 218–238. doi: 10.1016/j.asw.2013.05.002

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measure of implicit and explicit knowledge. *Stud. Second Lang. Acquisit.* 35, 423–449. doi: 10.1017/S0272263113000041

Haertel, E. H. (1999). Validity arguments for high-stakes testing: in search of the evidence. *Educ. Measur.* 18, 5–9. doi: 10.1111/j.1745-3992.1999.tb00276.x

Hall, W. E., and Robinson, F. P. (1945). An analytical approach to the study of reading skills. *J. Educ. Psychol.* 36, 429–442. doi: 10.1037/h0058703

Halliday, M. A. K., and Hasan, R. (1976). *Cohesion in English*. London: English Language Series, Longman.

Hambleton, R. K., and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Dordrecht: Kluwer Academic Publishers. doi: 10.1007/978-94-017-1988-9

Hamp-Lyons, L. (1991). “Scoring procedures for ESL contexts,” in *Assessing Second Language Writing in Academic Contexts*, ed L. Hamp-Lyons (New York, NY: Ablex Pub. Corp.), 241–276.

Harding, L., Alderson, J. C., and Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Lang. Test.* 32, 317–336. doi: 10.1177/0265532214564505

Harrington, M. and Sawyer, M. (1992). L2 Working memory capacity and L2 reading skill. *Stud. Second Lang. Acquisit.* 14:25. doi: 10.1017/S0272263100010457

Harsch, C. (2014). General language proficiency revisited: current and future issues. *Lang. Assess. Quart.* 11, 152–169. doi: 10.1080/15434303.2014.902059

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. New York, NY: Newberry House Publishers.

Hornberger, N. H., and Shohamy, E. (2008). *Encyclopedia of Language and Education Vol. 7: Language Testing and Assessment*. New York, NY: Springer.

Housen, A., and Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Appl. Linguist.* 30:amp048. doi: 10.1093/applin/amp048

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hulstijn, J. H. (2003). “Incidental and intentional learning,” in *The Handbook of Second Language Acquisition*, eds C. J. Doughty and M. H. Long (Blackwell Publishing), 349–381. (New Jersey: Blackwell handbooks in linguistics; No. 14) doi: 10.1002/9780470756492.ch12

In'nami, Y., and Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Lang. Test.* 26, 219–244. doi: 10.1177/02655322081006

Isbell, D. R. (2017). Assessing C2 writing ability on the certificate of English language proficiency: rater and examinee age effects. *Assess. Writing Int. J.* 34, 37–49. doi: 10.1016/j.asw.2017.08.004

Iwashita, N., Brown, A., McNamara, T., and O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: how distinct? *Appl. Linguist.* 29, 24–49. doi: 10.1093/applin/amm017

Jacobs, H. L. (1981). *Testing ESL Composition: A Practical Approach*. New York, NY: Newbury House.

Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for fusion model application to languedge assessment. *Lang. Test.* 26, 31–73. doi: 10.1177/0265532208097336

Jang, E. E. (2009b). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Lang. Assess. Quart.* 6, 210–238. doi: 10.1080/15434300903071817

Jones, K. (2004). Mission drift in qualitative research, or moving toward a systematic review of qualitative studies, moving back to a more systematic narrative review. *Q. Rep.* 9, 95–112.

Kane, M. T. (2006). “Validation,” in *Educational Measurement*, 4th Edn, ed R. L. Brennan (Westport, CT: American Council on Education/Praeger), 17–64.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *J. Educ. Measur.* 50, 1–73. doi: 10.1111/jedm.12000

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Lang. Assess. Quart.* 12, 239–261. doi: 10.1080/15434303.2015.1049353

Knoch, U. (2009). Diagnostic assessment of writing: a comparison of two rating scales. *Lang. Test.* 26, 275–304. doi: 10.1177/0265532208101008

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: what should they look like and where should the criteria come from? *Assess. Writing* 16, 81–96. doi: 10.1016/j.asw.2011.02.003

Knoch, U., Read, J., and von Randow, J. (2007). Re-training writing raters online: how does it compare with face-to-face training? *Assess. Writing* 12, 26–43. doi: 10.1016/j.asw.2007.04.001

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Lang. Test.* 19, 193–220. doi: 10.1191/0265532202lt227oa

Kormos, J., and Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32, 145–164. doi: 10.1016/j.system.2004.01.001

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests: A Teacher's Book*. Bristol, Inglaterra Longmans, Green and Company.

