FRONTIERS IN LANGUAGE ASSESSMENT AND TESTING

EDITED BY: Vahid Aryadoust, Thomas Eckes and Yo In'nami
PUBLISHED IN: Frontiers in Psychology and Frontiers in Communication

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-301-9
DOI 10.3389/978-2-88966-301-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

FRONTIERS IN LANGUAGE ASSESSMENT AND TESTING

Topic Editors:
Vahid Aryadoust, Nanyang Technological University, Singapore
Thomas Eckes, Ruhr University Bochum, Germany
Yo In'nami, Chuo University, Japan

Citation: Aryadoust, V., Eckes, T., In'nami, Y., eds. (2020). Frontiers in Language Assessment and Testing. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-301-9

Table of Contents

frontiers in Psychology

SYSTEMATIC REVIEW
published: 27 February 2020
doi: 10.3389/fpsyg.2020.00330

Assessing Speaking Proficiency: A Narrative Review of Speaking Assessment Research Within the Argument-Based Validation Framework

Jason Fan[1]* *and Xun Yan*[2]*

[1] *Language Testing Research Centre, The University of Melbourne, Melbourne, VIC, Australia,* [2] *Department of Linguistics, University of Illinois at Urbana-Champaign, Champaign, IL, United States*

This paper provides a narrative review of empirical research on the assessment of speaking proficiency published in selected journals in the field of language assessment. A total of 104 published articles on speaking assessment were collected and systematically analyzed within an argument-based validation framework (Chapelle et al., 2008). We examined how the published research is represented in the six inferences of this framework, the topics that were covered by each article, and the research methods that were employed in collecting the backings to support the assumptions underlying each inference. Our analysis results revealed that: (a) most of the collected articles could be categorized into the three inferences of *evaluation, generalization,* and *explanation*; (b) the topics most frequently explored by speaking assessment researchers included the constructs of speaking ability, rater effects, and factors that affect spoken performance, among others; (c) quantitative methods were more frequently employed to interrogate the inferences of *evaluation* and *generalization* whereas qualitative methods were more frequently utilized to investigate the *explanation* inference. The paper concludes with a discussion of the implications of this study in relation to gaining a more nuanced understanding of task- or domain-specific speaking abilities, understanding speaking assessment in classroom contexts, and strengthening the interfaces between speaking assessment, and teaching and learning practices.

Keywords: speaking assessment, speaking proficiency, argument-based validation framework, research methods, narrative review

OPEN ACCESS

Edited by:
Vahid Aryadoust,
Nanyang Technological University, Singapore

Reviewed by:
Stefan O'Grady,
Bilkent University, Turkey
Hongwen Cai,
Guangdong University of Foreign Studies, China
Alireza Ahmadi,
Shiraz University, Iran

***Correspondence:**
Jason Fan
jinsong.fan@unimelb.edu.au
Xun Yan
xunyan@illinois.edu

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

Received: *20 November 2019*
Accepted: *11 February 2020*
Published: *27 February 2020*

Citation:
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

INTRODUCTION

Speaking is a crucial language skill which we use every day to communicate with others, to express our views, and to project our identity. In today's globalized world, speaking skills are recognized as essential for international mobility, entrance to higher education, and employment (Fulcher, 2015a; Isaacs, 2016), and are now a major component in most international and local language examinations, due at least in part to the rise of the communicative movement in language teaching and assessment (Fulcher, 2000). However, despite its primacy in language pedagogy and assessment, speaking has been considered as an intangible construct which is challenging to conceptualize and assess in a reliable and valid manner. This could be attributable to the dynamic

and context-embedded nature of speaking, but may be also due to the various forms that it can assume (e.g., monolog, paired conversation, group discussion) and the different conditions under which speaking happens (e.g., planned or spontaneous) (e.g., Luoma, 2004; Carter and McCarthy, 2017). When assessing speaking proficiency, multiple factors come into play which potentially affect test takers' performance and subsequently their test scores, including task features, interlocutor characteristics, rater effects, and rating scale, among others (McNamara, 1996; Fulcher, 2015a). In the field of language assessment, considerable research attention and efforts have been dedicated to researching speaking assessment. This is evidenced by the increasing number of research papers with a focus on speaking assessment that have been published in the leading journals in the field.

This prolonged growth in speaking assessment research warrants a systematic review of major findings that can help subsequent researchers and practitioners to navigate the plethora of published research, or provide them with sound recommendations for future explorations in the speaking assessment domain. Several review or position papers are currently available on speaking assessment, either reviewing the developments in speaking assessment more broadly (e.g., Ginther, 2013; O'Sullivan, 2014; Isaacs, 2016) or examining a specific topic in speaking assessment, such as pronunciation (Isaacs, 2014), rating spoken performance (Winke, 2012) and interactional competence (Galaczi and Taylor, 2018). Needless to say, these papers are valuable in surveying related developments in speaking proficiency assessment and sketching a broad picture of speaking assessment for researchers and practitioners in the field. Nonetheless, they typically adopt the traditional literature review approach, as opposed to the narrative review approach that was employed in this study. According to Norris and Ortega (2006, p. 5, cited in Ellis, 2015, p. 285), a narrative review aims to "scope out and tell a story about the empirical territory." Compared with traditional literature review which tends to rely on a reviewer's subjective evaluation of the important or critical aspects of the existing knowledge on a topic, a narrative review is more objective and systematic in the sense the results are usually based on the coding analysis of the studies that are collected through applying some pre-specified criteria. Situated within the argument-based validation framework (Chapelle et al., 2008), this study is aimed at presenting a narrative review of empirical research on speaking assessment published in two leading journals in the field of language assessment, namely, *Language Testing* (LT) and *Language Assessment Quarterly* (LAQ). Through following the systematic research procedures of narrative review (e.g., Cooper et al., 2019), we survey the topics of speaking assessment that have been explored by researchers as well as the research methods that have been utilized with a view to providing recommendations for future speaking assessment research and practice.

THEORETICAL FRAMEWORK

Emerging from the validation of the revised Test of English as a Foreign Language (TOEFL), the argument-based validation framework adopted in this study represents an expansion of Kane's (2006) argument-based validation model, which posits that a network of inferences needs to be verified to support test score interpretation and use. A graphic display of this framework is presented in **Figure 1**. As shown in this figure, the plausibility of six inferences need to be verified to build a validity argument for a language test, including: *domain definition, evaluation, generalization, explanation, extrapolation,* and *utilization*. Also included in the framework are the key warrants that license each inference and its underlying assumptions. This framework was adopted as the guiding theoretical framework of this review study in the sense that each article collected for this study was classified into one or several of these six inferences in the framework. As such, it is necessary to briefly explain these inferences in **Figure 1** in the context of speaking assessment. The explanation of the inferences, together with their warrants and assumptions, is largely based on Chapelle et al. (2008) and Knoch and Chapelle (2018). To facilitate readers' understanding of these inferences, we use the TOEFL speaking test as an example to provide an illustration of the warrants, key assumptions, and backings for each inference.

The first inference, *domain definition,* links the target language use (TLU) domain to test takers' observed performance on a speaking test. The warrant supporting this inference is that observation of test takers' performance on a speaking test reveals the speaking abilities and skills required in the TLU domain. In the case of the TOEFL speaking test, the TLU domain is the English-medium institutions of higher education. Therefore, the plausibility of this inference hinges on whether observation of test takers' performance on the speaking tasks reveals essential academic speaking abilities and skills in English-medium universities. An important assumption underlying this inference is that speaking tasks that are representative of language use in English-medium universities can be identified and simulated. Backings in support of this assumption can be collected through interviews with academic English experts to investigate speaking abilities and skills that are required in English-medium universities.

The warrant for the next inference, *evaluation*, is that test takers' performance on the speaking tasks is evaluated to provide observed scores which are indicative of their academic speaking abilities. The first key assumption underlying this warrant is that the rating scales for the TOEFL speaking test function as intended by the test provider. Backings for this assumption may include: a) using statistical analyses (e.g., many-facets Rasch measurement, or MFRM) to investigate the functioning of the rating scales for the speaking test; and b) using qualitative methods (e.g., raters' verbal protocols) to explore raters' use of the rating scales for the speaking test. Another assumption for this warrant is that raters provide consistent ratings on each task of the speaking test. Backing for this assumption typically entails the use of statistical analyses to examine rater reliability on each task of the speaking test. The third assumption is that detectable rater characteristics do not introduce systematic construct-irrelevant variance into their ratings of test takers' performance. Bias analyses are usually implemented to explore whether certain rater

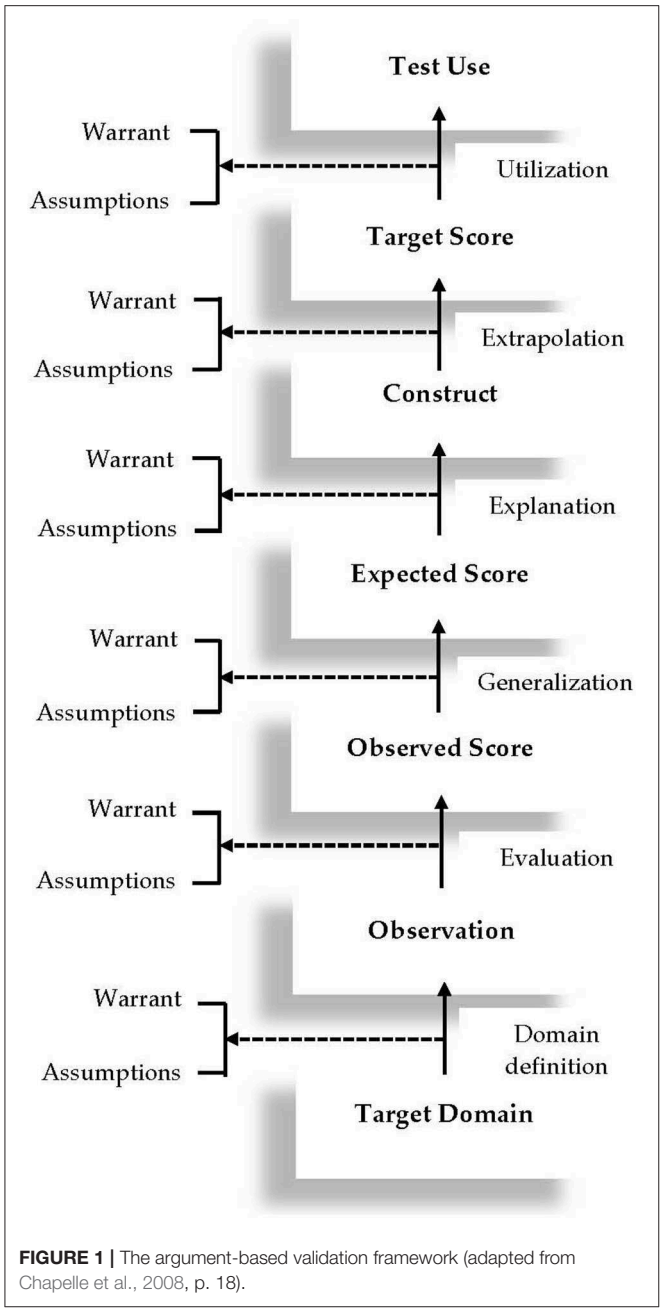


FIGURE 1 | The argument-based validation framework (adapted from Chapelle et al., 2008, p. 18).

characteristics (e.g., experience, L1 background) interact with test taker characteristics (e.g., L1 background) in significant ways.

The third inference is *generalization*. The warrant that licenses this inference is that test takers' observed scores reflect their expected scores over multiple parallel versions of the speaking test and across different raters. A few key assumptions that underlie this inference include: a) a sufficient number of tasks are included in the TOEFL speaking test to provide stable estimates of test takers' speaking ability; b) multiple parallel versions of the speaking test feature similar levels of difficulty and tap into similar academic English speaking constructs; and c) raters rate test takers' performance consistently at the test level. To support the first assumption, generalizability theory (i.e., G-theory) analyses can be implemented to explore the number of tasks that is required to achieve the desired level of reliability. For the second assumption, backings can be collected through: (a) statistical analyses to ascertain whether multiple parallel versions of the speaking test have comparable difficulty levels; and (b) qualitative methods such as expert review to explore whether the parallel versions of the speaking test tap into similar academic English speaking constructs. Backing of the third assumption typically entails statistical analyses of the scores that raters have awarded to test takers to examine their reliability at the test level.

The fourth inference is *explanation*. The warrant of this inference is that test takers' expected scores can be used to explain the academic English speaking constructs that the test purports to assess. The key assumptions for this inference include: (a) features of the spoken discourse produced by test takers on the TOEFL speaking test can effectively distinguish L2 speakers at different proficiency levels; (b) the rating scales are developed based on academic English speaking constructs that are clearly defined; and (c) raters' cognitive processes when rating test takers' spoken performance are aligned with relevant theoretical models of L2 speaking. Backings of these three assumptions can be collected through: (a) discourse analysis studies aiming to explore the linguistic features of spoken discourse that test takers produce on the speaking tasks; (b) expert review of the rating scales to ascertain whether they reflect relevant theoretical models of L2 speaking proficiency; and (c) rater verbal protocol studies to examine raters' cognitive processes when rating performance on the speaking test.

The fifth inference in the framework is *extrapolation*. The warrant that supports this inference is that the speaking constructs that are assessed in the speaking test account for test takers' spoken performance in English-medium universities. The first key assumption underlying this warrant is that test takers' performance on the TOEFL speaking test is related to their ability to use language in English-medium universities. Backing for this assumption is typically collected through correlation studies, that is, correlating test takers' performance on the speaking test with an external criterion representing their ability to use language in the TLU domains (e.g., teachers' evaluation of students' speaking proficiency of academic English). The second key assumption for *extrapolation* is that raters' use of the rating scales reflects how spoken performance is evaluated in English-medium universities. For this assumption, qualitative studies can be undertaken to compare raters' cognitive processes with those of linguistic laypersons in English-medium universities such as subject teachers.

The last inference is *utilization*. The warrant supporting this inference is that the speaking test scores are communicated in appropriate ways and are useful for making decisions. The assumptions that underlie the warrant include: (a) the meaning of the TOEFL speaking test scores is clearly interpreted by relevant stakeholders, such as admissions officers, test takers, and teachers; (b) cut scores are appropriate for making relevant decisions about students; and (c) the TOEFL speaking test has a positive influence on English teaching and learning. To collect the backings for the first assumption, qualitative studies

(e.g., interviews, focus groups) can be conducted to explore stakeholders' perceptions of how the speaking test scores are communicated. For the second assumption, standard setting studies are often implemented to interrogate the appropriateness of cut scores. The last assumption is usually investigated through test washback studies, exploring how the speaking test influences English teaching and learning practices.

The framework was used in the validation of the revised TOEFL, as reported in Chapelle et al. (2008), as well as in low-stakes classroom-based assessment contexts (e.g., Chapelle et al., 2015). According to Chapelle et al. (2010), this framework features several salient advantages over other alternatives. First, given the dynamic and context-mediated nature of language ability, it is extremely challenging to use the definition of a language construct as the basis for building the validity argument. Instead of relying on an explicit definition of the construct, the argument-based approach advocates the specification of a network of inferences, together with their supporting warrants and underlying assumptions that link test takers' observed performances to score interpretation and use. This framework also makes it easier to formulate validation research plans. Since every assumption is associated with a specific inference, research questions targeting each assumption are developed 'in a more principled way as a piece of an interpretative argument' (Chapelle et al., 2010, p. 8). As such, the relationship between validity argument and validation research becomes more apparent. Another advantage of this approach to test validation it that it enables the structuring and synthesis of research results into a logical and coherent validity argument, not merely an amalgamation of research evidence. By so doing, it depicts the logical progression of how the conclusion from one inference becomes the starting point of the next one, and how each inference is supported by research. Finally, by constructing a validity argument, this approach allows for a critical evaluation of the logical development of the validity argument as well as the research that supports each inference. In addition to the advantages mentioned above for test validation research, this framework is also very comprehensive, making it particularly suitable for this review study.

By incorporating this argument-based validation framework in a narrative review of the published research on speaking assessment, this study aims to address the following research questions:

RQ1. How does the published research on speaking assessment represent the six inferences in the argument-based validation framework?
RQ2. What are the speaking assessment topics that constituted the focus of the published research?
RQ3. What methods did researchers adopt to collect backings for the assumptions involved in each inference?

METHODS

This study followed the research synthesis steps recommended by Cooper et al. (2019), including: (1) problem formation; (2) literature search; (3) data evaluation; (4) data analysis; (5) interpretation of results; and (6) public presentation. This section includes details regarding article search and selection, and methods for synthesizing our collected studies.

Article Search and Selection

We collected the articles on speaking assessment that were published in *LT* from 1984[1] to 2018 and *LAQ* from 2004 to 2018. These two journals were targeted because: (a) both are recognized as leading high-impact journals in the field of language assessment; (b) both have an explicit focus on assessment of language abilities and skills. We understand that numerous other journals in the field of applied linguistics or educational evaluation also publish research on speaking and its assessment. Admittedly, if the scope of our review extends to include more journals, the findings might be different; however, given the high impact of these two journals in the field, a review of their published research on speaking assessment in the past three decades or so should provide sufficient indication of the directions in assessing speaking proficiency. This limitation is discussed at the end of this paper.

The PRISMA flowchart in **Figure 2** illustrates the process of article search and selection in this study. A total of 120 articles were initially retrieved through manually surveying each issue in the electronic archives of the two journals, containing all articles published in *LT* from 1984 to 2018 and *LAQ* from 2004 to 2018. Two inclusion criteria were applied: (a) the article had a clear focus on speaking assessment. Articles that targeted the whole language test involving multiple skills were not included; (b) the article reported an empirical study in the sense that it investigated one or more aspects of speaking assessment through the analysis of data from either speaking assessments or designed experimental studies.

Through reading the abstracts carefully, 13 articles were excluded from our analysis, with two special issue editorials and 11 review or position papers. A further examination of the remaining 107 articles revealed that three of them involved multiple language skills, suggesting a lack of primary focus on speaking assessment. These three articles were therefore excluded from our analysis, yielding 104 studies in our collection. Of the 104 articles, 73 (70.19%) were published in *LT* and 31 (29.81%) were published in *LAQ*. All these articles were downloaded in PDF format and imported into NVivo 12 (QSR, 2018) for analysis.

Data Analysis

To respond to RQ1, we coded the collected articles into the six inferences in the argument-based validation framework based on the focus of investigation for each article, which was determined by a close examination of the abstract and research questions. If the primary focus did not emerge clearly in this process, we read the full text. As the coding progressed, we noticed that some articles had more than one focus, and therefore should be coded into multiple inferences. For instance, Sawaki (2007) interrogated several aspects of an L2 speaking test that were considered as essential to its construct validity, including the

[1] *LT* and *LAQ* made their debut in 1984 and 2004, respectively.

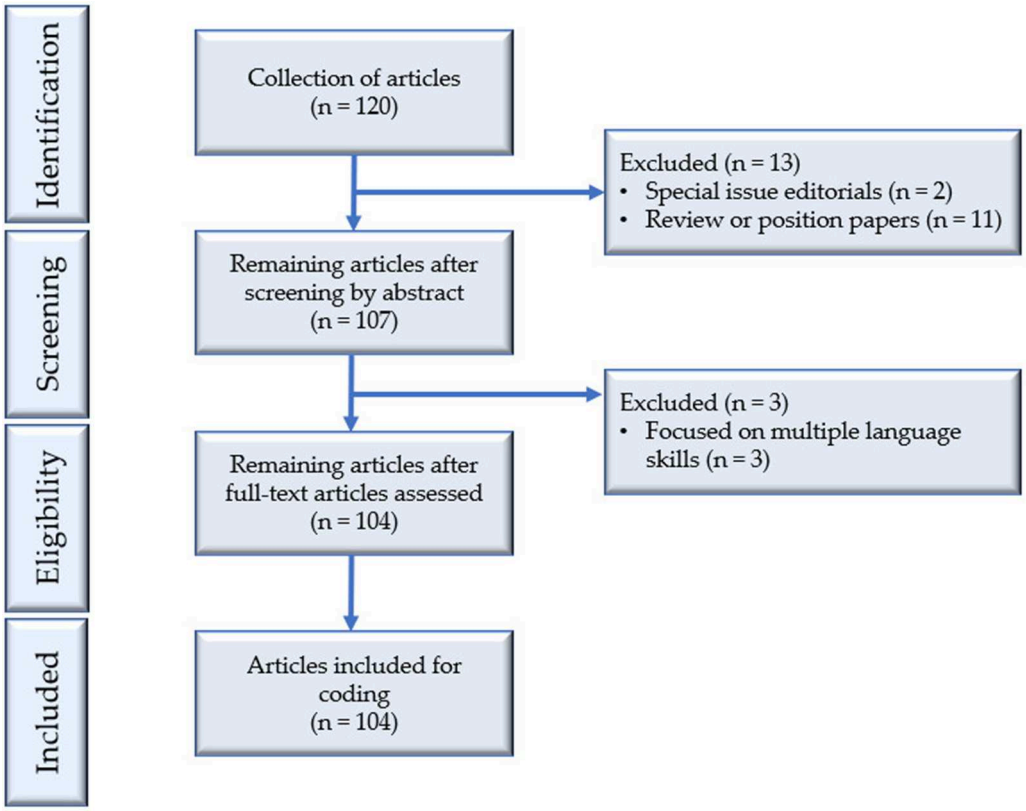


FIGURE 2 | PRISMA flowchart of article search and collection.

interrelationships between the different dimensions of spoken performance and the reliability of test scores. The former was considered as pertinent to the *explanation* inference, as it explores the speaking constructs through the analysis of test scores; the latter, however, was deemed more relevant to the *generalization* inference, as it concerns the consistency of test scores at the whole test level (Knoch and Chapelle, 2018). Therefore, this article was coded into both *explanation* and *generalization* inference.

To answer RQ2, the open coding method (Richards, 1999) was employed to explore the speaking assessment topics that constituted the focus of each article in our collection. This means that a coding scheme was not specified *a prior*; rather, it was generated through examining the abstracts or full texts to determine the topics and subtopics. RQ3 was investigated through coding the research methods that were employed by speaking assessment researchers. A broad coding scheme consisting of three categories was employed to code the research methods: (a) quantitatively oriented; (b) qualitatively oriented; and (c) mixed methods with both quantitative and qualitative orientations. Next, the open coding method was adopted to code the specific methods that were utilized under each broad category. Matrix coding analysis (Miles et al., 2014) was subsequently implemented in NVivo to explore the relationships between the speaking assessment topics, research methods and the six inferences in the argument-based validation framework. This would enable us to sketch the broad patterns of: (a) which topics on speaking assessment tended be investigated under each of the six inferences; (b) which research methods were frequently employed to collect the backings for the assumptions that underlie each inference.

The coding process underwent three iterative stages to ensure the reliability of the coding results. First, both authors coded 10 articles selected randomly from the dataset independently and then compared their coding results. Differences in coding results were resolved through discussion. Next, the first author coded the rest of the articles in NVivo, using the coding scheme that was generated during the first stage while adding new categories as they emerged from the coding process. Finally, the second author coded 20 articles (19.23%) which were randomly selected from the dataset, using the coding scheme that was determined during the second stage. Intercoder agreement was verified through calculating Cohen's kappa statistic in NVivo ($k = 0.93$), which suggested satisfactory coding reliability.

RESULTS AND DISCUSSION

Overall, our coding results indicate that a wide range of research was conducted of speaking assessment to interrogate the six inferences in the argument-based validation framework. These studies cover a variety of research topics, employing quantitative,

qualitative, and mixed research methods. In this section, we describe and discuss the analysis results through showcasing the broad patterns that emerged from our coding process. Illustrative studies are used as appropriate to exemplify the research that was undertaken in assessing speaking proficiency.

Representation of the Published Research in the Six Inferences

Table 1 presents the representation of the published research in the six inferences. As indicated in this table, most of our collected articles were categorized into the three inferences of *evaluation* ($n = 42$, 40.38%), *generalization* ($n = 42$, 40.38%), and *explanation* ($n = 50$, 48.08%); in contrast, a much smaller number of studies targeted the other three inferences of *domain description* ($n = 4$, 3.85%), *extrapolation* ($n = 7$, 6.73%), and *utilization* ($n = 5$, 4.81%). Despite the highly skewed representation of the published research in the six inferences, the findings were not entirely surprising. According to the argument-based validation framework (Chapelle et al., 2008), backings in support of the assumptions that underlie the three inferences of *evaluation, generalization,* and *explanation* relate to almost all key components in the assessment of speaking proficiency, including rater effects, rating scale, task features or administration conditions, interlocutor effects in speaking tasks such as paired oral, interview or group discussion, and features of produced spoken discourse. These components essentially represent the concerns surrounding the development, administration, and validation of speaking assessment (e.g., McNamara, 1996; Fulcher, 2015a). Take the inference of *evaluation* as an example. In the argument-based validation framework, this inference pertains to the link from the observation of test takers' performance on a speaking test to their observed scores. As mentioned previously (see section Theoretical Framework), backings in support of the key assumptions underlying this inference include an evaluation of rating scales as well as rater effects at the task level. Given the pivotal role that raters and rating scales play in speaking assessment (e.g., Eckes, 2011), it is not surprising to observe a reasonably high proportion of studies exploring the plausibility of this inference. Almost half of our collected articles ($n = 50$, 48.08%) interrogated the *explanation* inference. This finding can be interpreted in relation to the centrality of understanding the construct in language test development and validation (e.g., Alderson et al., 1995; Bachman and Palmer, 1996), which lies at the core of the *explanation* inference.

One possible explanation for the limited research on *domain description* is related to the journals that formed the basis for this review study. Both *LT* and *LAQ* have an explicit focus on language assessment, whereas in many cases, exploration of language use in TLU domains, which is the focus of *domain description*, might be reported as needs assessment studies in test development reports, which were beyond the purview of this study. Another plausible explanation, as pointed out by one of the reviewers, might lie in the lack of theoretical sophistication regarding this inference. The reason why few studies targeted the *extrapolation* inference might be attributable to the challenges in pinpointing the external criterion measure, or in collecting valid data to represent test takers' ability to use language in TLU domains. These challenges could be exacerbated in the case of speaking ability due to its intangible nature, the various forms that it may assume in practice, and the different conditions under which it happens. Similarly, very few studies focused on the *utilization* inference which concerns the communication and use of test scores. This could relate to the fact that test washback or impact studies have to date rarely focused exclusively on speaking assessment (Yu et al., 2017). Speaking assessment researchers should consider exploring this avenue of research in future studies, particularly against the backdrop of the increasingly extensive application of technology in speaking assessment (Chapelle, 2008).

TABLE 1 | Representation of the published research in the six inferences ($n = 104$).

Inferences	Number of articles	
	n	**%**
• Domain description	4	3.85
• Evaluation	42	40.38
• Generalization	42	40.38
• Explanation	50	48.08
• Extrapolation	7	6.73
• Utilization	5	4.81

Thirty-nine articles (37.50%) were coded into multiple inferences, of which 34 (32.69%) were coded into two inferences and five (4.81%) into three inferences.

Speaking Assessment Topics

Table 2 presents the matrix coding results of speaking assessment topics and the six inferences in the argument-based validation framework. It should be noted that some of the frequency statistics in this table are over-estimated because, as mentioned previously, some articles were coded into multiple inferences; however, this should not affect the general patterns that emerged from the results in a significant way. The topics that emerged from our coding process are largely consistent with the themes that Fulcher (2015a) identified in his review of speaking assessment research. One noteworthy difference is many-facets Rasch measurement (MFRM), a topic in Fulcher (2015a) but was coded as a research method in our study (see section Research Methods). In what follows, we will focus on the three topics which were most frequently investigated by speaking assessment researchers, namely, speaking constructs, rater effects, and factors that affect speaking performance, as examples to illustrate the research that was undertaken of speaking assessment.

Speaking Constructs

Table 2 shows that "speaking constructs" ($n = 47$) is the topic that was investigated most frequently in our collected studies. Matrix coding results indicate that this topic area appears most frequently under the inference of *explanation* ($n = 39$, 37.50%). The importance of a clear understanding of the construct cannot be overemphasized in language test development and

TABLE 2 | Matrix coding results of inferences and speaking assessment topics ($n = 104$).

Topics	Domain description		Evaluation		Generalization		Explanation		Extrapolation		Utilization	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
(1) Speaking constructs ($n = 47$)	0	0.00	8	7.69	11	10.58	39	37.50	5	4.81	0	0.00
(2) Rater effects ($n = 39$)	0	0.00	27	25.96	23	22.12	14	13.46	0	0.00	0	0.00
(3) Factors that affect test performance ($n = 30$)	0	0.00	9	8.65	19	18.27	13	12.50	0	0.00	0	0.00
(4) Speaking test design ($n = 14$)	2	1.92	9	8.65	4	3.85	8	7.69	1	0.96	0	0.00
(5) Test score generalizability ($n = 7$)	0	0.00	3	2.88	7	6.73	2	1.92	0	0.00	0	0.00
(6) Rating scale evaluation ($n = 6$)	2	1.92	4	3.85	2	1.92	2	1.92	0	0.00	0	0.00
(7) Test use ($n = 5$)	0	0.00	0	0.00	0	0.00	0	0.00	1	0.96	5	4.81

(1) The total number in the left column exceeds 104 because some articles were coded into multiple topic areas; (2) the total numbers of the rows exceed the numbers reported in the left column because some articles in these topic areas were coded into multiple inferences.

validation (e.g., Alderson et al., 1995; Bachman and Palmer, 1996). Indeed, construct definition forms the foundation of several highly influential test validation frameworks in the field (e.g., Messick, 1989; Weir, 2005). Our analysis indicates that considerable research has been dedicated to disentangling various speaking constructs. Two topics that feature prominently in this topic area are the analysis of spoken discourse and interactional competence.

A common approach to investigate the speaking constructs is through the analysis of produced spoken discourse (Carter and McCarthy, 2017), usually focusing on linguistic features that can distinguish test takers at different proficiency levels such as complexity, accuracy, and fluency (e.g., Iwashita, 2006; Gan, 2012; Bosker et al., 2013). Research in this area can provide substantial evidence concerning speaking proficiency. Iwashita (2006), for instance, examined the syntactic complexity of the spoken performance of L2 Japanese learners. Results reveal that learner' oral proficiency could be predicted significantly by several complexity indicators, including T-unit length, the number of clauses per T-unit, and the number of independent clauses per T-unit. In another discourse analysis study, Gan (2012) probed the syntactic complexity of test takers' spoken discourse and examined the relationship between syntactic complexity and task type in L2 speaking assessment. Gan's results show that, compared with the group interaction task, test takers' discourses on the individual presentation task featured longer T-units and utterances as well as significantly greater number of T-units, clauses, verb phrases and words. These discourse analysis studies have implications for understanding speaking proficiency as well as its development and maturity among L2 learners.

International competence (IC) is yet another topic which features prominently in this topic area. Despite the recognized need of including IC in speaking assessment (e.g., Kramsch, 1986, McNamara, 1997), how it should be conceptualized remains a contentious issue. Research has shown that this construct consists of multiple dimensions which is susceptible to the influence of a range of personal cognitive and contextual factors (Galaczi and Taylor, 2018). Our review suggests that IC was approached through analyzing test takers' spoken discourse as well as exploring raters' perspectives. Galaczi (2008), for instance, performed elaborate analyses of test takers' spoken discourse on the paired speaking task in the First Certificate in English (FCE) speaking test. The results led the researcher to conclude that test takers' interactions primarily featured three patterns on paired oral assessment tasks: collaborative, parallel and blended interaction (i.e., a mixture of collaborative/parallel or collaborative/asymmetric features). In a more recent study, Lam (2018) analyzed test takers' spoken discourse on a school-based group oral speaking assessment for the Hong Kong Diploma of Secondary Education (HKDSE) English Language Examination. Instead of exploring IC more broadly, as in Galaczi (2008), this study targeted a particular feature of IC, namely, producing responses contingent on previous speakers' contributions. The analyses pointed to three kinds of conversational actions that underpinned a response contingent on previous speaker's contributions: formulating previous speakers' contributions, accounting for (dis)agreement with previous speakers' ideas and extending previous speakers' ideas.

Some other studies explored the construct of IC from raters' perspectives. A typical study was reported by May (2011) who explored the features that were salient to raters on a paired speaking test. The study identified a repertoire of features which were salient to raters, and hence were potentially integral to the IC construct. Such features include, for example, the ability to manage a conversation, ask for opinion or clarification, challenge or disagree with an interactional partner, and demonstrate effective body language, and interactive listening. While suggesting that IC is a highly complex and slippery construct, these studies have significant implications for clarifying the IC construct and promoting its valid operationalization in speaking assessment. The findings are particularly meaningful in the context where interactive tasks are increasingly used in speaking assessment.

Rater Effects

Raters play a significant role in speaking assessment; their performance is affected by a host of non-linguistic factors, which are often irrelevant to the speaking constructs of interest, hence causing construct-irrelevant variance (Messick, 1989) or contamination (AERA et al., 2014). Not surprisingly, the next topic area that was most frequently explored by speaking assessment researchers is rater effects ($n = 39$). The studies

that focused on this topic were mostly classified into the two inferences of *evaluation* ($n = 27$, 25.96%) and *generalization* (n =23, 22.12%). Knoch and Chapelle (2018) applied the argument-based validation framework to the analysis of rater effects and rating processes in language assessment research. They observed that several important aspects of rater effects could be mapped onto *evaluation* and *generalization* inferences. The key assumptions of the *evaluation* inference relate to the raters' consistency at the task level, the bias that raters display against task types or other aspects of the assessment situation, and the impact of raters' characteristics on the ratings that they assign. When it comes to the *generalization* inference, the key assumptions largely concern raters' consistency at the whole test level and the number of raters that is required to achieve the desired level of consistency. Research on rater effects has significant implications for enhancing both the validity and fairness of speaking assessment (e.g., McNamara et al., 2019).

Two topics that feature prominently in the study of rater effects are the impact of raters' characteristics on their rating behaviors and rater cognition, that is, the cognitive processes that raters engage when assigning scores to a spoken performance. Raters' characteristics such as language background, experience and qualifications may have appreciable impact on their ratings. This topic has attracted considerable research attention as it has implications for test fairness and rater training programs. One such study was reported by Kim (2009) who examined and compared the rating behaviors of native and non-native English teachers when assessing students' spoken performance. The results indicate that native-speaker (NS) and non-native-speaker (NNS) teachers on the whole exhibited similar severity levels and internal consistency; however, in comparison with NNS teachers, NS teachers provided more detailed and elaborate comments on students' performance. The findings generally concur with Zhang and Elder (2011) who compared the rating behaviors of NS and NNS teachers in the context of the College English Test - Spoken English Test (CET-SET), a large-scale high-stakes speaking test in China. Instead of focusing on raters' L1 background, Winke et al. (2013) examined whether raters' accent familiarity, defined as their L2 learning experience, constituted a potential source of bias when they rated test takers' spoken performance. In other words, if a rater studies Chinese as his or her L2, is he or she biased toward test takers who have Chinese as their L1? Their findings indicate that the raters with Spanish or Chinese as their L2 were significantly more lenient toward L1 Spanish and Chinese test takers than they were toward those from other L1 backgrounds. However, in both cases, the effect sizes were small, suggesting that such effect had minimal impact in practice. The results are largely consistent with some other studies in our collection (e.g., Yan, 2014; Wei and Llosa, 2015), which explored a similar topic.

Rater cognition or rating processes constitute yet another important topic under the topic area of "rater effects". Studies along this line are typically implemented through analyzing raters' verbal protocols to explore their cognitive processes when applying the rating criteria or assigning scores to a spoken performance. Research into raters' cognitive processes can generate valuable insights into the validity of the rating scales as well as the speaking constructs that are being assessed in a speaking test. Findings from these studies have important implications for the revision of rating scales, improving rater training programs, and enhancing the validity and usefulness of the speaking test in focus. In a qualitative study, Kim (2015) explored the rating behaviors of three groups of raters with different levels of experience on an L2 speaking test by analyzing their verbal reports of rating processes. The study revealed that the three groups of raters exhibited varying uses of the analytic rating scales, hence suggesting that experience was an important variable affecting their rating behaviors. Furthermore, an analysis of their performance over time revealed that the three groups of raters demonstrated different degrees of improvement in their rating performance. It should be noted that several studies in our collection examined raters' rating processes with a view to either complementing or accounting for the quantitative analyses of speaking test scores. For instance, both Kim (2009) and Zhang and Elder (2011), two studies which were reviewed previously, investigated raters' rating processes, and the findings significantly enriched our understanding of the rating behaviors of raters from different backgrounds.

Factors That Affect Spoken Performance

The third topic area that emerged from our coding process is "factors that affect spoken performance" ($n = 30$). As shown in **Table 3**, most of the studies in this topic area were classified into the inference of *generalization* ($n = 19$, 18.27%). This is understandable as factors such as task features, administration conditions, and planning time might affect the generalizability of speaking test scores. Indeed, understanding factors that affect test performance has long since been one of the central concerns for language assessment research as a whole (e.g., Bachman, 1990; Bachman et al., 1995). Research along this line has implications for speaking test development and implementation, and for test score interpretation and use. Our coding analyses indicate that a range of factors have been explored by speaking assessment researchers, of which 'interlocutor effects' features most prominently. This could be related to the increasingly widespread use of interviews, paired oral or group discussion tasks to assess speaking ability in applied linguistics and language pedagogy. A notable advantage with these assessment formats lies in the unscripted and dynamic nature of the interactions involved, which is key to increasing the authenticity of speaking assessments. Nonetheless, interlocutor characteristics, such as gender, proficiency levels, personality, and styles of interaction might have considerable impact on test takers' spoken performance, thus impinging on the validity, fairness and overall usefulness of these tasks.

An earlier study on interlocutor effects was reported by McNamara and Lumley (1997) who examined the potential impact of interlocutor characteristics on test scores in the context of the Occupational English Test (OET), a high-stakes speaking test for health professionals in Australia. Their study indicated that interlocutor characteristics had some influence on the ratings that test takers received. For example, they found that raters tended to compensate for interlocutors' incompetence in conducting the speaking test; in other words, if an interlocutor

TABLE 3 | Matrix coding results of research methods and inferences ($n = 104$).

Methods	Domain description		Evaluation		Generalization		Explanation		Extrapolation		Utilization	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
QUAN ($n = 50$)	0	0.00	21	20.19	27	25.96	18	17.31	3	2.88	1	0.96
• ANOVA or regression ($n = 34$)	0	0.00	13	12.50	14	13.46	15	14.42	2	1.92	3	2.88
• Rasch ($n = 28$)	0	0.00	19	18.27	20	19.23	9	8.65	0	0.00	0	0.00
• Correlation ($n = 20$)	1	0.96	7	6.73	9	8.65	10	9.62	4	3.85	1	0.96
• G-theory ($n = 7$)	0	0.00	4	3.85	7	6.73	2	1.92	0	0.00	0	0.00
• EFA ($n = 5$)	0	0.00	4	3.85	3	2.88	3	2.88	0	0.00	1	0.96
• SEM ($n = 5$)	0	0.00	2	1.92	3	2.88	2	1.92	1	0.96	0	0.00
• Cluster analysis ($n = 2$)	0	0.00	1	0.96	0	0.00	1	0.96	0	0.00	1	0.96
QUAL ($n = 23$)	3	2.88	4	3.85	3	2.88	16	15.38	2	1.92	0	0.00
• Discourse analysis ($n = 25$)	1	0.96	6	5.78	6	5.78	20	19.23	2	1.92	0	0.00
• Interview/Focus group ($n = 11$)	4	3.85	6	5.78	2	1.92	4	3.85	1	0.96	0	0.00
• Written comments ($n = 11$)	0	0.00	5	4.81	6	5.78	5	4.81	0	0.00	2	1.92
• Verbal protocols ($n = 10$)	1	0.96	7	6.73	2	1.92	5	4.81	0	0.00	0	0.00
• Eye-tracking ($n = 1$)	0	0.00	0	0.00	0	0.00	1	0.96	0	0.00	0	0.00
MIXED ($n = 31$)	1	0.96	17	16.35	12	11.53	16	15.38	2	1.92	4	3.85

(1) QUAL, Quantitative; QUAL, Qualitative; G-theory, Generalizability theory; EFA, Exploratory factor analysis; SEM, Structural equation modeling; (2) the total number in the left column exceeds 104 because some articles used multiple methods; (3) the total numbers of the rows exceed the numbers reported in the left column because some articles using these methods were coded into multiple inferences.

was perceived as less competent, test takers tended to receive higher ratings than expected. In addition, they also observe that an interlocutor's ability to build rapport with test takers had a positive effect on the ratings that test takers received. In another study, Brown (2003) probed the effects of interlocutor characteristics on test takers' performance in the context of a conversational interview. She performed elaborate analyses of the interactions between the interviewers (i.e., interlocutors) and test takers, revealing that the interlocutors differed quite significantly in terms of: (a) how they structured topical sequences; (b) their questioning technique; and (c) how they provided feedback and built rapport with test takers. Further analyses uncovered that interviewer styles had quite significant impact on the ratings that test takers received. Resonating with McNamara and Lumley (1997), the findings of this study again call for the reconceptualization of speaking proficiency.

Several other studies focused on the effects of interaction partners in paired or group oral tasks on spoken performance. (Ockey, 2009), for instance, investigated the potential effects of group member's assertiveness levels on spoken performance on a group discussion task. Results confirmed that test takers' assertiveness levels had an impact on the scores that they received. Specifically, assertive test takers were awarded higher scores than expected when grouped with non-assertive test takers; this trend, however, was reversed when they were grouped with test takers with similar assertiveness levels. A plausible explanation could be that raters viewed assertive test takers more positively when other members in the groups were non-assertive, whereas more negatively when other group members, who were also assertive, competed to be the leaders in the interactions. This study reiterates the co-constructed nature of speaking proficiency. Despite the research that has been undertaken of interlocutor effects, controversy remains as to whether this variation is part of the speaking construct and therefore should be incorporated in the design of a speaking test or it should be controlled to such an extent that it poses minimal threat to the reliability and fairness of speaking test scores (Fulcher, 2015a).

In addition to the three topics above, researchers also explored speaking test design ($n = 14$) in terms of the task features (e.g., Wigglesworth and Elder, 2010; Ahmadi and Sadeghi, 2016) and the use of technology in speaking test delivery (e.g., Nakatsuhara et al., 2017; Ockey et al., 2017). The next topic is test score generalizability ($n = 7$), typically investigated through G-theory analysis (e.g., Lee, 2006; Sawaki, 2007; Xi, 2007). Furthermore, six studies in our collection evaluated the rating scales for speaking assessments, including comparing the effectiveness of different types of rating scales (e.g., Hirai and Koizumi, 2013), and examining whether a rating scale functioned as intended by the test developer (e.g., Isaacs and Thomson, 2013). Finally, five studies focused on the use of speaking assessments, mainly relating to test takers' perceptions of speaking assessments (e.g., Scott, 1986; Qian, 2009) and standard setting studies to determine the cut scores for certain purposes (e.g., Pill and McNamara, 2016).

Research Methods

Table 3 presents the matrix coding results of research methods and inferences. As indicated in this table, quantitative research methods were more frequently employed by speaking assessment researchers ($n = 50$), in comparison with qualitative methods ($n = 23$). It is worth noting that a number of studies ($n = 31$) utilized mixed methods design, which features a combination of both quantitative and qualitative orientations.

Table 3 indicates that quantitative methods were most frequently used to collect backings in support of the *evaluation* (n = 21, 20.19%) and *generalization* inferences (n = 27, 25.96%). This finding can be interpreted in relation to the key assumptions that underlie these two inferences (see section Theoretical Framework). According to the argument-based validation framework, the assumptions of these two inferences largely concern rater consistency at task and whole-test level, the functioning of the rating scales, as well as the generalizability of speaking test scores across tasks and raters. Understandably, quantitative methods are widely used to collect the backings to test these assumptions. In addition to the overall representation of quantitative methods in speaking assessment research, we also went a step further to examine the use of specific quantitative methods. As shown in **Table 3**, while traditional data analysis methods such as ANOVA or regression (n = 34) continued to be utilized, mainly in the interrogation of the inferences of *evaluation* (n = 13, 12.50%), *generalization* (n = 14, 13.46%), and *explanation* (n = 15, 14.42%), Rasch analysis methods were also embraced by speaking assessment researchers (n = 28). Note that Rasch analysis is an overarching term which encompasses a family of related models, among which the many-facets Rasch model (MFRM) is frequently used in speaking assessment (e.g., McNamara and Knoch, 2012). As an extension of the basic Rasch model, the MFRM allows for the inclusion of multiple aspects or facets in a speaking context (e.g., rater severity, task difficulty, difficulty of rating scales). Furthermore, compared with traditional data analysis methods such as correlation and ANOVA which can only provide results at the group level, the MFRM can provide both group- and individual-level statistics (Eckes, 2011). This finding concurs with Fulcher (2015a) who identified the MFRM as an important theme in speaking assessment. It also resonates with the observation of Fan and Knoch (2019, p. 136) who commented that Rasch analysis has indeed become "one of the default methods or analysis techniques to examine the technical quality of performance assessments." The power of Rasch analysis in speaking assessment research is best illustrated by studies such as Bonk and Ockey (2003), Eckes (2005), and Winke et al. (2013), among others, all of which examined rater effects on speaking assessments in different contexts. Finally, G-theory (n = 7) and structural equation modeling (n = 5), two complicated quantitative methods, were also utilized by speaking assessment researchers.

In terms of qualitative research methods, discourse analysis is the one which was most frequently employed by speaking assessment researchers (n = 25). Matrix coding results indicate that this method features most prominently under the inference of *explanation* (n = 20, 19.23%). This finding is aligned with the key assumptions that underlie the *explanation* inference, namely, (a) features of the spoken discourse produced by test takers can effectively distinguish L2 speakers at different proficiency levels, and (b) raters' cognitive processes are consistent with the theoretical models of L2 speaking, both entailing the use of discourse analysis method to explore test takers' spoken responses and raters' rating processes. Importantly, our analysis results indicate that conversation analysis (CA) was the method that appeared frequently under the category of "discourse analysis." This is best represented by studies such as Galaczi (2008), Lam (2018), and Roever and Kasper (2018), all endeavoring to elucidate the construct of interactional competence. As a data analysis method, CA provides speaking researchers with a principled and intricate approach to analyze the interactions between test takers and examiners in interview, paired oral, or group discussion tasks. **Table 3** shows that some other qualitative methods were also quite frequently used by speaking researchers, including interview/focus groups (n = 11), written comments (n = 11), and verbal protocol reports (n = 10). These research methods were typically adopted following the quantitative analyses of test takers' scores, which explains the increasingly widespread use of mixed methods in speaking assessment research (n = 31). The finding could find resonance in the observation that mixed method research has been gaining momentum in language assessment research more broadly (e.g., Turner, 2013; Jang et al., 2014; Moeller et al., 2016). As shown in **Table 3**, mixed-methods design is most frequently employed to collect backings in support of the inferences of *evaluation* (n = 17, 16.35%) and *explanation* (n = 16, 15.38%). For the *evaluation* inference, mixed method design was often utilized to research rater effects where quantitative and qualitative analyses were used sequentially to examine rating results and processes. When it comes to the *explanation* inference, researchers tended to use a combination of quantitative and qualitative analyses to explore the differences in test takers' speaking scores as well as the spoken discourse that they produced.

CONCLUSIONS AND IMPLICATIONS

In this study, we conducted a narrative review of published empirical research on assessing speaking proficiency within the argument-based validation framework (Chapelle et al., 2008). A total of 104 articles on speaking assessment were collected from *LT* (1984–2018) and *LAQ* (2004–2018), two highly influential journals in the field of language assessment. Following the coding of the collected articles, matrix coding analyses were utilized to explore the relationships between the speaking assessment topics, research methods, and the six inferences in the argument-based validation framework.

The analysis results indicate that speaking assessment was investigated from various perspectives, primarily focusing on seven broad topic areas, namely, the constructs of speaking ability, rater effects, factors that affect spoken performance, speaking test design, test score generalizability, rating scale evaluation, and test use. The findings of these studies have significantly enriched our understanding of speaking proficiency and how assessment practice can be made more reliable and valid. In terms of research methods, it was revealed that quantitative research methods were most frequently utilized by speaking assessment researchers, a trend which was particularly pronounced in the inferences of *evaluation* and *generalization*. Though traditional quantitative methods such as ANOVA, regression, and correlation continued to be employed, Rasch analysis played a potent role in researching speaking assessment.

In comparison, qualitative methods were least frequently used, mainly for the interrogation of the *explanation* inference. Mixed-methods design, recognized as "an alternative paradigm" (Jang et al., 2014, p. 123), ranked in the middle in terms of frequency, suggesting its increasingly widespread use in speaking assessment research. This is noteworthy when it comes to the *evaluation* and *explanation* inference.

Despite the abundance of research on speaking assessment and the variety of research topics and methods that emerged from our coding process, we feel that there are several areas which have not been explored extensively by language assessment researchers, and therefore warrant more future research endeavors. First, more studies should be conducted to interrogate the three inferences of *domain description*, *extrapolation*, and *utilization* in the argument-based validation framework. As indicated in our study, only a small fraction of studies have been dedicated to examining these three inferences in comparison with *evaluation*, *generalization*, and *explanation* (see **Table 2**). Regarding *domain description*, we feel that more research could be undertaken to understand task- and domain-specific speaking abilities and communicative skills. This would have significant implications for enhancing the authenticity of speaking assessment design, and for constructing valid rating scales for evaluating test takers' spoken performance. The thick description approach advocated by Fulcher et al. (2011) could be attempted to portray a nuanced picture of speaking ability in the TLU domains, especially in the case of Language for Specific Purposes (LSP) speaking assessment. When it comes to the *extrapolation* inference, though practical difficulties in collecting speaking performance data in the TLU domains are significant indeed, new research methods and perspectives, as exemplified by the corpus-based register analysis approach taken by LaFlair and Staples (2017), could be attempted in the future to enable meaningful comparisons between spoken performance on the test and speaking ability in TLU domains. In addition, the judgments of linguistic layperson may also be employed as a viable external criterion (e.g., Sato and McNamara, 2018). The *utilization* inference is yet another area that language assessment researchers might consider exploring in the future. Commenting on the rise of computer-assisted language assessment, Chapelle (2008, p. 127) argued that "test takers have needed to reorient their test preparation practices to help them prepare for new test items." As such, it is meaningful for language assessment researchers to explore the impact of computer-mediated speaking assessments and automated scoring systems on teaching and learning practices.

Next, though the topic of speaking constructs has attracted considerable research attention from the field, as evidenced by the analysis results of this study, it seems that we are still far from achieving a comprehensive and fine-grained understanding of speaking proficiency. The results of this study suggest that speaking assessment researchers tended to adopt a psycholinguistic approach, aiming to analyze the linguistic features of produced spoken discourse that distinguish test takers at different proficiency levels. However, given the dynamic and context-embedded nature of speaking, there is a pressing need for a sociocultural perspective to better disentangle the speaking constructs. Using pronunciation as an example, Fulcher (2015b) argued convincingly the inadequacy of a psycholinguistic approach in pronunciation assessment research; rather, a sociocultural approach, which aims to demystify rationales, linguistic or cultural, that underlie (dys)fluency, could significantly enrich our understanding of the construct. Such an approach should be attempted more productively in future studies. In addition, as the application of technology is becoming prevalent in speaking assessment practices (Chapelle, 2008), it is essential to explore whether and to what extent technology mediation has altered the speaking constructs and the implications for score interpretation and use.

We also found that several topics were under-represented in the studies that we collected. Important areas that received relatively limited coverage in our dataset include: (a) classroom-based or learning-oriented speaking assessment; (b) diagnostic speaking assessment; and (c) speaking assessment for young language learners (YLLs). The bulk of the research in our collection targeted large-scale high-stakes speaking assessments. This is understandable, perhaps, because results on these assessments are often used to make important decisions which have significant ramifications for stakeholders. In comparison, scanty research attention has been dedicated to speaking assessments in classroom contexts. A recent study reported by May et al. (2018) aimed to develop a learning-oriented assessment tool for interactional competence, so that detailed feedback could be provided about learners' interactional skills in support of their learning. More research of such a nature is needed in the future to reinforce the interfaces between speaking assessment with teaching and learning practices. In the domain of L2 writing research, it has been shown that simply using analytic rating scales does not mean that useful diagnostic feedback can be provided to learners (Knoch, 2009). Arguably, this also holds true for speaking assessment. In view of the value of diagnostic assessment (Lee, 2015) and the call for more integration of learning and assessment (e.g., Alderson, 2005; Turner and Purpura, 2015), more research could be conducted to develop diagnostic speaking assessments so that effective feedback can be provided to promote L2 learners' speaking development. Finally, young language learners (YLLs) have specific needs and characteristics which have implications for how they should be assessed (e.g., McKay, 2006). This is particularly challenging with speaking assessment in terms of task design, implementation and score reporting. This topic, however, has rarely been explored by speaking assessment researchers and therefore warrants more future research.

In terms of research methods, we feel that speaking assessment researchers should consider exploring more the potentials of qualitative methods which are well-suited to investigating an array of research questions related to speaking assessment. Our analysis results indicate that despite the quite frequent use of traditional qualitative methods such as interviews and focus groups, new qualitative methods that are supported by technology (e.g., eye-tracking) have only recently been utilized by speaking assessment researchers. For example, a recent study by Lee and Winke (2018) demonstrated the use of eye-tracking in speaking assessment through examining

test-takers' cognitive processes when responding to computer-based speaking assessment tasks. Eye-tracking is advantageous in the sense that as opposed to traditional qualitative methods such as introspective think-aloud protocols, it causes minimal interference of the test taking process. Our final comment concerns the use of mixed-methods design in speaking assessment research. Despite it being applied quite frequently in researching speaking assessment, it appears that only the sequential explanatory design (i.e., the use of qualitative research to explain quantitative findings) was usually employed. Speaking assessment researchers may consider other mixed methods design options (e.g., convergent parallel design or embedded mixed methods design, see Moeller et al., 2016) to investigate more complex research questions in speaking assessment.

We acknowledge a few limitations with this study. As mentioned previously, we targeted only two highly influential journals in the field of language assessment, namely, *LT* and *LAQ* while aware that numerous other journals in applied linguistics or educational evaluation also publish research on speaking and its assessment. As such, caution needs to be exercised when interpreting the relevant research findings that emerged from this study. Future studies could be undertaken to include more journals and other publication types (e.g., research reports, PhD dissertations) to depict a more representative picture of speaking assessment research. In addition, given the sheer volume of published research on speaking assessment available, our research findings can only be presented as indications of possible trends of the wider publishing context, as reflected in the specific articles we explored. Arguably, the findings might be more revealing if we zoomed in on a few key topics in speaking assessment (e.g., rater effects, speaking constructs), analyzed specific studies on these topics in detail, and compared their findings. Finally, it would be worthwhile to explore how the research on some key topics in speaking assessment has been evolving over time. Such analysis could have provided a valuable reference point to speaking assessment researchers and practitioners. Such a developmental trend perspective, however, was not incorporated in our analysis and could be attempted in future research.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

JF designed the study, collected and coded the data, and drafted the article. XY collected and coded the data, and drafted this article together with JF.

ACKNOWLEDGMENTS

The preparation of this manuscript was supported by the National Planning Office for Philosophy and Social Sciences (NPOPSS) of the People's Republic of China under the project title Reform of English speaking assessment and its impact on the teaching of English speaking (19BYY234). We would like to thank Angela McKenna and three reviewers for their insightful and perspicacious comments on the previous draft of this article.

REFERENCES

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Ahmadi, A., and Sadeghi, E. (2016). Assessing English language learners' oral performance: a comparison of monologue, interview, and group oral test. *Lang. Assess. Q.* 13, 341–358. doi: 10.1080/15434303.2016.1236797

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The interface between Learning and Assessment*. London: Bloomsbury.

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

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford England; New York, NY: Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K., and Choi, I.-C. (1995). *An Investigation into the Comparability of Two Tests of English as a Foreign Language, Vol. 1* (Cambridge: Cambridge University Press).

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

Bonk, W. J., and Ockey, G. J. (2003). A many-facet rasch analysis of the second language group oral discussion task. *Lang. Test.* 20, 89–110. doi: 10.1191/0265532203lt245oa

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., and De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Lang. Test.* 30, 159–175. doi: 10.1177/0265532212455394

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

Carter, R., and McCarthy, M. (2017). Spoken grammar: where are we and where are we going? *Appl. Linguistics* 38, 1–20. doi: 10.1093/applin/amu080

Chapelle, C. A. (2008). *Utilizing Technology in Language Assessment Encyclopedia of Language and Education, Vol. 7* (New York, NY: Springer), 123–134.

Chapelle, C. A., Cotos, E., and Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Lang. Test.* 32, 385–405. doi: 10.1177/0265532214565386

Chapelle, C. A., Enright, M. K., and Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educ. Meas. Iss. Pract.* 29, 3–13. doi: 10.1111/j.1745-3992.2009.00165.x

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; London: Routledge; Taylor & Francis Group.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis*. New York, NY:Russell Sage Foundation.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: a many-facet Rasch analysis. *Lang. Assess. Q: Int. J.* 2, 197–221. doi: 10.1207/s15434311laq0203_2

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement*. Frankfurt: Peter Lang.

Ellis, R. (2015). Introduction: complementarity in research syntheses. *Appl. Linguistics* 36, 285–289. doi: 10.1093/applin/amv015

Fan, J., and Knoch, U. (2019). Fairness in language assessment: what can the Rasch model offer? *Pap. Lang. Test. Assess.* 8, 117–142. Available online at: http://www.altaanz.org/uploads/5/9/0/8/5908292/8_2_s5_fan_and_knoch.pdf

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System* 28, 483–497. doi: 10.1016/S0346-251X(00)00033-6

Fulcher, G. (2015a). Assessing second language speaking. *Lang. teaching* 48, 198–216. doi: 10.1017/S0261444814000391

Fulcher, G. (2015b). *Re-Examining Language Testing: a Philosophical and Social Inquiry*. New York, NY: Routledge.

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

Galaczi, E., and Taylor, L. (2018). Interactional competence: conceptualisations, operationalisations, and outstanding questions. *Lang. Assess. Q.* 15, 219–236. doi: 10.1080/15434303.2018.1453816

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: the case of the *First Certificate in English examination*. *Lang. Assess. Q.* 5, 89–119. doi: 10.1080/15434300801934702

Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Lang. Assess. Q.* 9, 133–151. doi: 10.1080/15434303.2010.516041

Ginther, A. (2013). "Assessment of speaking," in *The Encyclopedia of Applied Linguistics,* ed C. A. Chapelle (New York, NY: Blackwell Publishing Ltd.), 1–7.

Hirai, A., and Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Lang. Assess. Q.* 10, 398–422. doi: 10.1080/15434303.2013.824973

Isaacs, T. (2014). "Assessing pronunciation," in *The Companion to Language Assessment, Vol. 1,* ed A. J. Kunnan (New York, NY: John Wiley & Sons), 140–155.

Isaacs, T. (2016). "Assessing speaking," in *Handbook of Second Language Assessment, Vol. 12,* eds D. Tsagari and J. Banerjee (Boston, MA; Berlin, Germany: De Gruyter), 131–146.

Isaacs, T., and Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: revisiting research conventions. *Lang. Assess. Q.* 10, 135–159. doi: 10.1080/15434303.2013.769545

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Lang. Assess. Q. Int. J.* 3, 151–169. doi: 10.1207/s15434311laq0302_4

Jang, E. E., Wagner, M., and Park, G. (2014). Mixed methods research in language testing and assessment. *Annu. Rev. Appl. Linguistics* 34, 123–153. doi: 10.1017/S0267190514000063

Kane, M. T. (2006). "Validation," in *Educational Measurement,* ed R. L. Brennan (Westport, CT: American Council on Education), 17–64.