Lallamode, S. P., Daud, N. M., and Abu Kassim, N. L. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assess Writing* 30, 44–62. doi: 10.1016/j.asw.2016.06.001

Lam, D. M. K. (2018). What counts as “responding? Contingency on previous speaker contribution as a feature of interactional competence. *Lang. Test.* 35, 377–401. doi: 10.1177/0265532218758126

Langsam, R. S. (1941). A factorial analysis of reading ability. *J. Exp. Educ.* 10, 57–63. doi: 10.1080/00220973.1941.11010235

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Appl. Linguist.* 27, 590–619. doi: 10.1093/applin/aml029

Laufer, B. (1992). “How much lexis is necessary for reading comprehension?,” in *Vocabulary and Applied Linguistics*, eds P. J. L. Arnaud, and H. Bejoing (New York, NY: Macmillan), 129–132. doi: 10.1007/978-1-349-12396-4_12

Laufer, B., and Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: the construct of task-induced involvement. *Appl. Linguist.* 22, 1–26. doi: 10.1093/applin/22.1.1

Laufer, B., and Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Read. Foreign Lang.* 22, 15–30.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Lang. Test.* 13, 151–172. doi: 10.1177/026553229601300202

Lee, Y. W., and Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: an overview. *Lang. Assess. Quart.* 6, 172–189. doi: 10.1080/15434300902985108

Lei, L., and Liu, D. (2019). The research trends and contributions of system's publications over the past four decades (1973e2017): a bibliometric analysis. *System* 80:1e13. doi: 10.1016/j.system.2018.10.003

Levett, W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Li, S. (2010). The effectiveness of corrective feedback in SLA: a meta-analysis. *Lang. Learn.* 60, 309–365. doi: 10.1111/j.1467-9922.2010.00561.x

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: a longitudinal study of new and experienced raters. *Lang. Test.* 28, 543–560. doi: 10.1177/0265532211406422

Linacre, J. M. (1994). *Many-Facet Rasch Measurement* (2nd Ed.). Chicago, IL: MESA.

Long, M. H. (2007). *Problems in SLA*. New Jersey: Lawrence Erlbaum Associates Publishers.

Long, M. H. (1991) "Focus on form: a design feature in language teaching methodology," in *Foreign Language Research in Cross-Cultural Perspective*, eds K. D., Bot, C. Kramsch, and R. Ginsberg. (Amsterdam: John Benjamins), 39–52. doi: 10.1075/sibil.2.07lon

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Lang. Test.* 34, 493–511. doi: 10.1177/0265532217710675

Lumley, T. (1993). The notion of subskills in reading comprehension tests: an EAP example. *Lang. Test.* 10:211. doi: 10.1177/026553229301000302

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters?. *Lang. Test.* 19, 246–276. doi: 10.1191/0265532202lt230oa

Lumley, T., and McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Lang. Test.* 12, 54–71. doi: 10.1177/026553229501200104

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511733017

Lynch, B., Davidson, F., and Henning, G. (1988). Person dimensionality in language test validation. *Lang. Test.* 5:206. doi: 10.1177/026553228800500206

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Stud. Second Lang. Acquisit.* 20:51. doi: 10.1017/S027226319800103X

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Stud. Second Lang. Acquisit.* 26, 399–432. doi: 10.1017/S0272263104263021

Lyster, R., and Ranta, L. (1997). Corrective feedback and learner uptake: negotiation of form in communicative classrooms. *Stud. Second Lang. Acquisit.* 19, 37–66. doi: 10.1017/S0272263197001034

Lyster, R., and Saito, K. (2010). Oral feedback in classroom SLA: a meta-analysis. *Stud. Second Lang. Acquisit.* 32:265. doi: 10.1017/S0272263109990520

Mackey, A., and Goo, J. (2007) "Interaction research in SLA: a meta-analysis and research synthesis," in *Conversational Interaction In Second Language Acquisition*, eds A. Mackey (Oxford: Oxford University Press), 407–453.