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

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral english performance: a mixed methods approach. *Lang. Test.* 26, 187–217. doi: 10.1177/0265532208101010

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

Knoch, U., and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Lang. Test.* 35, 477–499. doi: 10.1177/0265532217710049

Kramsch, C. (1986). From language proficiency to interactional competence. *Mod. Lang. J.* 70, 366–372. doi: 10.1111/j.1540-4781.1986.tb05291.x

LaFlair, G. T., and Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Lang. Test.* 34, 451–475. doi: 10.1177/0265532217713951

Lam, D. M. (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

Lee, S., and Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Lang. Test.* 35, 239–269. doi: 10.1177/0265532217704009

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Lang. Test.* 23, 131–166. doi: 10.1191/0265532206lt325oa

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Lang. Test.* 32, 299–316. doi: 10.1177/0265532214565387

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

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

May, L., Nakatsuhara, F., Lam, D. M., and Galaczi, E. (2018). "Learning-oriented assessment feedback for interactional competence: developing a checklist to support teachers and learners," *Paper presented at the Language Testing Research Colloquium* (Auckland).

McKay, P. (2006). *Assessing Young Language Learners*. Cambridge: Cambridge University Press.

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

McNamara, T., and Knoch, U. (2012). The Rasch wars: the emergence of Rasch measurement in language testing. *Lang. Test.* 29, 553–574. doi: 10.1177/0265532211430367

McNamara, T., Knoch, U., and Fan, J. (2019). *Fairness, Justice and Langauge Assessment*. Oxford: Oxford University Press.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: whose performance? *App. Linguistics* 18, 446–466. doi: 10.1093/applin/18.4.446

McNamara, T. F., and Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational I settings. *Lang. Test.* 14, 140–156. doi: 10.1177/026553229701400202

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

Miles, M. B., Huberman, A. M., and Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook, 3rd Edn*. Thousand Oaks, CA: Sage.

Moeller, A. K., Creswell, J. W., and Saville, N. (2016). Second *Language Assessment and Mixed Methods Research*. Cambridge: Cambridge University Press.

Nakatsuhara, F., Inoue, C., Berry, V., and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: a mixed-methods study. *Lang. Assess. Q.* 14, 1–18. doi: 10.1080/15434303.2016.1263637

Norris, J. M., and Ortega, L. (2006). *Synthesizing Research on Language Learning and Teaching, Vol. 13,* Amsterdam; Philadelphia, PA: John Benjamins Publishing.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Lang. Test.* 26, 161–186. doi: 10.1177/0265532208101005

Ockey, G. J., Gu, L., and Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Lang. Assess. Q.* 14, 346–359. doi: 10.1080/15434303.2017.1400036

O'Sullivan, B. (2014). "Assessing speaking," in *The Companion to Language Assessment, Vol. 1,* ed A. J. Kunnan (Chichester: Wiley and Sons Inc), 156–171.

Pill, J., and McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Lang. Test.* 33, 217–234. doi: 10.1177/0265532215607402

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: affective effects on test takers. *Lang. Assess. Q.* 6, 113–125. doi: 10.1080/15434300902800059

QSR (2018). *NVivo Qualitative Data Analysis Software*. Melbourne, VIC: QSR International Pty Ltd.

Richards, L. (1999). *Using NVivo in Qualitative Research*. Thousand Oaks, CA: Sage.

Roever, C., and Kasper, G. (2018). Speaking in turns and sequences: interactional competence as a target construct in testing speaking. *Lang. Test.* 35, 331–355. doi: 10.1177/0265532218758128

Sato, T., and McNamara, T. (2018). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Appl. Linguist.* 40, 894–916. doi: 10.1093/applin/amy032

Sawaki, Y. (2007). Construct validation of analytic rating scale in speaking assessment: reporting a score profile and a composite. *Lang. Test.* 24, 355–390. doi: 10.1177/0265532207077205

Scott, M. L. (1986). Student affective reactions to oral language tests. *Lang. Test.* 3, 99–118. doi: 10.1177/026553228600300105

Turner, C., and Purpura, J. (2015). "Learning-oriented assessment in the classroom," in *Handbook of Second Language Assessment,* eds D. Tsagari and J. Banerjee (Berlin; Boston, MA: DeGruyter Mouton), 255–273.

Turner, C. E. (2013). "Classroom assessment," in *The Routledge Handbook of Language Testing*, eds G. Fulcher and F. Davidson (London; New York, NY: Routledge), 79–92.

Wei, J., and Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Lang. Assess. Q.* 12, 283–304. doi: 10.1080/15434303.2015.1037446

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

Wigglesworth, G., and Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Lang. Assess. Q.* 7, 1–24. doi: 10.1080/15434300903031779

Winke, P. (2012). "Rating oral language," in *The Encyclopedia of Applied Linguistics,* ed C. A. Chapelle (New York, NY: Blackwell Publishing Ltd.).

Winke, P., Gass, S., and Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Lang. Test.* 30, 231–252. doi: 10.1177/0265532212456968

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Lang. Test.* 24, 251–286. doi: 10.1177/0265532207076365

Yan, X. (2014). An examination of rater performance on a local oral Englissh proficiency test: a mixed-methods approach. *Lang. Test.* 31, 501–527. doi: 10.1177/0265532214536171

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., et al. (2017). Preparing for the speaking tasks of the TOEFL iBT® test: an investigation of the journeys of Chinese test takers. *ETS Res. Rep. Ser.* 2017, 1–59. doi: 10.1002/ets2.12145

Zhang, Y., and Elder, C. (2011). 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

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 declared a past collaboration with one of the authors JF.

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fpsyg.2020.00494

Relating Lexical and Syntactic Knowledge to Academic English Listening: The Importance of Construct Representation

*Hongwen Cai**

Center for Linguistics and Applied Linguistics,Guangdong University of Foreign Studies, Guangzhou, China

This study aims to resolve contradictory conclusions on the relative importance of lexical and syntactic knowledge in second language (L2) listening with evidence from academic English. It was hypothesized that when lexical and syntactic knowledge is measured in auditory receptive tasks contextualized in natural discourse, the measures will be more relevant to L2 listening, so that both lexical and syntactic knowledge will have unique contributions to L2 listening. To test this hypothesis, a quantitative study was designed, in which lexical and syntactic knowledge was measured via partial dictation, an auditory receptive task contextualized in a discourse context. Academic English listening was measured via a retired IELTS listening test. A group of 258 college-level native Chinese learners of English completed these tasks. Pearson correlations showed that both lexical and syntactic measures correlated strongly with English listening (r = 0.77 and r = 0.67 respectively). Hierarchical regression analyses showed that both measures jointly explained 62% of the variance in the listening score and that each measure had its unique contribution. These results demonstrated the importance of considering construct representation substantially and using measures that well reflect constructs in practical research.

Keywords: lexical knowledge, syntactic knowledge, academic English, second language listening, construct

OPEN ACCESS

Edited by:
Vahid Aryadoust,
Nanyang Technological University, Singapore

Reviewed by:
Rie Koizumi,
Juntendo University, Japan
Matthew P. Wallace,
University of Macau, China

***Correspondence:**
Hongwen Cai
hwcai@gdufs.edu.cn

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

Received: *18 November 2019*
Accepted: *02 March 2020*
Published: *31 March 2020*

Citation:
Cai H (2020) Relating Lexical and Syntactic Knowledge to Academic English Listening: The Importance of Construct Representation. Front. Psychol. 11:494. doi: 10.3389/fpsyg.2020.00494

INTRODUCTION

It is not uncommon for researchers to report different or even contradictory findings when they try to address the same issue in second language (L2) studies. A case in point is the relative importance of lexical and syntactic knowledge in L2 listening comprehension, where mixed findings have been reported, some alluding to the sole significance of lexical knowledge while downplaying or masking the role of syntactic knowledge (Mecartty, 2000; Stæhr, 2009; Vandergrift and Baker, 2015; Cheng and Matthews, 2018; Matthews, 2018), others rendering the relative importance unclear (Oh, 2016; Wang and Treffers-Daller, 2017) or resorting to the more general construct of linguistic knowledge and avoiding the distinction between lexical and syntactic knowledge (Andringa et al., 2012).

The different findings and their relative generalizability may be attributed to various factors, such as the characteristics of the participant groups, the treatments delivered to the participants, the properties of the instruments used, and the settings of the studies. Among these factors, the instruments are of vital importance to the construct validity of the studies

(Shadish et al., 2002; Shadish, 2010). In the case of lexical and syntactic knowledge, the mixed findings may be attributed, at least partially, to the variety of instruments used in different studies, which are based on different theoretical underpinnings and construct representations. It is, therefore, important to understand the construct definitions and specifications related to the various instruments before the contradictions between findings can be resolved.

Following this reasoning, the relevant studies will be reviewed to compare the various construct representations of lexical and syntactic knowledge under a uniform framework, with a view to identifying the key features that are central to L2 listening. On the basis of these key features, an observational study will be designed in the academic English context to quantify the relative importance of lexical and syntactic knowledge in L2 listening, with a view to resolving the contradictions between findings from earlier studies.

Literature Review

Lexical and Syntactic Processes in Listening

To establish a uniform framework for comparing the construct representations of lexical and syntactic knowledge, a brief account of psycholinguistic theories of language comprehension is inevitable. Fortunately, descriptions of the key stages of comprehension are more or less the same across the rich variations of models, such that a "basic" model can be conceptualized, comprising word-, sentence-, and discourse-level processes (Fernández and Cairns, 2018). A variation of this basic model often cited in applied linguistics literature is the three-stage cognitive model of Anderson (2015), consisting of *perception*, *parsing*, and *utilization*. The division into three stages is supported by neurological evidence, such that psychologists are able to identify the different combinations of brain regions involved in the three stages (Anderson, 2015).

In the L2 listening literature, the three stages are sometimes rephrased as *decoding*, *parsing*, and *meaning construction* (Field, 2011). In brief, the listener converts the acoustic-phonetic signal into words, relates the words syntactically for a combined meaning, and enriches the meaning by integrating it with meaning derived from earlier text, context, and background. While the three-stage model deals with the cognitive processes of listening comprehension, these processes depend upon a multitude of sources, linguistic, contextual, and schematic, among which linguistic sources can be further classified into phonetic, phonological, prosodic, lexical, syntactic, semantic, and pragmatic processes (Lynch, 2010).

The interplay between lexical and syntactic processes is an essential part of the cognitive processes in L2 listening. For one thing, word-level processes, such as the identification of a single word, depend on both lexical-semantic and syntactic cues in the context (Buck, 1991, 1994; Anderson, 2015). Neurologically, the speech signal of a word needs to be combined with information about its acoustic-phonological, syntactic, and conceptual semantic properties before it is recognized (Hagoort, 2013). Similarly, parsing also draws on both syntactic and lexical-semantic cues (Anderson, 2015). Underlying this process are two classes of neural mechanisms—lower-order bottom–up mechanisms that enable the lexical-semantic and morphosyntactic categorizations of the speech input and higher-order bottom-up and predictive top–down mechanisms that assign the complex relations between the elements detected in a sentence and integrate them into a conceptual whole (Skeide and Friederici, 2018). There is also evidence that the lexical-semantic and morphosyntactic categorizations are parallel processes, as they occur within 50–80 and 49–90 ms, respectively, after the onset of the speech signal (Friederici, 2012). In general, the three stages of listening comprehension are described as partly parallel and partly overlapping (Anderson, 2015).

Findings in L2 Listening Research

Findings in L2 listening research have mirrored the interplay between lexical and syntactic processes, though to different degrees. For example, some studies on L2 English and French listening focused solely on the correlation between lexical knowledge and L2 listening (Stæhr, 2009; Vandergrift and Baker, 2015; Cheng and Matthews, 2018; Matthews, 2018), reporting significant correlations between 0.39 and 0.73. With regard to the psycholinguistic theories reviewed above, the emphasis on lexical knowledge may have masked the contribution of syntactic knowledge to L2 listening. For the purpose of this study, however, these findings can be regarded as an initial indication of how strong the correlation between lexical knowledge and L2 listening can be.

That being said, the wide range of correlation estimates from these studies points to a potential problem—the inconsistent measures of the same construct. In fact, Cheng and Matthews (2018) deliberately compared the correlations of three different measures of lexical knowledge to L2 English listening and found that the correlations ranged between 0.39 and 0.71. The measure of lexical knowledge may also be confounded with other measures. In the study of Wang and Treffers-Daller (2017) on L2 English listening, the measures included a general language proficiency test, a vocabulary size test, and a questionnaire of metacognitive awareness. However, the general language proficiency test included a large number of items targeting lexical knowledge. Although their results of hierarchical regression analyses showed that general language proficiency and vocabulary both contributed uniquely to the variance of listening, the size of these contributions is subject to this confounding effect.

Another problem arises in empirical studies when the masking of syntactic knowledge in L2 listening is so conspicuous that it may negate the interplay between lexical and syntactic processes. Mecartty's (2000) study on L2 Spanish learners found that both lexical and syntactic knowledge were significantly correlated with L2 listening (r = 0.38 and r = 0.26 respectively), but his hierarchical regression analysis showed that only lexical knowledge explained 13% of the variance in listening. Although the addition of syntactic knowledge to the model increased the percentage of explained variance to 14%, the R^2 change was not statistically significant, and Mecartty concluded that syntactic knowledge had no unique contribution, which contradicts the psycholinguistic theories that both syntactic and lexical-semantic

cues are necessary for listening comprehension. Interestingly, the correlation between lexical and syntactic measures in Mecartty's (2000) study was estimated at $r = 0.34$, which, though significant, could be considered weak. This practically rules out the possibility of substantial overlap between the two measures being the cause of the insignificant R^2 change.

Other studies that were related to the contribution of lexical and syntactic knowledge to L2 listening yielded findings that agreed more with psycholinguistic theories. A common methodological feature among these studies is that L2 listening was regressed on multiple independent variables. Oh's (2016) study on L2 English listening included four measures of processing speed, two measures of grammar, and three measures of vocabulary. While she found significant correlations between listening and all but one of the processing speed measures, she reported that none of the three groups of measures explained a unique portion of variance in listening when the other two groups of measures were already entered into the hierarchical regression model, which seems to suggest that while lexical and syntactic knowledge both contribute to listening, they were not distinct processes. The assumption of joint contribution of lexical and syntactic knowledge agreed with psycholinguistic theories, but the lack of distinction between the two processes may be considered as construct confounding for the purpose of this study.

Among the studies published so far, Andringa et al. (2012) have captured the psycholinguistic sophistication of L2 listening most faithfully. These authors constructed a structural equation model to explain L2 Dutch listening with a multitude of variables, including three measures of linguistic knowledge, five measures of processing speed, and six cognitive measures of intelligence. They found that the latent construct of linguistic knowledge indicated by vocabulary, grammatical processing, and segmentation (of speech stream into words) explained 90% of the variance in listening. As no distinction was made between lexical and syntactic knowledge in their original model, this result cannot be compared to those discussed above. For this purpose, a hierarchical multiple regression was run by the author of this paper on the R package "lavaan" version 0.6-2 (Rosseel, 2012), using the correlation matrix and standard deviations reported in the original paper in lieu of raw data. The R^2 was estimated at 0.46 when L2 listening was regressed on vocabulary only and at 0.59 on grammar only, but increased to 0.67 when both predictors were entered. This result was closest to psycholinguistic findings in that lexical and syntactic sources both had unique contributions to the variance in L2 listening, and that the joint contribution of the two sources had significantly stronger explanatory power than single sources. Moreover, the lexical and syntactic measures were moderately correlated with each other ($r = 0.60$), which ruled out the threat of multicollinearity.

The Importance of Measures

With regard to the relative importance of lexical and syntactic knowledge in L2 listening, the most notable contradiction was between the findings of Andringa et al. (2012) and those reported by Mecartty (2000). Andringa et al. (2012) themselves noted that linguistic knowledge explained a larger percentage of variance in their study than in Mecartty's (2000) study. This is an important observation, in that 90% was considerably greater than 14%, which merits much further investigation. A comprehensive search for possible reasons may cover experimental factors or treatments, classificatory factors or personal variables, situational variables or settings, and outcome measures or observations (Shadish et al., 2002), as there are differences between the two studies in all these aspects. A heuristic search, however, could be based on the explanations of the authors themselves, who know the details of their study best. The first possible reason given by Andringa et al. (2012) was that measurement error had been cleared in the latent variable model they used, but even in raw score terms, lexical and syntactic knowledge explained 67% of variance in L2 listening, as this author's reanalysis demonstrated. Another factor Andringa et al. (2012) postulated was restriction of range in L2 proficiency in Mecartty's study. This could have attenuated correlations as well, but a closer examination of the coefficients of variation (CVs) yielded comparable results: 0.24 for L2 listening, 0.33 for lexical knowledge, and 0.15 for syntactic knowledge in Andringa et al. (2012) and 0.35 for L2 listening, 0.24 for lexical knowledge, and 0.25 for syntactic knowledge in Mecartty (2000). Calculated as the ratio of the standard deviation to the mean, the CV is a standardized measure of dispersion such that it can be directly compared between two studies. It follows that the comparable results can be taken as evidence that restriction of range was not a key factor that attenuated correlations in Mecartty's study. Therefore, the more probable reason that underlies the different findings in the two studies may be that the linguistic knowledge tests in Andringa et al. (2012) were "more pertinent to listening," whereas "grammatical knowledge was measured in a production task in Mecartty" (p. 70).

A more common term for pertinence is "construct relevance," and the pertinence issue raised by Andringa et al. (2012) is essentially the issue of construct representation (Bachman, 1990), which takes the form of measures of L2 listening, lexical knowledge, and syntactic knowledge. Underlying the reasoning of Andringa et al. (2012) is the assumption of how lexical and syntactic knowledge should be measured when examining their role in L2 listening. Though the measure of L2 listening itself is also a construct representation issue of no less importance, this paper will be confined to the discussion of the independent variables. A closer examination of the above-mentioned reason reveals two basic conceptual dichotomies familiar to most researchers in applied linguistics, the dichotomy of visual and auditory modes and the dichotomy of receptive and productive skills. Take the syntactic measure used in Andringa et al. (2012); the underlying construct was knowledge of the "distributional and combinatorial properties" of the Dutch language, most notably word order and agreement. A judgment task was designed, which required the participants to judge whether a fragment presented aurally was a possible sentence-initial string in Dutch, e.g., *Die stad lijkt heel* ("That city seems very") and *Precies ik weet* ("Exactly I know"). In comparison, Mecartty (2000) used two syntactic measures, the first of which was a sentence completion task aiming to measure "local-level

understanding of the grammatical features" of Spanish and requiring the participants to complete Spanish sentences with function words, such as *Me gusta aquel automóvil; _____ me gusta el rojo* ("I like that car; I ____ like the red one"). The second task was grammaticality judgment and error correction, aiming to measure knowledge of the "underlying rules" of Spanish, which required the participants to identify grammatical errors in Spanish sentences and correct them, such as **Compró el carro y transportó lo a su garaje* ("He bought *the* car and transported it to his garage"). In terms of the two dichotomies, the syntactic measure used in Andringa et al. (2012) was an auditory receptive task, whereas Mecartty's (2000) syntactic measures were visual productive tasks. As listening is an auditory receptive language use activity, it is natural to expect the former to be more strongly correlated with listening than the latter. More specifically, difficulty in a productive task does not necessarily transfer to a receptive task. For example, an L2 Spanish learner may have difficulty in choosing the right word to complete the sentence *Me gusta aquel automóvil; _____ me gusta el rojo,* but no difficulty at all in understanding the sentence presented in auditory mode, even if the incomplete sentence is presented. In contrast, identifying *Die stad lijkt heel* as a sentence-initial string in Dutch is helpful for understanding the meaning of the whole sentence containing the string, as word order is important in Dutch syntax (Oosterhoff, 2015) and thus a key factor for parsing (Anderson, 2015). In sum, a relevant measure of syntactic knowledge in L2 listening should take the form of an auditory receptive task with a focus on the key processes in parsing.

The same features apply to relevant measures of lexical knowledge in L2 listening, as evidenced by the three measures of lexical knowledge in Cheng and Matthews (2018). Intended for receptive vocabulary, their first measure took a multiple-choice format after the Vocabulary Levels Test of Nation (2001). Their second measure, targeting productive vocabulary, was adapted from the controlled-production vocabulary levels test of Laufer and Nation (1999) and required the participants to complete a sentence with the target word, whose initial letters were provided. Both measures were presented visually. The third measure of receptive[1] vocabulary took the form of a partial dictation task and required the participants to complete each sentence they heard with a missing word. All three measures covered the first 5,000 frequency levels of word lists extracted from the British National Corpus (BNC, Leech et al., 2001). The researchers correlated these measures with the scores from a retired IELTS listening test and estimated Pearson correlation at 0.39 for the visual receptive measure, 0.55 for the visual productive measure, and 0.71 for the auditory receptive measure. This is evidence that auditory receptive measures of lexical knowledge are most relevant to L2 listening, due to similarity in task characteristics between the lexical measure and the L2 listening test. Another dimension that may have contributed to the relevance of lexical measures may be the context provided. The visual productive measure in Cheng and Matthews (2018) was contextualized in single sentences, whereas the visual receptive measure was decontextualized, which may explain why the former was more strongly correlated with L2 listening (r = 0.55) than the latter (r = 0.39). A similar pattern is uncovered when comparing the correlation with L2 listening of the sentence-based visual receptive measure in Andringa et al. (2012) and the correlation with L2 listening of the decontextualized visual receptive measure in Mecartty (2000). Correlation was higher when lexical knowledge was measured in sentential context (r = 0.68) but lower when the measure was decontextualized (r = 0.34).

In sum, construct representation is a key factor that affects the findings on the relative importance of lexical and syntactic knowledge in L2 listening. Different measures of lexical and syntactic knowledge may represent different features of the constructs, which affects their relevance to L2 listening. More specifically, the visual/auditory, receptive/productive, and contextualized/decontextualized dichotomies may be key considerations for examining the contribution of lexical and syntactic knowledge to L2 listening.

Research Questions

To examine the above understanding, and to demonstrate the importance of theoretical underpinnings in practical research, the findings of Andringa et al. (2012) and Cheng and Matthews (2018) need to be replicated, with regard to the relationship between lexical and syntactic knowledge and L2 listening. Following the relevance principle, it is hypothesized that when lexical and syntactic knowledge is measured in auditory receptive tasks contextualized in natural discourse, the measures will be more relevant to L2 listening, so that both lexical and syntactic knowledge will have unique contributions to L2 listening. To test this hypothesis, the replication study should include both lexical and syntactic measures, similar to Andringa et al. (2012), but will be set in the academic English context, similar to Cheng and Matthews (2018). Two key research questions (RQs) are:

(1) How do lexical and syntactic knowledge correlate with L2 listening in the academic English context?
(2) Do lexical knowledge and syntactic knowledge have unique contributions to L2 listening in the academic English context?

RQ1 aims to measure the degree of association between lexical and syntactic knowledge and L2 listening. It is hypothesized that with a high level of relevance, Pearson correlations around 0.70 may be expected for both measures, similar to the findings with regard to the sentence-based visual receptive measure in Andringa et al. (2012) and the auditory receptive measure in Cheng and Matthews (2018). RQ2 is based on the psycholinguistic theories reviewed above, assuming that lexical and syntactic processes are distinct but contribute jointly to listening. It is hypothesized that both lexical knowledge and syntactic knowledge have unique contributions to L2 listening, and that the joint contribution of the two sources has stronger explanatory power.

[1]Cheng and Matthews (2018) called their third measure productive because participants had to respond in words. However, the core construct of this task was word recognition, which is receptive in nature. Responding in words does not change the receptive nature of partial dictation as a listening task (Cai, 2013).

MATERIALS AND METHODS

Participants

The study was conducted on 258 native Chinese learners of academic English as a second language. At the time of the study, they were first-year English majors enrolled in a university in China. Their mean raw score on the academic English listening test used in this study (15.33) converted to a band score (5) according to the official conversion table[2] close to the mean band score (5.89) on IELTS listening of test-takers from China in 2018[3].

Instruments

The measure of L2 academic English listening was a retired IELTS listening test published by Cambridge University Press. No participants had had access to the material prior to this study. The input material included the recordings of two monologs and two conversations, with 40 printed questions in four different formats—multiple-choice questions with four options, matching questions, judgment questions with three options (yes/no/not given or true/false/not given), and fill-in-the-blank questions in the form of questionnaires or forms to be filled. The monologs and conversations were recorded by native English speakers and were set in a variety of everyday social and educational/training contexts. These were designed to measure the ability to understand the main ideas and detailed factual information, the opinions and attitudes of speakers, and the purpose of an utterance and the development of ideas[4].

The measures of lexical and syntactic knowledge were integrated into a partial dictation task. Eight minutes of recording of the IELTS listening test were selected as the auditory input of the partial dictation, so that the same level of naturalness in spoken English can be achieved (Cai, 2013). The selection was based on the requirement that at least 10 words could be found in the recording on each of the three frequency-based levels, i.e., the 1,001–2,000 frequency range, the 2,001–3,000 frequency range, and the 3,001–5,000 frequency range, of the BNC (Leech et al., 2001). This decision was based on the findings of Matthews (2018) that each of these three levels had unique contributions to L2 listening performance, and on the practice to include 10 items from each 1,000-word-family level for testing vocabulary size (Nation and Beglar, 2007). Each blank was produced by taking away a single word or a two-to-three-word phrase. To give the participant sufficient time to write down the words and phrases they heard, the blanks were set apart at intervals of at least nine words, as the underlined segments (17, 18, and 19) in the following excerpt exemplify.

> *. . . I'd like to say at this point that you shouldn't worry (17) if this process doesn't work all that quickly – I mean occasionally there are postal problems, but most often the (18) hold-up is caused by references – the people you give as (19) referees, shall we say, take their time to reply.*

The interval between blanks No. 18 and No. 19, which both involved single words, was the minimum nine words. The interval between a blank for a missing phrase and another blank was typically longer to allow more time for writing. For example, the interval between blanks No. 17 and No. 18 in the above excerpt was 17 words. This excerpt also exemplifies the items included in the lexical and syntactic scales. The lexical scale was made up of 30 single words, 10 from each of the three levels described above. For example, the words "referees" (blank No. 19) and "hold-up" (blank No. 18) were from the 2,001–3,000 and 3,000–5,000 levels respectively. Each correctly spelled word was worth 1 point, so that the maximum score was 30 for the scale.

The syntactic scale consisted of 15 two-to-three-word phrases, such as "if this process" for blank No. 17, which is the initial string of a subordinate clause, consisting of the subordinating conjunction "if" and the noun phrase "this process," which serves as the subject of the clause. Identifying this phrase involves knowledge of word order and subordination, which are both important syntactic cues for parsing (Anderson, 2015). The other syntactic features involved in the items included ellipses, noun conjunctions, pronouns, parentheses, emphatic expressions, etc. (see **Appendix** for details.) To avoid confounding with lexical processes, none of the phrases in the syntactic scale included words beyond the first 1,000-word-family level of the BNC (Leech et al., 2001). As word order is the key syntactic feature that influences parsing in English (Anderson, 2015), the participants' responses were scored according to the degree of conformity to the original word order. The maximum score for each segment was 2, for responses that retrieved the original phrase in its full form, for example, "if this process." A score of 1 was given to responses that retrieved only a semantically proper pairwise sequence, e.g., "this process," Otherwise the response would be given a score of 0, regardless of the number of words retrieved, e.g., "if process" or "process." To avoid inconsistent judgments, misspelt words were considered errors. The maximum score for each of the 15 segments was 2 points, and the maximum score for the full scale was 30.

As the lexical and syntactic measures both took the form of a partial dictation task, word recognition may be the key process underlying both measures, which poses a major threat to the validity of the syntactic measure. For preliminary evidence of validity, a homogeneity test by way of internal consistency (Anastasi and Urbina, 1997; Urbina, 2014) was conducted. The lexical scale was broken into three subscales, each consisting of 10 items from each of the three levels described above, i.e., the 1,001–2,000 frequency range, the 2,001–3,000 frequency range, and the 3,001–5,000 frequency range, of the BNC (Leech et al., 2001). Coefficient alpha was calculated at 0.85 for the three subscales (which coincided with the item-level estimate reported in **Table 1**) but dropped to 0.78 when the syntactic measure was included as the fourth subscale. As internal consistency is essentially a measure of homogeneity (Anastasi and Urbina, 1997; Urbina, 2014), this is evidence that the three lexical subscales constituted a more homogeneous scale, whereas

[2]https://www.ielts.org/ielts-for-organisations/ielts-scoring-in-detail, retrieved as of Nov. 16, 2019.

[3]https://www.ielts.org/teaching-and-research/test-taker-performance, retrieved as of Nov. 16, 2019.

[4]https://www.ielts.org/en-us/about-the-test/test-format, retrieved as of Nov. 16, 2019.

the syntactic measure was more heterogeneous to the lexical measure. Together with the content analysis presented above and detailed in the **Appendix**, this provides the preliminary evidence for interpreting the 15 phrases as a syntactic measure.

Data Collection Procedures

The IELTS listening test was administered in its paper-and-pen form in a computerized language lab as part of a mid-term test for the academic listening course. In accordance with the official IELTS administration procedures, the participants heard the recordings once only and responded to the questions in 30 min, after which they transferred their responses to a commercial web-based testing platform, which saved the responses as a downloadable Microsoft Excel file for scoring.

The partial dictation task was completed immediately after the participants submitted their listening test responses online, as another part of the mid-term test. The task was also administered in its paper-and-pen form. The participants heard the recordings once only, after which the participants submitted their responses to the same testing platform. The responses were also downloaded as a Microsoft Excel file for scoring.

Data Analysis

The scores used in the analyses were numbers of correct responses. The maximum score was 40 for the listening test and 30 for the lexical and syntactic scales. To answer RQ1, scores on the lexical and syntactic scales were correlated to the score on the listening test. To answer RQ2, the listening score was regressed on the lexical and syntactic scales in two sequential analyses. The first analysis started with the lexical scale in the first step, with the addition of the syntactic scale in the second step. The second analysis was conducted in the reverse order, starting with the syntactic scale. All analyses were conducted in SPSS18.

RESULTS

Correlations

Correlations between lexical and syntactic measures and L2 academic English listening were calculated to answer RQ1. **Table 1** reports the mean, standard deviation, and internal consistency reliability (coefficient alpha) for each of the three measures, as well as Pearson correlations between each pair of measures with their 95% confidence intervals.

Prior to discussing the descriptive statistics, the internal consistency reliability of the three scores should be examined. Coefficient alpha was estimated at 0.78 for the listening score, 0.85 for the lexical score, and 0.72 for the syntactical score. These were considered acceptable for the study. The coefficient of variation can be calculated for each measure from the mean and standard deviation reported in **Table 1**, i.e., 5.20/15.33 = 0.34 for listening, 5.07/9.38 = 0.54 for the lexical score, and 5.00/12.05 = 0.41 for the syntactical score. The CV for the listening score was comparable to the estimates calculated from the descriptive statistics reported in Mecartty (2000) and Andringa et al. (2012). However, the CVs for the lexical and syntactical scores were considerably greater than those calculated from the two previous studies. Taken together, these were evidence that restriction of range in the three scores did not attenuate the correlations seriously. The skewness and kurtosis estimates of the three scores are also reported in **Table 1**. None of these had an absolute value greater than 1, so the scores were considered to be approximately normally distributed, which supported the use of Pearson correlations to represent the bivariate relationships.

As **Table 1** shows, the three pairwise correlations were all close to 0.70, comparable to findings reported in Andringa et al. (2012) and Cheng and Matthews (2018). Considered separately, both lexical and syntactic scores were moderately correlated with the L2 listening score. The correlation between lexical and syntactic scores was also moderate, consistent with psycholinguistic theories that lexical and syntactic processes are distinct processes in listening.

Regression Analyses

To answer RQ2, two hierarchical regression analyses were conducted, both regressing L2 academic English listening on the lexical and syntactic scores, but with different predictors in each step. Prior to the analyses, the outlier and collinearity assumptions were examined. The maximum value of Cook's distance in the sample was 0.055, far less than the critical value of 1, indicating that there were no overly influential cases that warranted exclusion from the analyses (Cook, 1977). The tolerance was estimated at 0.504, indicating that around half of the variance in one predictor could be explained by the other predictor. The corresponding variance inflation factor was 1.984, and multicollinearity was not considered a serious threat to result interpretation. After the regression analyses, diagnostics were also run to examine the normality and homoscedasticity of the residuals. **Figure 1** displays the resulting plots.

The upper panel is the normal P-P plot of the standardized residuals from the regression model, which displays only minor deviations from normality. The lower panel is the scatterplot

TABLE 1 | Descriptive statistics, reliabilities, and correlations (n = 258).

	Mean	*SD*	Maximum	Minimum	Skewness	Kurtosis	Alpha	Correlation* (95% CI)	
								Lexical	Syntactic
Listening	15.33	5.20	29	5	0.55	−0.16	0.78	0.77 (0.71,0.81)	0.67 (0.60,0.73)
Lexical	9.38	5.07	25	1	0.78	0.42	0.85		0.70 (0.64,0.76)
Syntactic	12.05	5.00	25	1	0.34	−0.30	0.72		

**All correlations were significant at the 0.001 level.*

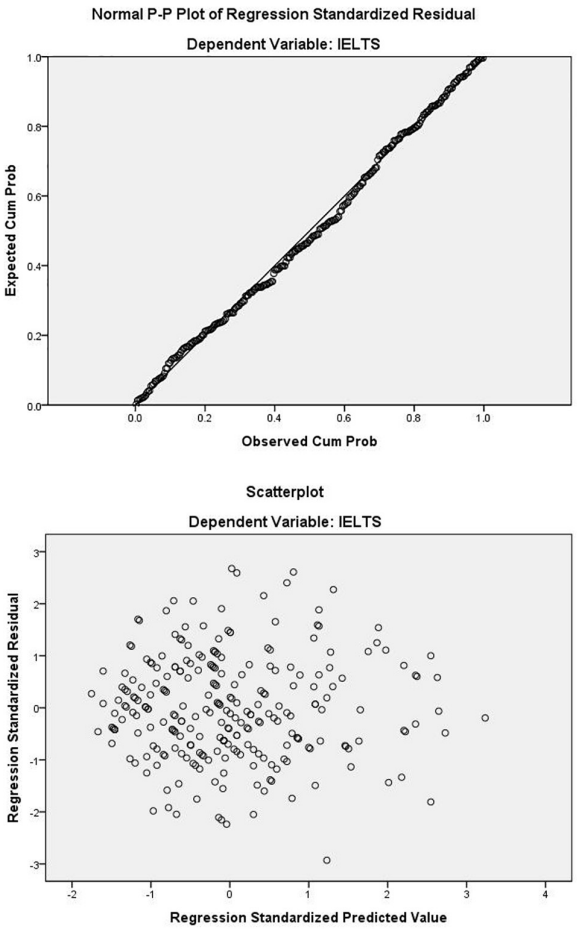


FIGURE 1 | Regression diagnostics.

with the standardized predicted value on the X-axis and the standardized residuals on the Y-axis. No obvious deviation from homoscedasticity is observed. Therefore, the two regression analyses were considered appropriate.

In the first analysis, the lexical score was entered as a sole predictor of L2 academic English listening in the first step, with the addition of the syntactic score in the second step. The regression with only the lexical score was significant, $R^2 = 0.59$, adjusted $R^2 = 0.59$, $F(1,256) = 369.76$, $p < 0.001$. The addition of the syntactic score produced a significant R^2 change, $R^2 = 0.03$, $F(1,255) = 22.24$, $p < 0.001$. These results showed that lexical score alone was a good predictor of L2 academic English listening, explaining 59% of the variance in the listening score. The addition of the syntactic score contributed 3% more to the variance in the listening score.

The second analysis reversed the order and started with the syntactic score as a sole predictor of L2 academic English listening, with the addition of the lexical score in the second step. The regression with only the syntactic score was significant, $R^2 = 0.45$, adjusted $R^2 = 0.45$, $F(1,256) = 208.46$, $p < 0.001$. The addition of the lexical score produced a significant R^2 change, $R^2 = 0.18$, $F(1,255) = 118.53$, $p < 0.001$. These results showed that syntactic score alone was a good predictor of L2 academic English listening, explaining 45% of the variance in the listening score. The addition of the lexical score contributed 18% more to the variance. In either order, both predictors were able to account for 62% of the variance in the listening score.

In answer to RQ2, the above results show that both lexical and syntactic processes had unique contributions to L2 listening in the academic English context.

DISCUSSION

Comparability to Earlier Studies

The correlation and regression analyses have yielded results that agree more with Andringa et al. (2012) and Cheng and Matthews (2018) than with Mecartty (2000). When considered separately, both lexical and syntactic measures correlated moderately with L2 academic English listening, with Pearson correlations close to 0.70. When considered jointly, both lexical and syntactic measures had unique contributions to the variance in the listening score. These results have confirmed the hypotheses stated earlier. More generally, they have provided evidence in support of the claim that different degrees of relevance in the measures will yield different results with regard to the relative importance of lexical and syntactic knowledge in L2 listening. More specifically, contextualized auditory receptive measures of lexical and syntactic knowledge are more similar to L2 listening tasks in terms of task characteristics and are considered more relevant to L2 listening in this sense, which explained the different results between Mecartty (2000) and Andringa et al. (2012). In particular, the lower correlations between lexical and syntactic measures and L2 listening in Mecartty (2000) may be attributed to the decontextualized visual feature of the lexical measure and the visual productive feature of the syntactic measure. The lack of unique contribution of syntactic knowledge to L2 listening in Mecartty (2000) may also be attributed to the same features.

It is also interesting to compare the findings of these studies to similar studies on L2 reading. Studies on the relative significance of lexical and syntactic knowledge in L2 reading have also yielded mixed results—some studies found a greater contribution of lexical knowledge (Bossers, 1992; Brisbois, 1995; Yamashita, 1999), while others reported heavier regression weights of syntactic knowledge (Shiotsu and Weir, 2007). Shiotsu and Weir (2007) emphasized the difference between a structural equation model and a regression model but also noted that sample size, test difficulty relative to the participants, characteristics of the participants, and the nature and reliabilities of the instruments used are important methodological factors that may explain the differences between studies. The commonality between the findings of the present study and those of Shiotsu and Weir (2007) is that both lexical and syntactic knowledge have a unique contribution to L2 English comprehension.

Importance of Theoretical Underpinnings

The comparison of results between this study and earlier studies also demonstrates the importance of theoretical underpinnings in practical research. For example, the findings that syntactic

knowledge does not contribute uniquely to the variance of L2 listening beyond lexical knowledge (Mecartty, 2000) are difficult to explain in light of psycholinguistic theories (Field, 2011; Anderson, 2015; Fernández and Cairns, 2018; Skeide and Friederici, 2018), whereas an emphasis on the joint contribution of lexical and syntactic knowledge (Andringa et al., 2012) agrees in principle with these theories and relative findings. This shows the importance of basing the measures on clear theoretical definitions of the constructs (Bachman, 1990).

The literature review has focused on psycholinguistic theories as the framework for depicting the partly parallel and partly overlapping relation between lexical and syntactic processes (Anderson, 2015). This coincides with findings in applied linguistics. For example, the verbal protocol studies of Buck (1991, 1994) found that L2 English listening tasks intended to test lexical knowledge turned out to involve higher-order processes, including syntactic processes. In turn, these findings also coincide with the lexico-grammatical approach to language studies in contemporary linguistics, which views lexis and syntax as the two ends of one continuum (Broccias, 2012; Sardinha, 2019). However, adopting a psycholinguistic framework offers the convenience of smooth transition to cognitive diagnostic assessment of listening, which is gaining increasing attention in L2 assessment (Lee and Sawaki, 2009; Aryadoust, 2018).

Another issue raised in the literature review is construct confounding, which reduces the relevance of results from Oh (2016) and Wang and Treffers-Daller (2017) to the issue under consideration in this study. The relative importance of lexical and syntactic processes in L2 listening was not distinguished in Oh's (2016) results, while lexical knowledge was intertwined with general language proficiency in Wang and Treffers-Daller (2017). It is a pity that these studies do not provide further evidence for examining the theoretical relationship between lexical knowledge, syntactic knowledge, and L2 listening.

In passing, it is worthwhile to mention that the relationship between lexical knowledge, syntactic knowledge, and L2 listening is not only of theoretical significance but also has practical implications. In practice, L2 listening is often assessed as a uniform skill for general purposes such as placement, certification, progress monitoring, and teaching evaluation (Bachman and Palmer, 2010). However, there is a growing need for diagnostic assessment that calls for more fine-grained understanding of the cognitive processes that underlie L2 listening activities, which invariably include lexical and syntactic processes (Field, 2009, 2013; Goh and Aryadoust, 2014; Alderson et al., 2015; Harding et al., 2015).

Generalizability Issues

Closely related to theoretical underpinnings is the idea of construct validity, which is a key requirement for making causal inferences in Campbell's validity framework (Shadish et al., 2002; Shadish, 2010). One of the key reasons given by Andringa et al. (2012) to account for the differences between their results and Mecartty's (2000) results was the different instruments used in the two studies. In a recent commentary, Schmitt et al. (2020) recommended argument-based approaches for vocabulary test development and validation, which "start with a clear and explicitly stated purpose and provide structured and comprehensive evidence for justifiable interpretations." Earlier, Read (2000) emphasized the important role of context in a vocabulary test and argued against presenting words in isolation. It is the hope of this author that the present study provides some guidelines on how to suit the specific characteristics of assessment tasks (such as the visual/auditory, receptive/productive, and contextualized/decontextualized dichotomies) to the purpose for researchers who need a vocabulary test as an instrument in their future studies.

The other two reasons provided by Andringa et al. (2012) in explanation of the differences between their results and those of Mecartty (2000), i.e., measurement error and attenuated correlation due to restriction of range, were both issues related to the statistical validity of the studies in Campbell's validity framework (Shadish et al., 2002; Shadish, 2010). While raw scores were used for replication purposes, restriction of range was not found to be a serious problem in this study. Together with relevance and theoretical underpinnings, both of which are construct validity issues in Campbell's framework, they form the foundations for the generalizability of findings of this study. The measures of lexical and syntactic knowledge in this study were not exactly the same as those used in Andringa et al. (2012) and Cheng and Matthews (2018) but were comparable to them with regard to features of relevance. This means that if similar relevant measures are used in future studies, the researcher may expect to obtain similar results.

As for the measure of L2 academic English listening, this study has used the IELTS listening test, which was also used in Cheng and Matthews (2018), albeit not the same version. There is some threat to generalizability here, as the IELTS listening test has been criticized for underrepresenting the listening construct by tapping only the ability to understand explicitly stated information and to make paraphrases (Geranpayeh and Taylor, 2008; Field, 2009; Aryadoust, 2013). More generally, the construct definition of L2 listening, i.e., the dependent variable, has not been compared across earlier studies, as it was only vaguely mentioned in Andringa et al. (2012). Furthermore, the task characteristics of L2 listening have not been compared between earlier studies, or between this study and earlier studies. The visual/auditory, receptive/productive, and contextualized/decontextualized dichotomies have been proposed as the key features, but other task characteristics such as topical knowledge, linguistic complexity, speed, and response format also play a key role in the listening process (Bachman and Palmer, 2010; Taylor and Geranpayeh, 2011; Révész and Brunfaut, 2013). Therefore, comprehensive studies that address variations in both the independent and dependent variables, with clear definition and operationalization, will provide much insight into the issue under consideration in this study. For this study, the construct of L2 listening should be understood with these limitations in mind.

The findings of this study could have been more convincing if multiple types of measures had been used, so that direct comparison could be made between the visual/auditory, receptive/productive, and contextualized/decontextualized

dichotomies, similar to what Cheng and Matthews (2018) did in their study. Furthermore, this study has used raw scores in regression models to enable comparison to Mecartty's (2000) results, but latent variable models would promise more stable results with measurement errors considered, as Andringa et al. (2012) have done.

It was mentioned in the literature review that personal variables may also constitute a significant source of difference across studies. The participants of this study were more similar to those in Cheng and Matthews (2018) but were from a single major. In contrast, for example, the participants in Andringa et al. (2012) were adults with more varied ages, first language backgrounds, and socioeconomic statuses. These factors have been treated as random in the regression model but may play a systematic role. This is a pending question before a more comprehensive study is conducted.

Furthermore, the field is moving fast ahead, with new technologies being added to the repertoire of research methods. The psycholinguistic studies reviewed earlier have used event-related potential to capture neural activity related to both sensory and cognitive processes in listening (Friederici, 2012; Hagoort, 2013; Skeide and Friederici, 2018). Recently, there are also scholars who attempt to use eye tracking to unveil the listening process. For example, Aryadoust (2019) and Holzknecht (2019) found that test-takers spend much time on reading the test items and answering them, thus confusing listening ability with reading ability. These studies have both theoretical and methodological significance. Theoretically, they shed light on the cognitive process of L2 listening comprehension; methodologically, they demonstrate the powerful potential of modern technologies. Therefore, future studies on L2 listening comprehension can benefit considerably from these technologies.

CONCLUSION

With regard to the causal relationship between lexical and syntactic knowledge and L2 listening, each study reviewed earlier has approached the issue by focusing on one particular combination of features, contributing to various degrees of relevance. As Shadish (2010) sees it, any single study sheds a little light on the nature of the causal relationship, but multiple studies on the same question are needed to find out which features are irrelevant to the causal knowledge and which are central. This study is just such an attempt. Built upon earlier studies, it helps find out the key features in lexical and syntactic knowledge that contribute to L2 listening. Using lexical and syntactic measures with similar task characteristics in terms of the visual/auditory, receptive/productive, and contextualized/decontextualized dichotomies, the study has replicated the findings in earlier studies that used similar relevant measures. The results showed that when lexical and syntactic knowledge is measured in auditory receptive tasks contextualized in natural discourse, both measures have unique contributions to L2 listening. The key message from these results is that research instruments should be designed to validly represent constructs if practical research is to yield consistent findings that agree with theory and with each other.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the author, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This study was funded by The National Social Science Fund of China (18BYY099).

REFERENCES

Alderson, J. C., Brunfaut, T., and Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: insights from professional practice across diverse fields. *Appl. Linguist.* 36, 236–260. doi: 10.1093/applin/amt046

Anastasi, A., and Urbina, S. (1997). *Psychological Testing*, 7th Edn. Upper Saddle River, NJ: Prentice Hall.

Anderson, J. R. (2015). *Cognitive Psychology and Its Implications*, 8th Edn. New York, NY: Freeman.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., and Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: an individual differences approach. *Lang. Learn.* 62, 49–78. doi: 10.1111/j.1467-9922.2012.00706.x

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

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education o-level: application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *Int. J. List.* 1–24. doi: 10.1080/10904018.2018.1500915

Aryadoust, V. (2019). Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: an eye-tracking study. *Comput. Assist. Lang. Learn.* doi: 10.1080/09588221.2019.1574267

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

Bachman, L. F., and Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Bossers, B. (1992). *Reading in Two Languages: a Study of Reading Comprehension in Dutch as a Second Language and in Turkish as a First Language*. Rotterdam: Drukkerij Van Driel.

Brisbois, J. E. (1995). Connections between first- and second-language reading. *J. Read. Behav.* 27, 565–584. doi: 10.1007/s11145-017-9791-8

Broccias, C. (2012). "The syntax-lexicon continuum," in *The Oxford Handbook of the History of English*, eds T. Nevalainen and E. C. Traugott (Oxford: Oxford University Press), 735–747.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Lang. Test.* 8, 67–91. doi: 10.1177/026553229100800105

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Lang. Test.* 11, 145–170. doi: 10.1177/026553229401100204

Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: evidence from confirmatory factor analysis. *Lang. Test.* 30, 177–199. doi: 10.1177/0265532212456833

Cheng, J., and Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Lang. Test.* 35, 3–25. doi: 10.1177/0265532216676851

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* 19, 15–18.

Fernández, E. M., and Cairns, H. S. (2018). "Overview," in *The Handbook of Psycholinguistics*, eds E. M. Fernández and H. S. Cairns (Hoboken, NJ: John Wiley & Sons), 185–192.

Field, J. (2009). The cognitive validity of the lecture-based question in the IELTS listening paper. *IELTS Res. Rep.* 9, 17–65.

Field, J. (2011). Into the mind of the academic listener. *J. Engl. Acad. Purposes* 10, 102–112. doi: 10.1016/j.jeap.2011.04.002

Field, J. (2013). "Cognitive validity," in *Examining Listening: Research and Practice in Assessing Second Language Listening*, eds A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 77–151.

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* 16, 262–268. doi: 10.1016/j.tics.2012.04.001

Geranpayeh, A., and Taylor, L. (2008). Examining listening: developments and issues in assessing second language listening. *Res. Notes* 32, 3–5.

Goh, C. C., and Aryadoust, V. (2014). Examining the notion of listening sub-skill divisibility and its implications for second language listening. *Int. J. Listen.* 29, 109–133. doi: 10.1080/10904018.2014.936119

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Front. Psychol.* 4:416. doi: 10.3389/fpsyg.2013.00416

Harding, L., Alderson, C., and Brunfaut. (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

Holzknecht, F. (2019). *Double Play in Listening Assessment*. Ph.D. thesis, Lancaster University, Bailrigg.

Laufer, B., and Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Lang. Test.* 16, 33–51. doi: 10.1177/026553229901600103

Lee, Y.-W., and Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Lang. Assess. Q.* 6, 239–263. doi: 10.1080/15434300903079562

Leech, G., Rayson, P., and Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. New York, NY: Routledge.

Lynch, T. (2010). "Listening: sources, skills, and strategies," in *The Oxford Handbook of Applied Linguistics*, 2nd Edn, ed. R. B. Kaplan (Oxford: Oxford University Press), 74–87.

Matthews, J. (2018). Vocabulary for listening: emerging evidence for high and mid-frequency vocabulary knowledge. *System* 72, 23–36. doi: 10.1016/j.system.2017.10.005

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Appl. Lang. Learn.* 11, 323–348.

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

Nation, I. S. P., and Beglar, D. (2007). A vocabulary size test. *Lang. Teach.* 31, 9–13.

Oh, E. (2016). Comparative studies on the roles of linguistic knowledge and sentence processing speed in L2 listening and reading comprehension in an EFL tertiary setting. *Read. Psychol.* 37, 257–285. doi: 10.1080/02702711.2015.1049389

Oosterhoff, J. (2015). *Modern Dutch Grammar: A Practical Guide*. New York, NY: Routledge.

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

Révész, A., and Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Stud. Second Lang. Acquis.* 35, 31–65. doi: 10.1017/s0272263112000678

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.3389/fpsyg.2014.01521

Sardinha, T. B. (2019). "Lexicogrammar," in *The Encyclopedia of Applied Linguistics*, ed. C. A. Chapelle (Hoboken, NJ: John Wiley & Sons), 1–5. doi: 10.1002/9781405198431.wbeal0698.pub2

Schmitt, N., Nation, P., and Kremmel, B. (2020). Moving the field of vocabulary assessment forward: the need for more rigorous test development and validation. *Lang. Teach.* 53, 109–120. doi: 10.1017/s0261444819000326

Shadish, W. R. (2010). Campbell and Rubin: a primer and comparison of their approaches to causal inference in field settings. *Psychol. Methods* 15, 3–17. doi: 10.1037/a0015916

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.

Shiotsu, T., and Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Lang. Test.* 24, 99–128. doi: 10.1177/0265532207071513

Skeide, M. A., and Friederici, A. D. (2018). "Neurolinguistic studies of sentence comprehension," in *The Handbook of Psycholinguistics*, eds E. M. Fernández and H. S. Cairns (Hoboken, NJ: John Wiley & Sons), 438–456. doi: 10.1002/9781118829516.ch19

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Stud. Second Lang. Acquis.* 31, 577–607. doi: 10.1017/S0272263109990039

Taylor, L., and Geranpayeh, A. (2011). Assessing listening for academic purposes: defining and operationalising the test construct. *J. English Acad. Purposes* 10, 89–101. doi: 10.1016/j.jeap.2011.03.002

Urbina, S. (2014). *Essentials of Psychological Testing*, 2nd Edn. Hoboken, NJ: John Wiley & Sons.

Vandergrift, L., and Baker, S. (2015). Learner variables in second language listening comprehension: an exploratory path analysis. *Lang. Learn.* 65, 390–416. doi: 10.1111/lang.12105

Wang, Y., and Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: the contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System* 65, 139–150. doi: 10.1016/j.system.2016.12.013

Yamashita, J. (1999). *Reading in a First and a Foreign Language: a Study of Reading Comprehension in Japanese (the L1) and English (the L2)*. Ph.D. thesis, Lancaster University, Bailrigg.

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

APPENDIX

List of phrases in the syntactical scale.

No.	Phrase	Syntactic features
1	1.6 or anything	Ellipsis ("1.6" = 1.6 liters) Noun conjunction ("or")
2	got one in	Phrase structure (verb + pronoun + adverb particle)
3	with the engine	Phrase structure (preposition + article + noun)
4	go for that	Phrasal verb ("go for") Pronoun ("that")
5	is that	Predicate after a parenthesis
6	what you do	Subordination Clause structure ("what" + clause)
7	if this process	Subordination Initial string of a clause Noun phrase as subject ("this process")
8	for this process	Clause structure (it is + adjective + for someone to do something)
9	if possible	Parenthesis Ellipsis (if. . . is possible)
10	If you decide	Subordination Initial string of a clause
11	to move on	Cohesive device Phrasal verb ("move on")
12	our student body	Compound noun
13	the better	Special structure ("The earlier. . . the better. . .")
14	What if	Question beginning
15	the very latest	Emphatic expression

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fpsyg.2020.00618

Structural Equation Modeling of Vocabulary Size and Depth Using Conventional and Bayesian Methods

Rie Koizumi[1]* *and Yo In'nami*[2]

[1] School of Medicine, Juntendo University, Chiba, Japan, [2] Faculty of Science and Engineering, Chuo University, Tokyo, Japan

In classifications of vocabulary knowledge, vocabulary size and depth have often been separately conceptualized (Schmitt, 2014). Although size and depth are known to be substantially correlated, it is not clear whether they are a single construct or two separate components of vocabulary knowledge (Yanagisawa and Webb, 2020). This issue has not been addressed extensively in the literature and can be better examined using structural equation modeling (SEM), with measurement error modeled separately from the construct of interest. The current study reports on conventional and Bayesian SEM approaches (e.g., Muthén and Asparouhov, 2012) to examine the factor structure of the size and depth of second language vocabulary knowledge of Japanese adult learners of English. A total of 255 participants took five vocabulary tests. One test was designed to measure vocabulary size in terms of the number of words known, while the remaining four were designed to measure vocabulary depth in terms of word association, polysemy, and collocation. All tests used a multiple-choice format. The size test was divided into three subtests according to word frequency. Results from conventional and Bayesian SEM show that a correlated two-factor model of size and depth with three and four indicators, respectively, fit better than a single-factor model of size and depth. In the two-factor model, vocabulary size and depth were strongly correlated (r = 0.945 for conventional SEM and 0.943 for Bayesian SEM with cross-loadings), but they were distinct. The implications of these findings are discussed.

Keywords: vocabulary size, vocabulary depth, factor structure, model testing, Bayesian structural equation modeling

OPEN ACCESS

Edited by:
Shuichi Takaki,
Fukushima University, Japan

Reviewed by:
Ahmed Masrai,
King Abdul Aziz Military Academy, Saudi Arabia
Stuart Webb,
University of Western Ontario, Canada

***Correspondence:**
Rie Koizumi
rkoizumi@juntendo.ac.jp

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

Received: *18 November 2019*
Accepted: *16 March 2020*
Published: *21 April 2020*

Citation:
Koizumi R and In'nami Y (2020) Structural Equation Modeling of Vocabulary Size and Depth Using Conventional and Bayesian Methods. Front. Psychol. 11:618. doi: 10.3389/fpsyg.2020.00618

INTRODUCTION

The structure of language ability is a focus of concern for second language (L2) assessment researchers. Research on this issue dates back to Oller (1983), who reported on the unitary (i.e., single-factor) structure of a university placement test comprised of sections on grammar, composition, vocabulary, phonology, and dictation or cloze tasks. The findings of Oller's study suggested that a single ability could be measured by a test consisting of multiple components of language ability. This finding was criticized by Bachman and Palmer (1983) and Farhady (1983), the whose results contradicted Oller and instead suggested that language ability consisted of multiple components. Research into the structure of language ability has continued, and numerous studies have made contributions to the issue (e.g., Shin, 2005; In'nami and Koizumi, 2012a; Sawaki and Sinharay, 2017; Yan et al., 2019).

Such investigations have for the most part focused on tests that assess skills (e.g., four skills in Sawaki and Sinharay, 2017; speaking in Sawaki, 2007; writing in Bae et al., 2016; listening and reading in Yamashita and Shiotsu, 2017). If the intended test constructs accord well with an observed factor structure, this constitutes one piece of evidence for the validity of interpretations based on test scores and evidence for an inference in the validity argument (e.g., Chapelle et al., 2008; Kane, 2013).

This line of research, which should also include vocabulary constructs (i.e., the vocabulary knowledge and ability that tests are intended to measure and what is actually measured; see Chapelle, 1998), has hitherto been limited. The necessity of investigating the quality of vocabulary tests and their constructs has been emphasized by Schmitt et al. (2020), who stated that L2 vocabulary fields need step-by-step test development and validation of vocabulary tests to allow for the meaningful interpretation and application of test scores.

While vocabulary knowledge has been conceptualized in various ways, vocabulary size and depth have often been separately conceptualized (Schmitt, 2014). Although the two have been shown to be substantially correlated, how size and depth should be conceptualized is not clear (Yanagisawa and Webb, 2020). Since they are strongly related to one another, should they be considered a single construct? Or should they be treated as two distinct constructs of vocabulary knowledge? These questions regarding the factor structure of size and depth can be better examined via structural equation models that take into account measurement error. Although structural equation modeling (SEM) has been used in language testing, models can be more flexibly tested using Bayesian estimation within the framework of SEM. The current study reports on the uses of conventional and Bayesian SEM to examine the factor structure of size and depth of L2 vocabulary knowledge of Japanese adult learners of English.

LITERATURE REVIEW

Defining Size and Depth

Many vocabulary researchers share the view that vocabulary knowledge can be classified into several components (e.g., Henriksen, 1999; Read, 2000; Meara, 2005; Daller et al., 2007; Milton, 2009; Schmitt, 2010; Nation and Webb, 2011; Webb and Nation, 2017; Nation, 2020). Of the several methods of classification, one in particular often used is the size and depth of vocabulary knowledge. It was proposed by Anderson and Freebody (1981) and is defined as follows: Size, or breadth, concerns a quantitative aspect related to knowledge of a word form and a primary meaning. This is also termed the form–meaning association. In contrast, depth involves a qualitative aspect associated with "how well a learner knows individual words or how well words are organized in the learner's mental lexicon" (Staehr, 2009, p. 579).

Size has garnered much more attention as a research target than depth (Schmitt, 2014; Qian and Lin, 2020; Yanagisawa and Webb, 2020). In contrast, depth covers a wide range of lexical dimensions and is difficult to define. According to Webb (2013), "there is no definition of vocabulary depth that is widely agreed upon" (p. 1657). One of the leading researchers in depth studies, Read (2004) classified depth into three aspects: precision of meaning, comprehensive word knowledge, and network knowledge. Schmitt (2014), in relation to relationships between size and depth, organized depth into seven aspects: "receptive versus productive mastery, knowledge of multiple word knowledge components, knowledge of polysemous meaning senses, knowledge of derivative forms (word family members), knowledge of collocation, the ability to use lexical items fluently, and the degree and kind of lexical organization" (p. 922). Nation (2013, 2020) offered a comprehensive list of vocabulary knowledge by using three categories (i.e., Form, Meaning, and Use), each of which is further classified into three aspects: (a) Form: spoken, written, and word parts; (b) Meaning: form and meaning, concept and referents, and associations; and (c) Use: grammatical functions, collocations, and constraints on use (e.g., register, frequency). Each aspect has receptive and productive dimensions. Among them, "form and meaning, concept and referents, and collocation," which Webb (2013) considers to be assessed by the Word Associates Format (WAF; Read, 1993, 1998), seem to be the aspects studied most.

While both size and depth are important for language use, size has been considered the primary aspect of vocabulary knowledge because of its importance in the form–meaning link for vocabulary use (e.g., Laufer et al., 2004; Webb, 2005; Schmitt, 2010). Given the centrality of size, indications of form–meaning knowledge are often interpreted as having the ability to use words in reading, listening, writing, and speaking, and even as having vocabulary depth such as derivatives and collocations. However, Kremmel and Schmitt (2016) argue that these interpretations are not justified based on their research.

Measuring Size and Depth

Size and depth have been measured using various formats. Size has been typically measured by means of a recognition (e.g., multiple-choice or matching) or recall (e.g., translation) format, in which the L2 target form or its meaning is presented and test takers select or supply the meaning or L2 form (e.g., Laufer et al., 2004). There exist many vocabulary size tests, such as the Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001) and the Vocabulary Size Test (Nation and Beglar, 2007). The following shows a sample item from the Vocabulary Size Test (Nation and Beglar, 2007, JALT2007, p. 2), in which test takers are asked to select the most appropriate meaning of the written target word out of four written choices.

poor: We are poor. (*answer)

a. have no money*
b. feel happy
c. are very interested
d. do not like to work hard

While the form–meaning association appears relatively simple to define and assess, research has shown that it is not: Size test scores are affected not only by the intended test construct but

also by various factors such as differences in item formats and test takers' test-taking strategies (Gyllstad et al., 2015; Kremmel and Schmitt, 2016; McLean et al., 2020).

Still, measuring size is less complicated than measuring depth. Depth is a multifaceted construct, ranging from various aspects of vocabulary to lexical organization, resulting in varied test formats. Yanagisawa and Webb (2020) grouped various approaches to measuring depth into three categories: a developmental approach, a lexical network approach, and a components approach. The developmental approach considers depth as something expanding from zero to full knowledge and attempts to test on what stage learners are located. An example can be found in the Vocabulary Knowledge Scale in Wesche and Paribakht (1993, p. 30), which uses self-assessment and some production items. Test takers are asked to indicate their degree of knowledge of each target word using the following scale.

I. I don't remember having seen this word before.
II. I have seen this word before, but I don't know what it means.
III. I have seen this word before, and I *think* it means.____ (synonym or translation)
IV. I *know* this word. It means.____ (synonym or translation)
V. I can use this word in a sentence:_____. (If you do this section, please also do Section IV.)

Scores vary according to the quality of written responses. For example, if the synonym or translation provided in III–IV by test takers is wrong, those who choose III gain a score of 2. Yanagisawa and Webb (2020) summarize validity issues related to the Vocabulary Knowledge Scale. These issues include the lack of empirical basis for the developmental scale structure and difficulty in interpreting total scores because the test assesses multiple aspects of vocabulary knowledge in different stages.

In the lexical network approach, depth is conceived as a lexical network in which words are associated in learners' mental lexicon, and indications of knowledge of word association are elicited in tests taking this approach. The WAF (Read, 1993, 1998) uses this approach and is possibly the most frequently used depth measure (Yanagisawa and Webb, 2020). In the following sample item from Read (1998, p. 46), test takers are asked to select four words related to the stimulus word out of eight options. In the box on the left, words that may have paradigmatic associations with the cue word (synonym or one element of the meaning) are presented, whereas the box on the right contains words that may have syntagmatic associations with the cue word (collocations). There are possibly one to three answers out of four in the left box and one to three answers in the right box, and four answers in total.

sudden

beautiful quick* surprising* thirsty	change* doctor noise* school

While the WAF is relatively easy to administer and score, there are limitations: For example, this format taps limited aspects of the lexical network; this format allows test takers to use guessing strategies; studies using the WAF have modified test formats and scoring methods according to their research orientations, so the scores are not always comparable across studies (Yanagisawa and Webb, 2020); it is also rather difficult to interpret what its total scores mean because multiple aspects of vocabulary depth are combined (Webb, 2013).

The components approach handles different aspects of depth separately. Webb (2013) recommended this approach, stating that creating tests assessing each aspect separately would bring the field forward for more precise depth assessment and research. Using this principle, multiple measures have been developed. For example, Webb (2005) developed 10 tests that focus on five aspects (i.e., written form, form and meaning, association, collocation, and grammatical functions), each of which was assessed with receptive or productive (i.e., recognition or recall) formats. Tests focusing on written form assessed size, whereas those focusing on the other four aspects assessed depth. Nguyen and Webb (2017) developed a collocation test in a multiple-choice format in which test takers were required to choose "the word that co-occurred most frequently with the node word from four options" (p. 306). An example is shown below (p. 309).

advantage a. get b. give c. have d. take*

Among the three approaches to measuring depth (i.e., the developmental, lexical network, and components approaches), Yanagisawa and Webb (2020) recommended the components approach most because of its transparency in what the test scores indicate. They suggested investigating a wider range of depth aspects by using separate tests. The current study responds to this call for research and develops tests separately focusing on three depth aspects: association, polysemy, and collocation.

Correlations Between Size and Depth

Numerous researchers have examined the relationship between size and depth in L2 vocabulary studies (e.g., Nurweni and Read, 1999; Mochizuki and Aizawa, 2000; Vermeer, 2001; Noro, 2002; Qian, 2002; Akase, 2005; Shimamoto, 2005; Ishii and Schmitt, 2009; Koizumi and In'nami, 2013; Kremmel and Schmitt, 2016; see Schmitt, 2014, for a comprehensive summary). They have been interested in exploring the degree to which size and depth are related and how constructs of size and depth can be conceptualized in L2 vocabulary assessment. In his seminal article on a critical review of studies on vocabulary size and depth, Schmitt (2014) posed the following questions: "Do size and depth behave as separate constructs," "or are they essentially the same construct?" (p. 941). These questions underlie the research conducted and discussions held thus far. For example, Akbarian (2010) reported a strong simple (zero-order) correlation ($r = 0.864$) between vocabulary size and depth among 112 Iranian learners of English. Size was measured using the Vocabulary Levels Test (Schmitt et al., 2001), whereas depth was measured using the WAF (Read, 1993). Strong correlations were also found in Vermeer (2001). He examined 25 L2 Dutch kindergarteners who took two size tests in which words were presented orally. In one test, they selected the picture option that showed the meaning of the word they heard; in the other, they described the meaning of the word presented. In the depth test,

they were asked to express what they knew about the target word by answering the following questions: "What is a . . .?" "What does a . . . usually look like?" "What can you do with a . . .?" "What do you feel when you touch a . . .?" and "Can you tell us some more about a . . .?" (p. 224). Their responses to the depth test were evaluated in terms of the quality of the word association network. It was found that the size test scores strongly correlated with the depth test scores (r = 0.72–0.76), which led him to state that "there is no conceptual distinction between the two" (p. 231). Schmitt (2014) attributed these high correlations to overlapping constructs. He argued that "the depth test only tapped into deeper semantic knowledge of a single meaning sense, so all tests (both size and depth) were essentially various types of meaning tests" (p. 921). He added that if he had used the measures that assess broad aspects of depth, the correlations would not have been so strong.

This hypothesis has been supported by previous studies such as Schmitt and Meara (1997), which examined the relationships between size and depth (i.e., both receptive and productive aspects of word association and suffix knowledge) among 88 Japanese learners of English. Size was assessed by the Vocabulary Levels Test (Nation, 1983). Receptive word association and suffix knowledge were assessed by requiring test takers to select the correct suffixes and words associated with a target word. Productive word association and suffix knowledge were assessed by requiring test takers to write every suffix that they thought could be added to the stimulus word as well as three word associations prompted by the stimulus. The simple correlations between size and depth aspects ranged from low to moderate (r = 0.27–0.62).

Findings from previous studies suggest that size and depth are correlated but that the strength of correlations varies from weak to strong across studies. Schmitt (2014) has attributed this variation mainly to different types of depth assessed and instruments used and to different L2 proficiency levels of test takers. He also pointed out that many depth tests may have problems related to reliability and validity. Since correlation coefficients are lowered in tests with low reliability, measurement error may partly explain the differing strengths of the relationships between size and depth across studies. One way to more accurately estimate correlation coefficients while addressing measurement error is to use SEM. SEM has been used to examine the factor structure of language ability by testing the fit of models to data. Ability and measurement error are modeled separately so that the relationships between abilities can be more precisely examined while separately estimating the impact of measurement error (see In'nami and Koizumi, 2011; Winke, 2014; Ockey and Choi, 2015, for SEM in an L2 assessment field).

In vocabulary studies, the factor structure of size and depth can be modeled using SEM in two ways. First, in a single-factor model, both size and depth measures (i.e., observed variables) are hypothesized to reflect one vocabulary factor (size and depth combined). If this model is the most likely, the distinction of size and depth is not very important, as size and depth assess the same vocabulary knowledge. Second, in a correlated factor model, size and depth factors are hypothesized to be correlated with one another. Even when they are correlated very highly, they should be treated separately, as doing so better explains the data.

A few previous studies examined a factor structure of the L2 vocabulary size and depth of L2 learners by modeling both size and depth as latent factors and comparing fit statistics across multiple models: Tannenbaum (2008) targeted first language (L1) users, and Kieffer and Lesaux (2012) targeted L1 and L2 users and analyzed a combined sample. To our knowledge, Vafaee (2016) is the only study that focuses on L2 learners' vocabulary factor structure. The authors are aware that several studies used SEM but did not model size or depth as a separate latent factor (Tseng and Schmitt, 2008; Zhang, 2012; Koizumi and In'nami, 2013), or one study (Tseng, 2011, as cited in Schmitt, 2014, p. 930–931) did not provide sufficient information for review.

Vafaee (2016) examined the relationship between size and depth of 263 lower-intermediate to advanced Persian learners of English. In the size test, test takers listened to a word and non-defining sentence once and selected from four choices of L2 meanings provided on the answer sheet (i.e., an aural version of the Vocabulary Size Test; Nation and Beglar, 2007). The test was divided into four sections according to the frequency of target words, and these four sections were used as indicators of vocabulary size (α = 0.67–0.84). In addition, an aural test of depth was created by adapting the WAF (Read, 1993, 1998). Test takers listened to the target word and options and were required to choose a synonym or collocation in relation to the target word. Results of synonym and collocation were separately scored, with synonym and collocation forming two indicators of depth (α = 0.92–0.93). There were moderate simple correlations between size and depth indicators (r = 0.64–0.77). A single-factor model with six indicators of size and depth was compared to another model (size and depth were separately modeled and correlated). The latter model (a correlated two-factor model) fit the data better than the single-factor model, with size and depth highly correlated (r = 0.94). However, the results of Vafaee (2016) may have been affected by (a) measures used to assess size and depth, (b) aspects assessed by depth tests, or (c) other features, such as participants' L1 and L2, or L2 proficiency levels. Regarding (a), Vafaee (2016) used aural versions of the Vocabulary Size Test and the WAF. Regarding (b), the research focused on synonym and collocation, as measured by the WAF. Although an aural version of the WAF was developed for the research, issues related to WAF test interpretation and use mentioned in the *Literature Review* apply to this research as well. Regarding (c), the participants were Persian learners of English at lower-intermediate to advanced levels. In order to know to what extent the findings of Vafaee can be generalized beyond contexts, further research is needed to examine a factor structure of size and depth with different types of measures addressing different aspects of the vocabulary knowledge of various target participants. Thus, this study examines a factor structure of size and depth, targeting beginner to intermediate Japanese learners of English, using diverse measures of depth.

CURRENT STUDY

To examine the relationship between the size and depth of L2 vocabulary knowledge, the following research question

is investigated in the context of L1 Japanese adult learners who studied English as a foreign language at beginner to intermediate levels.

Research question: Which factor structure of the size and depth of vocabulary knowledge explains the data better, a single-factor or a correlated two-factor model?

The current study expands on the findings of previous studies in five ways: First, we include three aspects of depth for analysis: word association, polysemy, and collocation. As described in the *Literature Review*, many previous studies, including Vafaee (2016), have used the WAF, which basically targets synonym and collocation; we increase the number of the aspects of vocabulary depth measured from two to three by employing more tests. We intentionally select word association, polysemy, and collocation, which are more closely related to a form–meaning link than are other depth aspects such as word parts, to rigorously examine the separability of size and depth constructs. Second, we use four separate depth tests by taking the components approach. Third, unlike studies that used simple correlations or regressions to investigate the relationship between size and depth (e.g., Vermeer, 2001; Akbarian, 2010), we use SEM to empirically identify the structure that best fits the data. The use of SEM should more clearly elucidate the relationship in question, with the measurement error of the instruments examined separately. Fourth, we explicitly compare a single-factor model with a correlated two-factor model using SEM. The identification of a best-fitting model of size and depth in comparison to competing models would have strong implications for vocabulary theory and practice. Fifth, we use both conventional and Bayesian SEM. In conventional SEM (and particularly in confirmatory factor analysis), the relationships between observed variables and factors are modeled by specifying paths between the two. Specifying no path indicates that no such relationship is hypothesized. According to Muthén and Asparouhov (2012), this is a very strong assumption and may not reflect researchers' theories or hypotheses since it is highly unlikely that no relationship exists between observed variables and factors. They have stated that it would be more sensible to model near-zero relationships with some variability between these observed variables and factors. Yet, conventional SEM does not allow researchers to specify such models. This is possible in Bayesian SEM, where degrees of a relationship can be specified using prior information (i.e., priors) based on theory and previous studies. This allows for more flexible testing of models by enabling researchers to specify major and minor loadings, namely those expressed as near-zero cross-loadings and correlations between residuals (i.e., measurement error).

METHOD

Participants

In 2012, a total of 255 adult learners (18 or older) took vocabulary tests as part of their L2 English courses or as volunteers. Originally, 257 students took the tests, but 2 were found not to have taken the test seriously, so these 2 students were not included. Of these 255 test takers, 239 were undergraduates at nine Japanese universities; 9 were graduate students at four Japanese universities; and 7 were professionals who used English frequently. The undergraduate and graduate participants attended national or private universities and majored in various subjects. Other information such as gender and age was not available, but it is reasonable to assume that most participants were 18–22 years old and studied English as a foreign language for at least 6 years at the secondary school level. This is because most undergraduates in Japan are in this age range and have similar English-learning experience.

They took five vocabulary tests (see *Instruments and Procedures* below) and provided scores obtained in 2012 or earlier for the TOEIC (Test of English for International Communication®) Listening and Reading Test. The distribution of participants' TOEIC scores (M = 514.84; SD = 181.17; Min = 205; Max = 985) resembled the distribution of all Japanese test takers for the TOEIC test (M = 520; SD = 180; reported in Educational Testing Service, 2019). Most participants were estimated to possess an A2 level proficiency of the Common European Framework of Reference (CEFR; Listening = 62.75%; Reading = 50.20%), based on their TOEIC Listening and Reading Test scores and a conversion table (Tannenbaum and Wylie, 2013).

Instruments and Procedures

The intent of the study's five vocabulary tests was to measure vocabulary size (one test) and vocabulary depth (four tests). We used a multiple-choice format with four or five options (see **Table 1** for examples and Appendix A in **Supplementary Material** for all the test items). All the tests employed a discrete, selective, context-independent format (Read, 2000). Words used in the size and depth tests were different across tests.

The tests were developed for research, using lemma as a basis of counting units (see Vilkaitė-Lozdienė and Schmitt, 2020, for its appropriateness). They were constructed using the JACET8000 vocabulary list, a word list specifically adapted to Japanese learners of English [JACET (Japan Association of College English Teachers) Basic Word Revision Committee, JACET Basic Word Revision Committee, 2003]. This list was compiled using the British National Corpus (BNC) and subcorpora based on material that Japanese learners of English are likely to encounter, such as in textbooks for secondary schools. We considered using the word list matching the target learners' learning context to be appropriate for measuring their vocabulary (Nation and Sorell, 2016). Readers can refer to Appendices B, C in **Supplementary Material** for information on word frequency. The JACET8000 vocabulary list was later updated (JACET Basic Word Revision Committee, 2016; see the older and latest version lists[1]). All the tests were piloted and revised before the validity of interpretations based on the scores of each test was examined and reported in Mochizuki et al. (2014).

The JACET8000 Vocabulary Size Test was intended to assess lexical knowledge of L2 written forms and the primary meanings (the first definition that appears in dictionaries) of up to 8,000 lemma. Test takers were required to select an L2 form that

[1] http://language.sakura.ne.jp/s/voc.html

TABLE 1 | Examples of the five vocabulary tests in order of administration.

Word Association Test [30 items]: Select the English word from options 1 through 4 that is most strongly associated with the given word.
Example 1. sky (1) blue* (2) yellow (3) white (4) black
Reason: Because of the expression *blue sky*.
Example 2. run (1) jog* (2) skip (3) sleep (4) throw
Reason: Because *run* and *jog* have similar meanings.
1. attack (1) action (2) defend* (3) guard (4) shout
Polysemy Test 1 [20 items]: Select the two Japanese meanings from options 1 through 5 that correspond to the English word provided.
35. introduce
(1) 導入する [introduce]* (2) 演奏する [perform] (3) 紹介する [introduce]*
(4) 反応する [respond] (5) 解説する [explain]
Polysemy Test 2 [20 items; 17 items were analyzed in the current study]: Select the English word from options 1 through 4 that corresponds to the Japanese meaning provided.
53. . . .を詰める [. . . wo tsumeru] (1) ban (2) attempt (3) stuff* (4) grasp
Collocation Test [20 items]: Select two words from options 1 through 4 that make a coherent meaning when each is combined with the word provided in English. Selected words appear in the place of ( ) displayed before or after the word provided.
Example. short ( ) (1) salt (2) shop (3) time* (4) supply*
Reason: Because of the expressions *short time* (mijikai jikan) and *short supply* (fusoku shiteiru kyokyu).
71. heavy ( ) (1) door* (2) mathematics (3) meal* (4) sunshine
Size Test (40 items): Select the English word from options 1 through 4 that best corresponds to the Japanese meaning provided.
91. 話に出す,言及する [hanashi ni dasu, genkyu suru]
(1) manipulate (2) mention* (3) minister (4) moderate

** = answer. [] = explanation to readers of this article, which did not appear on the tests. The instructions and explanations were written in Japanese to ensure participants fully understood the procedures.*

corresponded to a meaning provided in L1 Japanese. There were 40 items in total, with 5 items for each 1,000-lemma level. The 40 items were divided, according to word frequency, into three subtests of 15 (levels 1,000–3,000), 15 (4,000–6,000), and 10 items (7,000–8,000).

The second through the fifth vocabulary tests assessed three aspects of depth of vocabulary knowledge: word association, polysemy (two formats), and collocation. Stimulus words presented in each test were selected from the 1,000- to 3,000-lemma levels (Polysemy Test 2) or from the 1,000- to 2,000-lemma levels (the other three tests). All correct options but one (Word Association Test, No. 12) were within 3,000-lemma levels (see Appendix B in **Supplementary Material**).

In the 30-item Word Association Test, test takers were required to choose which L2 word was associated the most strongly with the L2 word provided. To construct this test, Mochizuki et al. (2014) asked Japanese learners of English with low to high proficiency to write three to five English words related to stimulus words (e.g., *sky*). They then selected (a) a word association as an answer that distinguished low- and high-level learners and (b) distractor word associations that did not distinguish between the two levels.

There were two polysemy tests. The first (Polysemy Test 1, 20 items) asked test takers to select two frequent meanings of an L2 polysemous word (including homographs). They were selected from the following lists of polysemous words (Gorfein et al., 1982; Twilley et al., 1994; Seto, 2007).

The other polysemy test (Polysemy Test 2) required test takers to choose an English word with the same meaning in Japanese. The stimulus words were selected from words that had at least three meanings displayed in the *Collins COBUILD Advanced Learner's English Dictionary*. The definition that appeared third in the dictionary was selected. Of the 20 items originally on the test, only 17 of the items were used for analysis because the remaining 3 items were found to assess knowledge of the first definition, which overlapped the concept of vocabulary size. It should be noted that the Size Test and Polysemy Tests 1 and 2 all assessed relationships between L2 form and L1 meaning but differed in their constructs in that the Size Test assessed primary meanings with higher frequency, Polysemy Test 1 assessed two frequent meanings, and Polysemy Test 2 assessed a less frequent meaning.

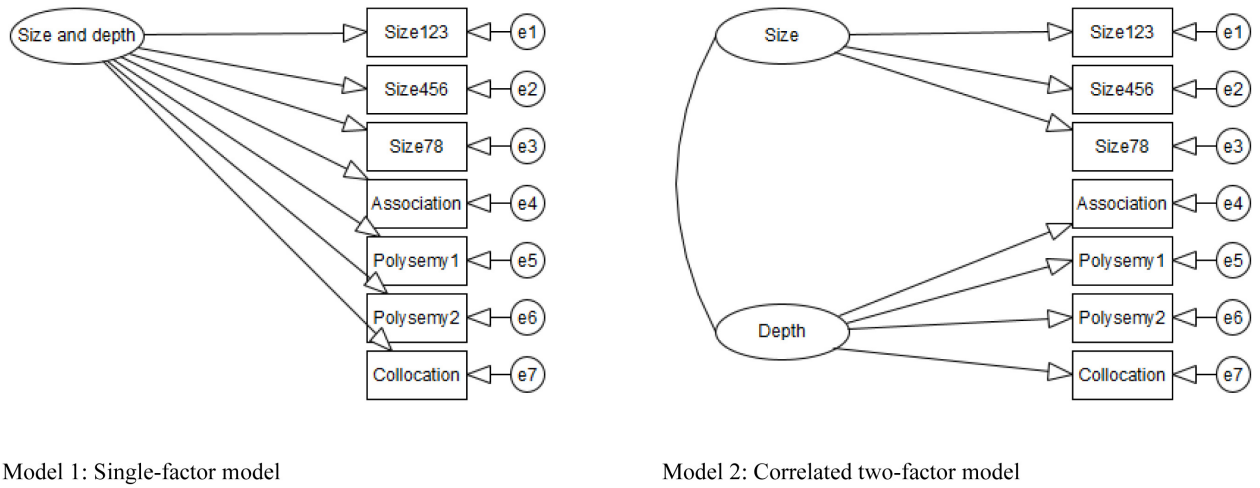
Finally, in the 20-item Collocation Test, test takers were required to select two L2 words that co-occurred with the L2 word provided. The collocation was either of an adjective + noun type or of a noun + noun type; these were selected from the *Longman Dictionary of Contemporary English* (5th ed.) and *Oxford Collocations Dictionary for Students of English*. Distractors were selected from words that least collocated with the stimulus word, and this assumption was confirmed by asking two experienced Japanese teachers of English and an English native speaker. We also examined the items by mutual information (MI) scores using the Corpus of Contemporary American English (COCA) and BNC, accessed through English-Corpora.org[2], and found no major problems (see Appendix C in **Supplementary Material** for details). As collocation is a part of word association, the Collocation Test and Word Association Test partially overlap the constructs. However, we intended to assess wider areas of depth of vocabulary knowledge instead of avoiding the overlaps.

For the Polysemy 1 and Collocation Tests, one point was awarded when two correct options were selected, whereas for the Size, Word Association, and Polysemy 2 Tests, one point was awarded when the correct option was selected.

Analysis

The structure of the size and depth of vocabulary knowledge was examined by testing two variants of models that hypothesized the relationships among variables as single-factor or correlated two-factor models. These models are presented in **Figures 1**, **2**. In each figure, the rectangles represent observed variables, the ovals represent latent factors, and the circles represent measurement errors or residuals. Models 1 and 2 were built based on the structures of vocabulary knowledge discussed in the literature (e.g., Vafaee, 2016). Model 1 had three indicators of size and four indicators of depth. Both size and depth

[2]https://www.english-corpora.org/corpora.asp



Model 1: Single-factor model

Model 2: Correlated two-factor model

FIGURE 1 | Models of vocabulary size and depth using conventional structural equation modeling (SEM). Size123 = size with 1,000–3,000 levels; Size456 = size with 4,000–6,000 levels; Size78 = size with 7,000 and 8,000 levels. All figures in this article were created using yEd Graph Editor (Version 3.19.1.1; yWorks GmbH, 2000–2019).

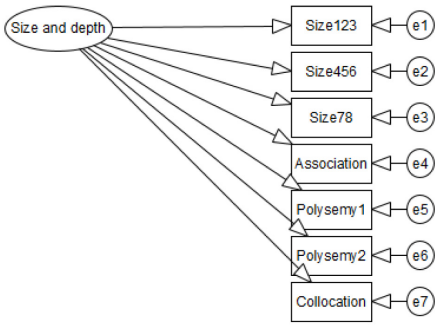
were hypothesized as a single factor of vocabulary knowledge. In Model 2, the same indicators were used to hypothesize correlated but separate factors of size and depth (see details of the models below).

The observed variables in this study were composite scores aggregated using item-level dichotomous data. The unidimensionality of each observed variable was examined and confirmed before the aggregation (e.g., Little et al., 2002, 2013; Meade and Kroustalils, 2006).

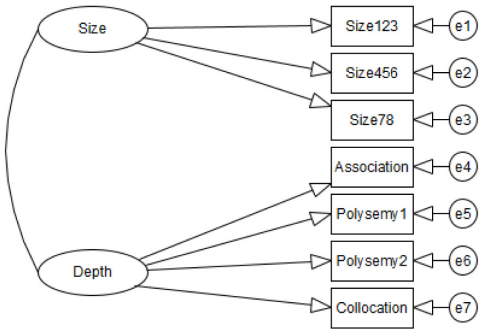
After a preliminary analysis of score distribution and reliability, conventional and Bayesian SEM was conducted using Mplus (Version 8.3; Muthén and Muthén, 1998–2019; see Appendices D–H in **Supplementary Material** for Mplus codes used). There were no missing values. For scale identification, one loading from a factor was fixed to 1.00. Observed variables were standardized to ease interpretation of priors (Muthén and Asparouhov, 2012). For conventional SEM, the data were univariately normally distributed, as judged by the skewness and kurtosis values of |3.30| (the z score at $p < 0.01$; e.g., Tabachnick and Fidell, 2014) and histograms. The data were multivariately non-normal according to Mardia's multivariate normality test available in an R package, MVN (Korkmaz et al., 2019). To account for such multivariate non-normality, a maximum likelihood estimation with a robust standard errors method was employed for estimation. Models were judged using fit indices: a comparative fit index (CFI) of 0.90 or above (Arbuckle and Wothke, 1995), a root mean square error of approximation (RMSEA) of 0.08 or below, and a standardized root mean square residual (SRMR) of 0.08 or below (Hu and Bentler, 1999). The Akaike information criterion (AIC) and chi-square difference tests were used to compare models (see Mplus, 2019). With statistical non-significance, a more parsimonious model with fewer parameters to estimate (usually, a model with a greater number of degrees of freedom) was selected. Model fit and statistical criteria were used with substantive interpretability to evaluate each model.

For Bayesian SEM, Models 1 and 2 were examined by specifying a series of priors (i.e., prior parameter distributions). Bayesian SEM can include two types of priors: non-informative and informative. They differ in the degree of specification imposed on the models, with informative priors specifying the particular distribution of parameters, as compared to non-informative priors, which do not specify such particular distributions. Our analyses using conventional and Bayesian SEM were also different in terms of cross-loadings and residual correlations. As mentioned above, Bayesian SEM can include not only major loadings but also cross-loadings and residual correlations that have small, non-major effects on the model (expressed as a dotted line) by specifying informative priors. Specifying approximate zeros is more realistic than specifying exact zeros (e.g., de Bondt and van Petegem, 2015).

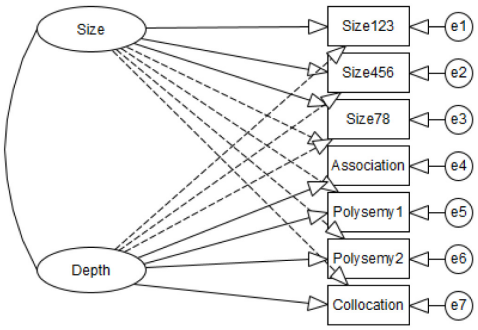
In Model 1a, non-informative priors were specified for factor loadings, with normally distributed priors with a mean of zero and infinite variance, and for observed variable variances, with inverse gamma distribution priors with infinite means and variances. In Model 2a, non-informative priors were additionally specified for factor (co)variance(s), with an inverse-Wishart distribution prior with a mean of zero and the degree of freedom of the model. These specifications were the software-default settings of Mplus. In Model 2b, informative priors were additionally specified for cross-loadings, with normally distributed priors with a mean of zero and a variance of 0.01. A variance of 0.01 results in 95% cross-loading limits of $\pm$ 0.20 (Muthén and Asparouhov, 2012). This means that factor loadings vary in size between $\pm$ 0.20, although their means are zero. For example, this permitted the modeling of the small, non-major effects of vocabulary depth on vocabulary size. In Models 1c and 2c, informative priors were



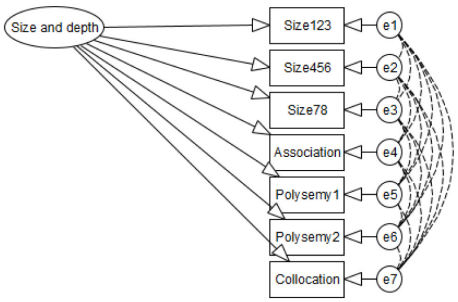
Model 1a: Single-factor model



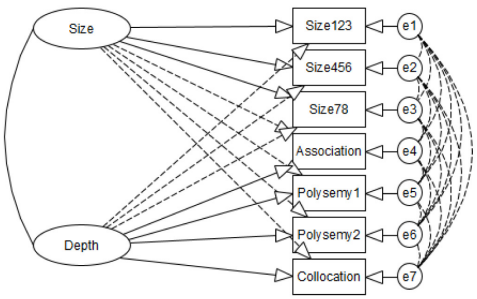
Model 2a: Correlated two-factor model



Model 2b: Correlated two-factor model (with cross loadings)



Model 1c: Single-factor model (with residual correlations)



Model 2c: Correlated two-factor model (with cross loadings and residual correlations)

FIGURE 2 | Models of vocabulary size and depth using Bayesian SEM. Undotted line = major loading; dotted line = minor loading. All residuals are correlated in Models 1c and 2c.

additionally specified for residual covariances, with inverse-Wishart distribution priors with a mean of zero and the degree of freedom of the model.

Model convergence was judged using (a) potential scale reduction (PSR) values and (b) Bayesian posterior parameter trace plots showing little change at each iteration. For (a), the value should be 1.0 at convergence, but values of less than 1.1 are considered acceptable. Model fit was assessed using posterior predictive p values of near 0.5, with 95% confidence intervals (CIs) that should be symmetric and center around 0. The models were compared using the deviance information criterion and Bayesian information criterion. For details on these criteria for convergence, model fit, and model comparison (see Muthén and Asparouhov, 2012; de Bondt and van Petegem, 2015; Norouzian et al., 2018).

RESULTS

Descriptive Statistics

Table 2 shows that internal consistency was high for all vocabulary tests ($\alpha = 0.74$–0.88), except for the Word Association

Test (α = 0.56). Pearson product-moment correlations between size and depth indicators ranged from small to moderate (0.297–0.793), which were lower than the criterion for concerns about multicollinearity ($r \geq 0.9$; Tabachnick and Fidell, 2014). Note that correlations between Polysemy 1 and Polysemy 2 and between Association and Collocation were not high (r = 0.498, 0.297). This suggests that these tests measure rather different, marginally overlapping constructs.

While detailed analysis is conducted using SEM, results of simple correlations between size and depth are reported for the sake of comparison with previous studies. Correlations of Size with Association, Polysemy 2, and Collocation were moderate, ranging from 0.422 to 0.530. These correlations were considered similar because their 95% CIs overlapped with each other. The correlation between Size and Polysemy 1 was strong (r = 0.814), and its 95% CI did not overlap with the CIs of the correlations of Size with Association, Polysemy 2, or Collocation. Differences between Polysemy 1 and 2 in relation to their correlations with Size arose mainly due to minor differences between constructs. Polysemy 1 assessed the knowledge of primary and secondary meanings, whereas Polysemy 2 assessed the knowledge of only the secondary meaning.

Conventional SEM

As seen in **Table 3**, Models 1 and 2 fit the data well (e.g., SRMR = 0.024 and 0.018, respectively). A comparison of these two models revealed Model 2 to be the best model to represent the structure of vocabulary knowledge for the current data, as shown by a lower AIC (4,113.674 vs. 4,107.574 for Models 1 and 2, respectively) and the significant result produced by a chi-square difference test (the chi-square difference between the two models was 5.806, exceeding the critical value of 3.841 at $p < 0.05$). The standardized parameter estimates [see the column "Conventional SEM" (Model 2) in **Table 4**] show that each vocabulary factor was, overall, well explained by the tests (vocabulary size: 0.832 for Size123 to 0.899 for Size456; vocabulary depth 0.496 for Collocation to 0.901 for Polysemy 1). The vocabulary size and depth factors were highly correlated (r = 0.945). Thus, size and depth are considered to be separate but closely related.

Bayesian SEM

Table 5 shows the results for Bayesian estimation. Models 1a, 2a, and 2b converged, whereas Models 1c and 2c did not. For example, Model 1a had a PSR value of 1.001, which was very close to 1.0 and less than 1.1. Bayesian posterior parameter trace plots, although not reported here, showed a stable, horizontal band for the parameter in question. These results suggest the convergence of the parameters in the model. On the other hand, Models 1c and 2c failed to converge. For example, Model 1c had a PRS value of 5.463, which considerably exceeded 1.1. Bayesian posterior parameter trace plots, although not reported here, showed a widely fluctuating, horizontal band for the parameter in question. These results suggest that Models 1c and 2c displayed poor convergence for their parameters.

TABLE 2 | Summary statistics for and correlations between the vocabulary tests.

	M	*SD*	Skewness	Kurtosis	α	Full mark	(1)	(1a)	(1b)	(1c)	(2)	(3)	(4)
(1) Size	25.25	8.57	0.00	−1.07	0.91	40	–						
(1a) Size123	11.61	2.84	−0.65	−0.51	0.74	15	0.891 [0.862, 0.913]	–					
(1b) Size456	8.37	3.86	0.10	−0.96	0.83	15	0.942 [0.927, 0.955]	0.743 [0.683, 0.794]	–				
(1c) Size78	5.29	2.64	0.10	−0.98	0.82	10	0.906 [0.881, 0.926]	0.724 [0.660, 0.778]	0.793 [0.742, 0.834]	–			
(2) Association	16.82	3.52	−0.03	−0.27	0.56	30	0.530 [0.435, 0.613]	0.485 [0.385, 0.573]	0.486 [0.386, 0.574]	0.486 [0.386, 0.574]	–		
(3) Polysemy 1	11.39	4.25	0.03	−0.94	0.78	20	0.814 [0.768, 0.852]	0.721 [0.656, 0.775]	0.769 [0.713, 0.815]	0.739 [0.678, 0.790]	0.514 [0.418, 0.599]	–	
(4) Polysemy 2	8.56	4.69	−0.37	−0.67	0.88	17	0.504 [0.406, 0.590]	0.451 [0.348, 0.544]	0.466 [0.364, 0.557]	0.467 [0.365, 0.558]	0.351 [0.239, 0.454]	0.498 [0.399, 0.585]	–
(5) Collocation	10.83	4.06	−1.06	1.54	0.77	20	0.422 [0.315, 0.518]	0.359 [0.247, 0.461]	0.403 [0.295, 0.501]	0.392 [0.283, 0.492]	0.297 [0.181, 0.405]	0.453 [0.350, 0.546]	0.368 [0.257, 0.469]

N = 255. Size123 = size with 1,000–3,000 levels; Size456 = size with 4,000–6,000 levels; Size78 = size with 7,000 and 8,000 levels. [] = 95% confidence intervals (CIs). All correlations were statistically significant, $p < 0.01$. (1) Size was a composite of (1a) Size123, (1b) Size456, and (1c) Size78 and was not analyzed using not analyzed using structural equation modeling (SEM). Instead, these three variable were analyzed using SEM.

TABLE 3 | Results from maximum likelihood estimation with robust standard errors estimation for the single-factor and the correlated two-factor models.

	χ^2	*df*	*p*	CFI	RMSEA [90% CI]	SRMR	AIC	Fit?	Best-fitting model?
Criteria	The smaller, the better	–	> 0.05	>0.90	< 0.08	<0.08	The smaller, the better	–	–
Model 1 Single-factor	16.153	14	0.304	0.997	0.023 [0.000, 0.068]	0.024	4,113.674	Yes	
Model 2 Correlated two-factor	8.391	13	0.817	1.000	< 0.001 [0.000, 0.039]	0.018	4,107.574	Yes	Yes

N = 255. CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; AIC, Akaike information criterion. The scaling correction factors for maximum likelihood estimation with robust standard errors estimation (MLR) were 1.0116 and 0.9821 for Models 1 and 2, respectively.

A comparison of converged models – Models 1a, 2a, and 2b – revealed that Models 2a and 2b were statistically equally likely, which is supported by similar values for model fit indices. Nevertheless, Model 2b was considered to best represent the structure of vocabulary knowledge for the current data, given that it was sensible to specify small-variance cross-loadings: The role of vocabulary size was very small but not zero in the vocabulary depth tests, and the role of vocabulary depth was, likewise, very small but not zero in the size test. The standardized parameter estimates [see the column "Bayesian SEM" (Model 2b) in **Table 4**] show that each skill factor was, overall, well explained by the tests (vocabulary size: 0.817 for Size123 to 0.900 for Size456; vocabulary depth: 0.518 for Collocation to 0.877 for Polysemy 1). The vocabulary size and depth factors were highly correlated (r = 0.943). Cross-loadings were very close to zero. This shows that the size and depth measures were successful in assessing separate constructs. In conclusion, as with the results from conventional SEM, Bayesian SEM showed that both size and depth are separately modeled but closely related.

DISCUSSION

To examine the relationship between vocabulary size and depth for low- to intermediate-level Japanese learners of English using conventional and Bayesian SEM, the following research question was addressed: Which factor structure of the size and depth of vocabulary knowledge explains the data better, a single-factor or a correlated two-factor model? Vocabulary knowledge was modeled as a single factor (i.e., vocabulary size and depth as one entity) and as two correlated factors (i.e., vocabulary size and depth as separately conceptualized), and model fit was compared. The results of conventional SEM showed that the correlated two-factor model best explained the data. The results from Bayesian SEM showed that the best-fitting model was the correlated two-factor model with very small cross-loadings. For both models, vocabulary size and depth factors were highly correlated (r = 0.945 for conventional SEM and 0.943 for Bayesian SEM). Thus, both size and depth are closely related yet two separate constructs. It is worth recalling that the structure of language ability has been an important research area, and the current results of having two lexical components strongly correlated to each other would add to the existing literature of the multicomponential nature of language ability.

The strong relationship between size and depth is consistent with a previous study using SEM (Vafaee, 2016). The adoption of the correlated two-factor model over the single-factor model in the current study as well as in Vafaee (2016) suggests that, even with very high correlations of 0.9 or above (r = 0.943–0.945 in the current study; r = 0.94 in Vafaee, 2016), distinguishing the two factors better explains the data than analyzing them as one factor. This means that size and depth should be distinguished conceptually and statistically. In other words, a person who knows more words (i.e., vocabulary size) tends to have a deeper vocabulary knowledge (i.e., vocabulary depth),

TABLE 4 | Parameter estimate of the correlated two-factor model.

	Conventional SEM (Model 2)				Bayesian SEM (Model 2b)			
	Size		Depth		Size		Depth	
	Unstandardized	Standardized	Unstandardized	Standardized	Unstandardized	Standardized	Unstandardized	Standardized
Size123	1.000	0.832			**1.000**	**0.817**	0.026	0.014
Size456	1.080	0.899			**1.102**	**0.900**	−0.002	−0.001
Size78	1.051	0.875			**1.074**	**0.877**	−0.007	−0.003
Association			1.000	0.582	0.024	0.019	**1.000**	**0.559**
Polysemy 1			1.548	0.901	0.027	0.022	**1.570**	**0.877**
Polysemy 2			0.971	0.565	−0.004	−0.003	**1.018**	**0.568**
Collocation			0.853	0.496	−0.028	−0.022	**0.929**	**0.518**
Correlation and covariance between size and depth	0.458	0.945			0.432	0.943		

N = 255. SEM = structural equation modeling. All unstandardized parameters except for those fixed to 1 for identification were statistically significant. For Bayesian SEM, values in bold indicate hypothesized major loadings.

but researchers and practitioners should still consider size and depth separately.

It should be recalled that Vafaee (2016) and the current study differ in (a) the measures used to assess size and depth, (b) the aspects assessed by depth tests, and (c) the participants' L2 proficiency levels and their L1: Vafaee (2016) used an aural size test of meaning recall and an aural depth test (the WAF) of selecting synonyms and collocations, whereas we used a written size of form recognition and four depth tests focusing on association, polysemy, and collocation. Vafaee's (2016) participants were Persian learners of English at lower-intermediate to advanced levels, whereas the current study's participants were Japanese learners of English at low to intermediate levels. These differences may suggest some degree of generalizability regarding the factor structure of size and depth.

However, both studies involved L2 adult learners of English as a foreign language and multiple-choice formats of assessing size and depth. The relatively similar type of learners and the use of the same formats for measuring size and depth may have produced similar results across studies. Additionally, there was an overlap in the assessed depth aspects, with collocation tested in both studies. Synonym in Vafaee (2016) and "association and polysemy" in the current study are also related to meaning and are more similar to size (defined as knowledge of a word form and a primary meaning) than are other aspects of depth such as word parts. Vafaee (2016) and the current study showed that size and depth can be separately modeled, even when depth is operationalized as something similar to size. Thus, we can assume that when depth is operationalized as something more different from size, size and depth can also be separately modeled, with different degrees of correlations expected between size and depth, but this requires empirical research. Note that the results in the final models suggest that association, polysemy, and collocation measures primarily assess depth, not size: In Model 2 in conventional SEM, the three depth aspects loaded on the depth factor only (β = 0.496–0.901); in Model 2b in Bayesian SEM, the three depth aspects loaded on the depth factor to a large degree (β = 0.518–0.877) and on the size factor to negligible degrees (β = -0.022–0.022).

The use of SEM can help clarify the latent relationships between size and depth, but it is usually difficult to compare the SEM results with previous studies using simple correlation. We suggest three ways for effectively using previous study results. First, it is possible to model relationships using SEM when previous studies report means and *SD*s of the variables, and all correlations between them (see In'nami and Koizumi, 2010, 2012b; Vafaee and Kachinske, 2019), to examine how the latent factors of size and depth are correlated. However, this is often difficult because of the lack of reports of necessary statistics (Larson-Hall and Plonsky, 2015; see also Kline, 2016, p. 65, for cases where summary statistics are not enough and raw data are required) and the research design of previous studies. For example, previous studies often used only one measure of size, and it is, therefore, difficult to model it as a size factor.

The second and third methods use simple correlations and other descriptive statistics. Simple comparisons between correlation coefficients are not very productive because they are often affected by measurement error and sample size. In the second method, using reliability coefficients reported by previous studies, researchers can estimate the strength of correlation coefficients in the case of perfect test reliabilities (i.e., with no measurement error), using a formula for correcting for attenuation (Glass and Hopkins, 1996):

(Correlation coefficient between the first and second tests)/($\sqrt{}$[reliability coefficient of the first test] × $\sqrt{}$[reliability coefficient of the second test]).

For example, the correlation between Size and Collocation was r = 0.422, with the reliability of the two tests being α = 0.91 and 0.77, in the current study. The corrected correlation is 0.575 (i.e., 0.422/[$\sqrt{}$0.91 × $\sqrt{}$0.77]), which is higher than the original value, 0.422. Thus, the use of this method of correcting for attenuation allows researchers to examine relationships while at the same time accounting for measurement error. This concept is similar to the one used in SEM (Hancock and Schoonen, 2015). The

TABLE 5 | Results from Bayesian estimation for the single-factor and the correlated two-factor models.

			Convergence			Model fit					
			PSR	**BPPTPs**	**Converged?**	**PP *p***	**95% CrI of PP *p***	**DIC**	**BIC**	**Fit?**	**Best-fitting?**
		Criteria	1.0; 1.1 or less	Changes little on each iteration	–	Near 0.5	Symmetric, centering around 0	The smaller, the better	The smaller, the better	–	–
Model 1a	Single-factor	Non-informative priors	1.001	Changes little on each iteration	Yes	0.392	−18.620, 24.364	4,113.585	4,188.459	Yes	
Model 1b	Single-factor	Informative priors (SVPs for CLs)	This model was not tested, since it had only one factor and cross-loadings were not applicable accordingly.								
Model 1c	Single-factor	Informative priors (SVPs for RVs)	5.463	Changes much on each iteration	No	0.535	-24.541, 22.141	4,123.138	4,295.869	–	
Model 2a	Correlated two-factor	Non-informative priors	1.007	Changes little on each iteration	Yes	0.661	-25.924, 16.367	4,106.968	4,186.052	Yes	
Model 2b	Correlated two-factor	Informative priors (SVPs for CLs)	1.001	Changes little on each iteration	Yes	0.670	-26.900, 16.065	4,103.706	4,225.306	Yes	Yes
Model 2c	Correlated two-factor	Informative priors (SVPs for CLs and RVs)	5.493	Changes much on each iteration	No	0.450	-22.548, 24.308	4,123.121	4,344.999	–	

N = 255. PSR, potential scale reduction; BPPTPs, Bayesian posterior parameter trace plots; PP p, posterior predictive p-value; CrI, credibility interval; DIC, deviance information criterion; BIC, Bayesian information criterion; SVPs, small-variance priors; CLs, cross-loadings; RVs, residual variances. Results are based on 10,000 iterations. Non-convergent model were further tested with 100,000 iterations. Models 1c and 2c still failed to meet convergence criterion in terms of PSR and BPPTPs.

TABLE 6 | Reliability and confidence intervals of correlation coefficients between size and depth in previous studies.

	n	Reliability	Original *r* reported in the article	95% CI of *r*
Vermeer (2001)	25	Size: NR Depth: α = 0.85	0.72–0.76	0.455, 0.868 0.522, 0.888
Akbarian (2010)	112	Size and depth: NR	0.864	0.808, 0.904
Schmitt and Meara (1997)	88	Size and depth: NR	0.27–0.62	0.065, 0.453 0.472, 0.734
Vafaee (2016)	263	Size: α = 0.67–0.84 Depth: α = 0.92–0.93	0.64–0.77	0.563, 0.706 0.716, 0.815
Current study	255	Size: α = 0.91 Depth: α = 0.56–0.77	0.422–0.814	0.315, 0.518 0.768, 0.852

NR, not reported. We used the website for calculating a "confidence interval for an observed (Pearson) correlation coefficient" (http://vassarstats.net/rho.html).

second method is relatively simple but is often difficult to execute because reliability results for all the variables are not always reported (Larson-Hall and Plonsky, 2015). **Table 6** summarizes the previous studies reviewed in the *Literature Review* as well as the current study. It shows that the three studies using simple correlations (i.e., Schmitt and Meara, 1997; Vermeer, 2001; Akbarian, 2010) did not report reliability sufficiently, and this hampers the use of the second method.

The third method uses 95% CIs of correlation coefficients. The use of CIs allows researchers to view sampling statistics (i.e., correlation coefficients) as values that fluctuate. In fact, researchers are encouraged to report CIs along with the point estimates (e.g., means and effect sizes; Norris et al., 2015; American Psychological Association, 2020). Although CIs are not always reported (Larson-Hall and Plonsky, 2015), CIs of correlation coefficients can be calculated using free online calculators[3]. The information of 95% CIs shows that if similar studies are conducted many times, 95% of those CIs will capture the population correlation. If 20 studies are conducted on the same topic, 19 (20*0.95) of those CIs will capture the population correlation. Considering CIs along with the point estimates allows researchers to interpret results more accurately. For example, correlations between Size and Collocation (r = 0.422) and between Size and Association (r = 0.530) appear different, but in fact they are not, considering the substantial overlap of their 95% CIs (0.315 and 0.518 for the former; 0.435 and 0.613 for the latter; see **Table 2**). **Table 6** shows 95% CIs of the correlations of previous studies and the current study, suggesting a similarity of relationships. The results can be quite a contrast when compared with the results using correlation coefficients only. For example, Vermeer (2001) has a wide CI (e.g., 0.455, 0.868) because the number of participants is small (n = 25), and the lower end of the CI (0.455) is very close to the upper end of the CI in Schmitt and Meara (1997; 0.065, 0.453). Schmitt (2014) suggested that different degrees of correlations between size and depth may be derived from different measures and different depth aspects measured, but different sample sizes and resulting measurement error may also be other factors. Future research should consider using the abovementioned three methods, especially CIs, for comparing previous studies. Additionally, other methods that would provide more precise estimates would be (a) a bootstrapping method (McLean et al., 2020), which is useful when primary data are available, and (b) meta-analysis, which can systematically integrate previous studies while taking sample size and measurement error into account (e.g., Plonsky and Oswald, 2015; In'nami et al., 2020). When researchers obtain a matrix of meta-analyzed correlation coefficients through (b), they can more rigorously examine relationships of size and depth using meta-analytic SEM (Cheung, 2015).

From a methodological viewpoint, it is important to note that for the best-fitting correlated two-factor model, the vocabulary size and depth factors were highly correlated (r = 0.945 for conventional SEM and 0.943 for Bayesian SEM with cross-loadings). Recall that Bayesian SEM was used in the current study to more flexibly examine the factor structure of vocabulary size and depth by specifying cross-loadings and residual correlations. Obtaining similar factor structures with similarly high correlations between vocabulary size and depth for conventional and Bayesian SEM indicates the robustness of such structures and correlations. This was revealed only after comparing the findings from conventional and Bayesian SEM approaches. It should be noted that this does not mean that conventional SEM is sufficient and the use of Bayesian SEM is redundant. Bayesian SEM allows for more varied specifications of parameter distribution that were not used in the current study. This advantage of Bayesian SEM should be best employed in future vocabulary studies.

CONCLUSION

We examined the factor structure of the size and depth of vocabulary knowledge with five tests focusing on size and depth (association, polysemy, and collocation). We found that a correlated two-factor model explained the data better than a single-factor model. As introduced in the *Literature Review*, Schmitt (2014) asked, "Do size and depth behave as separate constructs," "or are they essentially the same construct?" (p. 941). Our answer to these questions based on the findings of the current study is affirmative for the first question: Size and depth can be considered separate constructs, even when depth is measured by tests assessing aspects related to meaning and more similar to size.

Our study highlights the importance of distinguishing size and depth as two correlated but separate aspects of L2 vocabulary knowledge. This finding has implications for practice and theory. First, for L2 vocabulary assessment, if the purpose of the tests is to assess overall vocabulary knowledge, both size and depth should be included in tests to minimize construct underrepresentation of vocabulary knowledge. Given that vocabulary knowledge consists of size and depth, the inclusion of both aspects should maximize and best represent the construct of vocabulary knowledge (see Tseng and Schmitt, 2008; Zhang, 2012; Koizumi

[3]https://www.psyctc.org/stats/R/CI_correln1.html; http://vassarstats.net/rho.html

and In'nami, 2013; Vafaee and Suzuki, 2020; for examples). Further, in scoring and interpreting tests that include size and depth, separate scoring and interpretation of size and depth is justified, even though the two sections can be combined into a total score, given the high correlation between size and depth. When test developers or users conduct a validation study of their vocabulary tests that include size and depth, they should consider modeling these two when possible so as to examine the factor structure of their tests. For example, they can model a size factor using different frequency band scores and a depth factor using different section scores assessing depth. Moreover, in L2 instruction, teachers need to consider enhancing both size and depth as possible instructional goals and design, and allocate tasks for increasing these lexical components in a course (see Nation, 2013; Webb and Nation, 2017; Newton, 2020, for suggestions on effective learning and teaching in class).

Second, for theory building, the use of SEM for model construction and testing is helpful in empirically investigating relationships by considering measurement error. This allows researchers to examine the relationship between constructs of interest while separately estimating the impact of measurement error. This is how the close relationship (more than $r = 0.9$) between vocabulary size and depth was revealed in the current study. It should be recalled that simple correlations between size and depth measures were not that strong (0.422–0.793), which clearly shows one of the strengths of SEM. Taking this step further, minor cross-loadings and minor residual correlations were specified by Bayesian SEM. This would not have been possible with conventional SEM. These features of Bayesian SEM should help construct more realistic models to test specific hypotheses.

We have expanded previous studies and examined three depth aspects (association, polysemy, and collocation) by using separate tests, not the WAF. However, our results may be limited, as we targeted only Japanese adult learners of English at low to intermediate levels. What is needed are studies with different types of learners of diverse L1 and L2.

Further, we used only one measure of size (with three indicators at different levels of word frequency) and four measures of depth. There are other aspects that are important but were not examined (e.g., spoken forms, word parts, grammatical functions), and future research should include measures of size and depth by using various instruments (e.g., Godfroid, 2020, for offline and online measures to cover Nation, 2020, vocabulary elements) to examine their relationships. Tests should be developed, and test validation should be conducted, by following the principles summarized by Schmitt et al. (2020).

Specifically, to improve measures in the current study, the following three points are stated. First, we used only the multiple-choice format. A problem with this format is that it allows guesswork and overestimates the scores. Gyllstad et al. (2015) showed that takers of multiple-choice tests make educated (e.g., using the knowledge of word family) as well as blind guesses. For a more valid measurement, recall and other formats should also be employed. Second, the vocabulary size test may have had a low sampling rate. According to Gyllstad et al. (2015), 30 words (taken from a pool of 1,000 words) function with greater precision than 10 words in a vocabulary size test. In the current test, we included a maximum of 40 items in the test battery (a total of 130 items to be attempted in 80 min) keeping in mind the issue of test takers' concentration. Future research should consider how to manage the balance between the assessment need for including more items – so as to have a representative sample of vocabulary – and the practical need to reduce the items (Gyllstad, 2020). Third, some items in depth tests may need improvement. In particular, the collocation test should use a standard recently employed for selecting collocations, such as a minimum frequency of 10–50 in a corpus and an MI score of 3.00 or more (Kremmel and Schmitt, 2016; Nguyen and Webb, 2017).

In order to provide researchers with a comprehensive picture of relationships between size and depth, the following are required: a wider range of participants, and size and depth measures with larger item size, better quality, and wider focus. Comparisons of size–depth relationships across different L2 proficiency groups using multi-sample SEM (In'nami and Koizumi, 2012a; Zhang et al., 2014) would further elucidate intricate relationships. It would also be possible to model various aspects of depth separately as latent factors to specify more precise models of size and depth (see Stewart et al., 2013, for an attempt at modeling spoken vocabulary knowledge, polysemous word knowledge, and contextual word knowledge). The current study's insight into the highly correlated but distinctive nature of size and depth should help researchers advance the understanding of the structure of vocabulary knowledge.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RK conceptualized, designed, and conducted the study, organized the data, performed the conventional SEM, and wrote the first draft of the manuscript and revised it throughout. YI performed the Bayesian SEM, revised the manuscript and added sections to the manuscript. RK and YI contributed to manuscript revision and read and approved the submitted version.

FUNDING

This study was funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (B) and (C) Grant Nos. 22320110 and 26370737.

ACKNOWLEDGMENTS

We would like to thank Masamichi Mochizuki, Kazumi Aizawa, Toshihiko Uemura, Naoki Sugimori, Shin'ichiro Ishikawa, Tatsuo Iso, Kiwamu Kasahara, and Kenji Tagashira for their invaluable support for this research. We are also very thankful to the two reviewers for their appropriate and constructive feedback to improve our paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00618/full#supplementary-material

REFERENCES

Akase, M. (2005). The roles of breadth and depth of vocabulary knowledge in EFL reading comprehension: with a focus on English major students. *Ann. Rev. English Lang. Educ. Jp.* 16, 141–150. doi: 10.20581/arele.16.0_141

Akbarian, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System* 38, 391–401. doi: 10.1016/j.system.2010.06.013

American Psychological Association (2020). *Publication Manual of the American Psychological Association*, 7th Edn. Washington, DC: American Psychological Association, doi: 10.1037/0000165-000

Anderson, R. C., and Freebody, P. (1981). "Vocabulary knowledge," in *Comprehension and Teaching: Research Reviews*, ed. J. T. Guthrie (Newark, DE: International Reading Association), 77–117.

Arbuckle, J. L., and Wothke, W. (1995). *Amos 4.0 User's Guide*. Chicago: SmallWaters Corporation.

Bachman, L. F., and Palmer, A. S. (1983). "The construct validation of the FSI oral interview," in *Issues in Language Testing Research*, ed. J. W. Oller Jr. (Rowley, MA: Newbury House), 154–169.

Bae, J., Bentler, P. M., and Lee, Y.-S. (2016). On the role of content in writing assessment. *Lang. Assess. Q.* 13, 302–328. doi: 10.1080/15434303.2016.1246552

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. eds (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.

Cheung, M. W.-L. (2015). *Meta-Analysis: A Structural Equation Modeling Approach*. West Sussex: John Wiley & Sons.

Daller, H., Milton, J., and Treffers-Daller, J. (2007). "Editors' introduction: conventions, terminology and an overview of the book," in *Modelling and Assessing Vocabulary Knowledge*, eds H. Daller, J. Milton, and J. Treffers-Daller (Cambridge, MA: Cambridge University Press), 1–32.

de Bondt, N., and van Petegem, P. (2015). Psychometric evaluation of the Overexcitability Questionnaire–Two applying Bayesian structural equation modeling (BSEM) and multiple-group BSEM-based alignment with approximate measurement invariance. *Front. Psychol.* 6:1963. doi: 10.3389/fpsyg.2015.01963

Educational Testing Service (2019). *2018 Report on Test Takers Worldwide–TOEIC® Listening & Reading Test*. Princeton, NJ: Author.

Farhady, H. (1983). "On the plausibility of the unitary language proficiency factor," in *Issues in Language Testing Research*, ed. J. W. Oller Jr. (Rowley, MA: Newbury House), 11–28.

Glass, G. V., and Hopkins, K. D. (1996). *Statistical Methods in Education and Psychology*, 3rd Edn. Boston, MA: Allyn & Bacon.