May, L. (2011). Interactional competence in a paired speaking test: features salient to raters. *Lang. Assess. Quart.* 8, 127–145. doi: 10.1080/15434303.2011.565845

McNamara, D., Graesser, A., McCarthy, P., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. London: Cambridge University Press. doi: 10.1017/CBO9780511894664

McNamara, T. (2014). 30 Years on-evolution or revolution? *Epilogue. Lang. Assess. Quart.* 11, 226–232. doi: 10.1080/15434303.2014.895830

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T. F. (1990). *Assessing the second language proficiency of health professionals*. (Ph.D. thesis), Department of Linguistics and Language Studies, The University of Melbourne, Australia.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test1. *Lang. Test.* 8:139. doi: 10.1177/026553229100800204

Merton, R.K. (1988). The matthew effect in science, II: cumulative advantage and the symbolism of intellectual property. *ISIS* 79, 606–623. doi: 10.1086/354848

Merton, R. K. (1968). The matthew effect in science. *Science* 159, 56–63. Reprinted in: *The Sociology of Science: Theoretical and Empirical Investigations*. (Chicago: University of Chicago Press, 1973), p. 438–459.

Messick, S. (1989). "Validity," in *Educational Measurement*, 3rd Edn, ed R. L. Linn. (New York, NY: American Council on Education/Macmillan), 13–103.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educ. Res.* 23, 13–23. doi: 10.3102/0013189X023002013

Messick, S. (1996). Validity and washback in language testing. *Lang. Test.* 13, 241–256. doi: 10.1177/026553229601300302

Mingers, J., and Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *Eur. J. Operation Res.* 246, 1–19. doi: 10.1016/j.ejor.2015.04.002

Miyake, A., and Friedman, N. P. (1998). "Individual differences in second language proficiency: working memory as language aptitude," in *Foreign Language Learning: Psycholinguistic Studies on Training and Retention*. eds A. F. Healy and L. E. Bourne Jr (New Jersey: Lawrence Erlbaum Associates Publishers), 339–364.

Mostafa, M. M. (2020). A knowledge domain visualization review of thirty years of halal food research: themes, trends and knowledge structure. *Trends Food Sci. Technol.* 99, 660–677. doi: 10.1016/j.tifs.2020.03.022

Nalimov, V., and Mulcjenko, B. (1971). *Measurement of Science: Study of the Development of Science as an Information Process*. Washington, DC: Foreign Technology Division.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Can. Modern Lang. Rev.* 63, 59–82. doi: 10.3138/cmlr.63.1.59

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139858656

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York, NY: Newbury House Publishers.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524759

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Norris, J. M., and Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Lang. Learn. J. Res. Lang. Stud.* 50:417. doi: 10.1111/0023-8333.00136

Norris, J. M., and Ortega, L. (2003). "Defining and measuring SLA," in *The Handbook of Second Language Acquisition*, eds C. J. Doughty and M. H. Long (Malden, MA: Blackwell), 717–761

Norris, J. M., and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Appl. Linguist.* 30, 555–578. doi: 10.1093/applin/amp044

Oller, J. W. (1979). *Language Tests at School: A Pragmatic Approach*. London: Longman.

O'Malley, J. M., and Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 Writing. *Appl. Linguist.* 24, 492–518. doi: 10.1093/applin/24.4.492

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Lang. Test.* 19, 277–295. doi: 10.1191/0265532202lt205oa

Pae, C. U. (2015). Why systematic review rather than narrative review?. *Psychiat. Invest.* 12:417. doi: 10.4306/pi.2015.12.3.417

Papageorgiou, S., Stevens, R., and Goodwin, S. (2012). The relative difficulty of dialogic and monologic input in a second-language listening comprehension test. *Lang. Assess. Quart.* 9, 375–397. doi: 10.1080/15434303.2012.721425

Petticrew, M., and Roberts, H. (2006). *Systematic Reviews in the Social Sciences*. New Jersey: Wiley Blackwell. doi: 10.1002/9780470754887

Phakiti, A., and Roevert, C. (2011). Current issues and trends in language assessment in Australia and New Zealand. *Lang. Assess. Quart.* 8, 103–107. doi: 10.1080/15434303.2011.566397

Pica, T. (1994). Research on negotiation: what does it reveal about second-language learning conditions, processes, and outcomes? *Lang. Learn. J. Res. Lang. Stud.* 44, 493–527. doi: 10.1111/j.1467-1770.1994.tb01115.x

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assess. Writing* 13, 111–129. doi: 10.1016/j.aw.2008.07.001

Plakans, L., and Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assess. Writing* 39, 98–112. doi: 10.1016/j.aw.2016.08.005

Plakans, L., Liao, J.-T., and Wang, F. (2019). "I should summarize this whole paragraph": Shared processes of reading and writing in iterative integrated assessment tasks. *Assess. Writing* 40, 14–26. doi: 10.1016/j.aw.2019.03.003

Plonsky, L. (2013). Study quality in SLA: an assessment of designs, analyses, and reporting practices in quantitative L2 research. *Stud. Second Lang. Acquisit.* 35:655. doi: 10.1017/S0272263113000399