Godfroid, A. (2020). "Sensitive measures of vocabulary knowledge and processing: expanding Nation's framework," in *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb (London: Routledge), 433–453. doi: 10.4324/9780429291586-28

Gorfein, D. A., Viviani, J. M., and Leddo, J. (1982). Norms as a tool for the study of homography. *Mem. Cogn.* 10, 503–509. doi: 10.3758/BF03197654

Gyllstad, H. (2020). "Measuring knowledge of multiword items," in *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb (London: Routledge), 387–405. doi: 10.4324/9780429291586-25

Gyllstad, H., Vilkaitė, L., and Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: issues with guessing and sampling rates. *Int. J. Appl. Linguist.* 166, 278–306. doi: 10.1075/itl.166.2.04gyl

Hancock, G. R., and Schoonen, R. (2015). Structural equation modeling: possibilities for language learning researchers. *Lang. Learn.* 65(Suppl. 1), 160–184. doi: 10.1111/lang.12116

Henriksen, B. (1999). Three dimensions of vocabulary development. *Stud. Sec. Lang. Acquis.* 21, 303–317. doi: 10.1017/S0272263199002089

Hu, L.-T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Struc. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

In'nami, Y., and Koizumi, R. (2010). Can structural equation models in second language testing and learning research be successfully replicated? *Int. J. Test.* 10, 262–273. doi: 10.1080/15305058.2010.482219

In'nami, Y., and Koizumi, R. (2011). Structural equation modeling in language testing and learning research: a review. *Lang. Assess. Q.* 8, 250–276. doi: 10.1080/15434303.2011.565844

In'nami, Y., and Koizumi, R. (2012a). Factor structure of the revised TOEIC® test: a multiple-sample analysis. *Lang. Test.* 29, 131–152. doi: 10.1177/0265532211413444

In'nami, Y., and Koizumi, R. (2012b). "Reproduction of structural equation models in second language testing and learning research," *Reports of 2011 Studies in Japan Association for Language Education and Technology*, Kansai Chapter, Methodology Special Interest Groups (SIG), Vol 2, 15–40.

In'nami, Y., Koizumi, R., and Tomita, Y. (2020). "Meta-analysis in applied linguistics," in *The Routledge Handbook of Research Methods in Applied Linguistics*, eds J. McKinley and H. Rose (New York, NY: Routledge), 240–252. doi: 10.4324/9780367824471-21

Ishii, T., and Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC J.* 40, 5–22. doi: 10.1177/0033688208101452

JACET Basic Word Revision Committee (2016). *Nihonjin Daigakusei yo kihon goi shin JACET8000 [The new JACET list of 8000 Basic Words]*. Tokyo: Kirihara Shoten.

JACET Basic Word Revision Committee Ed. (2003). *JACET List of 8000 Basic Words*. Tokyo: Author.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Measure.* 50, 1–73. doi: 10.1111/jedm.12000

Kieffer, M., and Lesaux, N. (2012). Knowledge of words, knowledge about words: dimensions of vocabulary in first and second language learners in sixth grade. *Read. Writ.* 25, 347–373. doi: 10.1007/s11145-010-9272-9

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling*, 4th Edn. New York, NY: Guilford.

Koizumi, R., and In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *J. Lang. Teach. Res.* 4, 900–913. doi: 10.4304/jltr.4.5.900-913

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2019). *MVN*. Available online at: https://cran.r-project.org/web/packages/MVN/MVN.pdf

Kremmel, B., and Schmitt, N. (2016). Interpreting vocabulary test scores: what do various item formats tell us about learners' ability to employ words? *Lang. Assess. Q.* 13, 377–392. doi: 10.1080/15434303.2016.1237516

Larson-Hall, J., and Plonsky, L. (2015). Reporting and interpreting quantitative research findings: what gets reported and recommendations for the field. *Lang. Learn.* 65(Suppl. 1), 127–159. doi: 10.1111/lang.12115

Laufer, B., Elder, C., Hill, K., and Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Lang. Test.* 21, 202–226. doi: 10.1191/0265532204lt277oa

Little, T. D., Cunningham, W. A., and Shahar, G. (2002). To parcel or not to parcel: exploring the question, weighing the merits. *Struc. Equ. Model.* 9, 151–173. doi: 10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., and Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychol. Methods* 18, 285–300. doi: 10.1037/a0033266

McLean, S., Stewart, J., and Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: a bootstrapping approach. *Lang. Test.* doi: 10.1177/0265532219898380

Meade, A. W., and Kroustalils, C. W. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organ. Res. Methods* 9, 369–403. doi: 10.1177/1094428105283384

Meara, P. (2005). "Designing vocabulary tests for English, Spanish, and other languages," in *The Dynamics of Language Use*, eds C. S. Butler, M. A. Gómez-González, and S. M. Doval-Suárez (Amsterdam: John Benjamins), 271–285. doi: 10.1075/pbns.140.19mea

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.

Mochizuki, M., and Aizawa, K. (2000). An affix acquisition order for EFL learners: an exploratory study. *System* 28, 291–304. doi: 10.1016/s0346-251x(00)00013-0

Mochizuki, M., Uemura, T., Aizawa, K., Sugimori, N., Ishikawa, S., Iso, T., et al. (2014). *Goichishiki Sokutei Niyoru Eigo Noryoku no suitei: Goi saizu, kosei, akusesu sokudo no kantenkara [Estimation of English ability using measures of vocabulary size, organization, and access speed of vocabulary knowledge]. Report of the Grant-in-Aid for Scientific Research (B)(2010-2013), Supported by Japan Society for the Promotion of Science. Project No. 22320110*. Tokyo: Japan Society for the Promotion of Science. Available online at: http://mochvocab.sakura.ne.jp/img/file6.pdf

Mplus (2019). *Chi-square Difference Testing Using the Satorra-Bentler Scaled Chi-Square*. Los Angeles, CA: Mplus.

Muthén, B., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthén and Muthén (1998–2019). *Mplus (Version 8.3) [Computer software]*. Los Angeles, CA: Muthén & Muthén.

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language*, 2nd Edn. Cambridge, MA: Cambridge University Press.

Nation, I. S. P., and Beglar, D. (2007). A vocabulary size test. *Lang. Teach.* 31, 9–13.

Nation, I. S. P., and Webb, S. (2011). *Researching and Analyzing Vocabulary*. Boston, MA: Heinle, Cengage Learning.

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines* 5, 12–25.

Nation, P. (2020). "The different aspects of vocabulary knowledge," in *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb (London: Routledge), 15–29. doi: 10.4324/9780429291586-2

Nation, P., and Sorell, J. (2016). "Corpus selection and design," in *Making and using Word Lists for Language Learning and Testing*, ed. I. S. P. Nation (Amsterdam: John Benjamins), 95–105. doi: 10.1075/z.208.10ch10

Newton, J. (2020). "Approaches to learning vocabulary inside the classroom," in *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb (London: Routledge), 255–270. doi: 10.4324/9780429291586-17

Nguyen, T. M. H., and Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Lang. Teach. Res.* 21, 298–320. doi: 10.1177/1362168816639619

Noro, T. (2002). The roles of depth and breadth of vocabulary knowledge in reading comprehension in EFL. *Ann. Rev. English Lang. Educ. Jpn.* 13, 71–80. doi: 10.20581/arele.13.0_71

Norouzian, R., de Miranda, M., and Plonsky, L. (2018). The Bayesian revolution in second language research: an applied approach. *Lang. Learn.* 68, 1032–1075. doi: 10.1111/lang.12310

Norris, J. M., Plonsky, L., Ross, S. J., and Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Lang. Learn.* 65, 470–476. doi: 10.1111/lang.12104

Nurweni, A., and Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English Specific Purposes* 18, 161–175. doi: 10.1016/S0889-4906(98)00005-2

Ockey, G. J., and Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Lang. Assess. Q.* 12, 305–319. doi: 10.1080/15434303.2015.1050101

Oller, J. W. Jr. (1983). "Evidence for a general language proficiency factor: an expectancy grammar," in *Issues in Language Testing Research*, ed. J. W. Oller Jr. (Rowley, MA: Newbury House), 3–10.

Plonsky, L., and Oswald, F. L. (2015). "Meta-analyzing second language research," in *Advancing Quantitative Methods in Second Language Research*, ed. L. Plonsky (New York, NY: Routledge), 106–128. doi: 10.4324/9781315870908-6

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective. *Lang. Learn.* 52, 513–536. doi: 10.1111/1467-9922.00193

Qian, D. D., and Lin, L. H. F. (2020). "The relationship between vocabulary knowledge and language proficiency," in *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb (London, UK: Routledge), 66–80. doi: 10.4324/9780429291586-5

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Lang. Test.* 10, 355–371. doi: 10.1177/026553229301000308

Read, J. (1998). "Validating a test to measure depth of vocabulary knowledge," in *Validation in language assessment*, ed. A. J. Kunnan (Mahwah, NJ: Lawrence Erlbaum Associates), 41–60. doi: 10.4324/9780203053768

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

Read, J. (2004). "Plumbing the depths: how should the construct of vocabulary be defined?," in *Vocabulary in a Second Language: Selection, Acquisition, and Testing*, eds P. Bogaards and B. Laufer (Amsterdam: John Benjamins), 209–227. doi: 10.1075/lllt.10.15rea

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: reporting a score profile and a composite. *Lang. Test.* 24, 355–390. doi: 10.1177/0265532207077205

Sawaki, Y., and Sinharay, S. (2017). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Lang. Test.* 35, 529–556. doi: 10.1177/0265532217716731

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Hampshire: Palgrave MacMillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: what the research shows. *Lang. Learn.* 64, 913–951. doi: 10.1111/lang.12077

Schmitt, N., and Meara, P. (1997). Researching vocabulary through a word knowledge framework: word associations and verbal suffixes. *Stud. Sec. Lang. Acquis.* 20, 17–36. doi: 10.1017/S0272263197001022

Schmitt, N., Nation, P., and Kremmel, B. (2020). Moving the field of vocabulary assessment forward: the need for more rigorous test development and validation. *Lang. Teach.* 53, 109–120. doi: 10.1017/S0261444819000326

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

Seto, K. Ed. (2007). *Eigo Tagi Nettowaku Jiten. [Dictionary of English Lexical Polysemy]*. Tokyo: Shogakukan.

Shimamoto, T. (2005). Exploring lexical network systems of Japanese EFL learners through depth and breadth of word knowledge. *Ann. Rev. English Lang. Educ. Jpn.* 16, 121–130. doi: 10.20581/arele.16.0_121

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Lang. Test.* 22, 31–57. doi: 10.1191/0265532205lt296oa

Staehr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Stud. Sec. Lang. Acquis.* 31, 577–607. doi: 10.1017/S0272263109990039

Stewart, J., Fryer, L., and Gibson, A. (2013). Assessing the dimensionality of three hypothesized sub-skills of L2 vocabulary proficiency. *JACET J.* 56, 57–71.

Tabachnick, B. G., and Fidell, L. S. (2014). *Using Multivariate Statistics*, 6th Edn. Harlow: Pearson.

Tannenbaum, K. R. (2008). *Relationships Between Measures of Word Knowledge and Reading Comprehension in Third- and Seventh-Grade Children*. Unpublished Ph.D. dissertation, Florida State University, Florida.

Tannenbaum, R. J., and Wylie, E. C. (2013). "Mapping the TOEIC® and TOEIC Bridge™ test scores to the Common European Framework of Reference," in *The Research Foundation for the TOEIC tests: A Compendium of Studies*, Vol. II, ed. D. E. Powers (Princeton, NJ: Educational Testing Service), 6.1–6.10.

Tseng, W.-T. (2011). *Modeling Vocabulary Knowledge: A Mixed Model Approach*. Paper presented at 2011 Language Testing Research Colloquium. Ann Arbor: University of Michigan.

Tseng, W.-T., and Schmitt, N. (2008). Toward a model of motivated vocabulary learning: a structural equation modeling approach. *Lang. Learn.* 58, 357–400. doi: 10.1111/j.1467-9922.2008.00444x

Twilley, L. C., Dixon, P., Taylor, D., and Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Mem. Cogn.* 22, 111–126. doi: 10.3758/BF03202766

Vafaee, P. (2016). *The Relative Significance of Syntactic Knowledge and Vocabulary Knowledge in Second Language Listening Comprehension*. PhD dissertation, University of Maryland, Maryland, doi: 10.13016/M2B485 .

Vafaee, P., and Kachinske, I. (2019). The inadequate use of confirmatory factor analysis in second language acquisition validation studies. *Stud. Appl. Linguist. TESOL* 19, 1–18. doi: 10.7916/salt.v19i2.4184

Vafaee, P., and Suzuki, S. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Stud. Sec. Lang. Acquis.* doi: 10.1017/S0272263119000676

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Appl. Psycholinguist.* 22, 217–234. doi: 10.1017/S0142716401002041

Vilkaitė-Lozdienė, L., and Schmitt, N. (2020). "Frequency as a guide for vocabulary usefulness: high-, mid-, and low-frequency words," in *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb (London: Routledge), 81–96.

Webb, S. (2005). Receptive and productive vocabulary learning: the effects of reading and writing on word knowledge. *Stud. Sec. Lang. Acquis.* 27, 33–52. doi: 10.1017/S0272263105050023

Webb, S. (2013). "Depth of vocabulary knowledge," in *Encyclopedia of Applied Linguistics*, ed. C. Chapelle (Oxford: Wiley-Blackwell), 1656–1663. doi: 10.1002/9781405198431.wbeal1325

Webb, S., and Nation, P. (2017). *How Vocabulary is Learned*. Oxford: Oxford University Press.

Wesche, M., and Paribakht, T. S. (1993). Assessing second language vocabulary knowledge: depth versus breadth. *Can. Modern Lang. Rev.* 53, 13–40.

Winke, P. (2014). Testing hypotheses about language learning using structural equation modeling. *Ann. Rev. Appl. Linguist.*34, 102–122. doi: 10.1017/S0267190514000075

Yamashita, J., and Shiotsu, T. (2017). Comprehension and knowledge components that predict L2 Reading: a latent-trait approach. *Appl. Linguist.* 38, 43–67. doi: 10.1093/applin/amu079

Yan, X., Cheng, L., and Ginther, A. (2019). Factor analysis for fairness: examining the impact of task type and examinee L1 background on scores of an ITA speaking test. *Lang. Test.* 36, 207–234. doi: 10.1177/0265532218775764

Yanagisawa, A., and Webb, S. (2020). "Measuring depth of vocabulary knowledge," in *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb (London, UK: Routledge), 371–386.

yWorks GmbH (2000–2019). *yEd Graph Editor (Version 3.19.1.1) [Computer software]*. Germany: yWorks.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: a structural equation modeling study. *Modern Lang. J.* 96, 558–575. doi: 10.2307/23361716

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.* 11, 76–102. doi: 10.1080/15434303.2013.853770

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 April 2020
doi: 10.3389/fpsyg.2020.00617

Using Meta-Analysis and Propensity Score Methods to Assess Treatment Effects Toward Evidence-Based Practice in Extensive Reading

*Akira Hamada**

Department of English, Faculty of Languages and Cultures, Meikai University, Urayasu, Japan

This study aimed to depict the assessment process of treatment effects of extensive reading in a second language (L2) toward the establishment of an evidence-based practice. Although standardized mean differences between treatment and control groups have been applied to interpret the magnitude of treatment effects in observational studies on L2 teaching, individual effect sizes vary according to differences in learners, measures, teaching approaches, and research quality. Prior research on extensive reading has suffered from methodological restrictions, especially due to a lack of appropriate comparison between treatment and control groups. For these reasons, a retrospective meta-analysis including only studies that ensured between-group equivalence was conducted in Study 1 to estimate the effect sizes of extensive reading expected in specific teaching environments. When the focused skill of the one-semester program was reading comprehension, its effect size was predicted as d = 0.55. However, the moderator analysis showed that this treatment effect was overestimated due to selection bias in the analyzed studies and adjusted the effect size from 0.55 to 0.37. In Study 2, propensity score analysis was applied to minimize selection bias attributed to observed confounding variables in the comparison between non-randomized treatment and control groups. Data were collected from 109 Japanese university students of English who received in-class extensive reading for one semester and 115 students who attended another English class as the control group. Various types of matching were attempted, and in consideration of balancing the five covariates that might affect treatment effect estimation, the best solutions were nearest neighborhood matching without replacement, nearest neighborhood matching with replacement, and full matching. The results showed that the average treatment effects of extensive reading on all the participants (d = 0.24–0.44) and on the treated individuals (d = 0.32–0.40) were both consistent with the benchmark established in Study 1. Pedagogical implications and methodological limitations are discussed for decision-making regarding the implementation of L2 teaching practices based on research evidence.

Keywords: evidence-based practice, quantitative methods, treatment effect assessment, meta-analysis, propensity score analysis, extensive reading

OPEN ACCESS

Edited by:
Yo In'nami,
Chuo University, Japan

Reviewed by:
John Norris,
Educational Testing Service, United States
Atsushi Mizumoto,
Kansai University, Japan

***Correspondence:**
Akira Hamada
hamada.akira@meikai.ac.jp;
a.hamada.0218@gmail.com

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

Received: *18 November 2019*
Accepted: *16 March 2020*
Published: *22 April 2020*

Citation:
Hamada A (2020) Using Meta-Analysis and Propensity Score Methods to Assess Treatment Effects Toward Evidence-Based Practice in Extensive Reading. Front. Psychol. 11:617. doi: 10.3389/fpsyg.2020.00617

INTRODUCTION

Treatment effect assessment in second language (L2) teaching plays an important role in determining its efficacy and utility and in facilitating pedagogical decision-making. Theories and hypotheses of L2 pedagogy have been proposed based on the variety of scientific evidence available in this field. Regarding this evidence, L2 teaching research has reported that effect sizes consist of the magnitude of treatment effects estimated by comparing treatment and control groups (e.g., Mackey and Gass, 2015; Marsden et al., 2018a). However, effect sizes from individual studies are not always applicable to other cases for pedagogical decision-making because of differences in research quality (Lipsey and Wilson, 1993; Plonsky and Gass, 2011; Plonsky and Oswald, 2014). An additional factor is the differences in study conditions, including participants, measures, and teaching approaches (Norris and Ortega, 2000). Given that a practical concern of L2 teaching is determining the type of instruction most applicable to a given class (Sato and Loewen, 2019), it is essential that treatment effect assessment provide information that facilitates effective pedagogical decision-making.

The concept of evidence-based practice provides a useful reference for pedagogical decision-making. In evidence-based practice, evidence is graded based on the quality of individual studies' research design, validity, and applicability (Chambless and Ollendick, 2001). The present study, therefore, aimed to establish a system of treatment effect assessment founded on evidence-based practices regarding the use of extensive reading for teaching L2 reading. The treatment effect of extensive reading has been reproduced several times (Day, 2015; Waring and McLean, 2015; Yamashita, 2015) and has been synthesized as available research evidence by two meta-analyses (Nakanishi, 2015; Jeon and Day, 2016). However, prior studies on extensive reading have been problematic due to deficits in measurements (Al-Homoud and Schmitt, 2009; Beglar et al., 2012), design, and analysis (Nakanishi, 2015; Suk, 2017). To argue whether extensive reading is an evidence-based approach to teaching L2 reading, it is necessary to introduce improved methodologies for accurate assessment of its treatment effects.

LITERATURE REVIEW

Meta-Analysis for Evidence-Based Practice

Since the start of the movement toward medical evidence-based practice in the early 1990s, evidence-based practice has spread across intervention studies in psychology as well as in education. The APA Presidential Task Force on Evidence-Based Practice (2006) described it as the integration of the best research evidence with practitioners' expertise in making decisions about interventions for individuals. In applied linguistics, the concept has been interwoven with policy-level educational decision-making (Pachler, 2003). For example, Mitchell (2000) suggested that L2 researchers would be required to offer an interpretation of current research evidence while engaging in ongoing policy debates. More recently, Sato and Loewen (2019) discussed evidence-based L2 pedagogy from the perspective of transferability of L2 acquisition research for classroom-level pedagogical decision-making. This is consistent with the core idea of evidence-based practice in psychology: to make practical interventions more effective by applying empirically supported principles of treatments (Chambless and Ollendick, 2001).

TABLE 1 | Levels of evidence for practical interventions.

Research question: Does this intervention help?	
Level 1a:	Systematic review with homogeneity of randomized controlled trials
Level 1b:	Individual randomized controlled trials
Level 2a:	Systematic review with homogeneity of cohort studies
Level 2b:	Individual cohort study including low-quality randomized controlled trials
Level 3a:	Systematic review with homogeneity of case–control studies
Level 3b:	Individual case–control study
Level 4:	Case series and poor-quality cohort and case–control studies
Level 5:	Expert opinion without explicit critical appraisal

The guidelines are adapted from the therapy/prevention, etiology/harm column of Oxford Centre for Evidence-Based Medicine (2009). In this criterion, homogeneity refers to being free of worrisome degrees of results between individual studies, and studies displaying worrisome heterogeneity are tagged with a minus at the end of their designated level.

Evidence-based practice starts by determining which research evidence will assist individuals in achieving the best outcome. According to the APA Presidential Task Force on Evidence-Based Practice (2006), any practical intervention should be evaluated in terms of its efficacy and utility. Efficacy refers to the strength of research evidence for determining causal relationships between treatments and outcomes. Utility indicates the feasibility of treatments, including generalizability, acceptability of participants, costs, and benefits. Efficacy and utility are accepted as the basis of practical significance in L2 teaching research (Plonsky and Oswald, 2014). For example, evidence-based L2 pedagogy as proposed by Sato and Loewen (2019) emphasizes the importance of L2 teaching utility. To this end, they recommended using a quasi-experimental design to balance ecological validity and internal/external research validity to maximize the transferability of L2 research findings to classroom conditions.

Although multiple types of research evidence evaluate the efficacy and utility of interventions, pedagogical decisions should be made by considering a hierarchy of research evidence quality. **Table 1** summarizes the levels of evidence for interventions, developed by the Oxford Centre for Evidence-Based Medicine (2009). When addressing a research question such as, "Does this intervention help?" the highest quality evidence is the expected treatment effects obtained through a systematic review of the research outcomes of randomized controlled trials (Level 1a). L2 teaching research has also evaluated treatment effects and intervention utility from synthesized research outcomes considering factors such as differences in populations, interventions, and settings (e.g., Norris and Ortega, 2000; Plonsky and Gass, 2011; Sato and Loewen, 2019). In contrast, low-level evidence holds little priority in deciding whether an intervention is effective for learners (see Plonsky and Gass, 2011, for review). Power and precision of treatment effect estimates have been gradually accepted (Oswald and Plonsky, 2010) and, more

recently, required in L2 teaching research (Plonsky and Oswald, 2014; Marsden et al., 2018a,b).

There are two types of benchmarks for interpreting the magnitude of treatment effects in L2 teaching research. First, an L2-specific benchmark provides information on the general magnitude of treatment effects, as it is developed through the synthesis of whole domains of L2 instruction (Plonsky and Oswald, 2014). Second, treatment-specific benchmarks are based on specific domains of L2 instruction that have been separately synthesized, such as grammar teaching (Norris and Ortega, 2000), interaction (Plonsky and Gass, 2011), and extensive reading (Nakanishi, 2015; Jeon and Day, 2016). As these meta-analyses indicate that the effects of L2 teaching vary according to its approaches, treatment-specific benchmarks can be interpreted as the intrinsic effects of individual L2 instruction domains.

It is essential to refer to treatment-specific benchmarks when considering individual learners' differences. Evidence-based practice requires empirical data on what works for whom (Mitchell, 2000; Chambless and Ollendick, 2001; Pachler, 2003; APA Presidential Task Force on Evidence-Based Practice, 2006; Sato and Loewen, 2019). In meta-analysis, moderator variables are introduced to represent learner characteristics (e.g., proficiency, age, and gender), as well as teaching differences (e.g., purpose, approach, and time on task). For example, Jeon and Day (2016) and Nakanishi (2015) showed differences in the effects of extensive reading according to learner characteristics, focused skills, length of instruction, and the implementation format (see **Table 2**). This information is useful to predict what forms of extensive reading work for what kinds of learners. For example, the effect of extensive reading on reading comprehension is between $d = 0.54$ (Jeon and Day, 2016) and $d = 0.63$ (Nakanishi, 2015). In other words, meta-analysis of L2 teaching research has the potential to identify specific variables, settings, and samples prospectively to determine as yet unknown treatment effects (Oswald and Plonsky, 2010).

However, Seidler et al. (2019) criticized the retrospective nature of traditional meta-analysis because researchers' knowledge of individual study results would influence the study selection process. Inconsistencies across individual studies in measurement methods also make the integration of data difficult. To solve these issues, they claimed the advantage of prospective meta-analyses, in which "studies are included prospectively, meaning before any individual study results related to the [prospective meta-analysis'] research question are known" (p. 1). This methodology is applied to a high priority research question only when previous evidence is limited, and new studies are expected to be conducted in the future. For example, evidence regarding the treatment effect of extensive reading is limited because of a lack of an appropriate comparison between treatment and control groups (Nakanishi, 2015). Although extensive reading has been accepted as part of L2 reading instruction because its statistical significance has been consistently reproduced, its possible effects in non-randomized controlled trials in prior studies have not been accurately analyzed (McLean and Rouault, 2017). This perspective will be a new research question such as how accurately the treatment effect of extensive reading can be assessed when using a study design that approximates randomized controlled trials. After defining a research question that has not been analyzed in primary studies, a systematic literature research, a synthesis of evidence, and an interpretation and reporting of results are conducted similar to the methods used in traditional systematic reviews. During this process, planned and ongoing studies eligible for inclusion are continuously added into the meta-analysis until the results can answer the research question (Pogue and Yusuf, 1998). For a more detailed explanation of and options for prospective meta-analyses, see Watt and Kennedy (2017) and Seidler et al. (2019).

TABLE 2 | Different effects of extensive reading by moderator variables.

Moderators		Jeon and Day (2016)	Nakanishi (2015)	
		Between	Between	Pre-post
Participants	1. Middle school	0.35	−0.05	0.27
	2. High school		0.57	0.61
	3. University	0.70	0.48	1.12
	4. Adults		0.67	1.48
Focus skills	1. Reading speed	0.83	0.98	0.61
	2. Comprehension	0.54	0.63	0.72
	3. Vocabulary	0.47	0.18	1.25
Length	1. One semester	0.51	0.36	0.89
	2. Two semesters	0.59	0.52	0.74
	3. Over a year	0.60		1.92
Extensive reading format	1. Exclusive activity	0.24		
	2. Part of course	0.47		
	3. Part of curriculum	0.91		
	4. Extracurricular	0.67		

Jeon and Day, 2016 categorized participants' ages as adolescent (middle and high school level) and adults (university level and above).

In relation to the present study, one of the most critical problems with observational non-randomized data for the comparison of groups is selection bias or biased assignments of participants to treatment and control groups (Reichardt, 2009). This non-ignorable, non-randomized treatment assignment is likely to cause initial differences between the two groups in the assessment of treatment effects (Rubin, 1974). In the between-group design, therefore, we must confirm that selection bias in non-randomized data is reasonably ignorable to provide evidence that potential differences in outcome measures were not caused by selection differences extant before the treatment (e.g., Rubin, 1974; Rosenbaum and Rubin, 1983; Imai et al., 2008). Referring to descriptive statistics before adjusting outcome measures using any confounding variables may cause bias in the results of meta-analyses. For example, if control groups had higher L2 reading proficiency than treatment groups at the beginning of extensive reading, the differences between the two groups at the time of outcome measurements should be underestimated. Although some extensive reading research claimed between-group equivalence before the treatment (e.g., Beglar et al., 2012; Robb and Kano, 2013; Suk, 2017), the two meta-analyses on the topic (Nakanishi, 2015; Jeon and Day, 2016) did not examine how the primary studies attempted to reduce selection bias in between-group comparisons. Therefore, new studies that address

possible selection bias are expected to emerge in the framework of a prospective meta-analysis.

Propensity Score Analysis for Extensive Reading Research

Extensive reading is widely recognized as an effective approach to teaching reading in English as a foreign/second language (EFL/ESL) pedagogy. According to a systematic review (Day, 2015), the core principle of extensive reading is that L2 learners choose what they want to read and read as much as possible for pleasure, information, and general understanding. As criticized by Nakanishi (2015), there is no definition of extensive reading in terms of the number of books and words L2 learners read during the treatment. A variety of extensive reading formats have also been implemented according to teaching environments. For example, extensive reading is employed as an independent reading course, a part of reading course, a part of the curriculum, and an extracurricular activity (Nation and Waring, 2019). The most frequently used practice is supervised extensive reading, in which teachers help L2 learners choose reading materials and respond to their questions about the storyline, word and phrase meanings, and grammatical structures (Day, 2015). Jeon and Day's (2016) meta-analysis showed that each extensive reading format contributed to improving L2 learners' reading comprehension, fluency, and vocabulary knowledge except when it was implemented as an independent reading course.

Within the framework of evidence-based practice, however, empirical results from past extensive reading research have not been informative for theory development or pedagogical decision-making. Deficits in the assessment of treatment effects in this field have resulted in research bias and waste. L2 teaching research considers covariates possibly affecting treatment effect estimation using analysis of (co)variance and multiple regression analysis (see discussion in Plonsky and Gass, 2011). However, adjustment by means of these linear models constrains the number of confounding variables that can be controlled for because the inclusion of too many covariates in the models will make it difficult to estimate the treatment effect (e.g., Imai et al., 2008; Guo and Fraser, 2015). Instead, the current study applies a propensity score to adjust for variables that may confound the treatment effect estimation of extensive reading.

Propensity score matching – a method that has recently been adopted in medical, psychological, and educational research (Guo and Fraser, 2015; Leite, 2017), but not in L2 teaching research – is a statistical approach for reducing selection bias in treatment effect estimation by approximating complete randomized controlled trials (King and Nielsen, 2019). By definition, the treatment effect is the difference in the potential outcomes between individuals who are assigned to a treatment group and the same individuals who are assigned to a control group. However, this cannot be directly observed (Rubin, 1974). To solve this problem, Rosenbaum and Rubin (1983) developed the propensity score, or "the conditional probability of assignment to a particular treatment given a vector of observed covariates" (p. 41). This method is applied to balance the distribution of confounding variables between treatment and control groups by matching only those who have similar propensity scores.

Using the propensity score method, the average treatment effect (ATE; e.g., Imai et al., 2008) can be estimated as the effect of extensive reading on all treated and control individuals, similar to establishing the standardized mean differences between two groups. Schafer and Kang (2008) described the nature of the ATE as the average difference in potential outcomes between the groups in the following scenario: All participants are assigned to a treatment group, and then, they are assigned to a control group. Furthermore, by excluding students from a control group whose propensity score cannot be matched, the average effect on only those students who participated in the treatment can be estimated (Ho et al., 2011). This average treatment effect on the treated individuals (ATTs) is also important to consider in treatment effect assessment for pedagogical decision-making.

Learners' initial L2 reading proficiency, L2 vocabulary size, and academic performance can be regarded as the confounding variables that cause selection bias in research on extensive reading. Since Jeon and Yamashita's (2014) meta-analysis revealed that variances of L2 learners' reading comprehension can be largely explained by cognitive aspects of reading, students with higher L2 reading proficiency and larger vocabulary size at the beginning of extensive reading should gain higher scores on the outcome measures. Reciprocal causation, where the amount of L2 reading increases as a result of motivation for engagement in extensive reading (Yamashita, 2004, 2007), should also be considered. When an extensive reading program is implemented as part of a course curriculum, students will be more dedicated to extensive reading in order to get higher grades and, accordingly, more likely to be proficient in L2 reading. Moreover, students will not only engage in extensive reading but also learn to read in L2 through other learning modes, such as vocabulary and grammar exercises in the classroom. Therefore, the outcome measures should reflect the treatment effects of classroom activities in addition to those of extensive reading. These covariate effects must be reduced to evaluate the treatment effect of extensive reading on L2 reading development accurately.

Reporting treatment effects of extensive reading, adjusted by propensity score methods, will be a key element of the protocol of a prospective meta-analysis. To mitigate the methodological deficits of extensive reading research designs (Nakanishi, 2015; McLean and Rouault, 2017), new studies applying propensity score methods similar to the current study are expected to emerge. Following a guide to prospective meta-analyses (Pogue and Yusuf, 1998; Watt and Kennedy, 2017; Seidler et al., 2019), the present study attempted to harmonize the design, implementation, and outcome collection of the planned studies. In Study 1, a meta-analysis was conducted to assess the selection bias in existing research on extensive reading and to estimate the expected effect size of extensive reading practice. In Study 2, a planned study using propensity score methods was integrated with the meta-analysis results. This methodology is a nested prospective meta-analysis, which integrates prospective evidence from planned study results into existing retrospective meta-analyses (Seidler et al., 2019).

STUDY 1

Method

Study Retrieval

Two large-scale meta-analyses on extensive reading (Nakanishi, 2015; Jeon and Day, 2016) were used to obtain synthesized effect sizes. Nakanishi (2015) included 34 studies using three keywords: *extensive reading*, *pleasure reading*, and *graded readers*. Jeon and Day (2016) updated this database in terms of the self-selected reading principle of extensive reading, and six studies were excluded because they offered obligatory assigned reading. In their meta-analysis, 21 studies from 1980 through 2014 were newly added.

In the present study, we conducted a search for the latest studies, written in English and published from April 2014 to April 2019. Five databases (Education Resources Information Center, Google Scholar, Linguistics and Language Behavior Abstracts, PsycINFO, and Web of Science) were electronically searched to locate relevant studies using the same keywords as Nakanishi (2015). After periodicals had been searched, full texts of book chapters, monographs, and relevant reports were also searched by citation chasing. This literature search found 47 studies published in 15 international peer-reviewed journals such as *Reading Research Quarterly*, *Studies in Second Language Acquisition*, *TESOL Quarterly*, and *Reading in a Foreign Language*. These studies were examined to determine whether they included information necessary for the present meta-analysis.

Criteria for Inclusion and Coding

The purpose of the inclusion criteria was to examine selection bias and to recalculate expected effect sizes to represent the present teaching environment. In Study 2, university students receiving English instruction were engaged in extensive reading for one semester as part of the curriculum, to improve their reading comprehension abilities. Their initial L2 reading proficiency was low [A1 level of the Common European Framework of Reference for Languages (CEFR)] as measured by a standardized reading test, TOEIC Bridge (Educational Testing Service, 2007). To select identified studies for the meta-analysis that were similar in terms of teaching and learner characteristics, the inclusion criteria were defined as follows:

Criteria for inclusion

- Studies that target EFL and ESL learners in high school, university, or educational institutions for adults and include their L2 proficiency information.
- Studies that report a specific length of instruction.
- Studies that use tests to measure learners' reading comprehension abilities.
- Studies that implement extensive reading as part of the curriculum.
- Studies that report the numerical results obtained from between-group comparisons.
- (Prospectively, studies that apply propensity score methods to estimate the treatment effect of extensive reading.)

All classification was duplicated in accordance with Nakanishi (2015) and Jeon and Day (2016). The existing 49 studies and the 47 newly collected studies were independently coded as below by two L2 reading researchers, with an intercoder agreement ratio of 92%. Any disagreements were resolved by reexamining the primary studies. Nineteen of the existing studies and three of the newly collected studies met the inclusion criteria (the primary studies included in the Present Meta-Analysis are presented in **Supplementary Data Sheet 1**). Statistical information to be analyzed was recorded by the author and checked by the other coder.

The primary studies included in the meta-analysis operationalized their extensive reading practice according to their teaching environment. For example, Suk (2017) implemented a 15-week semester extensive reading, in which Korean EFL students received 70 min of class time for intensive reading instruction that was similar to that received by the control group and the remaining 30 min for extensive reading activities. Some activities, such as scaffolded silent reading and writing a short book report, were incorporated to facilitate their reading during the class. These instructional procedures were similar to the present study and other primary studies (e.g., Al-Homoud and Schmitt, 2009; Nakanishi and Ueda, 2011; Beglar et al., 2012; Shih, 2015). Although some primary studies systematically promoted out-of-class extensive reading (e.g., Robb and Kano, 2013; Huffman, 2014; McLean and Rouault, 2017), we did not require our students to read outside class time because they were not independent learners.

Meta-Analysis

Standardized mean differences for between-group comparisons of outcome measures were calculated as an effect size of *d*. A random-effect model was applied to synthesize the effect sizes because the treatment effect of extensive reading differed according to various moderators (Nakanishi, 2015; Jeon and Day, 2016). Since four studies conducted multiple experiments using different samples (Sims, 1996; Mason and Krashen, 1997; Lee, 2007; Robb and Kano, 2013), data from each study were included in the meta-analysis separately, resulting in the resynthesis of 33 datasets from 22 primary studies, which included 6,806 participants (treatment, n = 3,343; control, n = 3,462).

Further meta-analysis explored the variance of standardized mean differences in pretests between treatment and control

Coding of study reports

- Learner characteristics: EFL/ESL settings, school, and L2 reading proficiency self-labeled by each primary study[1] (terms such as *beginner* and *novel* were categorized as lower proficiency; terms such as *intermediate* and *advanced* were categorized as higher proficiency).
- Length of instruction: one semester, two semesters, and over a year (cf. *short, medium*, and *long*, Jeon and Day, 2016).
- Tests used: a reading comprehension test and others.
- Ways to implement extensive reading: an independent course, a part of a reading course, a part of a curriculum, an extracurricular activity, and others.
- Research design: between-group comparison and others.

[1] *How to define participants' proficiency and integrate the outcomes of participants defined by different measurements levels is a major challenge for meta-analysis of extensive reading (Nakanishi, 2015). Future research needs to use common measurements, and researchers should define participants' proficiency levels using a common scale such as CEFR at least.*

groups. A significant difference at the time of the pretests indicates selection bias related to inherent differences among participants. Eleven datasets from four primary studies did not include information on the descriptive statistics for the pretests; therefore, 22 datasets were submitted to meta-analysis (N = 1,998; treatment, n = 1,000; control, n = 998). For the moderator analysis, studies in which control groups had higher/lower L2 reading proficiency than treatment groups were labeled as "control" and "treated," respectively, in cases where the 95% confidence intervals (CIs) of d did not include zero. Studies where the 95% *CIs* of d included zero were classified as "equivalent," indicating that they used statistically equivalent groups for comparisons. Studies that did not include any information about pretest were categorized as "unspecified." For the calculation of d, the means of control groups were subtracted from the means of treatment groups. The meta-analyses were executed with the metafor package for R (Viechtbauer, 2010)[1].

Results and Discussion

Publication bias in the meta-analysis was assessed and found by a trim-and-fill method to estimate the number of missing studies because the number of published and unpublished studies was unequal (published = 18, unpublished = 3). Biased meta-analysis results lead to undesirable decisions about the treatment effect (e.g., Lipsey and Wilson, 1993; Plonsky and Oswald, 2014; Seidler et al., 2019). For the treatment effects (i.e., posttests), one missing study was added to adjust the underestimated effect size from 0.52 to 0.55. In the same way, six missing studies for the pretest data were added to recover the underestimated effect size from 0.02 to 0.18. **Figure 1** shows that these adjustments resulted in symmetrical funnel plots.

The meta-analysis results showed a large variance in standardized mean differences between treatment and control groups at the time of pretests: *Min* = −0.71, 1-quantile = −0.19, *Mdn* = −0.06, 3-quantile = 0.18, *Max* = 1.38. The variance was positively skewed (skewness = 1.05), indicating that the primary studies were more likely to use control groups with higher L2 reading proficiency than the treatment groups before treatment. The moderator analysis results showed that the treatment effects of extensive reading differed according to the selection bias (**Table 3**). As expected, studies that used control groups whose initial L2 reading proficiency was higher than that treatment groups produced the lowest treatment effect [d = −0.24, 95% CI (−0.53, 0.05)]. Studies using treatment groups whose initial L2 reading proficiency was higher than control groups obtained higher treatment effects than the other two categories [d = 0.57, 95% CI (0.26, 0.87)]. Looking at the studies using the equivalent groups [d = 0.37, 95% CI (0.24, 0.50)], it is highly possible that selection bias caused under- or overestimations of the treatment effect of extensive reading. Note that the studies with no information about pretests greatly overestimated the treatment effect [d = 0.94, 95% CI (0.82, 1.05)].

These findings suggest that the previous meta-analyses overestimated the treatment effect of extensive reading on L2 reading comprehension skills (see **Table 2**, Focus skills, Comprehension: d = 0.54 in Jeon and Day, 2016; d = 0.63 in Nakanishi, 2015). Accordingly, the treatment effects of extensive reading accumulated so far are minimally informative for theories and pedagogical decision-making within the framework of evidence-based practice. Although the use of between-group designs has been recommended due to an inflation effect caused by pre–posttest designs in L2 teaching research (e.g., Plonsky and Oswald, 2014; Mackey and Gass, 2015; Sato and Loewen, 2019), the findings of the present study further indicate the importance of ensuring between-group equivalence by controlling participant factors that may affect outcome measures.

Before considering selection bias, **Table 3** showed that the overall effect size was 0.55 [95% CI (0.39, 0.70)]. This treatment effect was expected to decrease when targeting beginner-level students [d = 0.30, 95% CI (0.12, 0.49)] and implementing one-semester extensive reading [d = 0.25, 95% CI (0.04, 0.47)]. In Study 2, we conducted a study using propensity score methods to compare the treatment effects with the benchmarks established in Study 1. The results of Study 2 were not known before defining the present inclusion criteria, and it was fully eligible for inclusion in the meta-analysis. It is the key feature of a prospective meta-analysis that studies are identified as eligible for inclusion before those results are known (Pogue and Yusuf, 1998; Seidler et al., 2019). By including such planned studies that adopt propensity score methods to estimate the treatment effect of extensive reading, a prospective meta-analysis can largely eliminate biased effect sizes.

STUDY 2

Method

Participants

We used a non-randomized controlled trial that included five intact EFL classes, and 224 Japanese EFL learners participated in Study 2 (age = 18–19 years). Two classes were assigned to a control group (n = 115; female = 77, male = 38), where the general aim of the course was to improve English speaking and writing skills. The other two classes – the treatment group – engaged in extensive reading (n = 109; female = 67, male = 42). Participants were first-year undergraduates majoring in nursing (treatment, n = 43; control, n = 46), physiotherapy (treatment, n = 66; control, n = 44), and child education (control, n = 25). By the beginning of this study, they had received 6 years of English instruction as part of their formal education in Japanese secondary schools and had not experienced any extensive reading activities. Before the treatment, informed consent was obtained, and the participants were notified of how the personal data collected would be used.

The participants were obligatorily enrolled in a weekly 90-min basic English skills course at their university. Their English reading proficiency was assessed using a 50-item standardized reading test, TOEIC Bridge (score range = 10–90; Educational Testing Service, 2007) before the treatment (at the beginning of the academic year). Their dichotomously marked reading test score showed that they were at the A1 level of the CEFR [M = 42.00, 95% CI (39.67, 44.33), SD = 17.70, Cronbach's

[1]All raw data and the R scripts used for the meta-analysis and propensity score analysis are available to readers in the IRIS digital repository (https://www.iris-database.org/iris/app/home/detail?id=york%3a937791&ref=search).

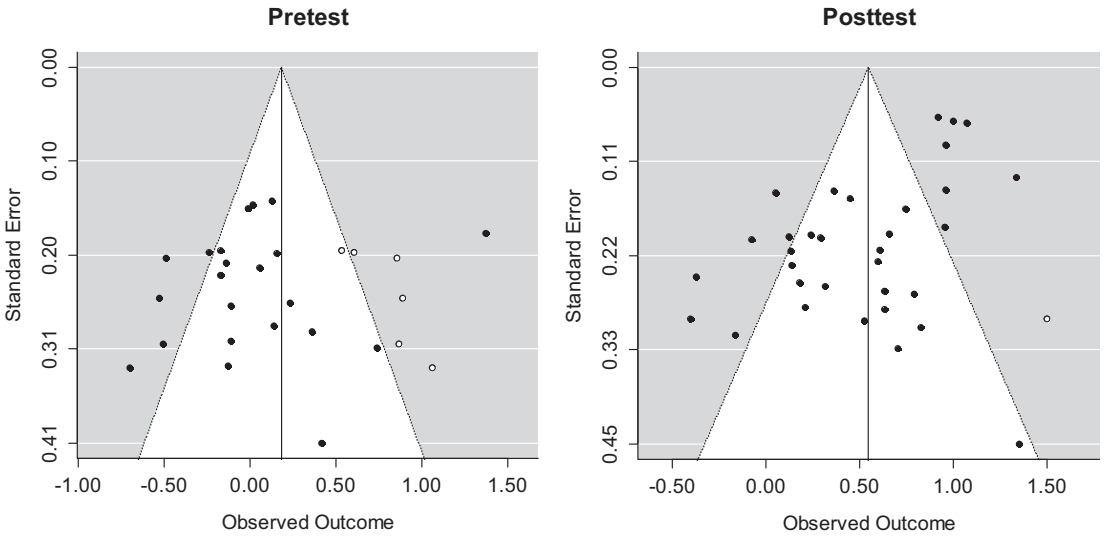


FIGURE 1 | Funnel plots after applying a trim-and-fill method to reduce the effects of the existing publication bias. Standard errors on the *y*-axis indicate the precision of each study; the largest *N*-size studies have the smallest standard error. Effect sizes *d* for each study are plotted on the *x*-axis. Diagonal lines show the expected 95% confidence intervals around the summary estimate. White dots indicate the missing studies estimated by the trim-and-fill method.

α = 0.83], indicating that the participants were not independent readers (Educational Testing Service, 2019).

Materials

The reading texts offered for extensive reading were derived from the short reading passages compiled by the Eiken Foundation of Japan. Although books such as graded readers are more appropriate for extensive reading, the length of these books may intimidate A1-level L2 readers. Nuttall (2005) recommended the use of short, appealing, varied, and easy passages for elementary readers. Accordingly, three positive reasons for using the EIKEN reading passages were as follows: (a) the reading texts were simplified in terms of word frequency and syntactic complexity, (b) the EIKEN grades were associated with the CEFR level, and (c) the text characteristics were synchronized with the Course of Study of English in Japan (see **Table 4**). Twenty-six different texts were prepared for seven grades, resulting in a total of 182 reading passages. Text genres included narrative, scientific expository, essay, and everyday language, such as emails, notices, and advertisements.

Two versions of standardized reading comprehension tests (Educational Testing Service, 2007) were used to measure participants' L2 reading proficiency at the beginning and end of the extensive reading. They consisted of 30 multiple-choice comprehension questions with 20 passages from various genres such as articles, emails, notices, forms, reports, and advertisements. To avoid testing and instrumentation effects (Reichardt, 2009), one treatment and one control group took the two tests in normal order (Test A for the pretest; Test B for the posttest), while the other two groups took them in reverse order (Test B for the pretest; Test A for the posttest). The reliability coefficients of the pretest (Cronbach's α = 0.83) and posttest (Cronbach's α = 0.89) were high.

TABLE 3 | Results of the meta-analysis for the treatment effects of extensive reading.

		Participants (*n*)		Effect sizes		
Moderators	***k***	**Treatment**	**Control**	***d***	**95% CI**	***SE***
Proficiency						
Higher	18	2,695	2,797	0.71	(0.56, 0.86)	0.08
Lower	15	648	666	0.30	(0.12, 0.49)	0.09
Instruction length						
One semester	9	368	321	0.25	(0.04, 0.47)	0.11
Two semesters	16	733	776	0.45	(0.30, 0.60)	0.08
Over a year	8	2,242	2,366	0.92	(0.74, 1.09)	0.09
Selection bias						
Control	4	141	114	−0.24	(−0.53, 0.05)	0.15
Equivalent	15	691	724	0.37	(0.24, 0.50)	0.07
Treated	3	132	123	0.57	(0.26, 0.87)	0.16
Unspecified	11	2,379	2,501	0.94	(0.82, 1.05)	0.06
Overall	33	3,343	3,463	0.55	(0.39, 0.70)	0.08

k, number of studies, CI, confidence interval, SE, standard error.

The 1,000- to 5,000-word level of a standardized vocabulary test (Koizumi and Mochizuki, 2011) was used to measure participants' L2 vocabulary size before the treatment. This test – 125 multiple-choice questions – was developed to assess the written receptive vocabulary knowledge of Japanese EFL learners. In each question, participants were given a Japanese word and instructed to select the most appropriate English translation from three options. The reliability coefficient was high (Cronbach's α = 0.95).

Participants' academic performance in a regular English class was evaluated using the average scores of two end-of-term tests prior to the treatment. The tests consisted of integrated

TABLE 4 | EIKEN grades and their Common European Framework of Reference for Languages (CEFR) level with text variables.

EIKEN grade	CEFR level	EIKEN benchmark	Mean standard words		Flesch–Kincaid grade level	
			M	*SD*	*M*	*SD*
Grade 2	B1	High school/graduates	367.45	12.56	9.25	1.02
Grade Pre-2	A2	High school/intermediates	307.38	8.81	8.31	0.90
Grade 3	A1	Junior high school/graduates	258.25	12.30	6.76	1.29
Grade 4	A1	Junior high school/intermediates	155.60	5.78	4.23	0.99

Each grade has 26 different kinds of passages.

TABLE 5 | Descriptive statistics for reading tests, L2 vocabulary size, and academic performance.

	Treatment (n = 109)			Control (n = 115)		
Measures	***M***	**95% CI**	***SD***	***M***	**95% CI**	***SD***
Pretest	7.19	(6.33, 8.06)	4.56	10.98	(10.28, 11.68)	3.67
Posttest	11.90	(10.98, 12.81)	4.97	11.91	(11.04, 12.79)	4.74
L2 vocabulary size	2704.45	(2562.82, 2846.07)	745.95	3311.38	(3193.46, 3429.31)	638.36
Academic performance	72.23	(70.39, 74.07)	9.69	79.80	(77.17, 82.43)	14.25

reading-to-writing task performance (50%), independent listening skills (15%), independent reading skills (15%), and spoken interaction (20%).

Procedure

Course work for the treatment group was broadly divided into two activities. For about 60 min in class, the treatment group relearned, through task-based language learning, grammatical and vocabulary items that had been introduced in junior and senior high schools. For the remaining 30 min, they engaged in the extensive reading at their own pace.

In the extensive reading segment, the participants were initially instructed to read EIKEN Grade 3 reading texts. After reading three texts from each grade, the participants were free to move on to higher levels; however, they were advised to read texts at lower levels if they had difficulty comprehending content. During class, they chose a reading text and engaged in sustained silent reading. Every time they finished reading a text, they briefly shared their thoughts about the contents by writing a short book report, then returned the text and took a new one for additional reading. To confirm that students had read the texts and to motivate extensive reading, a teacher monitored reading progress and answering any comprehension questions, writing brief comments after each class. Following Beglar et al. (2012), the total amount of reading by all participants was calculated using standard words comprising six characters as a nominal word.

Data Analysis

The main steps of propensity score analysis include propensity score estimation, matching and covariate balance evaluation, and treatment effect estimation (Leite, 2017). The included covariates should be true confounders that are measured before treatment assignment or are stable over time (e.g., Rosenbaum and Rubin, 1983; Imai et al., 2008; Ho et al., 2011). For propensity score estimation, this study considered as many variables as possible that could potentially determine students' participation in the treatment group. We included the following five covariates obtained before treatment: (a) initial L2 reading proficiency, (b) L2 vocabulary size, (c) academic performance, (d) gender, and (e) major in school. Although both gender and academic major were assumed not to be predictors of outcome, these were true confounders affecting the probability of treatment assignments in a non-randomized study. In other words, because the participants' gender and school faculty were not randomized when we assigned them into either treatment or control groups, both covariates were included in the analysis. Therefore, these five covariates were submitted to a stepwise logistic regression model, and propensity scores were estimated.

Propensity score matching was conducted for group participants with similar propensity scores. Since there are different matching methods, it is necessary to choose a method that shows the best balance of covariates and propensity scores. We employed and compared six different matching methods: nearest neighborhood matching without replacement, nearest neighborhood matching with replacement, genetic matching without replacement, genetic matching with replacement, optimal nearest neighborhood matching, and unconstrained full matching. For details about each matching method, see, for example, Leite (2017).

Next, both ATE and ATT were estimated. In this study, the ATE was the difference between the expected posttest values of all the participants in the treatment and control groups. The ATT was the difference between the expected posttest values of the participants in the treatment group only. The purpose of this study was to evaluate whether extensive reading was beneficial for those learners who were assigned to the treatment group (i.e., ATT) as well as whether, on average, extensive reading was beneficial for all the participants (i.e., ATE). The matching and treatment effect estimation were conducted with the MatchIt (Ho et al., 2011) and Matching (Sekhon, 2011) packages for R.

TABLE 6 | Differences in means of confounding variables by propensity score matching.

Matching methods	Treatment	Control	Standardized mean difference
Before matching (Treatment: *n* = 109, control: *n* = 115)			
Propensity score	1.58	−1.61	1.73
Initial L2 reading proficiency	7.19	10.98	1.03
L2 vocabulary size	2704.45	3311.38	0.81
Academic performance	72.23	79.80	0.78
Academic major	1.61	1.90	0.77
Nearest neighborhood matching without replacement (Treatment: *n* = 54, control: *n* = 54)			
Propensity score	0.17	−0.22	0.21
Initial L2 reading proficiency	9.43	9.83	0.09
L2 vocabulary size	3130.69	3138.89	0.01
Academic performance	73.67	75.89	0.23
Academic major	1.72	1.78	0.11
Nearest neighborhood matching with replacement (Treatment: *n* = 91, control: *n* = 41)			
Propensity score	1.17	1.11	0.03
Initial L2 reading proficiency	7.67	7.38	0.06
L2 vocabulary size	2817.62	3327.00	0.68
Academic performance	72.52	71.27	0.13
Academic major	1.67	1.41	0.54
Genetic matching without replacement (Treatment: *n* = 109, control: *n* = 109)			
Propensity score	1.58	−1.42	1.63
Initial L2 reading proficiency	7.19	11.47	0.94
L2 vocabulary size	2704.45	3297.82	0.80
Academic performance	72.23	79.22	0.72
Academic major	1.61	1.96	0.73
Genetic matching with replacement (Treatment: *n* = 109, Control: *n* = 34)			
Propensity score	1.58	1.05	0.26
Initial L2 reading proficiency	7.19	6.96	0.05
L2 vocabulary size	2704.45	2937.41	0.31
Academic performance	72.23	73.32	0.11
Academic major	1.61	1.68	0.15
Optimal nearest neighborhood matching (Treatment: *n* = 109, Control: *n* = 109)			
Propensity score	1.58	−1.39	1.61
Initial L2 reading proficiency	7.19	11.47	0.94
L2 vocabulary size	2704.45	3295.35	0.79
Academic performance	72.23	79.58	0.76
Academic major	1.61	1.94	0.67
Full matching (Treatment: *n* = 109, Control: *n* = 115)			
Propensity score	1.58	1.49	0.05
Initial L2 reading proficiency	7.19	6.63	0.12
L2 vocabulary size	2704.45	3399.81	0.93
Academic performance	72.23	67.78	0.46
Academic major	1.61	1.39	0.45

Finally, a sensitivity analysis was conducted to reveal how strongly the unidentified covariates would affect the significance test of the treatment effect. Evaluating sensitivity to the unidentified covariates is important because propensity score methods only reduce selection bias caused by observed covariates (Liu et al., 2013). The rbound package for R (Keele, 2014) was used for Rosenbaum's (2002) method to calculate *p*-values that showed how sensitive the results of treatment effect estimations were to the unidentified covariates.

Results and Discussion

Table 5 displays the descriptive statistics of the pre- and posttest results for the treatment and control groups. The treatment group read an estimated 25,000 standard words on average (*Min* = 11,630, 1-quantile = 18,235, *Mdn* = 23,505, 3-quantile = 26,865, *Max* = 42,985). A two-tailed *t*-test showed no significant difference in the posttest score between the two groups before applying the propensity score matching, $t(222) = 1.64$, $p = 0.103$, $d = 0.22$. This result can be attributed to the selection bias in this study because the control group was always better than the treatment group at initial L2 reading proficiency, L2 vocabulary size, and academic performance. These confounding variables affecting the treatment effect estimation complicated pedagogical interpretations, even though the pre–postgain score of the reading test was higher in the treatment group ($M = 4.71$) than in the control group ($M = 0.93$). These results suggest the necessity to control covariates by propensity score analysis.

For propensity score estimation, logistic regression results showed that initial L2 reading proficiency ($B = -0.198$, SE = 0.042, $p < 0.001$), L2 vocabulary size ($B = -0.001$, SE = 0.000, $p < 0.001$), academic performance ($B = -0.084$, SE = 0.016, $p < 0.001$), and academic major ($B = -1.973$, SE = 0.348, $p < 0.001$) explained 46% of variance of the treatment assignment probability. Participants' gender was not a strong predictor of the treatment assignment ($B = -0.402$, SE = 0.210, $p = 0.056$). The rank discrimination index showed that prediction by this logistic model was good [c-index = 0.89, 95% CI (0.85, 0.93)]. Thus, these four covariates were used in propensity score matching.

To select the best matching procedure, this study explored change in the absolute standardized mean differences of the propensity scores between before and after matchings. According to Leite (2017), when the absolute values of propensity scores are <0.10, covariate balances are strict, and when the absolute values are <0.25, covariate balances are lenient. **Table 6** shows that nearest neighborhood matching with replacement (0.03) and full matching (0.05) satisfied the criterion for "strict." Nearest neighborhood matching without replacement (0.21) satisfied the criterion for "lenient." **Figure 2** presents the propensity score distribution after six matching procedures, demonstrating whether there was sufficient propensity score overlap between the treatment and control groups. For example, nearest neighborhood matching with replacement, nearest neighborhood matching without replacement, and full matching all showed high overlap of the propensity scores for the matched treatment and control groups. By contrast, the other three matching procedures did not produce similarities between the matched groups. The treatment effect estimation was conducted based on these three matching procedures.

Tables 7, **8** summarize the ATEs and the ATTs of extensive reading on L2 reading improvement, estimated by the three matching procedures, respectively. Effect sizes were calculated based on the mean differences between the treatment and control

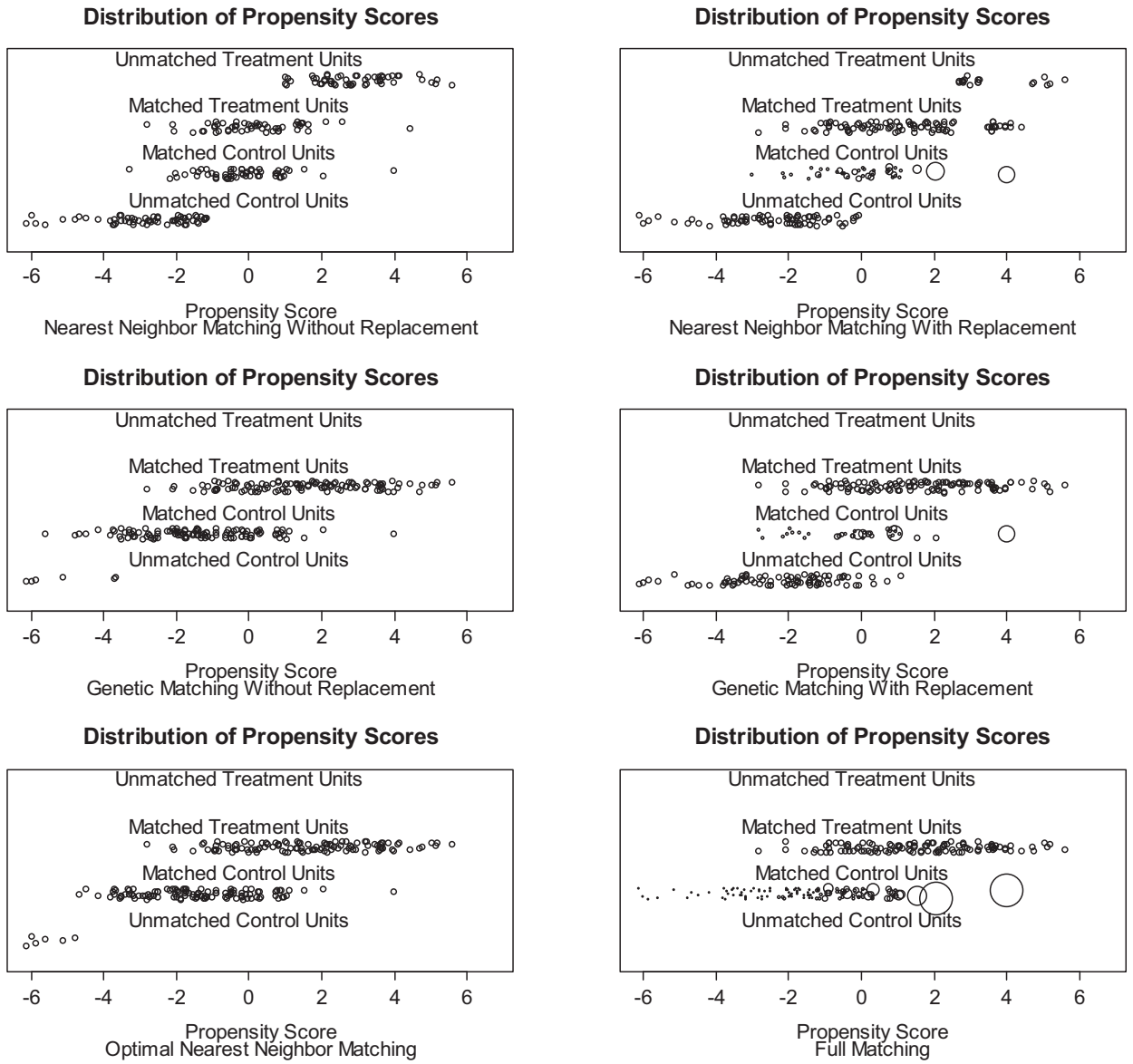


FIGURE 2 | Jitterplots displaying the distribution of propensity scores after six different matchings. Circle sizes indicate the assigned weights for group comparison regarding the treatment effect.

groups and the pooled standard deviations of the posttest. In **Table 7**, with regard to the ATE estimation after three matchings, the treatment effect increased from 0.22 (i.e., the effect size *d* calculated before ensuring between-group equivalence) to 0.24–0.44. More importantly, as shown in **Table 8**, the ATT results showed that, when matched on all covariates, the treated students' L2 reading proficiency improved significantly more than control students (*d* range = 0.32–0.40). These effect sizes were consistent with the results of the meta-analysis using the studies that ensured between-group equivalence (*d* = 0.37; see **Table 3**).

Finally, the results of the sensitivity analysis are shown in **Table 9**. According to Rosenbaum (2002), the value of gamma is interpreted as odds ratios of different probabilities of treatment assignment. If this value is close to 1, the estimated treatment effect is sensitive to unidentified covariates. In particular, a change in the lower and higher bounds of *p*-values from significant to insignificant (or vice versa) indicates the exact value of gamma to be discussed. Although this analysis can be generalized for matching procedures beyond one-to-one matching, it is not as easily implemented by the existing statistical software (Keele, 2014). Therefore, note that the sensitivity analysis with one-to-one greedy matchings (i.e., the nearest neighborhood matchings with and without replacement) was conducted but not with full matching. The results showed that, in both matching procedures, the higher bound estimates changed from significant to insignificant when gamma was 1.8. It is difficult to conclude

TABLE 7 | The average treatment effects (ATEs) of the extensive reading for different matching methods.

Matching methods	Treatment		Control			
	M	*SD*	*M*	*SD*	ATE	*d*
Nearest neighborhood matching without replacement	12.47	2.86	9.66	2.81	2.81	0.34
Nearest neighborhood matching with replacement	11.94	3.07	8.96	3.14	2.98	0.44
Full matching	12.87	2.66	10.17	2.67	2.61	0.24

whether the effects of unidentified covariates are present because the Rosenbaum's sensitivity analysis does not provide any objective criteria (e.g., Imai et al., 2008; Liu et al., 2013). However, the present results will be more robust against unidentified covariates if a large change in the odds ratio is needed by adding the covariates, theoretically affecting the treatment assignment of the extensive reading program.

GENERAL DISCUSSION

The purpose of this study was to propose the method of treatment effect assessment toward the establishment of an evidence-based practice in extensive reading. In Study 1, the existing two meta-analysis studies were reassessed for selection bias associated with primary studies to determine their quantitative reproducibility with regard to the practical significance of extensive reading. When including only the studies that ensured between-group equivalence, the effect size expected for the present extensive reading study was 0.37 [95% CI (0.24, 0.50)], indicating that the previous meta-analyses overestimated treatment effect. In Study 2, this estimation was validated by applying propensity score methods. By reducing the selection bias, this study produced ATEs and ATTs consistent with the meta-analysis results. These findings show that new primary studies should be planned for inclusion into prospective meta-analyses.

Systematic reviews and meta-analyses of the best available research evidence have the potential to inform pedagogical decision-making for L2 teaching. The current study, however, revealed that the retrospective nature of previous meta-analyses included biased interpretations regarding the treatment effect of extensive reading. The results showed significant differences in the effect sizes between studies that ensured between-group equivalence and those that did not. As many researchers have indicated that primary studies on extensive reading include methodological problems (e.g., Al-Homoud and Schmitt, 2009; Beglar et al., 2012; McLean and Rouault, 2017; Suk, 2017), the current status of existing extensive reading research is that it introduces bias and waste. In addition to future research including detailed descriptive statistics and control groups, as recommended by Nakanishi (2015), primary studies must ensure between-group equivalence by random assignment (McLean and Rouault, 2017) and by embedding propensity score adjustment in the planned research.

The current study adopted propensity score methods appropriate for addressing treatment effect estimation of extensive reading. Propensity score matching was conducted to reduce selection bias associated with possible confounding variables. The list of observed pretreatment covariates included the factors affecting outcome measures, typically considered by previous studies on extensive reading (Yamashita, 2004, 2007, 2015; Day, 2015; Waring and McLean, 2015). By matching the propensity scores between the treatment and control groups, the target population of students was defined in order to generalize causal inference about the effects of extensive reading in L2 settings. The results of the ATEs and ATTs both validated the causal inference that students who participated in extensive reading improved their L2 reading comprehension skills more than students who did not participate in the program. Following the L2-specific benchmark for effect sizes (Plonsky and Oswald, 2014), the treatment effect of extensive reading was small when the focused skill of the one-semester program for EFL students was reading comprehension (ATEs, $d = 0.24$–0.44; ATTs, $d = 0.32$–0.40). This is consistent with the primary studies that ensured the between-group equivalence (e.g., McLean and Rouault, 2017; Suk, 2017). Although the interpretation is disputable that empirical research ends in failure when the reproduced effect size is significantly lower than the meta-analyzed effect size, at least some pedagogical decision-making is necessary about why interventions are ineffective.

The robust results for meta-analyses of treatment effects are essential to implement evidence-based practice in L2 pedagogy. With respect to extensive reading, Beglar et al. (2012) pointed out that past research reporting treatment effects depended on null hypothesis significance testing. Marsden et al. (2018a) also demonstrated that the extent of reproducibility of primary L2 teaching research depended on a narrative comparison of the findings and dichotomous judgment based on null hypothesis significance testing. The present study showed the importance of considering the degree to which treatment effect would be expected in L2 teaching, based on meta-analysis. In particular, moderator analysis was used to inform variability and predictability of treatment effects of extensive reading (see also Nakanishi, 2015; Jeon and Day, 2016). This treatment effect assessment provides research evidence to interpret to what extent particular L2 teaching formats work successfully and for whom. As suggested by Oswald and Plonsky (2010), effect sizes predicted *a priori* must be used as criteria for interpreting the outcomes of L2 teaching. Research-based evidence will help reject over- or underestimates of the treatment effects reported in literature (Oswald and Plonsky, 2010).

The current extensive reading research was integrated in the two retrospective meta-analyses as part of the nested prospective meta-analysis suggested by Seidler et al. (2019). Given that new

TABLE 8 | The average treatment effects on the treated individuals (ATTs) of the extensive reading for different matching methods.

Matching methods	Estimate	*SE*	*t*	*p*	*d*
Nearest neighborhood matching without replacement	2.85	0.75	3.83	0.000	0.35
Nearest neighborhood matching with replacement	2.69	1.14	2.36	0.020	0.40
Full matching	3.64	0.82	4.47	0.000	0.32

TABLE 9 | Results of the Rosenbaum's sensitivity analysis for the Wilcoxon's signed rank test.

Gamma	Nearest neighborhood matching without replacement		Nearest neighborhood matching with replacement	
	Lower bound	Higher bound	Lower bound	Higher bound
1.0	0.0008	0.0008	0.0009	0.0009
1.1	0.0003	0.0019	0.0003	0.0023
1.2	0.0001	0.0040	0.0001	0.0048
1.3	0.0000	0.0076	0.0000	0.0089
1.4	0.0000	0.0130	0.0000	0.0150
1.5	0.0000	0.0206	0.0000	0.0236
1.6	0.0000	0.0307	0.0000	0.0349
1.7	0.0000	0.0434	0.0000	0.0491
1.8	0.0000	0.0589	0.0000	0.0661
1.9	0.0000	0.0770	0.0000	0.0860
2.0	0.0000	0.0977	0.0000	0.1085

Gamma values refer to odds ratios of differential assignment to treatment due to unidentified covariates. Lower and higher bounds mean the intervals of p-values based on the Wilcoxon's signed rank statistics for the outcome difference between treatment and control groups (Rosenbaum, 2002).

studies meeting the inclusion criteria are included in prospective meta-analyses until generalizability of findings is achieved (Pogue and Yusuf, 1998), prospective study registration is necessary to complete the current prospective meta-analysis. This approach can be useful in L2 teaching research because Marsden et al. (2018b) suggested participation in the open science movement by introducing registered reports of primary research in this field. L2 teaching researchers should therefore be encouraged to submit the full method and analysis protocol of their studies prior to data collection. Moreover, prospective meta-analyses encourage the inclusion of studies by providing information regarding the defined research question and eligibility criteria (Seidler et al., 2019). For example, the prospective meta-analysis proposed in this study requires more ongoing studies that use propensity score methods for treatment effect estimation of extensive reading. L2 teaching researchers can now plan their primary studies for prospective integration into the meta-analysis.

The present study had a limited quantitative focus on evidence-based practice. Moderator analysis will improve language teaching expertise because it provides information about what teaching methods work for whom. For example, the present results showed that the treatment effects of extensive reading changed according to participants' proficiency, focused skills, length of instruction, and implementation format (see also Nakanishi, 2015; Jeon and Day, 2016). However, a qualitative approach to decision-making on treatment effects is also necessary because sociocultural aspects, such as understanding the influence of individual and cultural differences on treatment (APA Presidential Task Force on Evidence-Based Practice, 2006), are often examined in qualitative studies, and these aspects should be examined as well in relation to extensive reading. Future studies should use a mixed-methods approach when examining the treatment effect of evidence-based practice in L2 pedagogy in conjunction with teacher cognition involved in pedagogical decision-making.

A statistical point that should be discussed is covariate selection. The pretreatment variables used as covariates in this study were mainly related to cognitive aspects in extensive reading. However, Yamashita (2004; 2007; 2015) highlighted the role of psychological aspects in L2 reading, such as reading attitude, motivation, and anxiety, affecting both participation in an extensive reading program and outcome measures. Hamada and Takaki (2019) also discussed the covariate effects of L2 reading anxiety on L2 reading proficiency. As the sensitivity analysis results implied that the assumption of ignorable treatment assignment (e.g., Rosenbaum and Rubin, 1983; Imai et al., 2008) was not fully applied in the current study, there is a need for further research that assesses all of the background variables relevant for treatment assignment. When selecting covariates in propensity score analysis, King and Nielsen (2019) recommend including (a) important covariates to cause selection bias, (b) information about how much imbalance caused by the covariates is left, and (c) a sample size still large enough after matching. Although the imbalance observed in the present study was adjusted by the propensity scores, the sample size for the treatment effect estimation consequently became smaller following the nearest neighborhood without and with replacement matchings (see **Table 6**). The thorough application of propensity score analysis is beyond the scope of this study; however, it will be necessary to replicate the results using the same research design.

In terms of implications for evidence-based practice in extensive reading in L2, the most essential contribution of this study is its attempt to advance the assessment theory of treatment effects for the integration of the best available research evidence into extensive reading activities in an intact class. Whereas Mitchell (2000) and Pachler (2003) critically discussed some difficulties in incorporating evidence-based practice in L2 teaching with educational policymaking, they suggested the applicability of research findings to classroom-based practice (see also Sato and Loewen, 2019). Among the many concerns regarding the implementation of evidence-based practice (see Pachler, 2003), it is important to synthesize and summarize existing research evidence (Chambless and Ollendick, 2001),

assess the levels of evidence quality (Oxford Centre for Evidence-Based Medicine, 2009), and acquire the best available research evidence as expertise (APA Presidential Task Force on Evidence-Based Practice, 2006).

Plonsky and Oswald (2014) recommended reviewing L2 teaching research to consider using meta-analysis as a procedure for pedagogical decision-making. In the case of extensive reading, Nakanishi (2015) and Jeon and Day (2016) provided the list of aggregated primary research coded by a well-structured scheme. In the same way, various L2 teaching researchers have published a bibliography with coding information, ranging from specific L2 instruction to educational programs. This research trend helps when moving from retrospective to prospective meta-analyses. In working toward evidence-based practice in L2 pedagogy, it is necessary to accumulate better quality research evidence by including planned, well-designed, and registered research in meta-analyses. While aggregated evidence in L2 teaching has so far been assessed by systematic review through retrospective meta-analysis, prospective meta-analyses require registered reports adhering to previously defined eligibility criteria. The best available research evidence obtained from prospective meta-analyses can be applied to pedagogical decision-making in individual classrooms. To this end, treatment effect assessment will strongly contribute to advancing L2 teaching research toward evidence-based practice.

CONCLUSION

This study focused on how to embed research evidence into classroom-based L2 teaching within the framework of evidence-based practice. The results showed that the effect sizes synthesized by moderator analysis could predict the treatment effects of L2 teaching for individual classrooms. The importance of research-based practice has been emphasized in foreign language education (Mitchell, 2000; Pachler, 2003; Sato and Loewen, 2019). To move toward evidence-based practice in L2 pedagogy, it is necessary to establish a virtuous cycle to (a) assess the levels of scientific evidence obtained from individual research, (b) acquire L2 teaching expertise from best available research evidence, and (c) apply it to other classrooms to provide further research evidence. This study suggests that planned and ongoing L2 teaching studies applying propensity score methods should be registered for inclusion into prospective meta-analyses. This methodological approach to treatment effect assessment helps reduce research bias and waste while also improving pedagogical decision-making based on efficient, adaptive, and collaborative use of educational data. The present findings provide strong support for this approach by demonstrating that the treatment effects of L2 teaching are reproducible when planning teaching procedures based on research evidence.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available in the IRIS digital repository (https://www.iris-database.org/iris/app/home/detail?id=york%3a937791&ref=search).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences of the University of Tsukuba. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This study was supported by the Grants-in-Aid for Young Scientists (B) No. 17K13512 from the Japan Society for the Promotion of Science.

ACKNOWLEDGMENTS

The author wishes to acknowledge YI, JN, and AM for their valuable comments to improve an earlier version of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00617/full#supplementary-material

REFERENCES

Al-Homoud, F., and Schmitt, N. (2009). Extensive reading in a challenging environment: A comparison of extensive reading and intensive reading approach in Saudi Arabia. *Lang. Teach. Res.* 13, 383–401. doi: 10.1177/1362168809341508

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *Am. Psychol.* 61, 271–285. doi: 10.1037/0003-066X.61.4.271

Beglar, D., Hunt, A., and Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Lang. Learn.* 62, 665–703. doi: 10.1111/j.1467-9922.2011.00651.x

Chambless, D. L., and Ollendick, T. H. (2001). Empirically supported psychological interventions: controversies and evidence. *Annu. Rev. Psychol.* 52, 685–716. doi: 10.1146/annurev.psych.52.1.685

Day, R. R. (2015). Extending extensive reading. *Read. Foreign Lang.* 27, 294–301.

Educational Testing Service (2007). *TOEIC Bridge Official Guide & Question Collection*. Tokyo: Institute for International Business Communication.

Educational Testing Service (2019). *Mapping the TOEIC Bridge Test on the Common European Framework of Reference for Languages*. Austin, TX: Educational Testing Service.

Guo, S., and Fraser, M. W. (2015). *Propensity Score Analysis: Statistical Methods and Applications*, 2nd Edn. London: SAGE Publications Ltd.

Hamada, A., and Takaki, S. (2019). Approximate replication of Matsuda and Gobel (2004) for psychometric validation of the foreign language reading anxiety scale. *Lang. Teach* 1–17. doi: 10.1017/S0261444819000296

Ho, D., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* 42, 1–28. doi: 10.18637/jss.v042.i08

Huffman, J. (2014). Reading rate gains during a one-semester extensive reading course. *Read. Foreign Lang.* 26, 17–33.

Imai, K., King, G., and Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Statist. Soc. A.* 171, 481–502. doi: 10.1111/j.1467-985X.2007.00527.x

Jeon, E.-H., and Yamashita, J. (2014). L2 reading comprehension and its correlates: a meta-analysis. *Lang. Learn.* 64, 160–212. doi: 10.1111/lang.12034

Jeon, E.-Y., and Day, R. R. (2016). The effectiveness of ER on reading proficiency: a meta-analysis. *Read. Foreign Lang.* 28, 246–265.

Keele, L. J. (2014). *Rbounds: An R pacKage for Sensitivity Analysis with Matched Data [R package]*. Available online at: https://cran.r-project.org/web/packages/rbounds/rbounds.pdf (accessed February 12, 2019).

King, G., and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Politic. Anal.* 27, 435–454. doi: 10.1017/pan.2019.11

Koizumi, R., and Mochizuki, M. (2011). Development and validation of the PC version of the mochizuki vocabulary size test. *JACET J.* 53, 35–55.

Lee, S.-Y. (2007). Revelations from three consecutive studies on extensive reading. *RELC J.* 38, 150–170. doi: 10.1177/0033688207079730

Leite, W. (2017). *Practical Propensity Score Methods Using R*. London: SAGE Publications Ltd.

Lipsey, M. W., and Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *Am. Psychol.* 48, 1181–1209. doi: 10.1037/0003-066X.48.12.1181

Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev. Sci.* 14, 570–580. doi: 10.1007/s11121-012-0339-5

Mackey, A., and Gass, S. (2015). *Second Language Research: Methodology and Design*. New York, NY: Routledge.

Marsden, E., Morgan-Short, K., Thompson, S., and Abugaber, D. (2018a). Replication in second language research: narrative and systematic reviews and recommendations for the field. *Lang. Learn.* 68, 321–391. doi: 10.1111/lang.12286

Marsden, E., Morgan-Short, K., Trofimovich, P., and Ellis, N. C. (2018b). Introducing registered reports at language learning: promoting transparency, replication, and a synthetic ethic in the language sciences. *Lang. Learn.* 68, 309–320. doi: 10.1111/lang.12284

Mason, B., and Krashen, S. (1997). Extensive reading in English as a foreign language. *Syst.* 25, 91–102. doi: 10.1016/S0346-251X(96)00063-2

McLean, S., and Rouault, G. (2017). The effectiveness and efficiency of extensive reading at developing reading rates. *Syst.* 70, 92–106. doi: 10.1016/j.system.2017.09.003

Mitchell, R. (2000). Applied linguistics and evidence-based classroom practice: The case of foreign language grammar pedagogy. *Appl. Linguist.* 21, 281–303. doi: 10.1093/applin/21.3.281

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Q.* 49, 6–37. doi: 10.1002/tesq.157

Nakanishi, T., and Ueda, A. (2011). Extensive reading and the effect of shadowing. *Read. Foreign Lang.* 23, 1–16. doi: 10.4324/9780367809256-1

Nation, I. S. P., and Waring, R. (2019). *Teaching Extensive Reading in Another Language*. New York, NY: Routledge.

Norris, J. M., and Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Lang. Learn.* 50, 417–528. doi: 10.1111/0023-8333.00136

Nuttall, C. (2005). *Teaching Reading Skills in a Foreign Language*. London: Macmillan Education.

Oswald, F. L., and Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annu. Rev. Appl. Linguist.* 30, 85–110. doi: 10.1017/S0267190510000115

Oxford Centre for Evidence-Based Medicine (2009). Levels of evidence. *Br. J. Urol. Int.* 103:566. doi: 10.1111/j.1464-410X.2009.08408.x

Pachler, N. (2003). Foreign language teaching as an evidence-based profession? *Lang. Learn. J.* 27, 4–14. doi: 10.1080/09571730385200031

Plonsky, L., and Gass, S. (2011). Quantitative research methods, study quality, and outcomes: the case of interaction research. *Lang. Learn.* 61, 325–366. doi: 10.1111/j.1467-9922.2011.00640.x

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

Pogue, J., and Yusuf, S. (1998). Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 351, 47–52. doi: 10.1016/S0140-6736(97)08461-4

Reichardt, C. S. (2009). "Quasi-experimental design," in *The SAGE Handbook of Quantitative Methods in Psychology*, eds R. E. Millsap and A. Maydeu-Olivares (London: SAGE Publications Ltd), 46–71.

Robb, T. N., and Kano, M. (2013). Effective extensive reading outside the classroom: A large-scale experiment. *Read. Foreign Lang.* 25, 234–247.

Rosenbaum, P. R. (2002). *Observational Studies*, 2nd Edn. New York, NY: Springer.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika.* 70, 41–55. doi: 10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701. doi: 10.1037/h0037350

Sato, M., and Loewen, S. (2019). *Evidence-Based Second Language Pedagogy: A Collection of Instructed Second Language Acquisition Studies*. New York, NY: Routledge.

Schafer, J. L., and Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychol. Methods.* 13, 279–313. doi: 10.1037/a0014268

Seidler, A. L., Hunter, K. E., Cheyne, S., Ghersi, D., Berlin, J. A., and Askie, L. (2019). A guide to prospective meta-analysis. *BMJ.* 367, 1–11. doi: 10.1136/bmj.l5342

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *J. Stat. Softw.* 42, 1–52. doi: 10.18637/jss.v042.i07

Shih, Y.-C. (2015). The impact of extensive reading on college business majors in Taiwan. *Read. Matrix.* 15, 220–233.

Sims, J. M. (1996). *A Comparative Study of Improvements in Reading Comprehension of Skill-Based Instruction and Extensive Reading for Pleasure with Taiwanese Freshmen University Students*. Dissertation, Florida State University, Tallahassee, FI.

Suk, N. (2017). The effects of extensive reading on reading comprehension, reading rate, and vocabulary acquisition. *Read. Res. Q.* 52, 73–89. doi: 10.1002/rrq.152

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.18637/jss.v036.i03

Waring, R., and McLean, S. (2015). Exploration of the core and variable dimensions of extensive reading research and pedagogy. *Read. Foreign Lang.* 27, 160–167.

Watt, C. A., and Kennedy, J. E. (2017). Options for prospective meta-analysis and introduction of registration-based prospective meta-analysis. *Front. Psychol.* 7:2030. doi: 10.3389/fpsyg.2016.02030

Yamashita, J. (2004). Reading attitudes in L1 and L2, and their influence on L2 extensive reading. *Read. Foreign Lang.* 16, 1–19.

Yamashita, J. (2007). The relationship of reading attitudes between L1 and L2: An investigation of adult EFL learners in Japan. *TESOL Q.* 41, 81–105. doi: 10.1002/j.1545-7249.2007.tb00041.x

Yamashita, J. (2015). In search of the nature of extensive reading in L2: Cognitive, affective, and pedagogical perspectives. *Read. Foreign. Lang.* 27, 168–181.

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 April 2020
doi: 10.3389/fpsyg.2020.00481

Developing Interpreting Competence Scales in China

Weiwei Wang[1], Yi Xu[1]*, Binhua Wang[2] and Lei Mu[1]*

[1] School of Interpreting and Translation Studies, Guangdong University of Foreign Studies, Guangzhou, China, [2] Centre for Translation Studies, University of Leeds, Leeds, United Kingdom

Tertiary-level interpreter training and education have developed rapidly in China, and over 200 undergraduate and over 200 postgraduate T&I programs have been launched over the past decade. Despite the rapid development, there has been no standardized framework allowing for the reliable and valid measurement of interpreting competence in China. Against this background, the China Standards of English (CSE), which are the Chinese counterpart to the Common European Framework of Reference (CEFR), were unveiled in 2018 after 4 years of government-funded research and validation. One vital component of the CSE is the descriptor-referenced interpreting competence scales. This article provides a systematic account of the design, development, and validation of the interpreting competence scales in China. Within the CSE, the construct of interpreting competence was defined according to an interactionist approach. It not only encompasses cognitive abilities, interpreting strategies, and subject-matter knowledge but also considers performance in typical communicative settings. Based on the construct definition, a corpus of relevant descriptors was built from three main sources, including: (a) interpreting training syllabuses, curricular frameworks, rating scales, and professional codes of conduct; (b) previous literature on interpreting performance assessment, competence development, and interpreter training and education; and (c) exemplar-generation data on assessing interpreting competence and typical interpreting activities, which were collected from interpreting professionals, trainers, and trainees. The corpus contains 9,208 descriptors of interpreting competence. A mixed-method survey was then conducted to analyze, scale, and validate the descriptors among 30,682 students, 5,787 teachers, and 139 interpreting professionals from 28 provinces, municipalities, and regions in China. The finalized set included 369 descriptors that reference interpreting competence. The CSE—Interpreting Competence Scales with theoretically and empirically based descriptors represent a major effort in research on interpreting competence and its assessment, and they have significant potential to be applied widely in interpreting training, research, and assessment.

Keywords: interpreting competence, assessment and scales, descriptors, China standards of english, scale development and validation

OPEN ACCESS

Edited by:
Vahid Aryadoust,
Nanyang Technological University, Singapore

Reviewed by:
Chao Han,
Southwest University, China
Thomas Eckes,
Ruhr University Bochum, Germany

****Correspondence:***
Weiwei Wang
wangweiwei96@hotmail.com
Yi Xu
xuyi@gdufs.edu.cn

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

Received: *17 November 2019*
Accepted: *02 March 2020*
Published: *23 April 2020*

Citation:
Wang W, Xu Y, Wang B and Mu L (2020) Developing Interpreting Competence Scales in China. Front. Psychol. 11:481. doi: 10.3389/fpsyg.2020.00481

INTRODUCTION

In the mid-twentieth century, universities began to offer programs designed to train conference interpreters (Pöchhacker, 2015), and the first group of programs was offered in Moscow (1930), Heidelberg (1933), Geneva (1941), and Vienna (1943). Since then, more universities have developed interpreter education programs. As of 2016, the International Association of Conference

Interpreters (AIIC) had listed 95 programs in its Interpreting Schools and Programs Directory. By 2019, the European Masters in Conference Interpreting (EMCI) had endorsed 16 programs through their member universities. Over the past decade, China has witnessed a rapid growth in translator and interpreter education. By March 2019, over 282 Chinese universities had Bachelor's degree programs in Translation and Interpreting (BTI), and 249 Chinese universities had Master's degree programs in Translation and Interpreting (MTI).

Chinese interpreting training programs differ from their European counterparts in a number of ways. First, Translation and Interpreting (T&I) programs in China offer both undergraduate and postgraduate training, while most European T&I programs offer postgraduate training. Second, Chinese T&I students need to work bi-directionally, including retour into their B languages (i.e. second language); this has long been a professional norm in the Chinese interpreting market. Western interpreters, meanwhile, often interpret into their A, or first, language (Seleskovitch and Lederer, 1989). Third, although students enrolled in T&I programs are expected to have a high level of general and cultural knowledge and adequate B language proficiency, i.e. Interagency Language Roundtable (ILR) Band 3 or 4 or Common European Framework of Reference for Languages (CEFR) C1 or C2 (Setton and Dawrant, 2016), experience in China shows that most students still need B language enhancement. The reason behind this deficit is that most students in China learn English through formal classroom teaching and are often deficient in conversational listening and speaking due to the limited opportunities for immersive English language learning with native speakers. In comparison, European T&I students tend to have higher B and C language (third language) proficiency (Szabari, 2002).

With over 200 undergraduate and over 200 postgraduate T&I degree programs being launched over the past decade in China, the lack of consistent teaching practices and common competence standards has become an urgent issue in the training and assessment of interpreters. To address the issue, interpreting educators and researchers have worked collaboratively on a national-level project led by the Ministry of Education: the China Standards of English (CSE). The purpose of CSE was to develop a national framework of interpreting competence that could support T&I students' professional development, and scales of interpreting competence to provide guidance for interpreting training and assessment in China.

Previous research on interpreting competence has focused on different areas, "most notably cognitive processes, education (including curriculum design, aptitude testing, and pedagogy) and certification programs" (Pöchhacker, 2015, 69). Subsequently, very few models or frameworks for interpreting competence have been put forward. Most of the available models examined the composition of interpreting competence. For instance, Pöchhacker (2000) proposed a multidimensional model of interpreting competence that highlighted language and cultural skills, translational skills, and subject-matter knowledge. In this model, linguistic transfer competence was regarded as a core element, complemented by cultural competence and interaction management skills. These elements were all supported by professional performance skills and ethical behavior (Pöchhacker, 2015). Albl-Mikasa (2013) referred to interpreting models suggested by Kalina (2002) and Kutz (2010) when proposing a detailed model that comprises five skill sets, each with a set of sub-skills: pre-process (language proficiency, terminology management, and preparation); in-process (comprehension, transfer, and production); peri-process (teamwork and ability to handle stress); post-process (terminology work and quality control); and para-process (business acumen, customer relations, and meta-reflection). Han (2015) applied an interactionist approach to construct the components of interpreting ability, including knowledge of languages, interpreting strategies, topical knowledge, and metacognitive process. Dong (2018) researched the development of students' interpreting competence through longitudinal empirical data and proposed a complex dynamic system to illustrate how self-organization among different key parameters results in interpreting competence.

When defining interpreting competence, bilingual linguistic competence and professionalism are frequently mentioned by scholars. For instance, Kalina (2000) defined interpreting competence from a psycholinguistic perspective, calling it the ability to process texts in a bilingual or multilingual communication environment. Zhong (2003) proposed that interpreting competence should include linguistic knowledge, encyclopedic knowledge, and skills related to both professional interpreting and artistic presentation. Wang (2007, 2012) defined interpreting competence as the underlying system of knowledge and skills required to accomplish the task of interpreting, including the necessary professional and physio-psychological qualities. Setton and Dawrant (2016) stated that interpreting competence is composed of four core elements: bilingual language proficiency, knowledge, skills, and professionalism.

The models or definitions of interpreting competence mentioned above indicate that researchers agree that interpreting competence goes beyond simple bilingual competence and includes skills of cross-cultural communication. They also demonstrate that, although there is no universally accepted model of interpreting competence, the previous discussions illustrate the composition of interpreting competence. It is also clear that little attention has been paid to the developmental stages of interpreting competence and that the different competence requirements for specific interpreting tasks have been overlooked.

When assessing interpreting competence, research in interpreting quality is highly relevant. The literature on the concept of interpreting quality, assessment, and evaluation is extensive (e.g. Barik, 1975; Berk-Seligson, 1988; Kurz, 1993; Moser-Mercer, 1996; Aís, 1998; Campbell and Hale, 2003; Napier, 2004a,b; Kalina, 2005a,b; Liu, 2008; Gile, 2011). For instance, Liu (2008) studied the differences in interpreting competence by comparing the performance of expert and novice interpreters. Research on certification also provides effective instruments and outlines potential problems for assessing interpreting competence (Pym et al., 2013). In terms of quality parameters, many scholars have proposed criteria including completeness, accuracy, intonation, voice projection, language use, and logical

cohesion (Kurz, 1993; Moser-Mercer, 1996; Garzone, 2003). Pöchhacker (2001) suggested four common criteria to cover the quality range from product to service: accurate rendition, adequate target language expression, equivalent intended effect, and successful communicative interaction. As Pöchhacker (2015, p.334) put it, "on a superficial level, quality relates to something that is good or useful, or to behavior that is sanctioned or expected." However, it is difficult to measure interpreting quality quantitatively given its complexity. Grbić (2008), in her conceptual study of interpreting quality, proposed that interpreting quality should be assessed based on actual settings. The past decades witnessed a distinct strand of research on Interpreting Quality Assessment (IQA) especially in the educational context (Yeh and Liu, 2006; Lee, 2008; Postigo Pinazo, 2008; Tiselius, 2009; Liu, 2013; Lee, 2015; Wang et al., 2015; Han, 2016, 2017, 2018, 2019; Han and Riazi, 2018; Lee, 2018, 2019). Among them, the design and application of rating scales that assess interpreting quality have been a priority for many interpreting researchers and trainers. Based on the prior literature, interpreting quality constructs and parameters have been operationalized into various rubric-referenced rating scales. For instance, Liu (2013) discussed the development of the rating scheme for Taiwan's interpretation certification exam. Lee (2015) has provided a detailed report on the process of developing an analytic rating scale for assessing undergraduate students' consecutive interpreting performances. Han (2017, 2018) probed into the application and validity of rating scales for students' English–Chinese consecutive interpreting performance. Despite the popularity of assessment rubrics and rating scales, competence-based scales that could describe the progressive development of interpreters at different levels and diagnose skills gaps have not been developed.

While research in IQA has been a prominent topic in interpreting studies, the construct and measurement of the progressive stages of interpreting competence have received limited attention. The ILR, as the earliest language proficiency scale developed by the United States government in 1955, is the only assessment scale that includes interpreting. ILR describes interpreting performance in three bands: Professional Performance (Levels 3 to 5), Limited Performance (Levels 2 and 2+), and Minimal Performance (Levels 1 and 1+). In ILR, only individuals performing at the Professional Performance levels are properly termed "interpreters" (Interagency Language Roundtable, 2002). Since then, several language proficiency scales have been proposed in Europe, Canada, Australia, and other countries and regions. These include the American Council on the Teaching of Foreign Languages (2019) Proficiency Guidelines, the International Second Language Proficiency Ratings (ISLPR), and the Canadian Language Benchmarks (CLB) (Centre for Canadian Language Benchmarks, 2019). The CEFR scale, jointly developed by more than 40 members of the Council of Europe, is widely used in countries around the world (Council of Europe, 2001). However, none of these frameworks seem to focus on interpreting or have accounted for students' progression from novice interpreter to expert interpreter.

Furthermore, understanding and describing the development of interpreting competence is even more pertinent due to the current challenges in education quality faced by the T&I degree programs in China. To this end, the Chinese Ministry of Education initiated the CSE Project to develop a national framework and a set of standards for Chinese–English interpreting students. This national project was supervised by the National Education Examinations Authority (National Education Examinations Authority, 2014).

In general, the CSE–Interpreting Competence Scales were developed with two broad aims: first, they were to act as a stimulus for reflection on current practice in the country; second, they were to provide a common reference for developing teaching syllabuses, curriculum guidelines, examinations, and textbooks for interpreting across China. The CSE–Interpreting Competence Scales were designed to contribute to educational reform and innovation in order to improve the efficiency of the teaching, learning, and assessment of interpreting.

This paper reports on the development process of the CSE–Interpreting Competence Scales, which involved research work in two major parts, divided into five stages, as follows (**Figure 1**):

> Part I: Drafting the scales: the creation of a descriptor pool.
> Stage 1: Defining interpreting competence with respect to the Chinese–English interpreting training context of China
> Stage 2: Developing an interpreting competence descriptive scheme
> Stage 3: Collecting descriptors with reasonable representativeness
> Part II: Validating the scales: scaling and refinement
> Stage 1: Quantitative validation: main data collection and scaling descriptors through teacher and student assessments
> Stage 2: Qualitative validation: consultation with teachers through focus group interviews and workshops

The remainder of this article is divided into four sections. The following two sections introduce the two stages that developed the CSE–Interpreting Competence Scales. The fourth section discusses the limitations and issues encountered in the design of the scales. Finally, the fifth section concludes with the possible application of the undertaken work.

DRAFTING THE SCALE: CREATION OF A DESCRIPTOR POOL

The CSE–Interpreting Competence Scales cover four aspects of descriptors: overall interpreting performance, typical interpreting activities, interpreting strategies, and self-assessment scales. In this section, we will illustrate the process of drafting the interpreting competence scales. This development process is divided into three stages: defining interpreting competence, developing the descriptive scheme, and collecting descriptors.

Defining Interpreting Competence

To fully account for the perceptions of different stakeholders (interpreting learners, trainers, testers, users, employers,

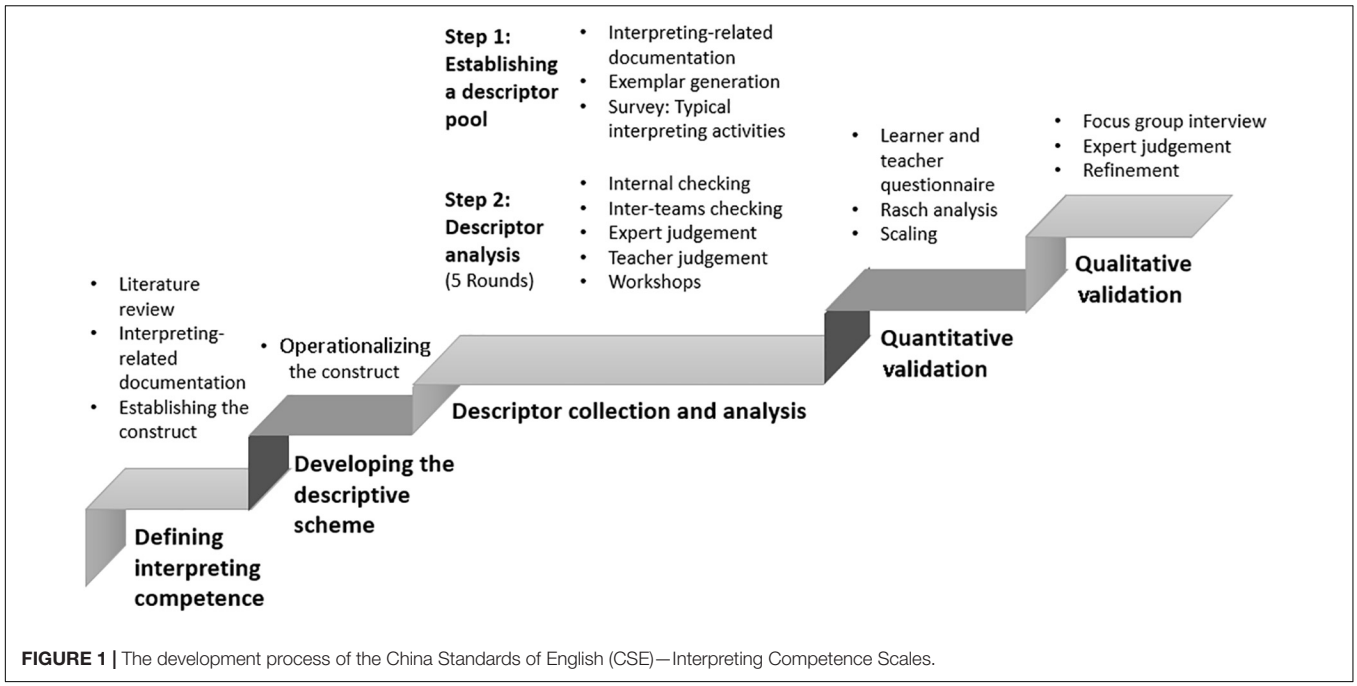


FIGURE 1 | The development process of the China Standards of English (CSE)—Interpreting Competence Scales.

policymakers, etc.), we constructed interpreting competence based on previous literature in interpreting studies and on Bachman's (1990) communicative language competence model. Interpreting competence is demonstrated as decisions made by the interpreter to purposefully perform an interpreting task in a given place at a given time. The competence involves cognitive processing, language proficiency, extra-linguistic knowledge, and interpreting strategies.

Drawing on existing literature, we define interpreting competence as the interlingual and intercultural mediation ability of instantaneously transferring utterances from a source language into a target language, using language proficiency, related world knowledge, and interpreting-specific strategies.

According to this definition, interpreting competence is, first and foremost, a comprehensive cognitive ability used in interpreting activities. It involves mechanisms and procedures of information processing (Bachman and Palmer, 1996). These activities include identifying the logic of the source text, retrieving memory, summarizing, and analyzing the structure of the source text. As Angelelli and Degueldre (2002) and many other scholars in interpreting studies have observed, bilingual proficiency is the prerequisite for interpreting. In this integrated process, the activity's basis is bilingual competence in Chinese and English. The interpreter's topic-specific and/or world knowledge plays a key role in the process of comprehension (Will, 2007; Díaz-Galaz, 2011; Díaz-Galaz et al., 2015; Fantinuoli, 2017). At the same time, interpreting strategies are used in both comprehension and reproduction (Kohn and Kalina, 1996; Bartłomiejczyk, 2006; Li, 2013, 2015; Arumí Ribas and Vargas-Urpi, 2017; Wu and Liao, 2018; Dong et al., 2019). The definition also includes professionalism, meaning that interpreters must abide by the code of conduct of the industry. They must be mentally ready to work under stress and make operational and ethical decisions aimed at optimizing interpretation in real life (Setton and Dawrant, 2016). The construct of interpreting competence is illustrated in **Figure 2**. As the cognitive process is invisible, our descriptions based on this construct focus on interpreting activities (i.e. interpreting modes, topics, and context) and products (performance).

Development of the Descriptive Scheme

The scheme of the description serves to create a link between real-life tasks and the construct of interpreting competence. The "can-do" principle of CEFR (North, 2014) suggests that descriptors of interpreting competence scales typically consist of three elements:

(1) Performance: the interpreting task (e.g. "*interpreting a speech consecutively*")
(2) Criteria: the intrinsic characteristics of the performance, involving a range of cognitive efforts or interpreting skills (e.g. "*actively anticipating speech information, with note-taking*")
(3) Conditions: any extrinsic constraint or condition defining the performance (e.g. "*moderate speech rate, high information density, and with no accent*")

The can-do principle describes the expected type of interpreting competence descriptors. The scheme of description determines the interpreting competence scale structure and reflects the interpreting competence construct defined above.

Figure 3 illustrates an operational descriptive scheme that covers overall interpreting performance and cognitive ability, interpreting strategies, knowledge, and professionalism; this is a practical application of the theoretical presentation in **Figure 2**. Interpreting-related cognitive ability, interpreting strategies, and

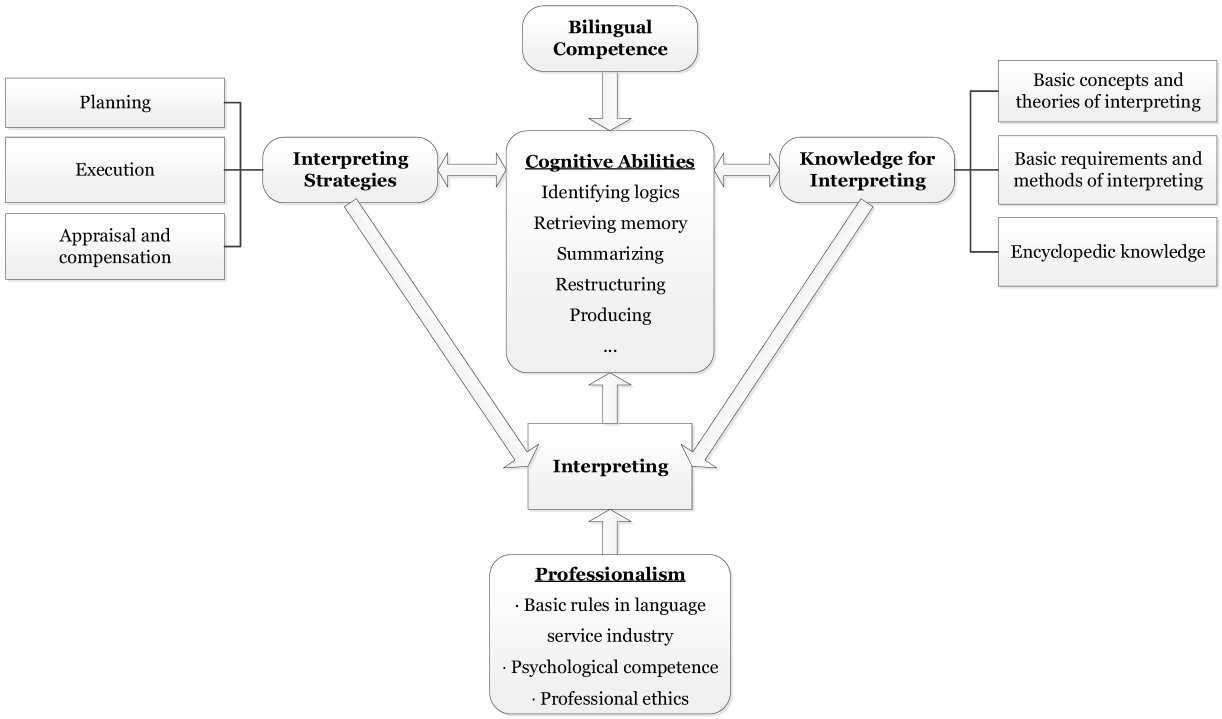


FIGURE 2 | Construct of interpreting competence.

subject-matter knowledge are identified in the construct as the three core elements of interpreting competence. It must be noted that bilingual competence is not included in the scheme, because the listening, speaking, writing, and reading scales of the CSE have already covered descriptions of English language proficiency.

Cognitive abilities are the key components of the interpreting competence construct and are described via interpreting tasks in both the scales of overall interpreting performance and sub-sets of typical interpreting activities. One of the cognitive abilities, for instance, is described as "*Can understand the content of an interview while analyzing the logical relationships in the source-language information during SI for a media interview.*"

As evident in **Figure 3**, interpreting strategies are described in different subscales, i.e. planning, execution, appraisal, and compensation. The strategy scales refer mainly to the skills, methods, and actions that aim to fulfill interpreting tasks and solve problems. For instance, "*Can use contextual information to anticipate upcoming content and information actively.*"

Typical interpreting activities include business negotiations, training, lectures, interviews, and conferences. For instance, "*Can interpret important information, such as research objectives, methodology, and conclusions, during SI for an academic talk*" and "*Can follow target-language norms to reflect source-language register and style during Consecutive Interpreting (CI) with note-taking for a foreign affairs meeting.*" The knowledge scales include descriptors about encyclopedic knowledge, basic methods, theories of interpreting, and so on.

Descriptor Collection and Analysis

Drawing on the theoretical concepts of interpreting competence and the operational descriptive scheme, we established a pool of interpreting descriptors based on documentation, exemplar generation, and surveys.

Step 1: Establishing a Descriptor Pool

Documentation: the analysis and editing of existing descriptions in interpreting training and research as well as the interpreting profession

Through documentation, we collected a wide range of materials such as existing scales related to interpreting (e.g. ILR); teaching syllabuses; curricular frameworks; textbooks; rating scales from the established interpreting programs such as the University of Leeds, Middlebury Institute of International Studies at Monterey, Shanghai International Studies University, and Guangdong University of Foreign Studies; test specifications and codes of conduct from established professional associations such as the AIIC and accreditation agencies such as the National Accreditation Authority for Translators and Interpreters (NAATI) in Australia and the China Accreditation Test for Translators and Interpreters (CATTI); and previous research on interpreting performance assessment, competence development, and interpreter training and education.

Due to the large number of interpreting-related documents, more than 110 postgraduate students in T&I majors were recruited as volunteers to help with the highly labor-intensive sorting and editing work. One of the authors of the article led

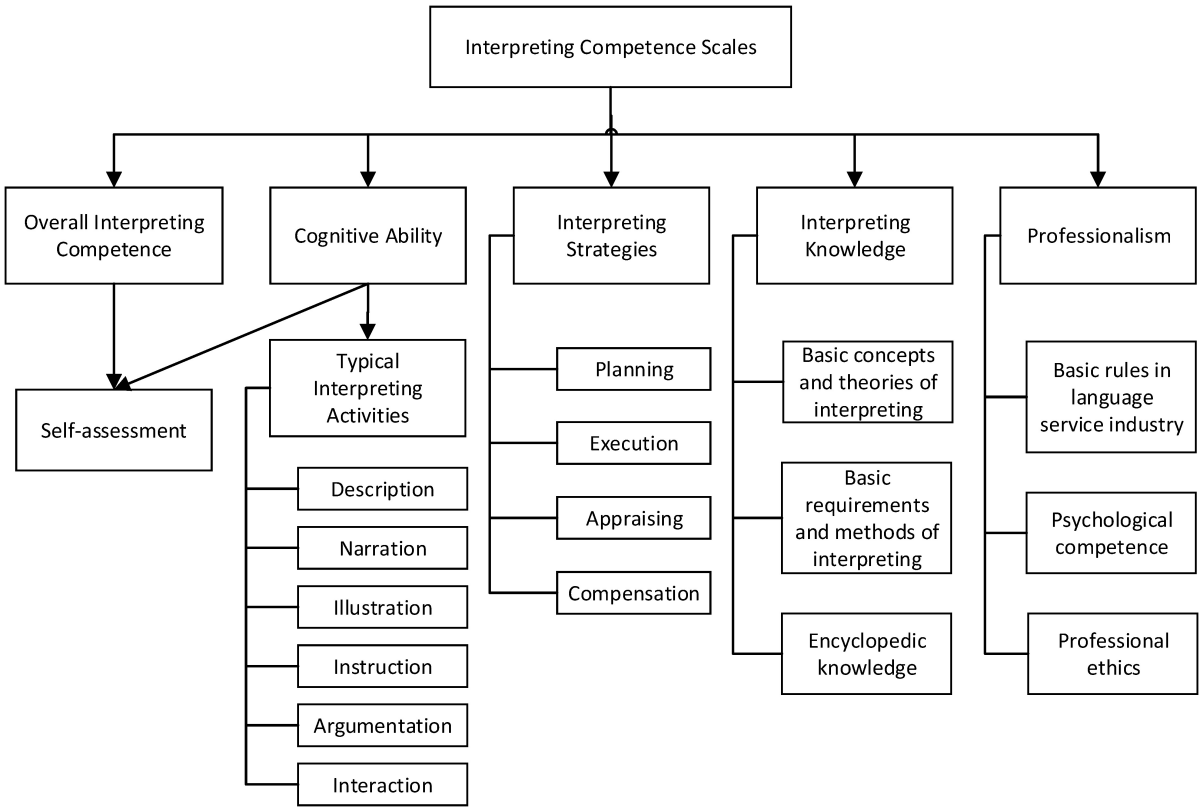


FIGURE 3 | Operational descriptive scheme of the Interpreting Competence Scales.

training with the volunteers and then worked with them to collect 8,937 descriptors from 1,048 documents.

Exemplar generation: writing new descriptors from video performance of selective learners

For the exemplar generation, 22 videos of interpreting performances by students at different levels were recorded; 69 professionals and trainers were invited to write descriptors based on the videos. After rigorous training by the authors on the descriptive scheme and according to the three-element can-do principle, 16 trainers were asked to describe the actual interpreting competence of their students in terms of cognitive abilities, interpreting strategies, and knowledge. In total, 271 descriptors were generated during this process.

Survey: collecting descriptors for typical interpreting tasks

We surveyed 53 professional interpreters and 150 student interpreters with online questionnaires to collect their opinions about typical interpreting tasks. We also surveyed and examined relevant textbooks used for different stages of interpreter training and T&I industry reports, including the annual reports of the language service industry in China by the Translators Association of China (TAC).

In this step, a corpus containing a total of 9,208 potential descriptors of interpreting competence was established. Each of the descriptors was assigned into relevant levels and categories of the descriptive scheme.

Step 2: Analyzing Descriptors

In this step, we carried out several rounds of analysis to remove the redundant and overlapping descriptors, reformulate the ambiguous ones, and rewrite those inconsistent with the style and quality requirement for descriptors. Moreover, a glossary of verbs and nouns frequently used in the three categories (cognitive ability, strategies, and knowledge) was generated from the corpus to ensure terminology consistency.

The first round

Four groups composed of members from the interpreting scales' author team checked the descriptors collected from different sources for redundancy and repetition. For instance, "accurate rendition," "accurate delivery," and "accuracy in interpretation" could be integrated. This round of sorting tasks consisted of an interactive process involving repetition checks in each set of descriptive categories; the redundancy analysis of each descriptor; and a series of workshops within each group to ensure the appropriate assignment of descriptors into relevant levels. The initial width of the interpreting scale was discussed within the author team. Level 6, which is the upper intermediate level of the nine levels of the CSE Project (Liu, 2019), was determined to be the beginner's level of interpreting scales for two reasons. First, bilingual proficiency is the basis for interpreting. Second, interpreting courses start in the third semester of the BTI programs and the fourth or fifth semester for English majors

in China. These student groups fell into the provisional Level 6 of CSE (Liu, 2019). As a result of the refinement, 3,259 descriptors remained.

The second round

Cross-group checking within the interpreting scales' author team was undertaken to review the quality of the descriptors in this round. A similar task to check repetition, redundancy, wording, and classification was conducted. For instance, the descriptor "*Can process the information in the source speech to find the logic structure, such as the main idea, supporting idea and details*" was too long and involved unnecessary examples. It was therefore revised into "*Can identify the logical threads of the source speech.*" At this point, each descriptor was reviewed by at least three team members. A total of 1,081 descriptors survived this round of scrutiny.

The third round

The remaining descriptors were cross-checked further by researchers in the translation and the speaking scales team of the CSE Project. Two external experts in interpreting education and research were also invited to identify ambiguities or inappropriate expressions. In this round, the provisional levels of the descriptors were based on the judgments made by these experts and were based on their in-depth understanding of interpreting theories and practices. The issue of whether or not sight interpreting/translation should be regarded as a distinct interpreting working mode was discussed. Considering that sight interpreting/translation and its variants were mostly used at earlier stages of interpreting training, as a deverbalization exercise, or as preparatory training for simultaneous interpreting (Setton and Dawrant, 2016), we deleted the descriptors on sight interpreting/translation. By the end of this round, we managed to further reduce the number of descriptors to about 700.

The fourth round

Over 5,000 descriptors that had survived the first three rounds of analysis from all teams of the CSE Project (listening, speaking, reading, writing, translation, interpreting, etc.) were used to construct over 100 online questionnaires of 50 to70 items (i.e. descriptors) each. All the researchers in the CSE Project and over 50 external experts took part in this online reviewing process to provide feedback on wording (e.g. explicitness, clarity, and appropriateness), descriptor structure (e.g. performance, criteria, and condition), and provisional levels and classification (e.g. representativeness). As a result, 673 descriptors were retained for the next round of analysis.

The fifth round

Based on the results from the previous round, the interpreting scales team revised the descriptors further. Two workshops with over 20 trainers and professionals were then conducted to elicit more feedback. The participants of the workshops were asked to examine each descriptor, identifying well-written ones and commenting on problematic ones. For instance, the contents of some descriptors were contradictory (e.g. "*Can understand the source speech but cannot monitor the target language quality*"), while others were incomprehensible to lower-level students (e.g. "*Can deverbalize during interpreting*"). These descriptors were then reworded or discarded upon further discussion within the CSE-Interpreting team. Finally, 548 descriptors on interpreting competence remained as the first draft.

The draft scales were then circulated among experts, teachers, and interpreters to verify the appropriateness of the descriptors, their categorization, and levels. The draft was then revised into the first edition.

VALIDATING THE SCALE

The next step was to scale the descriptors through quantitative and qualitative validation. Validation of descriptors plays a key role in the construction and development of language ability scales (Brindley, 1998; North, 2000), especially for a national framework of language competencies with high stakes. Two rounds of validation were carried out in 2 years. First, large-scale surveys were used to collect data from interpreting trainers, students, and potential users of the scales around China. The data collected by questionnaires were analyzed using statistical methods, including Rasch modeling, to scale the descriptors and to test the representativeness of each descriptor and the appropriateness of the descriptor levels. Second, expert judgment and focus group interviews were implemented to corroborate and contextualize the findings from the quantitative analysis.

Quantitative Validation: Rasch Scaling

A nationwide survey was first launched to collect quantitative data from potential users of the interpreting competence scales, including interpreting learners, trainers, and interpreters at various levels. This round of validation involved three steps:

Step 1: Questionnaire Design

The validation of interpreting competence descriptors was conducted within the context of the national project of CSE (Liu, 2019). In total, 5,046 descriptors from the eight CSE teams, including listening, speaking, reading, writing, and interpreting, were used to construct 80 sets of overlapping online/computerized adaptive questionnaires of 50 to70 items. Among them, 42 questionnaires contained the remaining 548 descriptors from the first draft of the CSE-Interpreting team. An excerpt from the sample questionnaire is provided in **Supplementary Appendix 1**.

As one of the aims of the Rasch methodology was to calibrate the descriptors onto a continuum, the questionnaires were linked though "anchor items" common to adjacent questionnaires. When covering a broad range of proficiency levels, this leads to an overlapping chain of questionnaires (targeted at successive levels), linked by the anchor items (North, 1995). Hambleton et al. (1991) recommended that anchor items comprise 20 to 25% of the total items. Following their suggestion, this project selected 20% of the total descriptors as anchor items. Using a similar approach to CEFR (North, 2000), we selected the anchor items from a larger pool of relevant items that were

considered (by external experts and the author team) the most clear and representative.

In the questionnaire, each descriptor was followed by a rating scale from 0 to 4 with the following statements (translated from Chinese):

> "0": Cannot do it at all. (Unable to execute the task in any circumstances. Their proficiency is obviously much lower than this level.)
> "1": Can do it with significant help. (Can execute the task in favorable circumstances. Their proficiency is a bit lower than this level.)
> "2": Can do it. (Can execute the task independently in normal circumstances. Their proficiency is at this level.)
> "3": Can do it well. (Can execute the task even in difficult circumstances. Their proficiency is a bit higher than this level.)
> "4": Can do it easily. (Can execute the task easily in any conditions. Their proficiency is clearly much higher than this level.)

Step 2: Questionnaire Distribution and Data Collection

The participants for the nationwide survey of the CSE Project included 29,167 teachers and 120,710 students from around 300 primary schools, 600 high schools, and nearly 300 universities in 28 provinces, municipalities, and regions in China.

For the 42 questionnaires with descriptors of the interpreting scales, 5,787 teachers from 259 colleges and universities responded by rating the descriptors against their students' actual Chinese–English interpreting performance, while 30,682 students from 215 colleges and universities took part by evaluating their own interpreting competence from June 20 to July 15, 2016. All teacher respondents were English teachers or T&I trainers. Most student respondents were English majors; only 3% of them were T&I majors. As the student population of T&I programs accounts for less than 5% of the large population of English language major students in China, 3% is considered sufficient.

To improve rating quality, the author team provided training to teachers (from the same institution) either in a half-day rating conference or via the CSE online working platform[1]. Several efforts were made to ensure the effectiveness of training. The half-day rating conference began with a brief presentation of the CSE Project and introduction of the rating procedures and was immediately followed by a mock rating session. During this session, teachers on site viewed video clips of three students' oral English performance. Next, sample descriptors were simultaneously read by the conference host and shown on a large screen that could be seen by all participants. The teachers were asked to rate the students by raising number cards (0 to 4) based on the video clips and the descriptor. The host and the volunteers checked the rating results. As an example, during a conference in a university in Chongqing municipality (located in Southwest China), 58 English teachers took part in the mock rating session. For a sample descriptor, 45 teachers chose "3" or "4" (which was considered to be consistent with the student's performance in the video), 11 teachers chose "1" or "2," and two teachers chose "0." The host invited the teachers who chose "1" or "0" to justify their ratings. The teachers who chose "0" said that they misunderstood "0" as "very easy." The rating scales were explained once again to all teachers. The host explained to the teachers why "3" or "4" was closer to the students' proficiency but also reminded them that there was no standard answer to the descriptors and that reaching consensus was not a requirement. The teachers were free to raise questions and discuss the rating results at any time. The teachers did not start answering the online questionnaire until all sample descriptors were evaluated in the mock session. Apart from the training on site, a mock rating session with further explanation was also provided in the CSE online working platform.