Plonsky, L., and Oswald, F. L. (2014). How big is "big?" interpreting effect sizes in L2 Research. *Lang. Learn. J. Res. Lang. Stud.* 64, 878–912. doi: 10.1111/lang.12079

Rakedzon, T., and Baram-Tsabari, A. (2017). To make a long story short: a rubric for assessing graduate students' academic and popular science writing skills. *Assess. Writ.* 32, 28–42. doi: 10.1016/j.asw.2016.12.004

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Santa Monica: Paedagogik Institute.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511732942

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Lang. Learn.* 63, 595–626. doi: 10.1111/lang.12010

Robinson, P. (2005). Cognitive complexity and task sequencing: studies in a componential framework for second language task design. *Int. Rev. Appl. Linguist. Lang. Teach.* 43, 1–32. doi: 10.1515/iral.2005.43.1.1

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Lang. Test.* 23, 229–256. doi: 10.1191/0265532206lt329oa

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: focus on the construct of speaking. *Lang. Test.* 34, 477–492. doi: 10.1177/0265532217711431

Rosenshine, B.V. (2017). "Skill hierarchies in reading comprehension," in *Theoretical Issues in Reading Comprehension: Perspectives From Cognitive Psychology, Linguistics, Artificial Intelligence and Education*, eds R. J. Spiro, B. C., Bruce, and W.F. Brewer (London: Taylor and Francis) 535–554. doi: 10.4324/9781315107493-29

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Sawaki, Y., Stricker, L. J., and Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Lang. Test.* 26, 5–30. doi: 10.1177/0265532208097335

Sawaki, Y., and Xi, X. (2019). "Univariate generalizability theory in language assessment," in *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques* eds V. Aryadoust and M. Raquel (London: Routledge) 30–53. doi: 10.4324/9781315187815-3

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Rev.* 11, 11–26.

Schmidt, R. (2001). "Attention," in *Cognition and Second Language Instruction*, eds P. Robinson (Cambridge: Cambridge University Press) 3–32. doi: 10.1017/CBO9781139524780.003

Schmitt, N., Schmitt, D., and Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Lang. Test.* 18, 55–88. doi: 10.1177/026553220101800103

Schmitt, N. (2008). Review article: instructed second language vocabulary learning. *Lang. Teach. Res.* 12, 329–363. doi: 10.1177/1362168808089921

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. London: Palgrave Macmillan. doi: 10.1057/9780230293977

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Lang. Teach. Res.* 8, 263–300. doi: 10.1191/1362168804lr146oa

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. *Stud. Second Lang. Acquisit.* 10:165. doi: 10.1017/S0272263100007294

Shohamy, E. G. (2001). The power of tests: a critical perspective on the uses of language tests. Harlow; New York, NY: Longman.

Skehan, P. (1988). State of the art article: language testing Part 1. *Lang. Teach.* 21, 211–221. doi: 10.1017/S0261444800005218

Skehan, P. (2009). Modelling second language performance: integrating complexity, accuracy, fluency, and lexis. *Appl. Linguist.* 30, 510–532. doi: 10.1093/applin/amp047

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford University Press. doi: 10.1177/003368829802900209

Small, H. (2004). On the shoulders of robert merton: towards a normative theory of citation. *Scientometrics* 60, 71–79. doi: 10.1023/B:SCIE.0000027310.68393.bc

Small, H., and Sweeney, E. (1985). Clustering the science citation index using co-citations: a comparison of methods. *Scientometrics* 7, 391–409. doi: 10.1007/BF02017157

Spada, N., and Tomita, Y. (2010). Interactions between type of instruction and type of language feature: a meta-analysis. *Lang. Learn.* 60, 263–308. doi: 10.1111/j.1467-9922.2010.00562.x

Spolsky, B. (1977). "Language testing: art or science," in *Proceedings of the Fourth International Congress of Applied Linguistics, Vol. 3*, ed G. Nickel (Stuttgart: Hochschulverlag), 7–28.

Spolsky, B. (1990). Oral examinations: an historical note. *Lang. Test.* 7, 158–173. doi: 10.1177/026553229000700203

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford: Oxford University Press.

Spolsky, B. (2017). "History of language testing," in *Language Testing and Assessment*, eds E. Shohamy, and N. H. Hornberger (New York, NY: Springer), 375–384. doi: 10.1007/978-3-319-02261-1_32

Swain, M. (1985). "Communicative competence: some roles of comprehensible input and comprehensible output in its development," in *Input in Second Language Acquisition*, eds S. Gass, and C. Madden (New York, NY: Newbury House), 235–253.

Swain, M. (1995). "Three functions of output in second language learning," in *Principle and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson*, eds G. Cook and B. Seidlhofer (Oxford: Oxford University Press), 125–144.

Swain, M. (2000). "The output hypothesis and beyond: mediating acquisition through collaborative dialogue," in *Sociocultural Theory and Second Language Learning*, eds Lantolf, J.P. (Oxford: Oxford University Press), 97–114.

Taylor, L. (2009). Developing assessment literacy. *Ann. Rev. Appl. Linguist.* 29, 21–36. doi: 10.1017/S0267190509090035

Upshur, J. A. (1971). "Productive communication testing: a progress report," in *Applications in Linguistics*, eds G. Perren and J. L. M. Trim. (Cambridge University Press). 435–442.

van Batenburg, E. S. L., Oostdam, R. J., van Gelderen, A. J. S., and Jong, N. H. (2018). Measuring L2 speakers' interactional ability using interactive speech tasks. *Lang. Test.* 35, 75–100. doi: 10.1177/0265532216679452

van Lier, L. (1989). Reeling, writhing, drawing, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Q.* 23:489. doi: 10.2307/3586922

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes*, eds M. Cole, V. John-Steiner, S. Scribner, and E. Souberman. Cambridge, MA: Harvard University Press.

Waring, R., and Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Read. Foreign Lang.* 15, 130–163.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Lang. Test.* 11:197. doi: 10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Lang. Test.* 15, 263–287. doi: 10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511732997

Weir, C. (1990). *Communicative Language Testing*. New Jersey: Prentice Hall.

Weir, C. J. (2005a). *Language Testing and validation :An Evidence-Based Approach*. London: Palgrave Macmillan.

Weir, C. J. (2005b). *Language Testing and Validation*. London: Palgrave Macmillan. doi: 10.1057/9780230514577

Weir, C. J., Vidakovics, I, and Galaczi, E. D. (2013). *Measured constructs. A history of Cambridge English Language Examinations 1913–2012. Studies in Language Testing* 37. Cambridge: Cambridge University Press.

Wilson, J., Roscoe, R., and Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assess. Writing* 34, 16–36. doi: 10.1016/j.asw.2017.08.002

Winke, P. (2011). Investigating the Reliability of the Civics Component of the U.S. naturalization test. *Language Assessment Q.* 8, 317–341. doi: 10.1080/15434303.2011.614031

Winke, P., and Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. Rubric: an eye-movement study. *Assess. Writing* 25, 38–54. doi: 10.1016/j.asw.2015.05.002

Wiseman, C. S. (2012). Rater effects: ego engagement in rater decision-making. *Assess. Writing* 17, 150–173. doi: 10.1016/j.asw.2011.12.001

Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Hawai'i: University Hawai'i Press.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511519772

Wright, B. D., and Stone, M. H. (1979). *Best Test Design*. Chicago, IL: Mesa Press.

Xi, X. (2010a). How do we go about investigating test fairness?. *Lang. Test.* 27, 147–170. doi: 10.1177/0265532209349465

Xi, X. (2010b). Automated scoring and feedback systems: where are we and where are we heading? *Lang. Test.* 27, 291–300. doi: 10.1177/0265532210364643

Zhang, L., Goh, C. C. M., and Kunnan, A. J. (2014). Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: a multi-sample SEM approach. *Lang. Assess. Q. Int. J.* 11, 76–102. doi: 10.1080/15434303.2013.853770

Zhang, Y., and Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: competing or complementary constructs?. *Lang. Test.* 28, 31–50. doi: 10.1177/0265532209360671

Zhang, Z., and Poucke, S. V. (2017). Citations for randomized controlled trials in sepsis literature: the halo effect caused by journal impact factor. *PLoS ONE* 12:e0169398. doi: 10.1371/journal.pone.0169398

Zhao, C. G. (2017). Voice in timed L2 argumentative essay writing. *Assess. Writing* 31, 73–83. doi: 10.1016/j.asw.2016.08.004

Zheng, Y., and Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from assessing writing (2000–2018). *Assess. Writing* 42:100421. doi: 10.1016/j.asw.2019.100421

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor is currently editing co-organizing a Research Topic with one of the author VA, and confirms the absence of any other collaboration.

Copyright © 2020 Aryadoust, Zakaria, Lim and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.