Step 3: Data analysis

The participants' questionnaire responses were analyzed using FACETS 3.71.0 (Linacre, 2013). As Bond and Fox (2015) suggested, the Rasch rating scale model (RSM) can establish patterns in the rating scale categories in order to yield a single rating scale structure common to all the items on the scale. In this project, RSM analysis was performed by the CSE statistics team to estimate the relative difficulty of each of the interpreting competence descriptors, as rated by students and by teachers, and to examine the quality of rating scale responses. To determine how well the items measured the underlying traits and to examine the overall rating quality, we adopted a relaxed fit analysis cutoff of between 0.5 and 1.5 (Wright and Linacre, 1994) to determine overfit and misfit to the Rasch model.

Due to the large sample size, the standard error (SE) of the estimated parameters was 0.2 for the teachers' evaluation and 0.08 for the students' self-assessment. The Rasch difficulty parameters of the two data pools ranged from -0.5172 to 3.7848 for the teachers' evaluation and from -0.9231 to 4.0254 for the students' self-evaluation. In addition, 4.38% (n = 24) and 3.28% (n = 18) of the items from the teachers' and students' questionnaires, respectively, displayed both infit and outfit mean square values that were outside the cutoff range (0.5–1.5 logits). These items are considered psychometrically problematic; the misfitting examples are presented in **Table 1**.

A second Rasch analysis was conducted after removal or revision of the misfitted items. This resulted in a second version of the CSE—Interpreting Competence Scales that included 304 items with both infit and outfit mean square values that fell between 0.5 and 1.5 logits.

The next analysis involved scaling. One way to check the acceptability and validity of the scale is to evaluate whether descriptors were calibrated in line with the original intentions of the design (North, 1995). To achieve that goal, appropriate cutoff points for scale levels need to be determined based on the logit scale. Setting pass/fail cutoff points requires precise conceptualization. There are many possible conceptualizations (North, 2000; Kolen and Brennan, 2004; Linacre, 2013). For this study, three factors were considered to locate the "zero" position

[1]http://cse.neea.edu.cn

TABLE 1 | Examples of Misfitting Items From China Standards of English (CSE)-Interpreting Questionnaires.

Level	Descriptor	Difficulty and misfit estimates			
		Difficulty estimate (SE)	Outfit MNSQ (z-std)	Observed average	Expected average
	在口译任务前，能利用网络资源查询发言人的相关资料和背景。				
6	Can search for material and background information pertinent to the speaker using internet resources prior to interpreting. 口译前，能借助互联网搜索相关话题的重要关键词。	−0.63 (0.14)	1.82 (6.3)	2.75	2.82
6	Can use the internet to search for keywords related to the topic of speech prior to interpr eting. 能尊重并支持其他职业译员。	−0.78 (0.15)	1.46 (3.8)	2.84	2.9
6	Can respect and support other interpreters. 能在口译中守时，并在无法守时的情况下，及时告知客户。	−1.16 (0.15)	2.75 (9)	2.93	2.98
6	Can be punctual for interpreting assignment and inform clients in a timely manner if the interpreter will be late. 在无笔记交传中，能使用手头的工具，如手机查询生词或专业词汇。	0.12 (0.13)	2.14 (8.4)	2.53	2.58
6	Can use hands-on tools, such as a mobile phone, to search for unfamiliar words or technical words during consecutive interpreting without notes. 在口译时，能使用翻译辅助工具迅速查找口译中重要的生词。	−1.27 (0.15)	1.81 (6.4)	2.97	3.01
7	Can use computer-aided tools to search for important unfamiliar words during interpreting. 在同声传译中，能在口译任务前检查相关设备，确保接收频道能接收发言人的声音，输出频道能传递自己的声音。	−0.99 (0.19)	1.64 (4.2)	2.95	2.98
8	Can inspect the relevant equipment to ensure the speaker's voice can be received through the input channel, and the interpreter's voice can be sent through the output channel prior to commencing simultaneous interpreting. 在口译任务前，能联系主办方或讲话人获取会议资料等相关信息。	0.31 (0.19)	2.34 (6.5)	2.66	2.75
8	Can contact event organizers or the speaker to collect pertinent information such as conference documents prior to interpreting. 在口译任务前，能借助网络和词典等工具准备专业术语表。	−1.04 (0.19)	1.93 (5.1)	2.94	2.97
8	Can use tools such as the internet and dictionaries to create a glossary prior to interpreting. 在同声传译中，能在听译的同时，利用网络资源查找相关术语。	−1.33 (0.19)	1.84 (4.7)	3.02	3.05
9	Can search for pertinent terminology using internet resources while listening and interpreting during simultaneous interpreting.	0.59 (0.2)	2.19 (6.2)	2.53	2.57

in the scale: logit values were used in an attempt to create a scale with more or less equal intervals, patterns with natural gaps on the vertical scale, and a comparison of current patterns with levels in real life.

As illustrated in **Table 2**, there are nine levels in CSE. Level 5 is the center point, and each level covers approximately 0.7 logits. Similar to the results presented by North (1995), the range was slightly narrower in the middle of the scale and wider at the ends. As an integral part of the national project, the CSE-Interpreting team adopted the same cutoff range from Level 6 to Level 9. Reviewing the CSE-Interpreting data according to the cutoff points in **Table 2**, we found that 40% of the descriptors

TABLE 2 | Scaling Results of CSE.

CSE level		Cutoff	Range on scale
Elementary	1	−2.39	
	2	−2.39 −1.65	0.74
	3	−1.65 −0.95	0.70
Intermediate	4	−0.95 −0.27	0.68
	5	−0.27 0.40	0.67
	6	0.40 1.08	0.68
Advanced	7	1.08 1.78	0.70
	8	1.78 2.52	0.74
	9	2.52	

were consistent in terms of their actual and original difficulty levels. Meanwhile, 52% of the descriptors displayed a discrepancy of one level between their measured levels and provisional levels. Based on this, the misfitted descriptors and descriptors with significant discrepancies were modified based on the statistics.

The data from the quantitative validation also demonstrated that the overall difficulty of interpreting competence descriptors was comparatively higher than that of the English language proficiency descriptors (such as listening, speaking, reading, and writing). This result, once again, seems to differentiate interpreting competence from pure linguistic or bilingual competence. However, according to the data from the teachers' and students' evaluation, some descriptors that were deemed to be at Levels 6 and 7 before the scaling process were considered to be easier than the newly determined cutoff points of Level 6. This result implies that the difficulty of some interpreting competence descriptors was lower than Level 6. Therefore, the beginner's level of CSE—Interpreting Competence Scales was revised to Level 5.

Qualitative Validation: Revision

The second round of validation was designed to re-validate the descriptors, especially those that had been modified and re-adjusted previously. The qualitative methods used included survey and focus group interviews among English teachers, interpreting trainers, and interpreters.

Step 1: Survey Design

In total, 49 interpreting competence descriptors were selected for the second round of validation in questionnaires. Of these, 13 were revised descriptors, 21 were re-calibrated, 10 had been newly written by external experts after the quantitative validation, and 5 descriptors were those with significant discrepancies from their original levels according to the result of quantitative validation.

Together with other descriptors selected by the CSE teams, these 49 descriptors were compiled into 10 questionnaires. For Levels 4 to 6, they were embedded into questionnaires B1, B2, B3, B4, and B5. For Levels 7 to 9, they were embedded into questionnaires C1, C2, C3, C4, and C5. An excerpt of a sample questionnaire (C-3) is provided in **Supplementary Appendix 2**.

Step 2: Focus Group Interviews and Workshops

With the collective support of members of the CSE Project, 260 participants from various groups, including high school teachers, English teachers, and T&I trainers from universities, took part in 26 focus group interviews from six regions and provinces (**Table 3**) from March to July 2017. The interviews were designed to obtain feedback on the representativeness and appropriateness of the 49 descriptors collected by the CSE-Interpreting team.

For each focus group interview, 10 English teachers of the targeted student population from at least three different schools or universities were recruited for a 3 h interview. One moderator led the interview with the help of two facilitators (all three were members of the CSE Project).

Before the focus group interview, written informed consent for participation was obtained. The use of the audio recorders was explained. Assurances of confidentiality and privacy in gathering, storing, and handling data were reiterated (Creswell, 2009), and participants were informed that they could withdraw from the interview at any time if they wished. In the interviews, the background and progress of this project was presented, and the purpose of the interview was explained in detail to the participants. The participants were also provided with an executive summary of the nine levels of CSE. Then, the teachers were divided randomly into smaller groups of three to four and worked for about 10 min to discuss and rate the sample descriptors with the guidance of the moderator in order to familiarize with the procedure. They were given the opportunity to ask any questions prior to the interview. The formal interview began with the moderator reading out each descriptor in the questionnaire. After 2 to 3 min of group discussion, the teachers were asked to show their scores by raising the number cards. They were then asked to explain their scoring and comment on the descriptors. They were asked to speak individually one at a time. If there were significant differences between the teachers' scores, or discrepancies between teachers' scores and provisional levels, the moderator could raise further questions. If any teacher had questions or comments

TABLE 3 | Questionnaire Allocation for Focus Groups.

Target population	Beijing	Guangdong	Hubei	Heilongjiang	Shandong	Yunnan
Grade 12 (high school)	B1	B5	B2	B3	B4	B5
EFL (English as a foreign language) course in Bachelor programs	B1		B2	B3	B4	B5
EFL course in Master's programs	B1		B2	B3	B4	B5
Bachelor programs for English majors and BTI programs	C1		C2	C3	C4	C5
Master's programs for English majors and MTI programs	C1		C2	C3	C4	C5

BTI, Bachelor's degree in Translation and Interpreting; MTI, Master's degree in Translation and Interpreting.

about the description, he or she could discuss it briefly. It was unnecessary for the teachers to reach consensus. The moderator's role was to ask questions and seek elaboration but stay neutral (Creswell, 2009). All 26 focus group sessions were simultaneously recorded in two ways: by a tape recorder, used with the permission of the participants, and by two facilitators who took notes during the session but did not participate in the discussion.

Besides the focus group interviews, the CSE-Interpreting team carried out two half-day workshops (4 to 5 h each) in July 2017. These workshops used all 49 descriptors and were conducted in the same format as shown in **Supplementary Appendix 2**. Four conference interpreting practitioners, who were also trainers, joined nine teachers of T&I programs (**Table 4**) in the workshops. While the same procedures were followed, in-depth discussion on the language and content of the descriptors was encouraged. Most participants were female and lecturers. About 55% of them were also practicing interpreters. Their professional background and teaching experience in T&I provided insights for descriptor refinement.

Step 3: Data Analysis

The audio recordings of the interviews and workshops were transcribed with reference to the notes taken during the sessions. For instance, in Questionnaire C3, an English teacher from Heilongjiang province (located in Northeast China) commented on Descriptor No. 45 ("*Can identify the general idea of vague information according to the context during simultaneous interpreting for foreign affairs*") as follows:

> *First of all, my students are unable to perform simultaneous interpreting even in the last semester of the program. Second, what do you mean by "vague information"? It seems to me that it is risky to explicate vague information in political settings, especially for diplomatic meetings.*

TABLE 4 | Workshops With Interpreting Practitioners and Translation and Interpreting (T&I) Teachers.

Total number of participants		***N* = 13**
Sex, *n* (%)	Female	9 (69)
	Male	4 (31)
Age, *years*	Mean	35.5
	Range	27–45
Position, *n* (%)	Professor	1 (8)
	Associate professor	2 (15)
	Lecturer	10 (77)
Highest level of education, *n* (%)	Master's degree	7 (54)
	Doctorate degree	6 (46)
Interpreting experience (years as an interpreter), *n* (%)	1–4 years	3 (23)
	5–10 years	1 (8)
	11–15 years	3 (23)
	16 years and above	3 (23)
Teaching experience, *n* (%)	1–4 years	5 (38)
	5–10 years	3 (23)
	11–15 years	1 (8)
	16 years and above	4 (31)

The provisional level of this descriptor was Level 8 (MTI program and above), which is above the level of the students taught by the teacher. This result supported the appropriateness of the provisional level. The teacher's second point reminded the research team to consider the representativeness and appropriateness of adjectives used in the descriptor.

For Descriptor No. 49 in Questionnaire B5 ("*Can accurately interpret daily conversation with normal speech during liaison interpreting*"), some English as a foreign language (EFL) teachers from universities in Yunnan province (in Southwest China) commented,

> *What does "liaison interpreting" mean? If I don't understand this term, I am not sure if my students can do it. How do you define normal speed?*

This comment indicates that "liaison interpreting" may be less familiar to some teachers and students. In terms of speech rate, the survey team explained the concept in detail (i.e. words per minute for slow, moderate, normal, and fast in the CSE—Interpreting Competence Scales).

In the workshops, most participants agreed with the level and content of the descriptors. Constructive suggestions were also offered to help refine the descriptors' wording. For instance, "*根据口译笔记* (*according to the note-taking*)" should be changed to "*借助口译笔记* (by note-taking)" for Descriptor No. 48 ("*在外事接见的有笔记交传中 能根据口译笔记 译出当地国民生产总值、主要产业、 未来发展方向等关键信息*"); all the "*翻译* (which could mean both 'translation' and 'interpreting')" should be changed to "*口译* (interpreting)"; and "*解决困难* (solve difficulties)" should be changed to "*应对困难* (overcome difficulties)."

Feedback from these verbatim transcripts was entered into Excel spreadsheets. Relevant metadata (e.g. questionnaire number, descriptor number, descriptor ID, category, provisional level) were also recorded. This feedback was then analyzed by both the CSE-Interpreting team and external experts to further revise the descriptors.

Results from both the focus groups and workshops showed that most of the participants agreed with the classification and descriptor levels. They felt that the descriptors were representative of typical interpreting activities. Most teachers, especially the interpreting trainers, agreed that interpreting competence descriptors were explicitly constructed and were generally easy to understand. Nevertheless, five types of problems associated with the 49 descriptors were identified and rectified through the qualitative validation, as shown in the following examples:

(1) Inconsistency. For example, "diplomatic interpreting" was phrased differently ("*外事会见口译*," "*外事接见口译*," and"*外事会面口译*") in Chinese, despite that they all can refer to the same setting. In addition, "*search*," "*collect*," and "*look for*" were found in different levels of subscales of interpreting strategy. Although these words were used to refer to the same action, they may indicate different

levels of difficulty. Therefore, revisions had to be made to ensure terminology consistency.

(2) Ambiguity. For example, "*在新闻报道的有笔记交传中，能监控译语并在发生逻辑错误时及时 与发言人确认并更正*." (*For consecutive interpreting using notes in a media setting, one can monitor their target language and confer with the speaker in a timely fashion when there is a logic error.*) In this descriptor, it is unclear whether the "*logic error*" is made by the speaker or the interpreter. The descriptor was revised by deleting the action initiator: "*For consecutive interpreting with notes in a media setting, one can monitor the logic error in their target language.*" In this way, the "*logic error*" could be made by either the speaker or the interpreter.

(3) Repetitiveness. Despite the five rounds of relevant analysis, some descriptors still seemed to be redundant. For instance, "*在商务接待的无笔记交传中，在能使用具体的表达形式，在译语表达中区分主要信息和次要信息*." (*For consecutive interpreting without notes in a business setting, one can use a specific expression to distinguish primary information and secondary information in the target language.*) Here, "*具体的表达形式* (*specific expression*)" and "*译语表达* (*target language*)" share the same meaning. The descriptor was then revised to "*在商务接待的无笔记交传中，能在译语表达中区分主要信息和次要信息*." (*For consecutive interpreting without notes in a business setting, one can distinguish primary information and secondary information in target language*).

(4) Descriptors with similar meanings within the same level. For example, "*能评估源语信息传递是，否出现错误 包括重要信息、观点、细节和重要例子* (*Can evaluate whether there is an error in delivering the source information, such as key information, opinions, details and important examples*)" and "*能评估核心信息遗漏、逻辑结构混乱、关键术语误译等重大错误* (*Can evaluate major errors such as core information loss, confusing logical structure, key terms mistranslation*)" were both found in the strategy subscale of Level 8. These two descriptors essentially touched on similar abilities. In this case, we revised the first descriptor into "*在同声传译中，我能评估并修正源语信息传递中出现的错误* (*In simultaneous interpreting, I can evaluate and correct major errors*)," by adding the specific interpreting mode to differentiate it from interpreting in general.

(5) Untypical activities. For example, in the descriptors "*在商务接待的无笔记交传中，能译出原 料价格信息之间的逻辑关系* (*For consecutive interpreting without notes in business receptions, one can identify the logical relationship between raw material prices*)," participants in the interviews felt that "raw material prices" were seldom mentioned in the scenarios of "business reception." Therefore, the descriptor was revised as "*在商务接待的无笔记交传中，能概括地译出接待方行 程安排等信息* (*For consecutive interpreting without notes in business receptions, one can interpret the itinerary and other information briefly*)."

Based on the results of the quantitative and qualitative validation, an external expert group consisting of researchers, trainers, and interpreters were invited to refine the descriptors. Eleven descriptors with typical interpreting activities were newly written by these experts. Upon the request of the CSE Project, the author team wrote descriptors to summarize the interpreting performance of each level. Sixteen descriptors were written for the Overall Interpreting Performance Scale, and 28 were written for the Self-assessment Scale for Interpreting Competence.

Upon final refinement by the Chinese editors, 12 scales with 369 descriptors and five levels were developed for interpreting competence: Overall Interpreting Performance (1 scale, 16 descriptors), Interpreting Competence in Typical Interpreting Activities (6 scales, 220 descriptors), Interpreting Strategy (4 scales, 105 descriptors), and Self-assessment for Interpreting Competence (1 scale, 28 descriptors). **Supplementary Appendix 3** provides two examples of CSE—Interpreting Competence Scales in English, and the full English version can be accessed on the National Education Examinations Authority (NEEA) website (see text footnote 1).

DISCUSSION

Description of Cognitive Ability

Cognitive ability is regarded as a core element of the interpreting competence construct. However, it is not feasible to operationalize it in the description stage. Based on several meetings and discussions among the project team members, the features of interpreting's cognitive process have been conceptualized in different settings. Cognitive activities are described through the process and the product of interpreting, such as identifying, retrieving, summarizing, analyzing, anticipating, and monitoring.

The description of cognitive ability sometimes appears to overlap with typical interpreting activities. For example, one may find descriptors with similar cognitive abilities in the subscales of Typical Interpreting Activities and scales for Overall Interpreting Performance. The two sets of scales differ because the first focuses on a few real-life interpreting settings and is of practical use in the workplace, while the second pertains to the core part of interpreting competence at each level.

Description of Interpreting Strategies

Interpreting strategies have, in some cases, turned out to be an unfamiliar or confusing concept for some teachers and students. This confusion may be related to how interpreting is taught and studied in the Chinese context. Compared with the product of interpreting (i.e. performance), the process of interpreting can easily be overlooked in interpreting training and learning. There has been very sparse coverage and minimal guidance in relevant training syllabuses on cognitive task analysis in interpreting, let alone the teaching of interpreting strategies (since most of the interpreting courses in China are skills-oriented). Through rigorous training and detailed illustration during the exemplar-generation stage, the teachers may consider demonstrating some useful strategies often used by their students in the classroom

or in their after-class interpreting exercises. When dealing with strategy descriptors, we grouped abstract descriptors concerning planning, monitoring, and evaluating into metacognitive strategies; meantime, we categorized specific and concrete descriptors involving inferencing, elaborating, summarizing, repeating, and note-taking into cognitive strategies. However, descriptors related to emotion or social interaction that might be categorized as socio-affective strategies (Vandergrift, 1997) were not yet included.

Interpreting Task Difficulty

As the scales are designed to be applied in the Chinese educational system, the descriptors are required to be explicit and internally consistent. For the interpreting competence descriptors, we usually used criteria such as delivery speed, length of the speech, and topic and lexical complexity to indicate the difficulty level of interpreting materials. However, in the case of delivery speed, what exactly is the difference between "delivered slowly" and "delivered at normal speed"? In our study, we distinguished between four levels of delivery speed—slow, moderate, normal, and fast—and then defined each level:

> *"Fast (in English): approximately 140–180 words/min; moderate speed (in English): approximately 100–140 words/min; fast (in Chinese): approximately 160–220 Chinese characters/min; moderate speed (in Chinese): approximately 120–160 Chinese characters/min"* (National Education Examinations Authority, 2018).

This level-defining approach could also be applied to other criteria, although decisions should be made carefully based on rigorous theoretical underpinning and sufficient evidential support. Similar to the validation of CEFR and other related scales, we should continue to collect relevant data in order to fine-tune interpreting competence descriptors.

Interpreting Modes and Levels in the Scales

To indicate the developmental stages of interpreting competence, five levels are used to represent the three classic stages of basic, intermediate, and advanced competence (**Figure 4**). For instance, Levels 5 and 6 are basic stages, at which one can complete liaison interpreting tasks. Typical interpreting activities at these levels are relatively simple and informal, with moderately slow speech rate and short segments. These could include a guided tour, guest reception, informal visit, or business escort. Levels 7 and 8 are the intermediate stages; student interpreters at these levels should be able to complete interpreting tasks with longer segments and in more formal settings. In particular, Level 8 involves the advanced stage of consecutive interpreting and the introductory stage of simultaneous interpreting. In other words, Level 8 represents the transition from beginner to advanced learners. Level 9 represents the most difficult tasks and the almost "perfect" performance of interpreting.

Although the scales were developed in a mixed-methods empirical approach, establishing the cutoff points between levels was, in part, a subjective procedure. While some students, teachers, researchers, or institutions may prefer broad levels, others prefer finer levels. The advantage of this three-stage-branching approach is that "a common set of levels and/or descriptors can be 'cut' into practical local levels at different points by different users to suit local needs and yet still relate back to a common system" (Council of Europe, 2001). According to specific purposes (teaching, learning, and testing, for example), users of the scale can introduce sub-levels to the scales to fit their specific needs.

LIMITATIONS AND APPLICATIONS

Despite the tremendous efforts involved in the CSE Project, there are still a number of limitations. First, because of the logistical constraints, only 2.54% of the participants in this study were T&I teachers or students. As the scales were designed for a wide range of potential T&I teachers, students, and staff working in corporations and government agencies, they warrant further revision and validation to ensure that they appropriately reflect professional practice. Second, although the Rasch-based results were quite encouraging, the relevant analyses were conducted collectively by the statistics team of the CSE Project based on the data of descriptors from all CSE teams. As a result, we did not obtain the Wright maps describing our data on interpreting competence. Finally, the 55 newly written descriptors created by the external experts and the author team after the qualitative

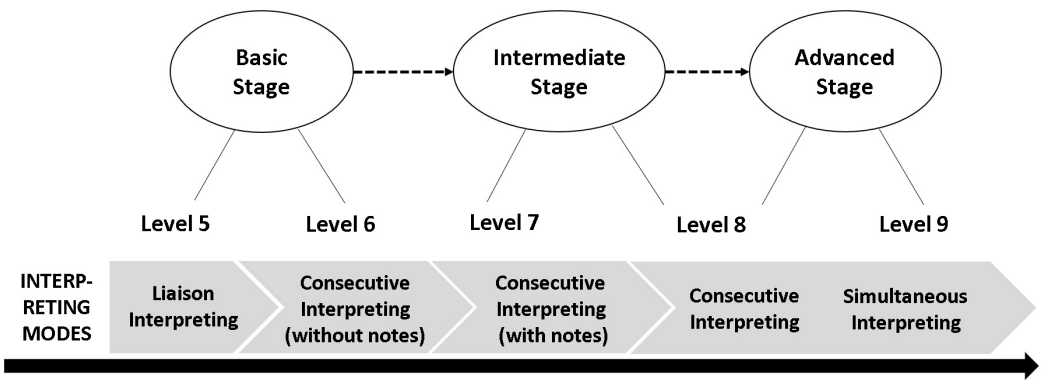


FIGURE 4 | Interpreting modes vs. levels in the Scales.

validation need to be validated based on the steps described in Section 3, *Validating the scale*.

In terms of application, the Interpreting Competence Scales can be operationalized for teaching, learning, and assessment purposes. First, interpreting trainers can make use of the levels and corresponding descriptors of the scale in their teaching plans, pedagogy, and teaching materials. For these applications, it will be useful to transform the descriptors into classroom tasks at different stages of interpreting training. This allows trainers to use descriptors to evaluate performance, develop teaching materials, and examine the appropriateness of the descriptors. Second, although students' self-study contributes to the development of interpreting competence, little guidance is available for material selection and performance assessment (Wang, 2015). The Self-Assessment Scale for Interpreting Competence can be used by students to self-diagnose and evaluate their learning outcomes. It potentially provides students with opportunities to understand their current level of interpreting competence, assess their performance, and set specific goals for further improvement. Further research is required to investigate the effectiveness and washback of the Self-Assessment Scale in students' self-directed practice.

Third, testing and assessment of interpreting competence is an important area in which our scales are expected to play a major role. The scales offer a window into interpretation aptitude by setting levels of baseline competence. They also provide detailed information about interpreting activities, strategies, and requirements for interpreting quality at different levels. Given the detailed descriptors, teachers and testers may be able to use the scales to inform the development of aptitude tests, diagnostic tests, and formative and summative assessments of interpreting. In addition, existing tests (e.g. NAATI and CATTI) should be aligned to the standardized descriptor scales. The focus of alignment should be on characteristics of practice domains (e.g. subject matter, interpreting activities), difficulty levels, and the rating methods. Such alignment would help achieve greater consistency and coherence in interpreting education and facilitate communication among interpreting trainers, learners, test developers, professionals, and policymakers.

DATA AVAILABILITY STATEMENT

The data sets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WW and YX wrote the initial draft of the manuscript. BW advised on the structure and revised the manuscript substantially. LM led the relevant project on the development of scales for interpreting competence.

FUNDING

This research was supported by the National Social Science Fund of China (19CYY053).

ACKNOWLEDGMENTS

We are grateful to all the students, teachers, and interpreters who participated in or contributed to this project. We would like to thank the reviewers and editors for their careful reading of our manuscript and insightful comments on this article. Our revisions reflect all reviewers' and editors' suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00481/full#supplementary-material

REFERENCES

Aís, A. ìC. (1998). *La Evaluacioìn de la Calidad en Interpretacioìn Simultaìnea: La Importancia de la Comunicacioìn Verbal In Spanish*. Granada: Comares.

Albl-Mikasa, M. (2013). Developing and cultivating expert interpreter competence. *Interpret. Newsl.* 18, 17–34. doi: 10.21256/zhaw-4081

American Council on the Teaching of Foreign Languages (2019). Available online at: https://www.actfl.org/publications/all (accessed December 9, 2019).

Angelelli, C. V., and Degueldre, C. (2002). "Bridging the gap between language for general purposes and language for work: an intensive superior level language/skill course for teachers, translators, and interpreters," in *Developing Professional-Level Language Proficiency*, eds B. L. Leaver and B. Shekhtman (Cambridge: Cambridge University Press), 91–110.

Arumí Ribas, M., and Vargas-Urpi, M. (2017). Strategies in public service interpreting. *Interpreting* 19, 118–141. doi: 10.1075/intp.19.1.06aru

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

Bachman, L. F., and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*, Vol. 1. London: Oxford University Press.

Barik, H. C. (1975). Simultaneous interpretation: qualitative and linguistic data. *Lang. Speech* 18, 272–297. doi: 10.1177/002383097501800310

Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting* 8, 149–174. doi: 10.1075/intp.8.2.03bar

Berk-Seligson, S. (1988). The impact of politeness in witness testimony: the influence of the court interpreter. *Multilingua* 7, 411–439. doi: 10.1515/mult.1988.7.4.411

Bond, G., and Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd Edn, Newbury Park: Routledge.

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

Campbell, S., and Hale, S. (2003). "Translation and interpreting assessment in the context of educational measurement," in *Translation Today: Trends and Perspectives*, eds G. Anderman and M. Rogers (Clevedon: Multilingual Matters), 205–224. doi: 10.1016/0167-9317(92)90091-5

Centre for Canadian Language Benchmarks (2019). *Centre for Canadian Language Benchmarks*. Available online at: https://www.language.ca/home/ (accessed December 9, 2019).

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

Creswell, J. W. (2009). *Research Design: Qualitative and Mixed Methods Approaches*. Thousand Oaks, CA: Sage Publications.

Díaz-Galaz, S. (2011). The effect of previous preparation in simultaneous interpreting: preliminary results. *Across. Lang. Cult.* 12, 173–191. doi: 10.1556/Acr.12.2011.2.3

Díaz-Galaz, S., Padilla, P., and Bajo, M. T. (2015). The role of advance preparation in simultaneous interpreting: a comparison of professional interpreters and interpreting students. *Interpreting* 17, 1–8. doi: 10.1075/intp.17.1.01dia

Dong, Y. (2018). Complex dynamic systems in students of interpreting training. Translation and interpreting studies. *J. Amer. Transl. Interpret. Stud. Assoc.* 13, 185–207. doi: 10.1075/tis.00011.don

Dong, Y., Li, Y., and Zhao, N. (2019). Acquisition of interpreting strategies by student interpreters. *Interpret. Transl. Tra.* 13, 408–425. doi: 10.1080/1750399X.2019.1617653

Fantinuoli, C. (2017). Computer-assisted preparation in conference interpreting. *Translat. Interpret.* 9, 24–37. doi: 10.12807/ti.109202.2017.a02

Garzone, G. E. (2003). "Reliability of quality criteria evaluation in survey research," in *La Evaluación de la Calidad en Interpretación: Investigación*, eds A. ìC. Aís, M. M. F. Sánchez, and D. Gile (Granada: Comares), 23–30.

Gile, D. (2011). "Errors, omissions, and infelicities in broadcast interpreting: preliminary findings from a case study," in *Methods and Strategies of Process Research: Integrative Approaches in Translation Studies*, eds C. Alvstad, A. Hild, and E. Tiselius (Amsterdam: John Benjamins), 201–218. doi: 10.1075/btl.94.15gil

Grbić, N. (2008). Constructing interpreting quality. *Interpreting* 10, 232–257. doi: 10.1075/intp.10.2.04grb

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.

Han, C. (2015). *Building the Validity Foundation for Interpreter Certification Performance Testing*. Doctoral thesis, Macquarie University, Sydney.

Han, C. (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: a generalizability theory approach. *Lang. Assess. Q.* 13, 186–201. doi: 10.1080/15434303.2016.1211132

Han, C. (2017). Using analytic rating scales to assess English–Chinese bi-directional interpreting: a longitudinal Rasch analysis of scale utility and rater behaviour. *Linguist. Antverpien. New Ser.* 16, 196–215.

Han, C. (2018). Using rating scales to assess interpretation: practices, problems, and prospects. *Interpreting* 20, 59–95. doi: 10.1075/intp.00003.han

Han, C. (2019). A generalizability theory study of optimal measurement design for a summative assessment of English/Chinese consecutive interpreting. *Lang. Test.* 36, 419–438. doi: 10.1177/0265532218809396

Han, C., and Riazi, M. (2018). The accuracy of student self-assessments of English-Chinese bidirectional interpretation: a longitudinal quantitative study. *Assess. Eval. Higher Educ.* 43, 386–398. doi: 10.1080/02602938.2017.1353062

Interagency Language Roundtable (2002). *About the Interagency Language Roundtable*. Available online at: https://www.govtilr.org/IRL%20History.htm# (accessed November 03, 2019).

International Second Language Proficiency (n.d.). *International Second Language Proficiency*. Available online at: https://www.islpr.org (accessed December 9, 2019).

Kalina, S. (2000). Interpretation competences as a basis and a goal for teaching. *Interpret. Newsl.* 10, 3–32.

Kalina, S. (2002). "Quality in interpreting and its prerequisites – a framework for a comprehensive view," in *Interpreting in the 21st Century: Challenges and Opportunities*, eds G. Garzone and M. Viezzi (Amsterdam: John Benjamins), 121–130. doi: 10.1075/btl.43.12kal

Kalina, S. (2005a). Quality assurance for interpreting processes. *Meta* 50, 769–784. doi: 10.7202/011017ar

Kalina, S. (2005b). "Quality in the interpreting process: what can be measured and how?," in *Directionality in Interpreting: The 'Retour' or the Native?*, eds R. Godijns and M. Hinderdael (Ghent: Communication and Cognition), 27–46.

Kohn, K., and Kalina, S. (1996). The strategic dimension of interpreting. *Meta* 41, 118–138. doi: 10.7202/003333ar

Kolen, M. J., and Brennan, L. R. (2004). *Test Equating, Scaling, and Linking*, 2nd Edn, Newbury Park: Springer.

Kurz, I. (1993). Conference interpretation: expectations of different user groups. *Interpret. Newsl.* 5, 13–21.

Kutz, W. (2010). *Dolmetschkompetenz: Was Muss Der Dolmetscher Wissen und Können?*. München: European University Press.

Lee, J. (2008). Rating scales for interpreting performance assessment. *Interpret. Transl. Tra.* 2, 165–184. doi: 10.1080/1750399X.2008.10798772

Lee, S. (2015). Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances. *Interpreting* 17, 226–254. doi: 10.1075/intp.17.2.04lee

Lee, S. (2018). Exploring a relationship between students' interpreting self-efficacy and performance: triangulating data on interpreter performance assessment. *Interpret. Transl. Tra.* 12, 166–187. doi: 10.1080/1750399X.2017.1359763

Lee, S. (2019). Holistic assessment of consecutive interpretation: how interpreter trainers rate student performances. *Interpreting* 21, 245–268. doi: 10.1075/intp.00029.lee

Li, X. (2013). Are interpreting strategies teachable? Correlating trainees' strategy use with trainers' training in the consecutive interpreting classroom. *Interpret. Newsl.* 18, 105–128.

Li, X. (2015). Putting interpreting strategies in their place: justifications for teaching strategies in interpreter training. *Babel* 61, 170–192. doi: 10.1075/babel.61.2.02li

Linacre, J. M. (2013). *A user's guide to FACETS. Program manual 3.71.0*. Available online at: https://www.winsteps.com/facets.htm (accessed May 20, 2018).

Liu, J. (2019). *The Development of the China Standards of English (CSE)*. Beijing: Higher Education Press.

Liu, M. (2008). "How do experts interpret? Implications from research in interpreting studies and cognitive science," in *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile*, eds G. Hansen, A. Chesterman, and H. Gerzymisch-Arbogast (Amsterdam: John Benjamins), 159–177.

Liu, M. (2013). "Design and analysis of Taiwan's interpretation certification examination," in *Assessment Issues in Language Translation and Interpreting*, eds D. Tsagari and R. Deemter (Frankfurt: Peter Lang), 163–178.

Moser-Mercer, B. (1996). Quality in interpreting: some methodological issues. *Interpret. Newsl.* 7, 43–55.

Napier, J. (2004a). Interpreting omissions: a new perspective. *Interpreting* 6, 117–142. doi: 10.1075/intp.6.2.02nap

Napier, J. (2004b). Sign language interpreter training, testing, and accreditation: an international comparison. *Am. Ann. Deaf.* 149, 350–359. doi: 10.1353/aad.2005.0007

National Education Examinations Authority (2014). The implementation plan of the national scale project. *Intern. Docu. Issued NEEA* 12:2014.

National Education Examinations Authority (2018). *China's Standards of English Language Ability*. Available online at: http://cse.neea.edu.cn/res/ceedu/1811/6bdc26c323d188948fca8048833f151a.pdf (accessed December 20, 2019).

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System* 23, 445–465.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency (Vol. 8)*. Switzerland: Peter Lang Pub Inc.

North, B. (2014). *The CEFR in Practice*. Cambridge: Cambridge University Press.

Pöchhacker, F. (2000). *Dolmetschen: Konzeptuelle Grundlagen und Deskriptive Untersuchungen (in German)*. Tübingen: Stauffenburg-Verlag.

Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta* 46, 410–425. doi: 10.7202/003847ar

Pöchhacker, F. (2015). *Routledge Encyclopedia of Interpreting Studies*. London: Routledge.

Postigo Pinazo, E. (2008). Self-assessment in teaching interpreting. *TTR Traduct. Terminol. Reidact.* 21, 173–209. doi: 10.7202/029690ar

Pym, A., Grin, F., Sfreddo, C., and Chan, A. L. J. (2013). *The Status of the Translation Profession in the European Union*. London: Anthem Press, doi: 10.2782/64819

Seleskovitch, D., and Lederer, M. (1989). *Pédagogie Raisonnée de l'Interprétation*. Paris: RID Publications.

Setton, R., and Dawrant, A. (2016). *Conference Interpreting: A Complete Course and Trainer's Guide (2 Vols). Amsterdam*. John Benjamins.

Szabari, K. (2002). "Interpreting into the B language," in *Teaching Simultaneous Interpretation into a "B" Language*, ed. EMCI Workshop (Paris: EMCI), 12–19.

Tiselius, E. (2009). "Revisiting carroll's scales," in *Testing and Assessment in Translation and Interpreting Studies*, eds C. V. Angelelli and H. C. Jacobson (Amsterdam: John Benjamins), 95–121. doi: 10.1075/ata.xiv.07tis

Vandergrift, L. (1997). The comprehension strategies of second language (French) listeners: a descriptive study. *Foreign Lang. Ann.* 30, 387–409. doi: 10.1111/j.1944-9720.1997.tb02362.x

Wang, B. (2007). From interpreting competence to interpreter competence –A tentative model for objective assessment of interpreting (in Chinese). *Foreign Lang. World* 120, 75–78.

Wang, B. (2012). From interpreting competence to interpreter competence: exploring the conceptual foundation of professional interpreting training (in Chinese). *J. Foreign Lang. Teach.* 267, 75–78.

Wang, B. (2015). Bridging the gap between interpreting classrooms and real-world interpreting. *Int. J. Interpret. Educ.* 1, 65–73.

Wang, J., Napier, J., Goswell, D., and Carmichael, A. (2015). The design and application of rubrics to assess signed language interpreting performance. *Interpret. Transl. Tra.* 9, 83–103. doi: 10.1080/1750399X.2015.1009261

Will, M. (2007). "Terminology work for simultaneous interpreters in LSP conferences: model and method," in *Proceedings of the MuTra – Multidimensional Translation Conference Proceedings*, eds H. Gerzymisch-Arbogast and G. Budin (Cham: Springer), 65–99.

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measur. Trans* 8:370.

Wu, Y., and Liao, P. (2018). Re-conceptualising interpreting strategies for teaching interpretation into a B language. *Interpret. Transl. Tra.* 12, 188–206. doi: 10.1080/1750399X.2018.1451952

Yeh, S., and Liu, M. (2006). A more objective approach to interpretation evaluation: exploring the use of scoring rubrics. *J. Nat. Inst. Compilat. Transl.* 34, 57–78.

Zhong, W. (2003). *Memory Training in Interpreting*. Availabe at: http://translationjournal.net/journal/25interpret.htm (accessed November 12, 2013).

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.

frontiers in Communication

ORIGINAL RESEARCH
published: 12 May 2020
doi: 10.3389/fcomm.2020.00025

The Development and Evaluation of a New ASL Text Comprehension Task

Patrick Rosenburg, Amy M. Lieberman, Naomi Caselli and Robert Hoffmeister*

Wheelock College of Education and Human Development, Boston University, Boston, MA, United States

Being able to comprehend a language entails not only mastery of its syntax, lexicon, or phonology, but also the ability to use language to construct meaning, draw inferences, and make connections to world knowledge. However, most available assessments of American Sign Language (ASL) focus on mastery of lower level skills, and as a result little is known about development of higher-order ASL comprehension skills. In this paper, we introduce the American Sign Language Text Comprehension Task (ASL-CMP), a new assessment tool to measure ASL text comprehension ability in deaf children. We first administered the task to a group of deaf children with deaf parents ($n = 105$, ages 8–18 years) in order to evaluate the reliability and validity of the task, and to develop norms. We found that the ASL-CMP has acceptable levels of internal consistency, difficulty, and discriminability. Next, we administered the task to an additional group of deaf children with hearing parents ($n = 251$, ages 8–18 years), and found that the ASL-CMP is sensitive to expected patterns: older children have better ASL text comprehension skills, literal questions are generally easier to answer than inferential questions, and children with early exposure to ASL generally outperform those with delayed exposure. We conclude that the ASL-CMP task is reliable and valid and can be used to characterize ASL text comprehension skills in deaf children.

Keywords: ASL, assessment, comprehension, deaf, sign language

Edited by:
Vahid Aryadoust, Nanyang Technological University, Singapore

Reviewed by:
Wolfgang Mann, University of Roehampton London, United Kingdom
Tobias Haug, Interkantonale Hochschule für Heilpädagogik (HfH), Switzerland
Joseph Bochner, Rochester Institute of Technology, United States

***Correspondence:**
Patrick Rosenburg
palen@bu.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *18 December 2019*
Accepted: *09 April 2020*
Published: *12 May 2020*

Citation:
Rosenburg P, Lieberman AM, Caselli N and Hoffmeister R (2020) The Development and Evaluation of a New ASL Text Comprehension Task. Front. Commun. 5:25. doi: 10.3389/fcomm.2020.00025

INTRODUCTION

Reading comprehension—the ability to extract meaning from a text, to evaluate that information, to draw inferences, and to make connections to outside information—is an essential skill for classroom learning, as well as for later academic, social, and occupational achievement (Duke and Pearson, 2002; Shanahan, 2005; Van den Broek and Espin, 2012; Ciullo et al., 2016). In 1994, The New London Group proposed a theory of multiliteracies (first published in 1996), which broadened the understanding of literacy to encompass the ability to engage with many forms of text. In a rapidly-evolving world of information and technology, they argued that texts encompass both traditional formats like essays, articles, or books, but should also consider forms such as speeches, blogs (Shema et al., 2012; Mackey and Jacobson, 2014), vlogs (Griffith and Papacharissi, 2009), graphic novels (Jimenez et al., 2017), and online reading (Leu et al., 2015). With a broadened definition of text, literacy can be considered as a constellation of skills through which a person can extract and construct meaning from these various forms.

ASL Texts

In parallel with these expanded definitions of text and literacy, some began to consider compositions in sign languages as a form of text, and the ability to engage with these compositions as a form of literacy (Kuntze, 2004; Kuntze et al., 2014; Wall, 2014). We embrace this reimagining, and use it as a framework to examine the complex linguistic and cognitive skills involved in engaging with passages composed in American Sign Language (ASL)[1], which we will refer to as ASL texts.

We define an ASL text as a composition expressed in ASL that is used to communicate information to others (Christie and Wilkins, 1997; Byrne, 2015). Although typically ASL is ephemeral, in the way that spoken language "disappears" once it is produced, signers can also of course record their own productions. ASL texts may be produced live, as in a lecture or presentation, or may be recorded by video or other medium (e.g., motion capture) or generated digitally (e.g., avatars). The form of ASL texts most analogous to a conventional understanding of written texts are signed videos that have been designed deliberately, often involving multiple iterations of editing and refining, and are recorded such that users can preview, review, and engage with them repeatedly. ASL texts can be classified into a host of literary genres, including poetry (Christie and Wilkins, 1997; Blondel et al., 2008), satire (e.g., Hearing Knows Best [https://youtu.be/MoxVdw6T0LA] by Malzkuhn and Bottoms, 2017), fiction, jokes, and stories (Bahan, 2006; Byrne, 2015). Non-fiction ASL texts have become prevalent in recent years with the establishment of several ASL news outlets that produce news stories of particular relevance to deaf people or about the world at large (see The Daily Moth [https://www.dailymoth.com] by Abenchuchan, 2019 and Sign1News [https://sign1news.com] by Jones, 2018). Additionally, some museums have installed ASL expository texts adjacent to each exhibit that offer ASL users access to self-guided tours (Martins, 2016). A more popular, generally less edited, example of an ASL text is the vlog, a short video message of one or two signers expressing an opinion or short narrative that is often shared through social media. Given the large and growing body of text available in ASL, it is critical to understand and evaluate how deaf children develop the ability to engage with this material (Snoddon, 2010).

Like all texts, ASL texts can be important sources of information through which people can expand their knowledge, skills, and experience. Additionally, by learning to comprehend an ASL text in their primary language, deaf students can gain familiarity with various genres, develop the ability to interpret explicit and implicit meaning, and make connections to prior knowledge or other texts (Kuntze, 1998, 2004; Kuntze et al., 2014), which in turn contributes to later reading comprehension (Duffy, 2009). These modality-general skills are important not only for engaging deeply with ASL texts, but many scholars have proposed that ASL texts provide an entry point to engaging with written English (Hoffmeister, 2000; Bailes, 2001; Kuntze, 2004; Cummins, 2006; DeLana et al., 2007; Kuntze et al., 2014). While comprehension of ASL text in deaf children has been, to our knowledge, underexplored, we expect that many of the same skills identified for written text comprehension underlie ASL text comprehension.

Text comprehension relies on a host of language and literacy skills. At a basic level, comprehending a text entails lower-level language skills including identifying words and parsing sentences (Perfetti and Stafura, 2014; Silva and Cain, 2015). In addition to these basic skills, higher-order skills are needed to integrate information explicitly stated in the text as well as information implied by the text (Pettit and Cockriel, 1974; Bishop and Adams, 1992; Cain and Oakhill, 2007). This requires the use of prior knowledge, and the ability to construct a coherent interpretation of the text including drawing conclusions and making predictions (Kintsch, 1998; Nassaji, 2003; Perfetti et al., 2005; Cromley and Azevedo, 2007; Landi, 2010).

Better understanding the development of ASL text comprehension is of particular interest for deaf children because the majority of deaf children are at risk of limited language proficiency and low literacy levels (Hrastinski and Wilbur, 2016). Deaf children do not have auditory access to all of the sounds of speech, and even with the best-available technology and interventions their spoken language outcomes are variable and often poor (Manrique et al., 2004; Bouchard et al., 2009; Niparko et al., 2010; Peterson et al., 2010; Ganek et al., 2011; Dettman et al., 2016; Kral et al., 2016; Szagun and Schramm, 2016; Humphries et al., 2017). At the same time, more than 90% of deaf children have hearing parents who do not know a sign language at the time the child is born (Hall, 2017; Hall et al., 2018), so in addition to incomplete access to spoken language, deaf children also often have limited exposure to a sign language during early childhood. For all these reasons, it is critical to assess whether or not children have developed the complex language and literacy skills involved in engaging with an ASL text.

Existing Assessments of ASL Comprehension

Despite the importance of higher-order text comprehension skills, existing ASL assessments generally focus on basic proficiency in ASL vocabulary and grammar, and there is currently no means of evaluating the more advanced skills that are necessary for ASL text comprehension. Currently available ASL tests include, for example, the American Sign Language Vocabulary Test [ASL-VT; (Mann et al., 2016)], the MacArthur Bates CDI for American Sign Language (Anderson and Reilly, 2002), the ASL-CDI 2.0 (Caselli et al., 2020), the ASL Phonological Awareness Test (ASL-PAT; McQuarrie and Spady, 2012), the American Sign Language Proficiency Assessment [ASL-PA; (Maller et al., 1999)], the ASL Receptive Skills Test (Enns and Herman, 2011), ASL and Non-linguistic Perspective Taking Comprehension Tests (Quinto-Pozos and Hou, 2015), and the Visual Communication and Sign Language Checklist [VCSL, (Simms et al., 2013)]. See Haug (2008) for an overview of available ASL assessment tests. These tests predominantly focus on lower-level language skills including phonology, vocabulary, morphology, and syntax, rather than higher-level text comprehension skills. One exception is the American Sign

[1]Our focus in this paper is on American Sign Language, though the approach would largely generalize to compositions in other sign languages.

Language Assessment Instrument (ASLAI; Hoffmeister et al., 2015), which includes sub-tasks that assess ASL analogical reasoning (Henner, 2015), and ASL complex syntax (Hoffmeister et al., 2015). Another exception is the Test of American Sign Language [TASL, (Prinz et al., 1994; Strong and Prinz, 1997)], which probes deaf children's comprehension of ASL text as a set of literacy skills, but has not been evaluated for psychometric quality nor are there developmental norms (Haug, 2008). To our knowledge there is no currently available normed assessment that evaluates deaf children's comprehension of ASL text.

The Current Study

In the current study, we present a new assessment of ASL text comprehension called the ASL Text Comprehension task (ASL-CMP). The goal of the ASL-CMP is to measure ASL text comprehension skills among deaf children. We first describe the development of the ASL-CMP, and present an evaluation of its psychometric properties in a sample of deaf children who had access to ASL from birth. Following the psychometric evaluation, we present results from a larger sample of deaf children that included both those with deaf parents and hearing parents. The goal of the larger sample was to test three primary predictions:

1) We expected that, because they generally have earlier exposure to language, deaf children who have deaf parents would outperform deaf children who have hearing parents in accuracy on the test (Hoffmeister, 2000; Goldin-Meadow and Mayberry, 2001; Berke, 2012; Henner et al., 2016). Because the age of onset of ASL acquisition is generally correlated with language proficiency (see Mayberry and Kluender, 2018 for a review), we also expected that age of entry into a school that uses ASL would be negatively correlated with ASL text comprehension among children who have hearing parents.
2) We predicted that accuracy on the ASL-CMP would increase during childhood and adolescence, as is generally found in studies of written text comprehension (Barnes et al., 1996; Cain and Oakhill, 1999; Nippold and Scott, 2010).
3) We predicted that accuracy would be higher for questions assessing literal comprehension than for those that required children to make inferences, as inferential comprehension is generally more difficult than explicit text comprehension (Pettit and Cockriel, 1974; Johnston, 1984; Miller and Smith, 1985; Bowyer-Crane and Snowling, 2005; Cain and Oakhill, 2007).

METHODS

Development of the Assessment

The ASL-CMP was created by a team of deaf native-signing linguists and educators and hearing linguists who are familiar with ASL. Deaf experts who have technical expertise as well as mastery of the language play a critical role in ensuring validity of ASL assessments (Hoffmeister, 1988; Paludneviciene et al., 2012; Hoffmeister et al., 2015; Enns et al., 2016; Henner et al., 2018). The ASL-CMP was developed as a subtest of the ASLAI, a large, comprehensive, norm-referenced ASL assessment. The ASLAI has been used to test receptive ASL skills in Deaf children from ages 4–18 years across the United States (Henner et al., 2018). The ASLAI evaluates a wide range of linguistic properties of ASL, such as vocabulary, syntax, and analogical reasoning skills (Hoffmeister et al., 2015). All tasks in the ASLAI, including the ASL-CMP, are administered via computer and are multiple-choice. All questions and answer choices are presented in ASL, and formatted with consideration of the linguistic demands of ASL, as described in the section Test Procedures.

Test Content of the ASL-Text Comprehension Task

The ASL-CMP consists of three ASL texts that were adapted–not translated–from texts in two different reading assessments: the Qualitative Reading Inventory-5 (QRI-5), an informal reading assessment used to identify students' reading levels (Leslie and Caldwell, 2011) and the Houghton Mifflin Reading Assessment (Houghton Mifflin, 2010), a research-based diagnostic reading assessment. In contrast to test translation where the goal is a sentence-by-sentence match between the original and translated version, our goal in adapting these tests was to create texts that had an overall conceptual match with the original but the words, sentences, and structure of the text were free to differ (Hambleton and Patsula, 1998; Van de Vijver and Poortinga, 2005).

The English texts that served as the models for the ASL texts were titled *Bridges, Photosynthesis*, and *Marva Finds a Friend* (Leslie and Caldwell, 2011). The English texts were originally designed for children ages 8–12 years. Two of the English texts (*Bridges* and *Photosynthesis*) are expository, non-fiction texts, and the third (*Marva Finds a Friend*) is fiction. Texts were selected based on the target age range, and because they contained a straightforward sentence structure, which enabled adaptation to ASL (e.g., no passive voice and simple sentence structure). The three adapted ASL texts and English translations of those texts are available at https://osf.io/dwhba/. The length of the ASL texts were 2 min, 39 sec (*Bridges),* 1 min, 36 sec (*Photosynthesis*), and 2 min, 58 sec (*Marva Finds a Friend).* Each ASL text was followed by five multiple-choice questions. Three of the questions were related to information that was explicitly mentioned in the text (literal questions) and two of the questions were related to information that was implied by the text but not explicitly stated (inferential questions). Further, each set of five questions was consistent in structure such that there were two WHAT questions (one literal, one inferential), two WHY questions (one literal, one inferential), and one WHICH question (literal). The foils for each question were all ASL signs and consisted of two related but incorrect answers, and one unrelated answer. For literal questions, the related foils differed from the correct answer in either verb or subject in ASL. For example, if the correct answer was GIRL *WALK SEE* OLD HOUSE[2], related but incorrect answers used the verb *RUN* or *BIKE* instead of *WALK.* For inferential questions, the correct answer included information that must be deduced from the text. For example, in one of the ASL texts a girl sees a ghost and runs away. One of the questions asked why the girl ran away and the correct answer

[2]Since ASL is not a written language, we use standard glossing conventions (i.e., capital letters) to represent ASL signs.

can be translated to, "She is scared." This is a plausible inference based on the text, but not explicitly stated. The three foils are less plausible explanations for her behavior (e.g., "she escapes because she is late for school," "she likes to run," or "because a dog chases after her").

The first draft of the ASL-CMP was piloted with a group of seven deaf, linguistically-trained, ASL-English bilingual adults who were not part of original task development. Target accuracy for the adult participants was 85% or higher (i.e., at least six out of seven participants selected the correct answer) for each question. Three questions (one literal and two inferential) did not meet this criterion, suggesting they were either unclear or too difficult. The pilot participants were also asked to evaluate the quality of the ASL texts for clarity and grammaticality of signing production. In this process, one video was identified that was not appropriately edited (i.e., it had extended pauses and jump cuts). The problematic questions and text were then modified: the questions that did not yield high accuracy were replaced with new questions and one video was re-filmed for fluidity. We then re-tested the same participant group, at which point all questions were answered with 85% accuracy or higher. Finally, to confirm that questions were appropriately labeled as literal and inferential, all of the questions were evaluated by three teachers of deaf students with a master's degree in either deaf education or ASL who were unfamiliar with the test. There was 100% agreement in the classification of the questions as literal and inferential.

Test Procedures

Participants were recruited to take the ASL-CMP as part of a large-scale study involving the ASLAI assessment battery (Hoffmeister et al., 2015). All of the language tasks in the ASLAI, including ASL-CMP, were self-administered by participants on a computer. Prior to each of the sub-tests, participants watched an instructional video in ASL (see Henner, 2015; Hoffmeister et al., 2015). The instructions encouraged children to try their best when answering the questions on the test. The students then began a practice section that included one short ASL text and three questions (two literal questions and one inferential question). The students were given feedback on the practice trials. The ASL-CMP test questions immediately followed this practice. For each text, children first viewed the ASL text, and then saw a screen with the first question. Each question screen contained six different small videos consisting of the ASL text on the bottom left, the question on the top left, and the four different answer choices on the right in a two-by-two grid (**Figure 1**). The participants were instructed to watch the question, click on each of the four answer videos, and then select whichever video they thought best answered the question by clicking on the relevant video screen. To reduce working memory load, the question screen and four answer screens showed a carefully selected image as a frozen frame when the videos were not playing. Each frozen frame contained a salient feature of an ASL sign that could help the participant remember the contents of the video (Hoffmeister et al., 2015). For example, the question screen might contain a frozen frame of a wh-question, and the answer choices might contain an image of a critical sign. The ASL text was included on the screen to allow the participants to review the ASL text if needed. In addition to the frozen frames, there was no time limit and participants could re-watch the ASL text, the questions, and possible responses as many times as needed. The ability to review the entire text at will is an important feature that distinguishes the current task from a listening comprehension task, in which the information "disappears" after it is presented. In the current task, akin to a reading comprehension task, participants could refer back to parts or all of the story as they were determining their responses to the questions. All of the participants' responses were automatically scored and saved on a server. Scoring was dichotomous: participants received one point for a correct response and zero points for an incorrect response.

Participants

All of the participants in the present study were recruited through Boston University's Center for the Study of Communication & the Deaf (CSCD). All participants were deaf children attending schools for the deaf where ASL was the primary language of instruction. Participants varied with regard to when they were first exposed to ASL, as well as their ethnicity, hearing ability, IQ and age of entry to school. All participants that were able to complete the test were included in the sample.

For the psychometric evaluation of the ASL-CMP, only participants that had deaf parents were included ($n = 105$). These participants were chosen because of their homogeneity of age of exposure to ASL (i.e., all were exposed to ASL from birth). These participants had an age range of 8–18 years ($M = 11.2$ years).

The second set of analyses include an initial evaluation of the ASL-CMP among a wider group of deaf children. For these analyses, participants included the above sample of deaf children who have deaf parents ($n = 105$), plus an additional group of deaf children with hearing parents ($n = 251$) between the ages of 8–18 years ($M = 12.6$; see **Figure 2**). The sample was racially and ethnically diverse: of the 356 participants, there were 185 White, 49 Hispanic/Latino, 26 African American, 16 Micronesian, 19 Filipino, 15 Asian, 22 other, and 24 did not report. Information about age of entry into a school for the deaf was available for a subset of participants ($n = 202$). Of these, children with deaf parents ($n = 48$) entered school between birth (i.e., via early intervention) and 9-years-old ($M = 3.62$ years), and children with hearing parents ($n = 154$) entered school between 1 year and 18 years ($M = 7.12$ years).

RESULTS

Psychometric Analysis of the Normative Sample

All analyses were conducted with the statistical software R. Psychometric analysis focused on the consistency and reliability of the test questions. We first used item response theory (IRT) to determine discrimination (how well an item differentiates between high- and low-skilled participants) and the level of difficulty of each question in a standardized test (Yang and Kao, 2014). In contrast to classic test theory, IRT considers both individual participants and individual items which provides greater sensitivity about the items in relation to

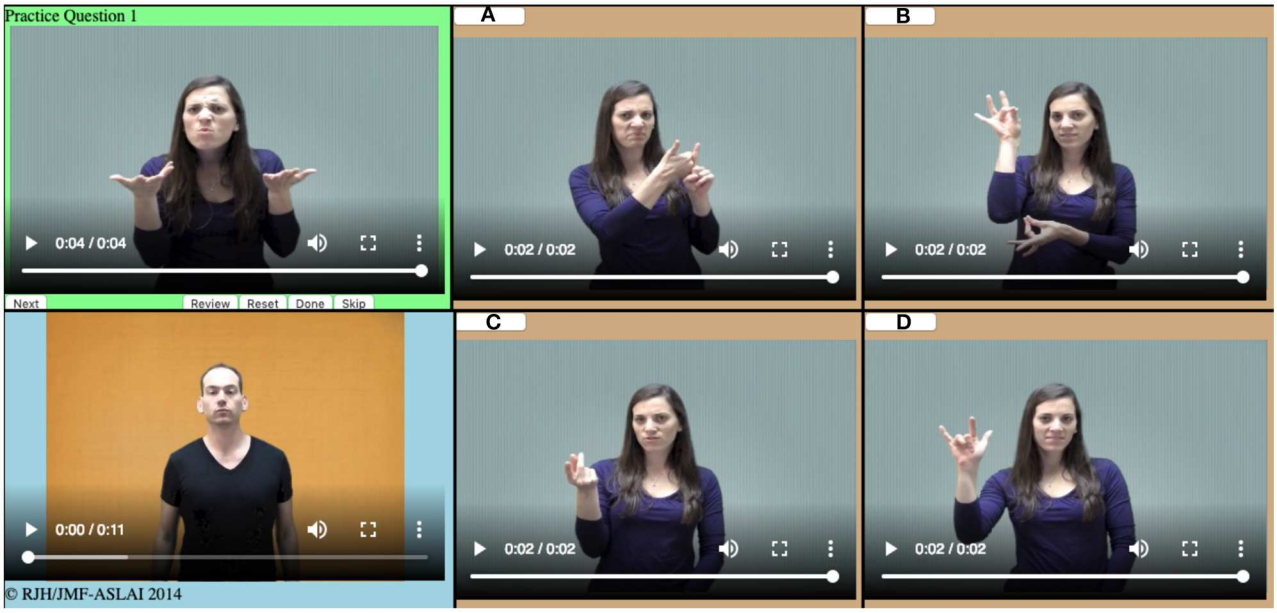


FIGURE 1 | Sample screenshot of one test question. The top left panel is the comprehension question, with a still image of the sign WHAT. The bottom left panel shows the ASL text, which participants will have already reviewed, but is available for review as students make their answer. The four panels on the right are each of the answer choices, with a button labeled with a letter that corresponds to their answer choice. **(A)** SELF; **(B)** GHOST; **(C)** DOG; **(D)** AIRPLANE. Written informed consent was obtained from the individuals in this image.

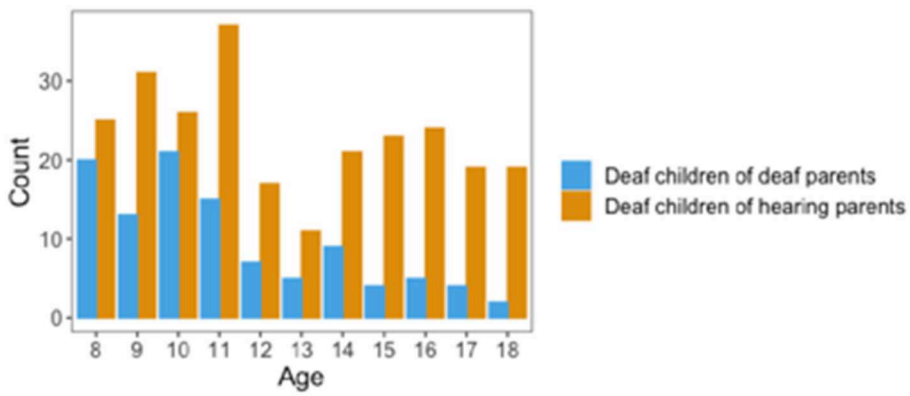


FIGURE 2 | Number of participants with deaf parents ($n = 105$) and hearing parents ($n = 251$) at each age.

individual abilities. Items with a discrimination value of 0.20 or above are considered acceptable, while values below the 0.20 threshold do not sufficiently discriminate between the skilled participant and the unskilled participant (Baker, 2001; Taib and Yusoff, 2014). The acceptable range of difficulty for each question is 0.20 and 0.80 (Baker, 2001). Values below 0.20 indicate that the question is too difficult, and above 0.80 indicate that the question is too easy. In general, questions that do not meet the criteria for both discrimination and difficulty should be revised or deleted (Ebel, 1954; Baker, 2001). As presented in **Table 1**, results from the IRT analysis indicated that all of the questions in the ASL-CMP test except for two literal questions had acceptable discrimination power and appropriate range of difficulty. These questions were removed.

In addition to item response and discrimination, we assessed internal consistency among questions on the task. We initially computed Cronbach's alpha of the ASL-CMP across all questions, which revealed an acceptable internal consistency of alpha 0.80. To determine consistency within each type of question, we also computed Cronbach's alpha separately for questions that assessed literal and inferential comprehension as two different, but related, constructs. We used a criterion of an alpha of 0.70 or greater, which indicates that the items are measuring the same construct (Santos, 1999; Tavakol and Dennick, 2011). We removed the two

TABLE 1 | Item difficulty and discrimination of the questions in ASL-Text Comprehension Task.

Question #	Type of question	Mean (*sd*)	Item difficulty	Item discriminability	A if deleted
1	Inferential	0.61 (0.49)	0.61	0.56	0.62
2	Literal	0.73 (0.44)	0.73	0.30	0.68
3	Literal	0.57 (0.50)	0.57	0.34	0.67
4	Literal	0.71 (0.45)	0.71	0.45	0.65
5	Inferential	0.47 (0.50)	0.47	0.30	0.71
6	**Literal**	**0.50 (0.50)**	**0.50**	**0.19**[a]	**0.70**
7	**Literal**	**0.22 (0.42)**	**0.22**	**0.04**[a]	**0.72**
8	Inferential	0.73 (0.44)	0.73	0.53	0.64
9	Literal	0.71 (0.45)	0.71	0.42	0.66
10	**Inferential**	**0.47 (0.50)**	**0.47**	**0.25**	**0.73**[b]
11	Literal	0.76 (0.43)	0.76	0.55	0.63
12	Literal	0.72 (0.45)	0.72	0.56	0.63
13	Literal	0.66 (0.48)	0.66	0.51	0.64
14	Inferential	0.71 (0.45)	0.71	0.61	0.61
15	Inferential	0.72 (0.45)	0.72	0.42	0.67

[a] *Denotes unacceptable discriminability value.*
[b] *Denotes change in alpha when removed.*
Bold row denotes omission in the final analysis.

literal questions in addition to one inferential question that did not meet the criteria (described above). The Cronbach's alpha for the final set of seven literal questions was 0.75 and for the five inferential questions was 0.72. Thus, the final version of the ASL-CMP, consisting of 12 questions, had acceptable levels of internal consistency (alpha = 0.85), discriminability, and difficulty.

Next, we evaluated concurrent validity by determining the relationship between the ASL-CMP and two other ASL vocabulary tests from the ASLAI, ASL Antonyms (Novogrodsky et al., 2014a) and ASL Synonyms (Novogrodsky et al., 2014b). Both of these tests used the same format as the ASL-CMP, and both tests asked students to select from a set of four different signs that best matches the given sign, synonymously or antonymously. We conducted Pearson correlation analyses for performance on the ASL-CMP and the two ASL vocabulary tasks in the ASLAI (Hoffmeister et al., 2015). Scores on both vocabulary tests were positively and significantly correlated with scores on ASL-CMP (antonyms: $r = 0.76$, $p < 0.001$; synonyms: $r = 0.74$, $p < 0.001$).

Finally, we used quantile regression to create growth charts of deaf children with deaf parents on the ASL-CMP (**Figure 3**). There was an increase in accuracy on the ASL-CMP with age, and an apparent ceiling effect at 12 years.

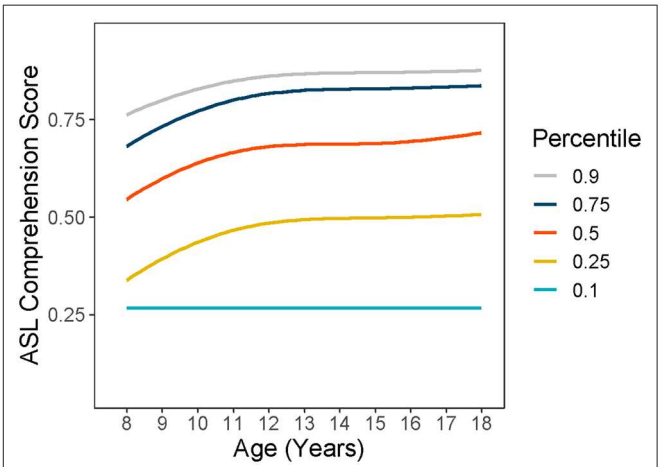


FIGURE 3 | The relationship between age and accuracy on the ASL-CMP for deaf children with deaf parents ($n = 105$). Lines indicate the 10th, 25th, 50th, 75th, and 90th percentiles, and were generated using the gcrq function in the R package quantregGrowth. The graph is not intended to be used to classify children's performance as within/above/below the normal range.

Evaluation of the ASL-CMP in Deaf Children With Deaf Parents and Deaf Children With Hearing Parents

Following the initial psychometric analysis, we assessed performance on the revised ASL-CMP on a larger group of participants, including children with deaf parents and those with hearing parents ($n = 356$). If the test is sensitive to differences in age and amount of language exposure, then we would expect to see higher accuracy in deaf children who have deaf parents vs. deaf children who have hearing parents, higher accuracy in children with hearing parents who entered school early vs. those who entered school late, and higher accuracy in older vs. younger children. We also predicted that accuracy would be higher for literal than inferential questions. **Figure 4** illustrates overall performance by age and participant group. Performance for deaf children with hearing parents shows greater change with age than for deaf children with deaf parents.

To analyze performance, we conducted a mixed-effects logistic regression using accuracy as the dependent variable (correct = 1, incorrect = 0; **Table 2**). In our initial model (Model 1), the fixed effects were participant group (deaf children who have deaf parents, deaf children who have hearing parents), age

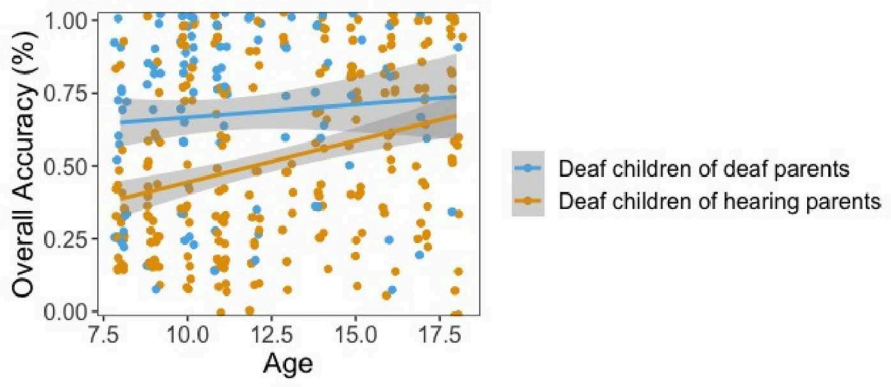


FIGURE 4 | The proportion of the questions answered correctly as a function of age and parental hearing status. Points were jittered slightly to avoid overlap.

(continuous), and type of question (literal, inferential). Random effects were included for story, participants, and items. Analysis revealed significant effects of participant group and question type: deaf children with deaf parents had higher accuracy than deaf children with hearing parents ($M_{\text{deafparents}} = 0.68$, $sd = 0.28$; $M_{\text{hearingparents}} = 0.52$, $sd = 0.30$), and literal questions were answered more accurately than inferential questions ($M_{\text{literal}} = 0.58$, $sd = 0.32$; $M_{\text{inferential}} = 0.55$, $sd = 0.33$). Age was also a positive and significant predictor of performance.[3] Children who have deaf parents appear to reach ceiling at about 12-years-old, which aligns with the target age range for this instrument (see **Figure 3**).

To investigate possible interaction effects, we ran a second regression model (Model 2) in which we added an interaction between parent hearing status and age, and an interaction between parent hearing status and question type. This analysis revealed no significant interaction effects. Further, Akaike's information criterion (AIC) revealed that adding the interaction terms to the model did not improve model fit: Model 2 (AIC = 4875.4) did not improve the model fit as compared to Model 1 [AIC = 4874.5; $\chi_2 = 3.07$, $p = 0.22$]. There were no significant differences in the developmental trajectories of ASL text comprehension in deaf children with deaf vs. hearing parents, and no interaction between question type and participant group.

For many deaf children, age of entry to school marks the time they are first immersed in ASL as a language of communication and instruction. For the subset of participants for whom we had information about age of entry to school ($n = 202$), we investigated the relationship between age of entry and performance on ASL-CMP by parental hearing status. We performed a mixed-effects logistic regression that was the same as the base model described above but also included an interaction between the participant group and age of school entry. We found a significant interaction between age of entry and parent hearing status ($\beta = 0.18$, $SE = 0.09$, $z = -1.98$, $p = 0.047$). *Post-hoc* analyses indicated that, as predicted, there was a significant, positive correlation between age of entry and performance for the deaf children of hearing parents ($n = 154$; $\beta = -0.10$, $SE = 0.03$, $z = 3.17$, $p = 0.002$), but not for the deaf children of deaf parents ($n = 48$; $\beta = 0.08$, $SE = 0.09$, $z = -0.93$, $p = 0.35$). This suggests that children who may have limited exposure to ASL at home show an increase in performance as a function of the amount of time they have spent in a school where ASL is the primary language of instruction.

DISCUSSION

In this study, we presented the development and validation of the ASL-CMP, a new ASL text comprehension task. We piloted the task on a group of native deaf signing adults, and then conducted a validation study with over 100 deaf children with deaf parents. This led to subsequent adjustments to ensure the task had high internal consistency and concurrent validity. We then analyzed performance in a group of more than 300 deaf children. Our findings suggest that the ASL-CMP is sensitive enough to detect patterns that are expected based on existing reports of deaf children's academic development, and is an appropriate measure of ASL text comprehension skills in children younger than 12 years of age. Below we discuss the primary findings, along with limitations and areas for further research.

As expected, deaf children of deaf parents, who were more likely to be exposed to ASL from birth, outperformed deaf children with hearing parents, who had more variable ages of exposure to ASL (Kuntze et al., 2014; Mitchiner, 2014; Henner et al., 2016; Hrastinski and Wilbur, 2016; Hall, 2017). Children with deaf parents are likely to be exposed to ASL from a wider range of individuals and in a broad range of contexts. This may lead to increased opportunities to develop inference-making skills, in which they need to extract information from ASL that is not explicitly stated. In contrast, deaf children with hearing parents may have had fewer opportunities to use ASL in these ways. Despite later exposure to ASL among the deaf children who have hearing parents, as a group they still showed evidence of development of higher-level comprehension skills in ASL

[3]A spearman correlation between age and ASL-CMP score was also significant ($r_s = 0.19$, $p < 0.01$).

TABLE 2 | Mixed effects logistic regression of factors predicting accuracy on the ASL-CMP.

Predictors	Model 1			Model 2		
	Odds ratios	CI	*p*	Odds ratios	CI	*p*
(Intercept)	0.58	0.26–1.28	0.179	1.26	0.31–5.06	0.742
Age	1.15	1.09–1.22	**<0.001**	1.07	0.95–1.20	0.275
Type of Question (Literal)	1.23	1.06–1.42	**0.006**	1.39	1.05–1.84	**0.021**
Parent hearing status (hearing)	0.33	0.22–0.48	**<0.001**	0.12	0.02–0.57	**0.008**
Age * Parent hearing status (hearing)				1.10	0.97–1.26	0.145
Parent hearing status (hearing) * Types of Question (Literal)				0.84	0.61–1.17	0.309
Random effects						
σ^2	3.29			3.29		
τ_{00}	$2.21_{StudentID}$			$2.19_{StudentID}$		
	0.09_{Story}			0.09_{Story}		
ICC	0.41			0.41		
N	3_{Story}			3_{Story}		
	$356_{StudentID}$			$356_{StudentID}$		
Observations	4,296			4,296		
Marginal R^2/Conditional R^2	0.063/0.448			0.065/0.448		
AIC	4874.5			4875.4		

Model one demonstrates original factors, while model two also includes two interaction terms.

over time. Further, it is important to note that not all deaf children with hearing parents performed below those with deaf parents. We speculate that many hearing parents who learn ASL likely provide a similarly rich environment for learning ASL as that provided by many deaf parents. This is additionally revealed in our analysis of age of school entry, which was a significant predictor of performance on the ASL-CMP for children with hearing parents. This provides promising evidence that exposure to ASL, even if it begins at school entry, can support students' acquisition of higher level ASL comprehension skills.

Our data revealed developmental patterns in deaf children's ASL text comprehension. Specifically, we found that older children had higher scores on the ASL-CMP than younger children. This pattern was particularly evident for children of deaf parents between the ages of 8 and 12 years and for children with hearing parents. This parallels findings from studies of literacy development in written language which show that text comprehension develops over a similar age range (Pettit and Cockriel, 1974; Cain et al., 2001; Silva and Cain, 2015). Many of the older children, particularly those with deaf parents, appeared to have already developed the ability to comprehend the ASL texts used in the task by 8-years-old. In future studies, it will be important to include deaf children who have deaf parents younger than 8 years, to better understand when comprehension skills are first developed among deaf children with early language exposure.

Lastly, there was a small but significant difference in accuracy on the ASL-CMP task based on the type of question, with higher overall accuracy on literal questions than on inferential questions. This is also compatible with previous studies showing literal comprehension is acquired prior to inferential comprehension (Pettit and Cockriel, 1974; McCormick, 1992; Basaraba et al., 2013). This suggests that literal comprehension may serve as a precursor to the ability to make inferences about information that is not explicitly stated in the text. Importantly, our findings are based on a small number of items, and the magnitude of the difference in performance between literal and inferential questions was small. We speculate that these differences would hold across a larger sample, but this must be borne out in future research.

Educational Application of the ASL-CMP

The ASL-CMP is a tool for measuring ASL text comprehension in deaf children ages 8 to 12, and will be useful for a range of purposes. First, the ASL-CMP provides a broad-strokes understanding of how ASL text comprehension develops over childhood. Since this task has been normed using a relatively large group of deaf children of deaf parents, it can be useful for clinicians and practitioners in determining whether a child has age-appropriate ASL text comprehension skills. Teachers may use this task to adapt their instruction to support the development of higher-level thinking skills, and to assess the quality and effectiveness of their ASL instructional approaches. Importantly, although the task has been normed, the ASL-CMP is not intended to diagnose deaf children with language delays. Instead, we recommend that this assessment be used to complement existing ASL assessments in that it measures more advanced language skills than are currently possible.

The ASL-CMP is a computer-based test that is automatically scored. No expertise or training is needed to administer the task. Scores at the individual and school level can be delivered rapidly. This is especially important for deaf children who

attend classrooms in which there are no professionals who are fluent in ASL (Hoffmeister, 1988; Hrastinski and Wilbur, 2016; Henner et al., 2018). Inquiries about using the ASL Text Comprehension can be directed to The Learning Center for the Deaf Center for Research and Training at CRT@tlcdeaf.org or to their website (www.ASLEducation.org).

Theoretical Implications of the ASL-CMP

While text comprehension was previously conceived of primarily as the comprehension of a written composition (e.g., a book, article, essay, poem), a broader conception of literacy makes it possible to see that higher-level thinking skills underlie the ability to consume compositions of a wide range of forms. Because these different forms of literacy may share a common underlying proficiency (Mackey and Jacobson, 2014), developing literacy skills through engagement with one type of text may generalize and benefit children's ability to comprehend additional text types (Mayer and Sims, 1994; Mayer, 2009), both within and across languages. It is important to consider how ASL text comprehension might then support children's development of other skills, both in ASL and other languages such as English. Specifically, one might expect those with strong ASL text comprehension skills to also develop strong English literacy skills (Bailes, 2001; Cummins, 2006; Kuntze et al., 2014; Hrastinski and Wilbur, 2016). With this novel way of assessing ASL text comprehension, we can begin to empirically test these questions.

Limitations and Areas for Further Research

The data here show a clear ceiling at around 12 years of age, but children as young as eight already achieve above-chance performance, so more data is needed to determine if the test is appropriate for children younger than eight. The sample size, although larger than many studies of deaf children, is relatively small compared to most normative samples. In a larger sample we may expect to see more robust interactions between participant group and age, as well as more fine-grained development of literal and inferential comprehension skills. Another limitation is that, because we did not have full demographic information on all of the participants in our sample, we were not able to tease out individual differences and how they impacted performance on the ASL-CMP. Due to the small number of questions, seven literal and five inferential, the ASL-CMP cannot reliably distinguish literal and inferential comprehension as two independent constructs, but rather it provides a measure of overall ASL comprehension. Finally, in the current analysis we looked at correct responses only. In future work we hope to carry out an analysis of incorrect responses to determine whether children are more likely to choose distractors of a specific type.

CONCLUSION

In summary, development of text comprehension skills in ASL is an important component of language and literacy development among deaf children. The newly developed ASL-CMP task is a first step in understanding how high-level text comprehension skills develop in children learning ASL. Our task is sensitive to ASL text comprehension in children from a wide range of backgrounds, and suggests that ASL text comprehension improves as children are exposed to ASL both at home and at school. The ASL-CMP makes it possible to evaluate children's ASL text comprehension skills, and identify children who may need support in developing such skills. Further, with a direct assessment of deaf children's text comprehension skills in ASL, we can begin to identify strategies to improve text comprehension skills in deaf children across languages.

DATA AVAILABILITY STATEMENT

The dataset is available on OSF along with the transcripts of the ASL texts in the ASL-CMP (https://osf.io/dwhba/?view_only=None).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by IRB approval was provided by the Boston University Charles River Campus Institutional Review Board. Consent for data collection was provided via Blanket Consent procedures. Parents were required to opt their children out of assessment. Information about the assessment was provided in both print and via ASL videos. Adults over the age of 18 who were included in assessment were also provided text or video consent documents. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

PR contributed to the study design, carried out data collection, analysis, and manuscript preparation. AL and NC contributed to data analysis and manuscript preparation. Lastly, RH encouraged PR to design the ASL-CMP and supervised the study design and data collection. All authors discussed the results and contributed to the final manuscript.

FUNDING

This research was provided by the National Center for Special Education Research, Grant Award No. R234A100176. Preparation of this manuscript was supported by the Ann S. Ferren Research Scholarship from the Wheelock College of Education and Human Development at Boston University.

ACKNOWLEDGMENTS

We thank the members of the Center for the Study of Communication in the Deaf at Boston University, specifically Sarah Fish, Rachel Benedict, and Jon Henner for assistance with the development of this test, and Anne Wienholz for her comments and feedback on earlier drafts of the manuscript. Also, we extend our thanks to the Center of Research and Training (CRT) at the Learning Center. Lastly, we thank all the Deaf schools and Deaf children who participated in this study.

REFERENCES

Abenchuchan, A. (2019, November). *About Alex Abenchuchan*. Retrieved from: https://www.dailymoth.com

Anderson, D., and Reilly, J. (2002). The MacArthur communicative development inventory: normative data for American Sign Language. *J. Deaf Stud. Deaf Educ.* 7, 83–106. doi: 10.1093/deafed/7.2.83

Bahan, B. (2006). "Face-to-face tradition in the American deaf community: dynamics of the teller, the tale, and the audience," in *Signing the Body Poetic: Essays on American Sign Language Literature*, eds H. Bauman, J. Nelson, and H. Rose (Berkeley, CA: University of California Press), 21–50.

Bailes, C. (2001). Integrative ASL/english language arts: bridging paths to literacy. *Sign. Lang. Stud.* 1, 147–174. doi: 10.1353/sls.2001.0002

Baker, F. (2001). *The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation*. Maryland, MD: University of Maryland College Park.

Barnes, M., Dennis, M., and Haefele-Kalvaitis, J. (1996). The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. *J. Exp. Child Psychol.* 61, 216–241. doi: 10.1006/jecp.1996.0015

Basaraba, D., Yovanoff, P., Alonzo, J., and Tindal, G. (2013). Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist? *Read. Writ.* 26, 349–379. doi: 10.1007/s11145-012-9372-9

Berke, M. (2012). *Deaf parents with deaf children and hearing parents with hearing children: a comparison of shared reading experiences* (Speech, Language, and Hearing Sciences Graduate theses and dissertations), 13. Available online at: https://scholar.colorado.edu/slhs_gradetds/13

Bishop, D., and Adams, C. (1992). Comprehension problems in children with specific language impairment: literal and inferential meaning. *J. Speech Hear. Res.* 35, 119–129. doi: 10.1044/jshr.3501.119

Blondel, M., Bauman, D., Nelson, J., and Rose, H. (2008). Signing the body poeti. *essays on American Sign Language literature. Appl. Linguist.* 29, 723–725. doi: 10.1093/applin/amno41

Bouchard, M., Ouellet, C., and Cohen, H. (2009). Speech development in prelingually deaf children with cochlear implants. *Lang. Linguist. Compass,* 3, 1–18. doi: 10.1111/j.1749-818X.2008.00079.x

Bowyer-Crane, C., and Snowling, M. (2005). Assessing children's inference generation: what do tests of reading comprehension measure? *J. Educ. Psychol.* 75, 189–201. doi: 10.1348/000709904X22674

Byrne, A. (2015). *American Sign Language (ASL) literacy and ASL literature: a critical appraisal* (Ph.D. dissertations), York University, Toronto, ON, Canada.

Cain, K., Barnes, M., Braynt, P., and Oakhill, J. (2001). Comprehension skill, inference making ability and their relation to knowledge. *Memory Cogn.* 29, 850–859. doi: 10.3758/BF03196414

Cain, K., and Oakhill, J. (1999). Inference making ability and its relation to comprehension failure. *Read. Writ.* 11, 489–503. doi: 10.1023/A:1008084120205

Cain, K., and Oakhill, J. (2007). "Reading comprehension difficulties: correlates, causes, and Consequences," in *Children's Comprehension Problems in Oral and Written Language: A Cognitive Perspective*, eds K. Cain and J. Oakhill (New York, NY: Guilford Press), 41–76.

Caselli, N., Lieberman, A., and Pyers, J. (2020). The ASL-CDI 2.0: an updated, normed adaptation of the Macarthur bates communicative development inventory for American Sign Language. *Behav. Res. Methods. doi: 10.3758/s13428-020-01376-6*

Christie, K., and Wilkins, D. (1997). A feast for the eyes: ASL literacy and ASL literature. *J. Deaf Stud. Deaf Educ.* 2, 57–59. doi: 10.1093/oxfordjournals.deafed.a014310

Ciullo, S., Ortiz, M., Otaiba, S., and Lane, K. (2016). Advanced reading comprehension expectations in secondary school? Considerations for students with emotional or behaviour disorders. *J. Disabil. Policy Stud.* 27, 54–64. doi: 10.1177/1044207315604365

Cromley, J., and Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *J. Educ. Psychol.* 99, 311–325. doi: 10.1037/0022-0663.99.2.311

Cummins, J. (2006). "The relationship between American Sign Language proficiency and English academic development: a review of the research," in *Paper Presented at the Conference Challenges, Opportunities, and Choices in Educating Minority Group Student* (Norway: Hamar University College).

DeLana, M., Genry, M., and Andrews, J. (2007). The efficacy of ASL/English bilingual education: considering public schools. *Am. Ann. Deaf* 152, 73–87. doi: 10.1353/aad.2007.0010

Dettman, S., Dowell, C., Richard Choo, J. D., Arnott, J. W., Abrahams, J. Y., et al. (2016). Long- term communication outcomes for children receiving cochlear implants younger than 12 months: a multicenter study. *Otol. Neurotol.* 37, e82–e95. doi: 10.1097/MAO.0000000000000915

Duffy, G. (2009). *Explaining Reading, 2nd Edn*. New York, NY: Guilford.

Duke, N., and Pearson, P. (2002). "Comprehension instruction in the primary grades," in *Comprehension Instruction: Research-Based Best Practices,* eds C. Collins Block and M. Pressley (New York, NY: Guilford Press), 247–258.

Ebel, R. (1954). Procedures for the analysis of classroom tests. *Educ. Psychol. Meas.* 14, 352–364. doi: 10.1177/001316445401400215

Enns, C., Haug, T., Herman, R., Hoffmeister, R., Mann, W., and McQuarrie, L. (2016). "Exploring signed language assessment tools in Europe and North America," in *Diversity in Deaf Education*, eds M. Marschark, V. Lampropoulou, and E. Skordilis (New York, NY: Oxford University Press), 171–218.

Enns, C., and Herman, R. (2011) Adapting the assessing British sign language development: receptive skills test into American Sign Language. *J. Deaf Stud. Deaf Educ.* 16, 362–374. doi: 10.1093/deafed/enr004

Ganek, H., Mcconkey Robbins, A., and Niparko, J. K. (2011). Language outcomes after cochlear implantation. *Otolaryngol. Clin. North Am.* 45, 173–185. doi: 10.1016/j.otc.2011.08.024

Goldin-Meadow, S., and Mayberry, R. (2001). How do profoundly deaf children learn to read? Learning disabilities research and practice. *Curr. Status Res. Direct.* 16, 221–228. doi: 10.1111/0938-8982.00022

Griffith, M., and Papacharissi, Z. (2009). Looking for you: an analysis of video blogs. *First Monday* 15. doi: 10.5210/fm.v15i1.2769

Hall, W. (2017). What you don't know can hurt you: the risk of language deprivation by impairing sign language development in deaf children. *Matern. Child Health J.* 21, 961–965. doi: 10.1007/s10995-017-2287-y

Hall, W., Smith, S., Sutter, E., DeWindt, L., and Dye, T. (2018). Considering parental hearing status as a social determinant of deaf population health: insights from experiences of the "dinner table syndrome" *PLoS ONE* 13:e0202169. doi: 10.1371/journal.pone.0202169

Hambleton, R., and Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Soc. Ind. Res.* 45, 153–171. doi: 10.1023/A:1006941729637

Haug, T. (2008). "Review of sign language assessment instruments," in *Sign Language Acquisition,* eds A. Baker and B. Woll (Amsterdam: John Benjamins), 51–86. doi: 10.1075/bct.14.04hau

Henner, J. (2015). *The relationship between American Sign Language vocabulary and the development of language-based reasoning skills in deaf children* (Doctoral dissertation). Boston, MA: Boston University; ProQuest Dissertations and Theses Global.

Henner, J., Caldwell-Harris, C.L., Novogrodsky, R., and Hoffmeister, R. (2016). American Sign Language syntax and analogical reasoning skills are influenced by early acquisition and age of entry to singing schools for the deaf. *Front. Psychol.* 7:1982. doi: 10.3389/fpsyg.2016.01982

Henner, J., Novogrodsky, R., Reis, J., and Hoffmeister, R. (2018). Recent issues in the use of signed language assessments for diagnosis of language disorders in signing deaf and hard of hearing children. *J. Deaf Stud. Deaf Educ.* 23, 307–316. doi: 10.1093/deafed/eny014

Hoffmeister, R. (1988). "Cognitive assessment of deaf preschoolers," in *Assessment of Developmentally Disabled Children,* eds T. Wachs and R. Sheehan (Boston, MA: Plenum Publishing Corporation), 109–126. doi: 10.1007/978-1-4757-9306-2_7

Hoffmeister, R. J. (2000). "A piece of the puzzle: ASL and reading comprehension in deaf Children," in *Language Acquisition By Eye,* eds C. Chamberlain, J. P. Morford, and R. I. Mayberry (Mahwah, NJ: Lawrence Erlbaum Associates), 143–163.

Hoffmeister, R., Henner, J., Fish, S., Benedict, R., Rosenburg, P., and Caldwell Harris, C. (2015). *The American Sign Language Assessment Instrument (ASLAI), v3.0*. Center for the study of communication and the deaf, School of Education, Boston University, MA.

Houghton Mifflin (2010). Houghton Mifflin reading: Comprehensive Screening Assessment. Orlando, FL: Houghton Mifflin.

Hrastinski, I., and Wilbur, R. (2016). Academic achievement of deaf and hard-of-hearing students in an ASL/English bilingual program. *J. Deaf Stud. Deaf Educ.* 21, 156–170. doi: 10.1093/deafed/env072

Humphries, T., Kushalnagar, P., Mathur, G., Napoli, D., Padden, C., Rathmann, C., et al. (2017). Discourses of prejudice in the professions: the case of sign languages. *J. Med. Ethics* 43, 1–5. doi: 10.1136/medethics-2015-103242

Jimenez, L., Roberts, K., Brugar, K., Meyer, C., and Waito, K. (2017). Moving our can(n) ons: Towards an appreciation of multimodal texts in classroom. *Read. Teacher* 71, 363–368. doi: 10.1002/trtr.1630

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Read. Res. Quart.* 19, 219–239. doi: 10.2307/747364

Jones, C. (2018, August 14). [Sign1News]. *Meet Candace.* Retrieved online from: https://sign1news.com

Kintsch, W. (1998). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163

Kral, A., Kronenberger, W. G., Pisoni, D. B., and O'Donoghue, G. M. (2016). Neurocognitive factors in sensory restoration of early deafness: a connectome model. *Lancet Neurol.* 15, 610–621. doi: 10.1016/S1474-4422(16)00034-X

Kuntze, M. (1998). Literacy and deaf children: the language question. *Top. Lang. Disord.* 18, 1–15. doi: 10.1097/00011363-199808000-00003

Kuntze, M. (2004). *Literacy Acquisition and Deaf Children: A Study of the Interaction between ASL and Written English*. Stanford University, CA: Social Science Premium Collection.

Kuntze, M., Golos, D., and Enns, C. (2014). Rethinking literacy: broadening opportunities for visual learners. *Sign. Lang. Stud.* 14, 203–224. doi: 10.1353/sls.2014.0002

Landi, N. (2010). An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Read. Writ.* 23, 701–717. doi: 10.1007/s11145-009-9180-z

Leslie, L., and Caldwell, J. (2011). *Qualitative Reading Inventory, 5th Edn*. Boston, MA: Pearson/Allyn and Bacon.

Leu, D., Forzani, E., Rhoads, C., Maykel, C., Kennedy, C., and Timbrell, N. (2015). The new literacies of online research and comprehension: rethinking the reading achievement gap. *Read. Res. Quart.* 50, 37–59. doi: 10.1002/rrq.85

Mackey, T., and Jacobson, T. (2014). *Metaliteracy: Reinventing Inform Ation Literacy to Power Learners.* Chicago, IL: ALA Neal-Schuman, an imprint of the American Library Association.

Maller, S., Singleton, J., Supalla, S., and Wix, T. (1999). The development and psychometric properties of the American Sign Language proficiency assessment (ASL-PA). *J. Deaf Stud. Deaf Educ.* 4, 249–269. doi: 10.1093/deafed/4.4.249

Malzkuhn, M., and Bottoms, A. (2017, May 3). *Hearing Knows Best*. Retrieved online from: https://www.youtube.com/watch?v=MoxVdw6T0LA

Mann, W., Roy, P., and Morgan, G. (2016). Adaption of a vocabulary test from British sign language to American Sign Language. *Lang. Test.* 33, 3–22. doi: 10.1177/0265532215575627

Manrique, M., Cervera-Paz, F. J., Huarte, A., and Molina, M. (2004). Advantages of cochlear implantation in prelingual deaf children before 2 years of age when compared with later implantation. *Laryngoscope* 114, 1462–1469. doi: 10.1097/00005537-200408000-00027

Martins, P. (2016). Engaging the d/Deaf audience in museums: a case study at the calouste gulbenkian museum. *J. Museum Educ.* 41, 202–209. doi: 10.1080/10598650.2016.1193316

Mayberry, R. I., and Kluender, R. (2018). Rethinking the critical period for language: new insights into an old question from American Sign Language. *Biling. Lang. Cogn.* 21, 886–905. doi: 10.1017/S1366728917000724

Mayer, R. (2009). *Multimedia Learning, 2nd Edn*. Cambridge University Press. doi: 10.1017/CBO9780511811678

Mayer, R., and Sims, V. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *J. Educ. Psychol.* 86, 389–401. doi: 10.1037/0022-0663.86.3.389

McCormick, S. (1992). Disabled readers' erroneous responses to inferential comprehension questions. *Read. Res. Quart.* 27, 54–77. doi: 10.2307/747833

McQuarrie, A., and Spady, S. (2012). "American Sign Language phonological awareness: test development and design," eds *Proceedings of the 10th annual Hawaii International Conference on Education* (Honolulu), 1–17.

Miller, S., and Smith, D. (1985). Differences in literal and inferential comprehension after reading orally and silently. *J. Educ. Psychol.* 77, 341–348. doi: 10.1037/0022-0663.77.3.341

Mitchiner, J. (2014). Deaf parents of cochlear-implanted children: beliefs on bimodal bilingualism. *J. Deaf Stud. Deaf Educ.* 20, 51–66. doi: 10.1093/deafed/enu028

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *Modern Lang. J.* 87, 261–276. doi: 10.1111/1540-4781.00189

Niparko, J. K., Tobey, E. A., Thal, D. J., Eisenberg, L. S., Wang, N.-Y., Quittner, A. L., et al. (2010). Spoken language development in children following cochlear implantation. *J. Am. Med. Assoc.* 303, 1498–1506. doi: 10.1001/jama.2010.451

Nippold, M., and Scott, C. (Eds.). (2010). *Expository Discourse in Children, Adolescents, and Young Adults: Development and Disorders*. New York, NY: Psychology Press/Taylor and Francis.

Novogrodsky, R., Caldwell-Harris, C., Fish, S., and Hoffmeister, R. (2014a). The development of antonym knowledge in American Sign Language (ASL) and its relationship to reading comprehension in english. *Lang. Learn.* 64, 749–770. doi: 10.1111/lang.12078

Novogrodsky, R., Fish, S., and Hoffmeister, R. (2014b). The acquisition of synonyms in American Sign Language (ASL): toward a further understanding of the components of ASL vocabulary knowledge. *Sign. Lang. Stud.* 14, 225–249. doi: 10.1353/sls.2014.0003

Paludneviciene, R., Hauser, P. C., Daggett, D., and Kurz, K. B. (2012). "Issues and trends in sign language assessment," in *Measuring Literacy and its Neurocognitive Predictors Among Deaf Individuals: an Assessment Toolkit*, eds D. Morere, and T. Allen (New York, NY: Springer), 191–207.

Perfetti, C.A., and Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Sci. Stud. Read.* 18, 22–37. doi: 10.1080/10888438.2013.827687

Perfetti, C. A., Landi, N., and Oakhill, J. (2005). "The acquisition of reading comprehension skill," in *The Science of Reading: A Handbook,* eds. M. Snowling, and C. Hulme (Oxford: Blackwell), 227–253. doi: 10.1002/9780470757642.ch13

Peterson, N. R., Pisoni, D. B., and Miyamoto, R. T. (2010). Cochlear implants and spoken language processing abilities: review and assessment of the literature. *Restor. Neurol. Neurosci.* 28, 237–250. doi: 10.3233/RNN-2010-0535

Pettit, N., and Cockriel, I. (1974). A factor study of the literal reading comprehension test and the inferential reading comprehension test. *J. Literacy Res.* 6, 63–75. doi: 10.1080/10862967409547078

Prinz, P., Strong, M., and Kuntze, M. (1994). *The Test of ASL*. San Francisco, CA: Unpublished Test, San Francisco State University.

Quinto-Pozos, D., and Hou, L. (2015). *ASL Assessment Toolkits: ASL and Nonlinguistic Perspective Taking Comprehension Tests*. Retrieved online from: http://www.signlang-assessment.info/index.php/asl-perspective-taking-comprehension-test.html

Santos, A. (1999) Cronbach's alpha: a tool for assessing the reliability of scales. *J. Exten.* 37, 1–4.

Shanahan, T. (2005). *The National Reading Panel Report: Practical Advice for Teachers*. Learning Point Associates/North Central Regional Educational Laboratory (NCREL).

Shema, H., Bar-Ilan, J., and Thelwall, M. (2012). Research blogs and the discussion of scholarly information. *PLoS ONE* 7:e35869. doi: 10.1371/journal.pone.0035869

Silva, M., and Cain, K. (2015). The relations between lower and higher-level comprehension skills and their role in prediction of early reading comprehension. *J. Educa. Psychol.* 107, 321–331. doi: 10.1037/a0037769

Simms, L., Baker, S., and Clark, M. (2013). The standardized visual communication and sign language checklist for signing children. *Sign. Lang. Stud.* 14, 101–124. doi: 10.1353/sls.2013.0029

Snoddon, K. (2010). Technology as a learning tool for ASL literacy. *SiIgn. Lang. Stud.* 10, 197–213. doi: 10.1353/sls.0.0039

Strong, M., and Prinz, P. M. (1997). A study of the relationship between American Sign Language and english literacy. *J. Deaf Stud. Deaf Educ.* 2, 37–46. doi: 10.1093/oxfordjournals.deafed.a014308

Szagun, G., and Schramm, S. A. (2016). Sources of variability in language development of children with cochlear implants: Age at implantation, parental language, and early features of children's language

construction. *J. Child Lang.* 43, 505–536. doi: 10.1017/S0305000915000641

Taib, F., and Yusoff, M. (2014). Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. *J. Taibah Univ. Med. Sci.* 9, 110–114. doi: 10.1016/j.jtumed.2013.12.002

Tavakol, M., and Dennick, R. (2011). Making sense of cronbach's alpha. *Int. J. Med. Educ.* 2, 53–55. doi: 10.5116/ijme.4dfb.8dfd

The New London Group (1996). A pedagogy of multiliteracies: designing social futures. *Harv. Educ. Rev.* 66, 60–93. doi: 10.17763/haer.66.1.17370n67v22j160u

Van de Vijver, F., and Poortinga, Y. H. (2005). "Conceptual and methodological issues in adapting tests," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* eds R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum Associates), 39–64.

Van den Broek, P., and Epsin, C. (2012). Connecting cognitive theory and assessment: measuring individual differences in reading comprehension. *School Psychol. Rev.* 41, 315–325. doi: 10.1080/02796015.2012.12087512

Wall, L. (2014). *From the hands into the eyes: an analysis of children's American Sign Language story comprehension* (thesis). University of Toronto, Toronto, ON, Canada.

Yang, F., and Kao, S. (2014). Item response theory for measurement validity. *Shanghai Arch. Psychiatry* 26, 171–177. doi: 10.3969/j.issn.1002-0829.2014.03.010

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.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 01 June 2020
doi: 10.3389/fpsyg.2020.01116

Converging Development of English as Foreign Language Listening and Reading Comprehension Skills in German Upper Secondary Schools

Christian Spoden[1], Jens Fleischer[2] and Michael Leucht[3]*

[1] German Institute for Adult Education – Leibniz Centre for Lifelong Learning, Bonn, Germany, [2] Department of Instructional Psychology, University of Duisburg-Essen, Essen, Germany, [3] IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Receptive skills in English as a second language are important for students on the verge of entering higher education as this student group (aged 17–19) is expected to apply English regularly in their later life. Previous research in this age group in Germany already implied an increasing overlap between both skills in this age group, although robustness of this effect across student groups with different learning experiences was not tested. We used language assessment data collected from upper secondary schools (i.e., from 17 to 19-year-old students) in Germany to compare correlations at the beginning and the end of upper secondary education in groups of students from (1) language-related versus non-language-related study profiles and (2) from students with frequent versus less frequent self-reported English-language out-of-school learning activities. In all of these groups, correlations were increasing, indicating converging skills in upper secondary education. The results are discussed in terms of implications for current theories of language research.

Keywords: listening comprehension, reading comprehension, receptive language skills, upper secondary education, cross-lagged panel design

OPEN ACCESS

Edited by:
Vahid Aryadoust,
National Institute of Education,
Nanyang Technological University,
Singapore

Reviewed by:
Tugba Elif Toprak,
Izmir Democracy University, Turkey
Shangchao Min,
Zhejiang University, China

****Correspondence:***
Christian Spoden
spoden@die-bonn.de

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

Received: *18 November 2019*
Accepted: *30 April 2020*
Published: *01 June 2020*

Citation:
Spoden C, Fleischer J and Leucht M (2020) Converging Development of English as Foreign Language Listening and Reading Comprehension Skills in German Upper Secondary Schools. Front. Psychol. 11:1116. doi: 10.3389/fpsyg.2020.01116

INTRODUCTION

In a globalized world, there are few doubts about the importance of English-language skills for students who accomplish upper secondary and possibly higher education. English-language skills are crucial in higher education as a large amount of documented research is only available in English. There is also an increasing number of degree programs, lectures, and training courses held in English outside of English-speaking countries. For example, the number of "English-only" degree programs rose from 391 in 2005 to 1017 in 2015 in Germany (Balzer, 2015). English conversational skills are equally important for students who enter a profession directly after upper secondary school, as jobs addressing the more highly trained most often require advanced to excellent (business) skills in English as lingua franca and common corporate language for internal communication (e.g., Nickerson, 2005; Swift and Wallace, 2011). Advanced skills in English as a second language (ESL) and especially receptive skills, namely listening and reading comprehension, are thus counted among the key competencies necessary for educational and vocational achievement. Age groups on the verge of entering higher education have still been widely overlooked in language assessment research. The effects of different learning opportunities might

become noticeable in this age group as schools offer different "study profiles" in upper secondary education (e.g., Leucht et al., 2015), which imply a different course selection in preparation for the final secondary-school examinations (A levels). Students in this age group also engage differently in English-language-related activities outside school. Thus, this study investigated the impact of various in- and out-of-school learning opportunities that arise from different study profiles and different English-language-related activities (i.e., primarily media consumption) on the associations between English-language listening and reading comprehension skills at the beginning and the end of upper secondary schooling.

THEORETICAL BACKGROUND

Listening and Reading Comprehension as Substantially Associated Skills in L1 and L2 Language Acquisition

Most theories of comprehension recognize differences in the modality of the input, although the importance of such differences in the comprehension process is not accentuated and little information is given on how the comprehension is differently affected by input modalities. McNamara and Magliano (2009) reviewed seven comprehension theories and stated that none of these theories had described differences between modalities in a detailed manner or explained the integration of modalities. Aryadoust (2017) recently proposed an integrated theory of comprehension where modality-specific processes were incorporated as "pre-comprehension." The theory acknowledges that in these pre-comprehension processes of word recognition (processes of segmentation and selection, recognition and lexical access) listeners are disadvantaged as the auditory input fades away in few seconds while readers can always regress to previous parts of the text. The theory also assumes that later process of comprehension, which involve generation of mental meaning and evolving mental imagery (selection) as well as the following stage of integration, where inferencing and elaboration are activated and meaning is assigned to the textual representations, are taken to be common between listening and reading comprehension. Although processes of perception and recognition were not explicitly integrated in earlier theories of comprehension, all comprehension theories still notice that there is a substantial level of overlap (i.e., a correlation) between listening and reading comprehension; this was verified in several psychometric analyses (e.g., Jeon and Yamashita, 2014; Wolf et al., 2019).

The development of listening and reading comprehension in a second language (L2) differs from the development in the mother tongue (L1), where auditory comprehension skills are naturally acquired in early childhood and the acquisition of reading skills usually originates in elementary school. Still, theories like the simple view of reading explain (early) reading comprehension in L1 and L2 as a product of decoding and listening comprehension; a larger number of studies provided evidence for the theory in L2 learning (e.g., Yagoub-Zadeh et al., 2012; Gottardo et al., 2018). On the other hand, L2 listening comprehension, especially in early stages of L2 learning, involves different tasks for the language learner and requires high levels of attention and working memory, as the speed and pronunciation of an authentically spoken foreign language is hardly controllable by the listener (e.g., Vandergrift, 2007; Vandergrift and Baker, 2015). In contrast, the training of early L2 reading comprehension is facilitated by the persisting availability of the text. At later stages of language acquisition, listening comprehension potentially benefits from overlapping subskills (e.g., an enlarged range of vocabulary and grammar knowledge). This might explain higher gains in listening comprehension compared to reading comprehension at the end of secondary schooling (Leucht et al., 2015). However, out-of-school learning opportunities may also contribute to these differences. It has also been shown for some time that out-of-school training of reading fosters reading comprehension (e.g., Watkins and Edwards, 1992; Pfost et al., 2013). Listening comprehension in later adolescence notably benefits from implicit training through out-of-school English-language media consumption in closed-captioned television (Huang and Eskey, 1999), gaming (Sylvén and Sundqvist, 2012), or new educational technology for leisure activities (Liu et al., 2017).

The Development of Listening and Reading Comprehension in German Upper Secondary Schools

Leucht et al. (2015) analyzed the development of both L2 (English) competencies in upper secondary schools in Germany by means of language assessment data. They found that the development of reading comprehension was slowing down (d = 0.15 to d = 0.38, depending on different study profiles) but high learning gains were identified in listening comprehension (d = 1.04 to d = 1.35). The results were moderated by school profiles, with stronger learning gains found in schools with a language-related study profile, which involves additional instruction courses in (foreign) languages. In the studied German federal state of Schleswig-Holstein, schools can offer up to five study profiles (languages, aesthetics with music and arts, sport, science, and social sciences). In the language-related profile, L1 and L2 are usually complemented by a third language (second foreign language) taught with 4 h of instruction a week and two additional foreign languages taught with 3 h a week. In contrast, in the science study profile, as an example for a non-language-related profile, three sciences are taught with increased expenditure of time, but no other languages are taught in non-language-related study profiles besides German classes and English classes. Correlations between listening and reading comprehension in the study by Leucht et al. (2015) increased over a two-year period up to the end of upper secondary education. Investigating this finding in more detail was beyond the scope of their analyses and, thus, Leucht et al. did not study how these skills developed in different student groups and whether the convergence of both skills was a consistent pattern in all groups. Yet, the finding is in need of some further investigation.

RESEARCH QUESTIONS

Considering previous results of Leucht et al. (2015) on the converging development of listening and reading comprehension in this age group, the robustness of this effect across students with different in- and out-of-school learning opportunities was investigated in this study. Varying associations between learning-process related variables (e.g., cognitive ability) or other individual characteristics (e.g., gender, cultural capital) and receptive skills depending on different structural assumptions on language skills have already been demonstrated in younger age groups (Hartig and Höhler, 2008). The structure and the associations between both skills were also analyzed within and between classrooms (Höhler et al., 2010) and across different age groups (Tilstra et al., 2009), as examples for different learning settings. Thus far, this research did not involve the relevant group of students from upper secondary schools, although the potential learning trajectories actually grows especially in this age group with study profiles, English language media exposure or other types of possible in- and out-of-school learning opportunities. It is rather plausible to assume that different learning opportunities not only affect achievement in ESL learning (see above) but also the associations between language skills. Thus, the robustness across different learning-related variables becomes important in studying the development of the correlation of listening and reading skills in these different groups.

Information on study profiles was used to analyze effects of different curricular activities and subjective student ratings on activities related to the English language (e.g., English-language media consumption, engagement in English-language conversation on holidays, etc.) in self-reports were used as a proxy variable for the various out-of-school learning opportunities of the students. Based on this information, the study addressed the following two research questions:

1. Do L2 listening and reading comprehension skills in upper secondary education converge in groups of students from non-language-related and language-related study profiles (RQ1)?
2. Do L2 listening and reading comprehension skills in upper secondary education converge in groups of students with and without language-related out-of-school learning experiences (RQ2)?

Quantitative analyses were carried out, making use of the data from Leucht et al. (2015) from a language assessment administered in the German federal state of Schleswig-Holstein, to study these research questions.

METHOD

Sample

To investigate the first research questions, longitudinal data from Grades 11 (T1) and 13 (T2) provided by N = 1171 students (n = 228 with a language-related study profile) nested in 68 classes from 17 schools in the German federal state of Schleswig-Holstein were analyzed. To investigate the second research question, a subsample (n = 550; including 20 students with missings on the scale) from the original sample that had responded to a supplementary survey involving questions on out-of-school learning opportunities (see below) was analyzed.

The interval between the measurement points was 27 months; the measurement points marked the beginning and the end of upper secondary education in Germany ("Oberstufe"). Population weights were computed for each student, so that the real number of students in this grade in the state of Schleswig-Holstein could be approximated in both analyses.

Instruments

Listening comprehension and reading comprehension skills were measured by means of standardized test instruments. The assessment framework for English listening comprehension and reading comprehension was based on the German Educational Standards (see Rupp et al., 2008), which were themselves based on the Common European Framework of Reference for Languages (Council of Europe, 2001). The items were designed by trained item developers, piloted and optimized by item elimination before the assessment took place. The listening comprehension test comprised 118 items at T1 (rel_{PV} = 0.75) and 32 items at T2 (rel_{PV} = 0.82). The reading comprehension test comprised 133 items at T1 (rel_{PV} = 0.77) and 42 items at T2 (rel_{PV} = 0.82), administered in complex test designs. All items administered at T2 had previously been used at T1 in order to facilitate the linking of both measurement points on the same scale.

An additional questionnaire was administered to assess out-of-school learning activities as part of a supplementary survey. The questionnaire comprised eight items concerned with either listening to or reading English content in leisure time. Sample items are "I listen to audio books in English." and "I read books, newspapers, or magazines in English." (translation by authors). The students responded on a five-point scale with response options ranging from *never* to *more than five times*. The items on English-language-related activities obtained an internal consistency of α = 0.84, indicating that these activities are part of consistent behavioral patterns.

Data Analysis

In the first step of analysis, the language assessment data were scaled according to a Rasch model (e.g., Spoden and Fleischer, 2019), and plausible values were estimated for both measurement occasions. Plausible values (PVs; e.g., Wu, 2005) estimation is a statistical technique to approximate population characteristics in assessments by random draws from an empirically derived ability distribution. The Rasch model involved a latent regression with several factorized covariates incorporated (e.g., cognitive ability, HISEI, and gender). Item parameters were constrained to the same values for common items at T1 and T2 in order to establish a common metric. The PVs were also rescaled to the metric of the German Educational Standards (M = 500, SD = 100). The ConQuest software package (Adams et al., 2015), Version 4.0, was used to estimate the latent regression Rasch models.

To investigate the first research question, correlations of both skills at T1 and T2 (PVs) were computed in the

two groups of students from language- and non-language-related study profiles. To investigate the second research question, the continuous variable of out-of-school learning experiences was dichotomized by means of a median split. Correlations were computed for students with below and above median scores on this variable. It was also checked in each of these groups separately whether constraining the correlations of listening and reading comprehension at T1 and at T2 to be equal deteriorated model fit. Afterward, it was checked whether constraining the T1 and T2 correlations in both groups of analysis to equal parameters yielded a more parsimonious model fit. The PV correlations were estimated in the Mplus 8.0 software package (Muthén and Muthén, 1998–2017) with sampling weights incorporated into the model estimation and results pooled according to the rules by Rubin (1987).

RESULTS

Descriptive Measures on Learning Gains

The descriptive measures on gains in English-language receptive skills over the two-year period in the five groups are given in **Figure 1**. Gains were stronger for listening comprehension compared to reading comprehension, in language-related study profiles compared to non-language related study profiles and with students with higher levels of self-reported out-of-school learning experiences compared to students with lower levels.

Results for Research Question 1

The results on the convergence of both competencies in language- and non-language-related study profiles over the course of about two years are given in **Table 1** (upper part). These results revealed in both groups that constraining the correlations between listening and reading comprehension at T1 and T2 to equal parameter values did not provide a more parsimonious model. Comparing an unconstrained multiple-group model (AIC = 51,623.16, BIC = 51,770.06, BIC_{adj} = 51,677.94) with a constrained model, where correlations between listening and reading comprehension were constrained to be equal across both groups of students from language- and non-language-related study profiles (AIC = 51,625.16, BIC = 51,761.93, BIC_{adj} = 51,676.17; $\chi^2(2) = 6.00$, $p < 0.05$), gave mixed results. Following AIC and the likelihood ratio, equally growing correlations of different size in the two groups of students from language-related study profiles from 0.52 at T1 to 0.69 at T2 and the group of non-language-related study profiles from 0.56 at T1 to 0.77 at T2 were to be assumed.

Results for Research Question 2

The results on correlations of both competencies in language- and non-language-related study profiles over the course of about two years are given in **Table 1** (lower part). The results supported the (unconstrained) model with varying correlations at T1 and T2 in both groups. Comparing models with and without constrained correlations between listening and reading comprehension to equal parameter values across both studied groups favored the constrained model (AIC = 23,587.90, BIC = 23,703.25, BIC_{adj} = 23,617.54; vs. AIC = 23,590.61, BIC = 23,714.52, BIC_{adj} = 23,622.47; $\chi^2(2) = 1.27$, $p = 0.47$). Lower correlations of the PVs were again found at T1 with 0.55 compared to T2 with 0.74, giving evidence that listening and reading comprehension skills converge in upper secondary education in Germany in all studied groups.

DISCUSSION

In this study, the effects of in- and out-of-school learning opportunities on the associations between both skills in upper secondary education were analyzed from English-language assessment data, which offers a valid and reliable data source to study the interplay of these skills. The results revealed converging receptive skills from the beginning to the end of secondary education in Germany. The finding was robust across different student groups (language-related versus non-language-related study profiles, and students with different levels of self-reported English-language-related extracurricular activities), which were analyzed in order to test the influence of different in- and out-of-school learning activities.

The finding of a growing overlap of listening and reading scores extends results by Tilstra et al. (2009), who examined this effect up to ninth grade (i.e., nearly the end of lower secondary education) as part of a study on the simple view of reading. Substantial differences in the learning gains between both skills were also noticeable in this study. In a contemporary theory of text comprehension Aryadoust (2017) described that listeners are disadvantaged in low-level processes of (word) recognition compared to readers due to the transitory auditory input. Large gains in listening comprehension over two years of upper secondary education, low stability of the proficiency scores (results not presented here; see Leucht et al., 2015) and converging skills illustrate a different level of competence emerging at the end of upper secondary education. The results suggest that modality specificity becomes a less important factor to affect comprehension test scores at the end of secondary education in Germany. In line with the theory of Aryadoust (2017), summarizing these findings may also indicate that difficulties with perceptually earlier, modality-specific processes in ESL learning were simply overcome by a larger group of students over the course of secondary education.

More research is obviously needed to verify this assumption and to trace the effects back to underlying subskills, in particular with experienced ESL learners in higher tracks of education. Recently, Gottardo et al. (2018) "unpacked" listening comprehension by examining the contribution of subcomponents of the skill (vocabulary, morphological awareness, syntax knowledge) to reading comprehension. A growing number of studies also

FIGURE 1 | Listening comprehension and reading comprehension for students in the language-related and the non-language-related study profiles, for students with above and below median English language-related extracurricular activities, and for all students at two measurement occasions (T1 = Measurement Occasion 1, T2 = Measurement Occasion 2).

TABLE 1 | Model fit in group-specific analyses with unconstrained correlations and correlations at T1 and T2 constrained to be equal.

	Unconstrained correlations			Constrained correlations			$\chi^2(1)$
	AIC	BIC	BIC_{adj}	AIC	BIC	BIC_{adj}	
Language-related study profiles	9761.73	9809.74	9765.37	9769.15	9813.73	9772.53	9.43**
Non-language-related study profiles	40, 719.00	40, 786.89	40, 742.43	40, 797.18	40, 860.22	40, 818.93	80.18***
Below median out-of-school Learning experiences	10, 735.50	10, 784.69	10, 740.31	10, 752.33	10, 798.01	10, 756.80	18.83***
Above median out-of-school Learning experiences	12, 120.89	12, 171.88	12, 127.49	12, 132.12	12, 179.46	12, 138.24	13.22***

$^{**}p < 0.01$. $^{***}p < 0.001$.

differentiated subcomponents underlying listening and reading comprehension skills (e.g., Song, 2008; Goh and Aryadoust, 2014) by means of psychometric approaches. Research on the development of receptive language skills clearly benefits from a closer look on common subcomponents.

Still, increasing correlations between receptive skills shifts the focus of research more on general text comprehension skills. Previous research demonstrated that different conclusions on the test scores need to be drawn for a general text comprehension dimension compared to modality-specific scores, as text comprehension is, for example, conceptually closer to general cognitive abilities compared to modality-specific processing (Hartig and Höhler, 2008). Even reversed effects of covariates like gender occur when modality-specific processes are partialed out (Hartig and Höhler, 2008). The results presented

here revealed a consistent effect across different language-learning groups (with marginally different correlation levels depending on study profiles), which simplifies the interpretation toward a general development in upper secondary education in Germany. Thus, in this age group emphasis may be placed on didactic and educational settings in the future that focus on fostering general comprehension skills instead on modality-specific aspects.

Obviously, the sampling of students from upper secondary education still needs to be considered when interpreting the results. This educated group of students certainly owns specific individual skills such as high (working) memory capacity, which is known to be related to receptive language skills (e.g., Vandergrift and Baker, 2015), but assumedly also differs in terms of individual modality preferences in text comprehension (e.g., Kürschner et al., 2005). Thus, it seems reasonable to additionally study the development of associations between both skills in samples from less institutionalized learning, such as vocational education, adult education, or advanced business English training. It should also be noted that several major transformations have become obvious in terms of media consumption in recent years in Germany. A growing number of online media service providers rely predominantly on younger users and make English-language media with entertainment and academic content more popular in this age group. The effects of these changes in English-language-related activities on the development and the associations of listening and reading comprehension might become fully apparent only over a longer period, studied with multiple cohorts.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available. Contact information to request access to the dataset is: IPN – Leibniz Institute for Science and Mathematics Education.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

CS performed substantial contribution to the conception of the study and the interpretation of the results, conducted the statistical analyses, drafted the manuscript, and approved the submitted version. JF performed substantial contribution to the conception of the study and the interpretation of the results, reviewed the manuscript critically for important intellectual content, and approved the submitted version. ML performed substantial contribution to the conception of the study and the interpretation of the results, reviewed the manuscript critically for important intellectual content, and approved the submitted version.

REFERENCES

Adams, R. J., Wu, M. L., and Wilson, M. R. (2015). *ACER ConQuest: Generalised Item Response Modelling Software Version 4*. Camberwell: Australian Council for Educational Research.

Aryadoust, V. (2017). An integrated cognitive theory of comprehension. *Int. J. Listen.* 33, 71–100. doi: 10.1080/10904018.2017.1397519

Balzer, I. (2015). *Wissenschaft? Science! "English Only" an Deutschen Hochschulen [Science? "Science"! English only at German universities]*. Berlin: Tageszeitung.

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

Goh, C. C. M., and Aryadoust, V. (2014). Examining the notion of listening sub-skill divisibility and its implications for second language listening. *Int. J. Listen.* 29, 109–133. doi: 10.1080/10904018.2014.936119

Gottardo, A., Mirza, A., Koh, P. W., Ferreira, A., and Javier, C. (2018). Unpacking listening comprehension: the role of vocabulary, morphological awareness, and syntactic knowledge in reading comprehension. *Read. Writ. Interdiscip. J.* 31, 1741–1764. doi: 10.1007/s11145-017-9736-2

Hartig, J., and Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *J. Psychol.* 216, 89–101.

Höhler, J., Hartig, J., and Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychol. Test Assess. Model.* 52, 323–340.

Huang, H.-C., and Eskey, D. E. (1999). The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students. *J. Educ. Technol. Syst.* 28, 75–96. doi: 10.2190/RG06-LYWB-216Y-R27G

Jeon, E. H., and Yamashita, J. (2014). L2 reading comprehension and its correlates: a meta-analysis. *Lang. Learn.* 64, 160–212. doi: 10.1111/lang.12034

Kürschner, C., Schnotz, W., Eid, M., and Hauck, G. (2005). Individuelle Modalitätspräferenzen beim Textverstehen: präferenzen für auditive oder visuelle Sprachverarbeitung in unterschiedlichen Bevölkerungsgruppen [Individual modality preferences in text comprehension: preferences for auditory or visual language processing in different populations]. *Z. Entwicklungspsychol. Pädagog. Psycho.* 37, 2–16. doi: 10.1026/0049-8637.37.1.2

Leucht, M., Retelsdorf, J., Pant, H. A., Möller, J., and Köller, O. (2015). Effekte der Gymnasialprofilzugehörigkeit auf Leistungsentwicklungen im Fach Englisch [Study profile effects on English as a first foreign language development]. *Z. Pädagog. Psychol.* 29, 77–88. doi: 10.1024/1010-0652/a000153

Liu, G.-Z., Cheng, J.-Y., and Hwang, G.-J. (2017). Mobile-based collaborative learning in the fitness center: a case study on the development of English listening comprehension with a context-aware application. *Br. J. Educ. Technol.* 49, 305–320.

McNamara, D. S., and Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychol. Learn. Motivat.* 51, 297–384. doi: 10.1016/S0079-7421(09)51009-2

Muthén, L. K. and Muthén, B. O. (1998–2017). *Mplus User's Guide*, 8th Edn. . Los Angeles, CA: Muthén & Muthén.

Nickerson, C. (2005). English as a lingua franca in international business contexts. *English Specific Purposes* 24, 367–380. doi: 10.1016/j.esp.2005.02.001

Pfost, M., Dörfler, T., and Artelt, C. (2013). Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. *Learn. Individ. Differ.* 26, 89–102. doi: 10.1016/j.lindif.2013.04.008

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.

Rupp, A. A., Vock, M., Harsch, C., and Köller, O. (2008). *Developing Standards-Based Assessment Tasks for English as a First Foreign Language – Context, Processes, and Outcomes in Germany*. Münster: Waxmann.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Lang. Test.* 25, 435–464. doi: 10.1177/0265532208094272

Spoden, C., and Fleischer, J. (2019). "Multidimensional Rasch models in first language listening tests," in *Quantitative Data Analysis for Language Assessment, Vol. II: Advanced Methods*, eds V. Aryadoust and M. Raquel (London: Routledge), 33–55.

Swift, J. S., and Wallace, J. (2011). Using English as the common corporate language in a German multinational. *J. Eur. Ind. Train.* 35, 892–913. doi: 10.1108/03090591111185574

Sylvén, L., and Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL* 24, 302–321. doi: 10.1017/S095834401200016X

Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., and Rapp, D. (2009). Simple but complex: components of the simple view of reading across grade levels. *J. Res. Read.* 32, 383–401. doi: 10.1111/j.1467-9817.2009.01401.x

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Lang. Teach.* 40, 191–210. doi: 10.1017/S0261444807004338

Vandergrift, L., and Baker, S. (2015). Learner variables in second language listening comprehension: an exploratory path analysis. *Lang. Learn.* 65, 390–416. doi: 10.1111/lang.12105

Watkins, M. W., and Edwards, V. A. (1992). Extracurricular reading and reading achievement: the rich stay rich and the poor don't read. *Read. Improv.* 29, 236–242.

Wolf, M. C., Muijselaar, M. M. L., Boonstra, A. M., and de Bree, E. H. (2019). The relationship between reading and listening comprehension: shared and modality-specific components. *Read. Writ. Interdiscip. J.* 32, 1747–1767. doi: 10.1007/s11145-018-9924-8

Wu, M. (2005). The role of plausible values in large-scale surveys. *Stud. Educ. Eval.* 31, 114–128. doi: 10.1016/j.stueduc.2005.05.005

Yagoub-Zadeh, Z., Farnia, F., and Geva, E. (2012). Toward modeling reading comprehension and reading fluency in English language learners. *Read. Writ. Interdiscipl. J.* 25, 163–187. doi: 10.1007/s11145-010-9252-0

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 June 2020
doi: 10.3389/fpsyg.2020.01122

Examining Second Language Listening, Vocabulary, and Executive Functioning

Matthew P. Wallace[1]* *and Kerry Lee*[2]

[1] Department of English, University of Macau, Macau, China, [2] Department of Early Childhood Education, The Education University of Hong Kong, Tai Po, Hong Kong

Performance on second language (L2) listening tests is influenced by individual differences in listener characteristics (e.g., executive functioning and vocabulary size) and characteristics of the listening measure (e.g., text length or skills measured). For listeners, the amount of linguistic knowledge is most important for comprehension outcomes. As language proficiency increases, non-linguistic factors, like the executive functions (EF) of working memory, purportedly begin to exert influence on listening performance. EF represents the range of functions performed by the central executive (the processing component) of the working memory system and have largely been studied in the context of updating (revising information held in temporary storage) and shifting (switching attentional focus among mental representations). To test these theoretical claims, the relationship among L2 listening, vocabulary size, updating, and shifting was examined. This included a moderation analysis to investigate whether the relationship between EF and listening was dependent upon vocabulary size. The relationships among the variables were also examined for varied test characteristics to see if contributions from EF and vocabulary differed according to text length or skill measured. In total, 209 Japanese senior high school EFL learners completed a standardized listening test and tests measuring updating, shifting, and vocabulary size. Results from structural equation modeling showed that only vocabulary was predictive of listening performance, regardless of text length or skill measured on the test. Results also showed that vocabulary size did not moderate the relationship between EF and listening, suggesting that the non-linguistic factors were not important for listening regardless of vocabulary size. The findings support claims that linguistic knowledge is most important for listening and that non-linguistic factors are less important for low-level listeners. The findings also contribute empirical evidence for the relationship between L2 listening and EF, a novel conceptualization of the working memory construct.

Keywords: second language listening, executive functioning (EF), second language vocabulary, updating, shifting

OPEN ACCESS

Edited by:
Yo In'nami,
Chuo University, Japan

Reviewed by:
Ruslan Suvorov,
University of Western Ontario, Canada
Yasuyo Sawaki,
Waseda University, Japan

***Correspondence:**
Matthew P. Wallace
mpwallace@um.edu.mo

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

Received: *15 October 2019*
Accepted: *30 April 2020*
Published: *03 June 2020*

Citation:
Wallace MP and Lee K (2020) Examining Second Language Listening, Vocabulary, and Executive Functioning. Front. Psychol. 11:1122. doi: 10.3389/fpsyg.2020.01122

INTRODUCTION

Comprehension of second language (L2) speech is a complex cognitive process that involves mental processing and the use of knowledge resources to interpret what is said. Listening tests measuring comprehension are designed to gauge how efficiently test takers utilize these cognitive resources to accomplish listening tasks, like identifying specific information from speech. Performance

on listening tests may therefore be attributed to individual differences in characteristics of the listener (e.g., vocabulary size and working memory) or those of the listening task (e.g., response format) (Buck, 2001). Research investigating listening assessment has mainly focused on how listener characteristics influence performance (e.g., Andringa et al., 2012). The current study was designed to contribute to that literature by examining how individual differences in executive functioning and vocabulary knowledge contribute to variance in L2 listening performance. Executive functioning represents the range of functions performed by the central executive (the processing component) of the working memory system that are responsible for revising information held in temporary storage as needed for task accomplishment, switching attentional focus among mental representations generated from information processing, and suppressing distractions from influencing task performance (Miyake et al., 2000). It is domain-general, meaning that it is involved in the performance of a wide range of tasks, including language comprehension. Research has shown that individual differences in executive functioning affect L1 performance (Cantin et al., 2016), though it has yet to receive much attention in the L2 literature. To address this scarcity in research, the present study examined executive functioning in the context of L2 listening comprehension.

L2 Listening Comprehension

L2 listening comprehension is operationalized similarly to Buck's (2001) definition of the construct. He explains that L2 listening involves being able to "process extended samples of realistic L2 speech, automatically and in real time, to understand linguistic information that is included within a text, and to make inferences based on information that are implicated by the content of the passage" (p. 114). Listening tests assessing comprehension that are operationalized this way measure the ability to identify information explicitly stated within listening texts, and comprehend information implicitly provided in speech (Wagner, 2004). These instruments focus on evaluating higher-level listening skills, so it is important to understand the process listeners go through to arrive at their interpretations of L2 speech.

Imhof (2010) conceives listening comprehension as a recursive structure-building process that places working memory at the center of the sequence. Listeners first select information by filtering out recognizable sounds from irrelevant noise. These sounds are then grouped into meaningful units. Linguistic knowledge plays an important role in these early stages of processing when the mental lexicon is accessed to identify and attach meaning to words which are subsequently organized into a text model of the utterance. The text model represents the information provided within a text and serves as the basis for developing a situation model of the speech (Kintsch, 1998). The situation model represents what the speech is about and is based on inferences drawn from the text model. These inferences provide additional information inherent in the speech, but are not explicitly stated in it. The later stages of processing are happening in working memory, where mental representations of the speech are generated and revised based on their relevance for goal accomplishment by means of an executive function called updating. Imhof (2010) notes that the challenge for listeners is to store representations long enough to be accessed for further processing, while continually updating them when incoming utterances are processed. Further complicating the matter is the potential for interference from inappropriately activated schemas in building structures of the speech. Accurate structures are built when listeners are able to efficiently switch among schemas that are relevant to the input while inhibiting irrelevant schemas. The presence of irrelevant schemas slows the switching function and harms the quality of the situation model being developed.

Throughout the processing sequence, executive functioning plays a central role because it controls what information is selected for attention, aids in the organization of the information by switching among activated representations to generate a text model, and finally facilitates the information-integration process by updating incoming information for goal relevance.

Despite its theoretical significance, executive functioning has been labeled as a peripheral factor as it relates to language ability. Describing how individual listener factors influence language performance, Hulstijn (2015) proposes a core-peripheral model stating that linguistic knowledge, comprised of vocabulary, grammar, and phonological knowledge and the speed at which this knowledge is accessed, explains most variance in language performance for language users at all levels of proficiency. All other factors, including general cognitive abilities, like executive functioning, are peripheral and not as important as linguistic knowledge for language performance. However, the peripheral factors purportedly can contribute more to listening performance for high-proficiency learners than low-proficiency learners. This theory aligns with Cummins's (1979) threshold hypothesis, which states that language performance is mainly influenced by linguistic knowledge, but that non-linguistic factors become influential as proficiency increases. The limited literature that has examined executive functioning in comprehension appears to support this proposed relationship, though a direct observation has yet to be reported. The current study addressed this gap to examine the relationship among listening comprehension, vocabulary knowledge, and executive functioning.

Executive Functioning

Executive functioning is operationalized the same as Miyake and Friedman (2012), that is, as updating and shifting. Updating refers to processing representations of input, maintaining them, and revising them as needed for task completion. For L2 listening, new representations are created when an utterance of L2 speech is processed through the language comprehension process. As subsequent utterances are processed, new representations either combine with existing representations being maintained or replace representations based on their relevance to the current task (Morris and Jones, 1990). Shifting refers to switching attentional focus from one schematic representation to another, while inhibiting interference from influencing task performance. This interference includes representations that may have been previously activated from long-term memory to complete an earlier task. Earlier conceptions of executive functioning separate shifting from inhibition (Miyake et al., 2000), but because efficient shifting involves being able to suppress irrelevant

representations while switching to those needed for a new task, they are represented as one construct. For L2 listening, completing comprehension tasks requires listeners to switch among the representations generated from language processing as needed to accomplish listening goals (e.g., listening for specific information, listening for gist).

It is unclear how strong the relationship is between updating and comprehension because research has reported mixed results when examining the relationship. This inconsistency in findings may be attributed to differences in linguistic proficiency. Supporting the threshold hypothesis and core-peripheral model, it has been shown that updating is more strongly related to listening performance when listeners have more linguistic resources. For example, Andringa et al. (2012) reported that updating for L1 users was associated with linguistic knowledge (inclusive of vocabulary knowledge, grammatical processing, and segmentation processing), and that both updating and linguistic knowledge explained variance in listening comprehension. In contrast, updating for intermediate level language learners did not correlate with linguistic knowledge and had a weaker relationship with L2 listening comprehension. These findings indicate that listeners with greater linguistic resources are more efficient in updating information and comprehending what they hear than those with less knowledge. In other words, updating explains some variance in listening comprehension when listeners are more proficient language users. Another explanation for the mixed findings may be that the reliability estimates for working memory measures have rarely been reported in these studies (e.g., Brunfaut and Révész, 2014; Vandergrift and Baker, 2015, 2018; Wolfgramm et al., 2016). Because it is unclear if the measures were internally consistent or not, it is possible that the items on the working memory tests may not have consistently measured the same construct, which calls into question the validity of the results.

Similarly, the literature examining the shifting-L2 comprehension relationship has suggested that language users with greater linguistic resources tend to be more skilled at switching (Costa and Santesteban, 2004; Kroll et al., 2008; Bialystok, 2015). Having more knowledge of the target language leads to higher quality representations generated from the input as a result of the language processing cycle. Because the quality of the representations is better, being able to switch among them takes less effort and there are fewer representations competing for attentional focus. In contrast, listeners with limited linguistic resources may be forced to cope with a larger number of incomplete or irrelevant representations remaining from decoding. Navigating among these representations consumes cognitive resources, thus causing representations generated from the input that do receive attentional focus to decay, and ultimately harm comprehension. Because shifting has yet to be explicitly examined along with updating in the L2 listening context, it is unclear how it may relate to listening performance.

Auditory Vocabulary Size

In addition to executive functioning, auditory vocabulary size was examined to control for language knowledge that purportedly correlates strongly with language performance. The language knowledge construct is more comprehensive than vocabulary, but the current study focused solely on auditory vocabulary size because it accounts for breadth of vocabulary and phonological knowledge. Not including other factors (e.g., grammatical knowledge and access speed) is acknowledged as a limitation of this study. Auditory vocabulary size is operationalized as the ability to recognize target language vocabulary from speech. In many L2 listening studies, vocabulary is measured with vocabulary size tests that use the written format. However, it is important to examine vocabulary size through the same mode as the outcome variable, which in this study is listening comprehension. Doing so allows for phonological knowledge to be accounted for within the vocabulary construct, as opposed to orthographic knowledge that is inherently measured in written tests. Empirical research has consistently reported that auditory vocabulary size shares a relationship with L2 listening comprehension, and that it explains most variance in listening performance when measured alongside other factors. For example, Vandergrift and Baker (2015) reported that auditory vocabulary size shared the strongest relationship with L2 listening performance when measured with auditory discrimination, working memory, metacognition, and L1 vocabulary size for teenage, beginner-level L2 French learners. A similar pattern of results was reported by Vandergrift and Baker (2018), who showed that auditory vocabulary size was the strongest predictor of L2 listening comprehension when modeled along with the same variables as the 2015 article. In both of these studies, auditory vocabulary size explained the most variance in L2 listening performance for the low-level participants, lending support for the core-peripheral model. The present study aims to further test the validity of the core-peripheral model by examining differences in the relationships among L2 listening comprehension, vocabulary size, and executive functioning and whether the vocabulary size may moderate the relationship between executive functioning and listening performance.

Characteristics of L2 Listening Measures

Characteristics of the listening measures may also influence the relationship among listening comprehension, executive functioning, and vocabulary. Brunfaut and Révész (2014) explain that when listening tests utilize longer listening tracks, it can be expected that executive resources would be more heavily taxed because listeners would need to store large amounts of information from the extended input. This should manifest itself in a correlation between updating and listening measures, but this has yet to be examined. The listening test used in the current study contained longer tracks (68 s to 2 min), which were expected to exceed the short-term memory capacity of the listeners.

The skills measured on the test may also influence the executive functioning and listening comprehension relationship. Listening tests used in empirical studies have typically mirrored Wagner's (2004) model of listening assessment, where assessments measure the ability to identify information explicitly stated within a spoken text (inclusive of main ideas

and details) and to comprehend information implicit in speech (e.g., Tsuchihira, 2007; Andringa et al., 2012; Brunfaut and Révész, 2014; Vandergrift and Baker, 2015, 2018). Of the two, it is expected that items measuring comprehension of implicit information would tax executive resources more since doing so requires listeners to build a mental model of the speech and hold onto it while making connections to what is already known in existing memory to fill in gaps not provided from the input. This has yet to be investigated since most studies have examined listening comprehension using tests that have combined both skills within the same tasks (e.g., Vandergrift and Baker, 2015, 2018).

THE PRESENT STUDY

The current study examined the relationships among L2 listening performance, updating, shifting, and auditory vocabulary size. Data used to examine the relationships among these factors were taken from a larger study that investigated whether domain-specific knowledge (vocabulary knowledge and topical knowledge) mediated the relationship between L2 listening performance and domain-general cognitive abilities (metacognitive awareness [awareness of (1) oneself as a listener, (2) of a listening task, and (3) of listening strategies], short-term memory [recall of information from temporary memory], and attentional control [shifting]) (Wallace, in press). Specifically, the current study aimed to answer the following research questions.

1. What are the relative contributions of updating, shifting, and vocabulary size to L2 listening performance?
2. Do the contributions of updating and shifting differ for shorter and longer texts?
3. Do the contributions of updating and shifting differ for tasks requiring identification of information explicitly stated within texts and for tasks requiring comprehension of information implicit in texts?
4. In a L2 environment where vocabulary size may be small, does oral vocabulary size moderate the relationship between executive functioning and L2 listening performance?

Supported by the threshold hypothesis (Cummins, 1979) and the core-peripheral model (Hulstijn, 2015), it was expected that vocabulary size would be the strongest predictor of L2 listening performance. Regarding the task characteristics, because executive functions are expected to be more heavily recruited for longer listening texts than shorter, it was expected that updating and shifting would be predictive of listening comprehension for longer texts. The study also expected updating and shifting to be more predictive of tasks requiring comprehension of implicit information than tasks requiring listeners to identify information explicitly stated within texts. Understanding implicit information is more cognitively demanding because it recruits the executive functions to deal with the processing demands of generating a situation model, whereas identifying information within a text relies more on storage of information. Finally, because the relationship between executive functioning and listening performance may depend on vocabulary size, it was expected that vocabulary knowledge would moderate the relationship between listening comprehension and executive functioning, even for low-proficient listeners in this study.

MATERIALS AND METHODS

Participants

In total, 240 first and second year EFL students (aged 15–16) from a private senior high school in Tokyo were invited to participate in the study. The students were arranged in six in-tact classes of 40 students. Of the students asked to participate, 14 elected to withdraw at some point during the data collection and another 17 were eliminated through the data screening process (incomplete data or outliers), leaving 209 (53% female, 47% male) in total. All participants had undertaken at least 3 years of compulsory English education in junior high school (ages 12–14), where they received 4 h of instruction on average per week (MEXT, 2008). In senior high school, the participants received up to 8 h of English instruction per week. Two hours were devoted to explicit grammar instruction, while the remaining 6 h comprised reading, writing, listening, and speaking under an integrated skills syllabus. Students attending this school are typically within a higher socio-economic status than most senior high school students studying in Tokyo. They were expected to be around the Common European Framework of Reference for Languages (CEFR) A2 level. The results from the TOEFL Junior listening test measuring CEFR A2-B2 levels showing they scored an average of 45% (18 out of 40) indicate that they were on the lower end of that scale.

Instruments

L2 Listening

In line with the operational definition of L2 listening, the listening section of a pilot version of the Test of English as a Foreign Language (TOEFL) Junior Standard Test served as the L2 listening performance measure. This paper-based test was designed to measure the language proficiency of English-language learners ranging from below CEFR level A2 to CEFR B2 (ETS, 2018). Content analysis of the 40-item multiple-choice test by a content area expert and the researcher identified half of the items as measuring the ability to identify information provided explicitly in the text and half measuring the ability to comprehend information implicit in the text. Each item and its associated input were coded for whether the answer could be found directly within the text or not. The rater agreement was above 90% and disagreements were discussed until there was full agreement. The first section of the test (17 items) consisted of short monologs and conversations (8–40 s) between school staff members and students and among students themselves. One item was associated with each listening text. Tracks for the second section (23 items) consisted of longer monologs and conversations (68 s to 2 min), with multiple items (three to five) per listening track. Participants could see the

questions and answer choices for each associated listening track throughout the test.

Updating

Updating was measured using three widely used tests: the Keep-track test (KTU) (Yntema, 1963), the Letter-memory test (LMU) (Morris and Jones, 1990) and the Figural-Spatial 3-back test (FS3B) (Kirchner, 1958). The format of the KTU and LMU were consistent with how they were used in Miyake et al. (2000) and the FS3B in Schmiedek et al. (2009). The language was changed to Japanese to suit the present study's participants. The other characteristics of the measures mirror those used in Miyake et al. (2000) and Schmiedek et al. (2009).

The KTU required participants to recall the last word for a particular semantic category. Participants saw a sequence of 15 words presented serially. At the same time, two to four semantic categories (countries, clothes, animals, sports) were listed on the bottom of the screen. After all of the words from the trial were presented, participants wrote the last word for each category from the list on answer sheets. The tests included four practice trials (two trials with seven stimuli words and one semantic category, and two trials with 15 stimuli and two semantic categories) and 12 experimental trials (three trials each at two semantic groups, three semantic groups, and four semantic groups with 15 stimuli each).

The LMU required participants to recall only the last four Japanese characters from a sequence of characters. Japanese katakana characters (e.g., ス、ア、イ、ン、マ、etc.) were presented serially for 2000 ms in the middle of the computer screen, with a 500 ms pause between each character presentation. The final four characters did not form meaningful words or phrases in Japanese. The test included three practice trials (two 5-character sequences and one 7-character sequence) and 12 experimental trials (three trials each at 5, 7, 9, and 11 character lengths).

The FS3B required participants to recall the most recent position of boxes on a grid. Participants were presented with a 4 × 4 grid of white boxes in the middle of the screen. One box on the grid turned black for 500 ms and then turned white again for 1500 ms before another box turned black. Participants assessed whether the position of the box that turned black matched the position of the box that turned black three turns before (or three-back). Participants completed two practice trials (10 box positions needing matching judgment) and three experimental trials (21 boxes requiring judgment each trial).

After the experimental trials were completed, a score representing each test was calculated by summing the total number of correct responses for every possible response on the test.

Shifting

Shifting was measured using three well-established tests: Number-letter test (NLT) (Rogers and Monsell, 1995), Plus-minus test (PMT), and Global-local test (GLT) (Miyake et al., 2000). The test was administered on computers to collect response and response-time data. The language was changed to Japanese to suit the present study's participants and the characteristics of the tests are consistent with Miyake et al. (2000).

The NLT asked participants to indicate whether the number of a number-character pair (e.g., 2キ) was even or odd when presented on the top of the screen, and whether the character was a vowel (ア、イ、ウ、エ、オ) or a consonant (カ、キ、ク、ケ、コ) when presented on the bottom of the screen. The test consisted of six trials: number-only trial with pairs shown only at the top of the screen, character only trial with pairs only on the bottom of the screen, and two switch trials with pairs presented clockwise from top left quadrant of the screen to top right, bottom right, and bottom left.

The PMT required participants to switch between adding "two" to a number and subtracting "two" from a number. When numbers were presented in black on the computer screen, they added, and when it was gray, they subtracted. Participants indicated their response using the keyboard. The test consisted of four trials: add only with 34 black numbers, subtract only with 34 gray numbers, and two switch trials with 17 black and gray numbers presented alternatively. The GLT required participants to switch between features of large and small sized figures. Large (global) geometric figures (circles, cross, triangle, square) were presented on screen with lines composed of the same geometric figures (local). Depending on the color of the figure presented, participants counted the number of lines (1 for circle, 2 for cross, 3 for triangle, 4 for square) that composed either the "global" figure (if it was black) or the "local" figure (if it was blue). The test consisted of four trials: global only with 24 black figures, local only with 24 blue figures, and two switch trials with 12 black and blue figures presented alternatively.

After the experimental trials were completed, a shifting efficiency score was calculated for each test by dividing the total number of correct responses for each trial by the mean reaction time of correct trials (Ellefson et al., 2017). This allowed for speed-accuracy tradeoffs to be taken into account. For the purposes of analysis, the efficiency scores were converted to whole numbers by multiplying 100 to them.

Auditory Vocabulary Size

Auditory vocabulary size was measured using two sections of the Listening Vocabulary Levels Test (LVLT) (McLean et al., 2015). The words used on the LVLT came from Nation's (2012) word lists comprising the most frequently used headwords from the British National Corpus and Corpus of Contemporary American English. Nation compiled word lists based on these corpus databases, reduced word families to headwords, and divided them into levels (1000 words per level) based on frequency of occurrence. Only the first 2,000 word level sections of the test were used because a profile of the listening test texts showed that they contained over 94% of words from this level. It was expected that this level would be needed to have sufficient lexical coverage for the listening test. In terms of format, the test consisted of two sections: one section each for the first two 1000 word levels, with 24 words per section. Each word was spoken once, followed by a sentence that did not reveal the meaning of the word. Participants

matched the English word they heard to the corresponding word in Japanese (the L1). After the test, a total score for both sections was calculated.

Data Collection Procedures

After receiving ethical clearance and permission to conduct the study from the high school administration, students were recruited from their English classes by one of the researchers and a teacher. Students who provided parental consent and agreed to participate in the study completed the instruments after school on four separate days over a 3 week span. Each test was administered in groups of up to 40 students. The listening and vocabulary tests were delivered in their paper-and-pencil format in a classroom and took 40 and 20 min to complete, respectively. For the listening test, following recommendations by Educational Testing Service, the instrument developer, participants heard each audio once and recorded their responses on their corresponding answer sheet. Similarly, as recommended by McLean et al. (2015), participants heard each vocabulary word and corresponding sentence once and indicated their response on their answer sheet. The responses were inputted into SPSS version 24 (IBM, 2016) for subsequent analysis. A research assistant verified the accuracy of the data entry by manually checking the match between test responses and data input into SPSS.

The executive functioning tests were administered in a computer lab. Groups of up to 40 participants completed the three shifting tests on 1 day and the updating tests on a different day. The researcher led a demonstration of each test before directing the participants to complete them. It took 40 min to complete all three updating and all three shifting tests. After completing each test, participants took a 5 min break. All six tests were delivered on computers using E-Prime 2.0 (Schneider et al., 2002). For the shifting tests and FS3B, responses and response times were collected. For the KTU and LMU updating tests, participants indicated their responses on an answer sheet. The responses and their associated times were exported to SPSS for subsequent analysis. A research assistant verified the accuracy of these responses by matching the test responses with the input response in SPSS.

Data Analysis

Variables from each test were created for analysis. For the listening test, five variables were computed. One variable consisted of the total score on the TOEFL Junior listening test. Two variables divided the listening items by text length. One measured short texts (17 items) and another measured long texts (23 items). Two other variables divided the listening items by skill measured. One measured the ability to comprehend explicitly stated information (20 items) and another measured comprehension of implicit information (20 items). Descriptive statistics and reliability estimates were calculated for each measure to provide evidence of normality and internal consistency. Outliers were identified by examining the inter-quartile range of scores. Z-scores were calculated for the variables and if their values were larger than the absolute value of 2.68, they were considered an outlier and removed from the analysis. The skewness and kurtosis values were inspected after the outliers were removed. Variables with values smaller than 2.0 were considered normally distributed (Field, 2009). Multivariate outliers also were inspected by calculating the Mahalanobis distances for the variables in the study and comparing them to a chi-square distribution with the same degrees of freedom. If the *p*-value of the right tail of the chi-square distribution was below 0.001, then multivariate outliers would be present and subsequently removed. To inspect multivariate normality, Mardia's coefficient (Mardia, 1970) was calculated. Values outside the absolute value of 3.0 are considered non-normal (Bentler, 2006). To verify the unidimensionality of the variables, they were subjected to Principal Components Analysis of Residuals, the statistical procedure in Rasch Modeling that identifies the difference in the amount of variance that is explained by the Rasch model with variance left unexplained in the model, called Rasch residuals. To determine the difference in variance, Winsteps (Linacre, 2016) produces Eigenvalues and percentage of variance explained by both the Rasch model and Rasch residuals (called Contrasts in Winsteps). Larger Eigenvalues (above 2.0) with large percentages of variance explained by Contrasts would indicate the instrument was multidimensional. However, if the Eigenvalues of the Rasch model are up to three times in excess to that of the Contrast Eigenvalues, the instrument can still be considered unidimensional.

To test the dimensionality of the updating and shifting factors, and to verify that the executive functions are separate, confirmatory factor analysis (CFA) was conducted using MPlus (version 8.4) (Muthen and Muthen, 1998-2019). Two measurement models were examined: Single factor and Two-factor model. One factor was regressed onto all six executive function variables for the Single factor model (see **Figure 1**).

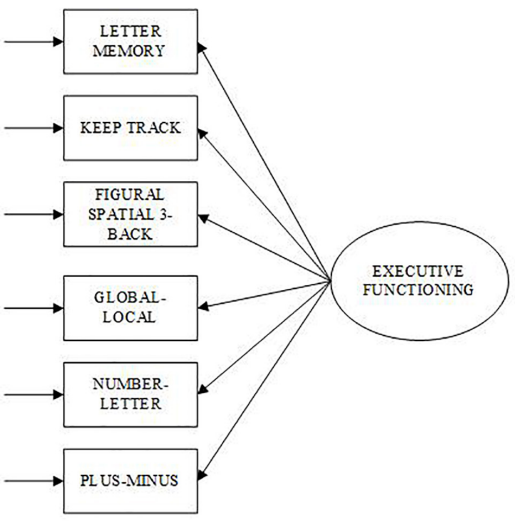


FIGURE 1 | Single factor confirmatory model of Executive Functioning. Ovals represent latent variable and rectangles represent observed variables.

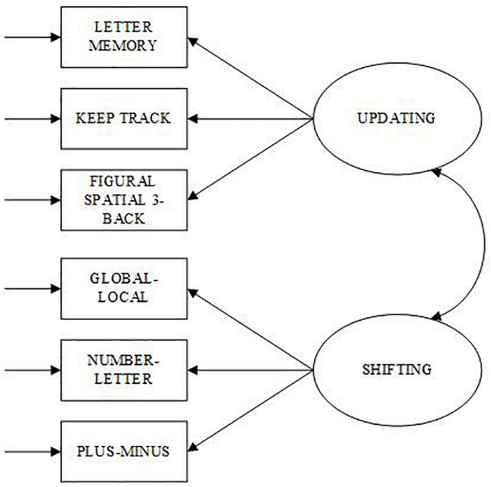


FIGURE 2 | Two-factor confirmatory model of Executive Functioning. Ovals represent latent variables and rectangles represent observed variables.

For the Two-factor model (see **Figure 2**), an updating factor was regressed onto three updating variables (KTU, LMU, FS3B) and a shifting factor was regressed onto three shifting variables (GLT, NLT, PMT). A correlation parameter was set between updating and shifting factors. The Maximum Likelihood estimation method was used for identification and the factor variances for the latent variables were set to 1.0, allowing the path coefficients to be freed.

To test which model fit the data better, the model fit statistics were compared and a chi-square difference test was run. Kline (2016) suggests that model fit is considered good when the Comparative Fit Index (CFI) is above 0.900, the Root Mean Square Error of Approximation (RMSEA) is below 0.05 and the Standardized Root Mean-Square Residual (SRMR) is below 0.08. The Bayesian Information Criteria (BIC), a statistic that is used to compare models that is sensitive to degrees of freedom, sample size, and model complexity, was also consulted. Lower BIC values indicate more parsimony, and therefore, better fitting model. Vocabulary size was added to the better fitting confirmatory model to confirm the factor structures of the predictors (EF-VS model). In the model, vocabulary size was correlated with updating and shifting factors.

To answer the first research question, L2 listening was regressed onto the updating, shifting, and vocabulary size factors. To answer the second and third questions, the listening factor was divided into two different subsets of the listening construct: one subset for length of text and the other for type of information requiring comprehension on the test. For length, variables for short and long texts were regressed onto the updating, shifting, and vocabulary factors. To answer the third research question examining comprehension of information type, the variables representing comprehension of explicit information (20 items) and implicit information (20 items) items were regressed onto the updating, shifting, and vocabulary size factors. Fit statistics were consulted to evaluate how closely the data fit the models. To answer the final research question, two moderator variables consisting of vocabulary size and updating and vocabulary size and shifting were created. The structural model was re-run twice with the moderator variables included, respectively. If the moderator variable explained variance in listening performance, then an interaction would be present.

RESULTS

The descriptive statistics, reliability estimates, and Principal Components Analysis of Residuals estimates are presented in **Table 1**. The skewness and kurtosis values of the variables show that they all were within the absolute value of 2.0 and the Mardia coefficient was within the absolute value of 3.0, indicating the data was approximately normal. Coefficient alpha for each of the measures indicates an acceptable level of internal consistency for the variables. Principal Components Analysis of Residuals indicated that the variables were unidimensional. Though the vocabulary, LMU, KTU, NLT, and PMT variables had Eigenvalues above 2.0, the percentage of variance explained by the Rasch model was over three times that explained by the first contrast.

Intercorrelations among the variables show that not all of them were correlated with one another (**Table 2**). The listening and vocabulary variables were associated with each other, but the strength of the correlations was weaker than anticipated. None of the updating variables correlated with the listening variables, and only two shifting variables (NLT and GLT) correlated with listening variables.

Confirmatory Factor Analysis

Results presented in **Table 3** indicated that the Two-factor model fit the data better than the Single factor model. This was confirmed by the chi-square difference test showing that the Two-factor model was statistically different from the Single factor model ($\Delta x^2 = 18.78$, $\Delta df = 1$, $p < 0.01$), and therefore fits the data better. The results also showed that the two executive functions shared a moderate relationship ($r = 0.368$, $p < 0.01$). In line with Lee et al. (2013), these results support the expectations that the two executive functions were separable for the mid-adolescent participants in this study. They also support Miyake and Friedman's (2012) contention that updating and shifting are unified (in that they shared a relationship) yet diverse (the relationship was not strong). Vocabulary size was then added to the Two-factor model and the results show good fit to the EF-VS model of the predictors.

Structural Equation Modeling

Table 4 presents results from the SEM analyses. The results for the first research question show that the data fit the L2L model well (see **Figure 3**): non-significant $x^2(16) = 7.449$, $CFI = 1.00$, $RMSEA = 0.00$, $BIC = 7607.667$, and $SRMR = 0.026$. Of the three variables, vocabulary size was the only one that was predictive of L2 listening performance ($\beta = 0.410$, $p < 0.01$). Vocabulary also correlated with updating ($r = 0.281$, $p < 0.01$), but not shifting.

TABLE 1 | Descriptive statistics, reliability estimates, and Principal Components Analysis of Residuals of Rasch dimension and Unexplained Variance for all variables (n = 209).

Measure	Mean	*SD*	Max value	Skewness	Kurtosis	Reliability	Rasch dimension (EV)	Unexplained variance: first contrast (EV)
L2 listening	18.07	6.77	40	0.552	−0.438	0.823	20.6% (10.40)	4.5% (2.28)
Explicit	9.01	3.77	20	0.392	−0.452	0.705	20.5% (5.17)	7.5% (1.88)
Implicit	9.05	3.58	20	0.317	−0.338	0.690	25% (6.67)	6.2% (1.66)
Short	8.80	3.49	17	0.174	−0.655	0.718	24.5% (5.39)	7.2% (1.60)
Long	9.28	4.01	23	0.498	−0.329	0.711	20.5% (5.92)	6.2% (1.79)
KTU	18.88	3.62	27	−0.587	0.830	0.643	21.8% (7.54)	5.4% (1.87)
LMU	38.91	6.36	48	−0.708	0.292	0.845	22.3% (13.77)	5.5% (3.40)
FS3B	42.18	12.38	72	−0.931	0.670	0.925	18.7% (14.53)	3.4% (2.66)
NLT	3.65	0.817	12	0.384	0.476	0.757	47.6% (32.67)	3.7% (2.56)
PMT	1.35	0.299	7	0.525	0.787	0.836	30.2% (14.68)	4.9% (2.37)
GLT	3.06	0.629	16	0.321	0.149	0.774	44.0% (18.88)	4.5% (1.94)
VS	38.99	3.81	48	−0.489	0.626	0.640	34% (21.02)	3.8% (2.99)

Explicit, explicit information items; Implicit, implicit information items; Short, items associated with short texts; Long, items associated with long texts; KTU, keep-track test; LMU, letter-memory test; FS3B, figural-spatial 3-back test; NLT, number-letter test; PMT, plus-minus test; GLT, global-local test; VS, vocabulary size; Max score, maximum possible score; EV, eigenvalues.

TABLE 2 | Correlation matrix for the variables (n = 209).

Variable	1	2	3	4	5	6	7	8	9	10	11
L2 listening	1										
Explicit	0.926**	1									
Implicit	0.918**	0.700**	1								
Short	0.888**	0.763**	0.878**	1							
Long	0.917**	0.901**	0.787**	0.631**	1						
KTU	0.085	0.048	0.111	0.102	0.057	1					
LMU	0.130	0.126	0.112	0.118	0.114	0.290**	1				
FS3B	0.118	0.115	0.102	0.113	0.104	0.183**	0.082	1			
NLT	0.159*	0.163*	0.129	0.151*	0.135	0.180**	0.105	0.051	1		
PMT	0.107	0.086	0.111	0.142*	0.057	0.095	0.057	0.148*	0.249**	1	
GLT	0.154*	0.137*	0.147*	0.137*	0.142*	0.163*	0.073	0.107	0.450**	0.311**	1
VS	0.439**	0.388**	0.423**	0.437**	0.362**	0.147*	0.159*	0.116	0.094	0.097	0.083

**$p < 0.05$; **$p < 0.01$; Explicit, explicit information items; Implicit, implicit information items; Short, items associated with short texts; Long, items associated with long texts; KTU, keep-track test; LMU, letter-memory test; FS3B, figural-spatial 3-back test; NLT, number-letter test; PMT, plus-minus test; GLT, global-local test; VS, vocabulary size.*

TABLE 3 | Fit indices for Single factor, Two-factor, and EF-VS measurement models.

Model	x^2	*df*	*p*-value	*CFI*	*RMSEA*	*BIC*	*SRMR*
Single factor	22.945	9	0.006	0.856	0.086	5089.427	0.058
Two-factor	4.173	8	0.841	1.000	0.000	5075.998	0.025
EF-VS	6.064	12	0.913	1.000	0.000	6239.489	0.025

x^2, chi squared statistic; df, degrees of freedom; CFI, comparative fit index; RMSEA, root mean square error of approximation; BIC, Bayesian Information Criteria; SRMR, standardized root mean-square residual.

The SEM results for research question two show good fit to the Short-Long model (see **Figure 4**): $x^2(20) = 9.792$, *CFI* = 1.00, *RMSEA* = 0.00, *BIC* = 8420.937, and *SRMR* = 0.026. Vocabulary was a stronger predictor of shorter texts (β = 0.405, $p < 0.001$) than longer texts (β = 0.340, $p < 0.001$). The SEM results for the third research question also show good fit to the Explicit-Implicit model (see **Figure 5**): $x^2(20) = 10.045$, *CFI* = 1.00, *RMSEA* = 0.00, *BIC* = 8377.739, and *SRMR* = 0.027. Vocabulary was the only predictor of the listening variables, explaining more variance in scores for implicit information items (β = 0.390, $p < 0.01$) than explicit information items (β = 0.368, $p < 0.01$). In every model, neither updating nor shifting were predictive of L2 listening comprehension after controlling for vocabulary size. However, updating and shifting shared a moderate relationship with one another. Regarding the final research question, the results showed that the vocabulary size did not moderate

TABLE 4 | Fit indices for structural models.

Model	x^2	*df*	*p*-value	*CFI*	*RMSEA*	*BIC*	*SRMR*
L2 listening	7.449	16	0.964	1.000	0.000	7607.667	0.026
Short-long	9.792	20	0.972	1.000	0.000	8420.937	0.026
Explicit-implicit	10.045	20	0.967	1.000	0.000	8377.739	0.027

L2L, L2 listening as latent variable; Explicit-Implicit, L2 listening as explicit and implicit item observed variables; Short-long, L2 listening as short and long text observed variables; x^2, chi squared statistic; df, degrees of freedom; CFI, comparative fit index; RMSEA, root mean square error of approximation; BIC, Bayesian Information Criteria; SRMR, standardized root mean-square residual.

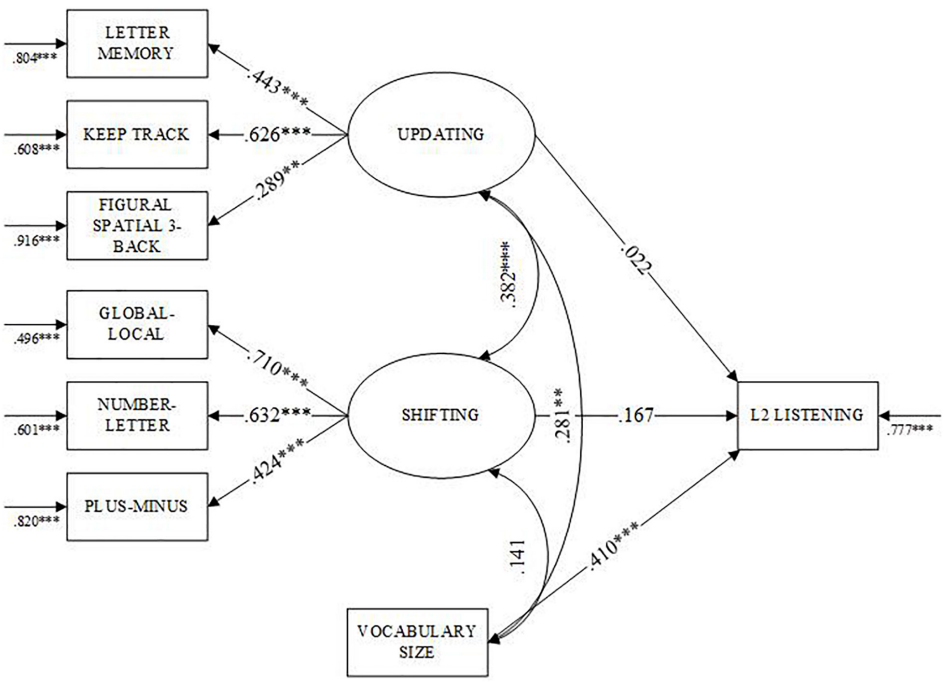


FIGURE 3 | Standardized parameters of the SEM model of L2 listening comprehension. $^{**}p < 0.01$, $^{***}p < 0.001$. Ovals represent latent variables and rectangles represent observed variables.

the relationship between listening performance and updating (β = 0.112, $p = 0.240$) or shifting (β = 0.018, $p = 0.805$).

DISCUSSION

Listener Characteristics

The first research question aimed to examine the relationship among L2 listening performance, updating, shifting, and vocabulary size. The results showed that only vocabulary size was associated with better listening performance, and that neither updating nor shifting were. These results support the core-peripheral model of language proficiency, which states that language knowledge is most important for language performance, and peripheral factors, like executive functioning, are less important (Hulstijn, 2015). The findings align with earlier studies showing that individual differences in working memory, of which executive functioning largely comprises, fails to predict L2 listening comprehension, but that linguistic knowledge in general (Andringa et al., 2012), and vocabulary size in particular (Vandergrift and Baker, 2015, 2018; Wolfgramm et al., 2016) does. For example, Vandergrift and Baker (2015) reported that vocabulary size was predictive of L2 listening performance, but working memory was not for teenage French immersion students with a limited vocabulary size (38% on a vocabulary size test). Explaining similar results for younger participants, Vandergrift and Baker (2018) speculated that the low vocabulary prevented executive resources from aiding in comprehension. They characterized the relationship between working memory and vocabulary as being developmentally linked, stating that efficiency in using executive resources improves alongside increases in language proficiency and as these two increase, comprehension improves. A similar explanation may be offered for the current study, since the young participants were of limited vocabulary size. van Zeeland and Schmitt (2012) set criteria for good comprehension of L2 spoken texts at knowledge of around 90–95% of the vocabulary. However, the mean scores of the vocabulary test (**Table 1**) show that the participants knew only about 81% of words at the two-thousand vocabulary level, which is well below the threshold required for good comprehension of

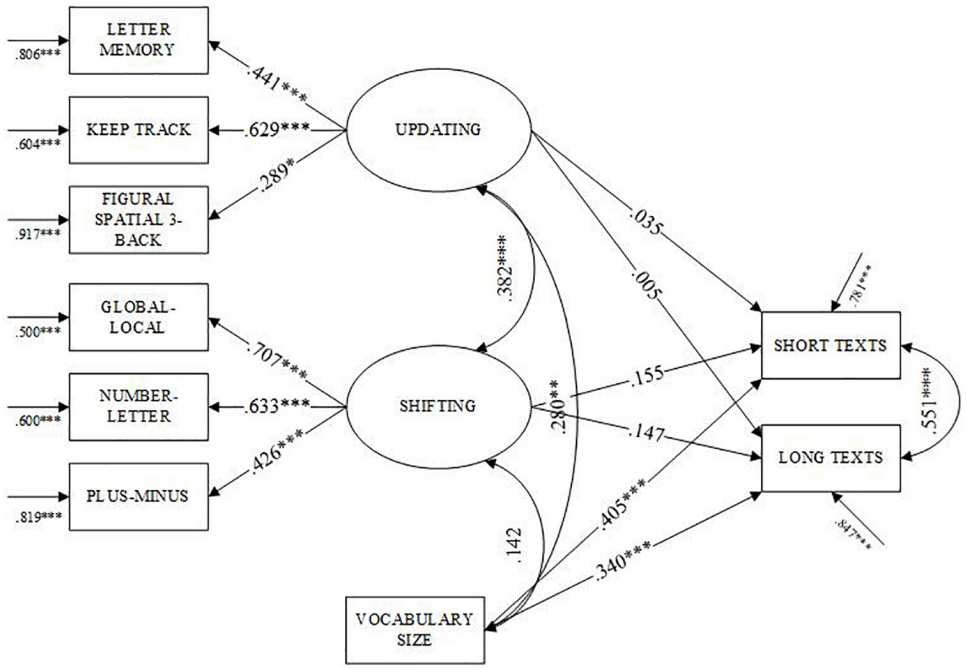


FIGURE 4 | Standardized parameters of the SEM model of short and long texts. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Ovals represent latent variables and rectangles represent observed variables.

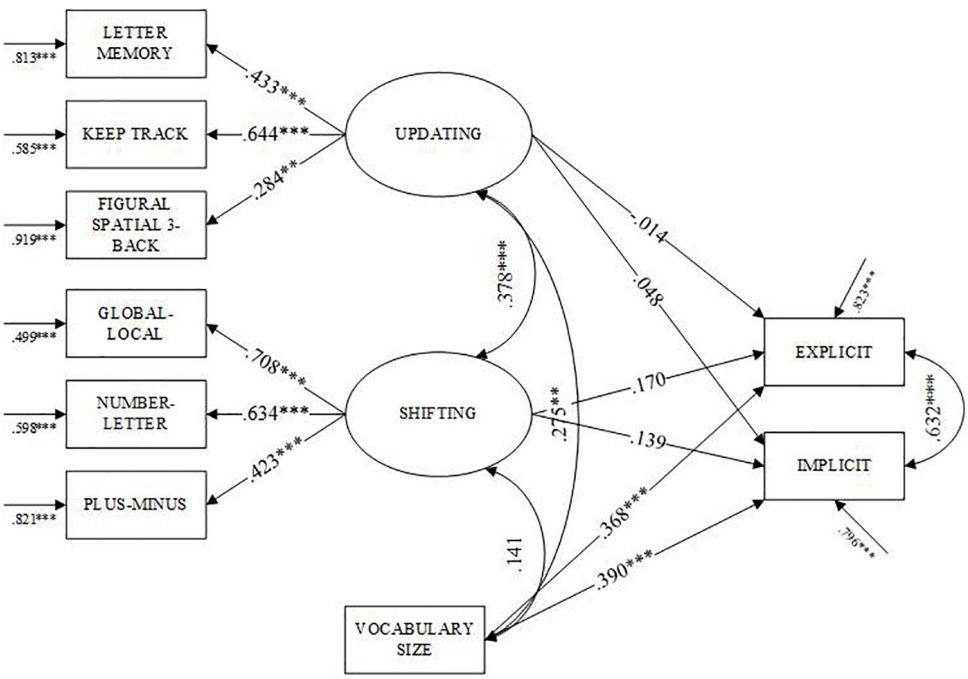


FIGURE 5 | Standardized parameters of the SEM model of explicit and implicit comprehension. **$p < 0.01$, ***$p < 0.001$. Ovals represent latent variables and rectangles represent observed variables.

the listening test texts containing 98% of words from that level. Because the participants were below the threshold of knowledge needed for adequate comprehension, most of their cognitive resources were likely spent in early stages of language processing working out what words they heard. This would limit how useful the executive resources would be in later stages of processing

when mental representations were switched among and revised to generate a mental model of the speech. If the listeners were unable to accurately or completely decode words, the executive processes would not be very useful for comprehension because switching among and updating low quality and quantity representations would not generate an accurate or complete mental model of speech.

Another explanation is that the participants' executive functions were not sufficiently developed to be of use during the listening tasks. It has been claimed that executive functioning may not come to maturity until adulthood (Rose et al., 2011), so it is possible that for the young teenage participants in this study, their updating and shifting abilities may have been undeveloped. The descriptive statistics show that this may have been the case for shifting, since the participants were not very skilled in shifting their attentional focus from one task to another. They were only 30% efficient at switching between numbers and characters, 19% efficient at switching between adding and subtracting numbers, and 19% efficient at switching between shape sizes. These results can be interpreted to mean that when the participants accurately completed tasks, they were slow in switching their attentional focus from one task to another and that when they quickly switched to new tasks, they were less accurate in completing them. This slow and effortful shifting would have posed challenges for listeners because they did not control the pace of speech or duration of the tasks. The TOEFL Junior had 23 items that were associated with six audio tracks (three to four items per track). Participants had to listen to the audio and shift their attentional focus between reading and answering the multiple-choice items and listening for information. Being slow in answering questions about information early in the audio track (first two out of three/four items) may have helped them accurately answer those questions, but they likely would have missed important information given later in the audio that was needed to answer other questions. Results from a paired-samples *t*-test conducted on the TOEFL Junior test items supports this claim and showed that for the multiple-item tasks, the listeners more accurately answered the first two items within a single track (12 items: $M = 5.49$, $SD = 2.56$) than the last items (11 items: $M = 3.79$, $SD = 2.04$), $t(208) = 10.678$, $p < 0.01$. Had shifting resources been more efficient, it is possible that the participants' listening performance would have been better. However, it appears that their shifting resources were too limited to be of much help for these listeners.

Updating also failed to share a relationship with listening comprehension, but it appears to have been sufficiently developed. The listeners performed moderately well on the updating tests, recalling 70% of the word stimuli, 81% of the character stimuli, and 67% of the figural-spatial stimuli. This means that they were somewhat accurate in being able to revise varied types of information in their short-term memory. It is therefore curious as to why updating was not important for listening performance, especially when doing so was expected to play a key role in the listening comprehension process (Imhof, 2010). The nature of the representations that are being updated may be a reason. In order for updating to aid in comprehension, listeners may need to be efficient in updating representations of the target language. When the representations are different from the target language, like numbers, first language characters, and figures that were measured by the updating tasks, they may not be as helpful for lower-level listeners in comprehending speech. Because the current study did not measure the ability to update target language speech, this was not observed and can be treated as a limitation of the study.

Listening Test Lengths and Skills

The second research question investigated if the contributions of updating and shifting would differ for longer or shorter text lengths. It was expected that longer listening tracks would engage the executive functions more than shorter tracks since listeners would need to revise more information from the input and switch among more mental representations to generate a mental model of the text. However, similar to the results for L2 listening performance overall, only vocabulary size explained variance in listening comprehension for both lengths (long texts: $\beta = 0.340$, $p < 001$; short texts: $\beta = 0.405$, $p < 0.001$). These results suggest that when vocabulary is controlled for, executive functions do not influence listening performance, regardless of text length. It has been reported that working memory (memory and executive functions) failed to explain variance in listening comprehension for short texts when controlling for vocabulary size (Vandergrift and Baker, 2018) and linguistic knowledge (grammar and vocabulary) (Andringa et al., 2012). The results of the present study indicate that a similar pattern of relationships may exist for the executive functions of working memory as well, since the updating and shifting executive functions failed to explain variance in comprehension beyond what was explained by vocabulary size.

It may be that the shorter texts did not extend beyond the participants' memory capacity, meaning that they could remember all of the information without having to revise what they held in their memory and limiting how many mental representations were needed to be switched among. Nearly half of the items of the TOEFL Junior Standard test (17 of the 40 items) utilized short listening tracks that were around 12–40 s long and involved little discourse beyond three to four sentences. For these texts, it is possible that the participants remembered everything that was said. However, this may not explain the results for the longer texts. The other half (23 of the 40 items) of the items were associated with texts ranging from 68 s to 2 min and it would be challenging to remember all of the information provided in these longer pieces of discourse. The executive functions may not have been engaged during these longer texts because having the answer sheet with the questions and answer options available throughout the test reduced the cognitive load required for listening. Participants could have written key points from the texts down on their answer sheets as they listened and/or marked key terms in the question and answer choices as they followed along with the audio. This would have eliminated the need to hold all of the information provided in the texts in their temporary memory and essentially exported the information from the memory system to the paper. Future studies may consider addressing this by examining if the executive

functions share a relationship with listening performance when the answer choices are provided and when they are not.

The third research question examined if the contribution of updating and shifting would differ depending on the skills measured on the listening test. It was expected that requiring listeners to comprehend implicit information provided within a text would tax the executive functions. However, updating and shifting did not explain variance in the explicit or implicit listening item scores beyond that explained by vocabulary size. These findings are consistent with Vandergrift and Baker (2015, 2018) who also measured the ability to understand explicit and implicit information in L2 speech for teenage language learners. The multiple-choice response format on the listening tests in these earlier studies and the present one may have contributed to the consistent findings. It has been suggested in the literature that multiple-choice response format may overload cognitive resources and limit comprehension because it introduces the construct-irrelevant factor of reading comprehension by requiring test takers to read and comprehend the questions and answer choices in addition to holding information in memory as they listen (Brunfaut and Révész, 2014). However, the opposite is proposed here as an explanation for the executive functions not sharing a relationship with listening performance. The premise of providing listeners with the goals of a listening task beforehand in order to signal what they should focus on as they listen is consistent with real-world listening events, where implicit or explicit listening goals drive what is attended to in speech. On assessments, this is taken in the form of providing the questions and answer choices before the listening begins. However, when these are provided, it alerts listeners to the specific language they should be listening for in addition to the goal of listening. In this way, the key words given in the questions and answer choices likely activate their prior knowledge before the listening track begins, thus reducing the amount of new representations the listeners needed to generate from the input and overall cognitive load. Executive functioning would therefore be of limited use because the relevant representations needed for comprehension have been pre-activated before the listening started. Listeners would simply narrow their focus on key terms as they listened and link what they heard to what was already activated.

Moderation

The final research question examined if the influence of executive functioning on listening performance is dependent upon language knowledge for learners of low proficiency levels. The results showed that vocabulary size did not moderate the relationship between listening comprehension and either updating or shifting. It is likely that the vocabulary size was too low for the moderation effect to be detected. For the data to exhibit interaction, participants who are well above the threshold of vocabulary size needed for comprehension of the texts would need to be included in the sample. The participants were well below that threshold. In order to detect this possible interaction, future studies are encouraged to include participants with a larger vocabulary size. Another explanation for this may be that auditory vocabulary size alone did not moderate the relationship. Though Mecartty (2000) reported that vocabulary explained much more unique variance in L2 listening comprehension than grammar, it is possible that not including grammar or other aspects of linguistic knowledge (e.g., vocabulary depth, grammar, speed of accessing language knowledge) may have limited the extent to which language knowledge influenced the executive functioning and listening comprehension relationship. The results may therefore be interpreted to mean that auditory vocabulary knowledge failed to moderate the relationship between executive functions and listening performance. This should be understood as a limitation of the study and future research is encouraged to include grammar, depth of vocabulary knowledge, and access speed when examining whether language knowledge moderates the effects of peripheral factors for listening comprehension.

CONCLUSION

To conclude, for the Japanese EFL participants in this study, having a larger auditory vocabulary size was most important for comprehending the L2 speech. Features of the listening test, namely the text lengths and skills measured, did not affect the contributions that updating and shifting made to L2 listening performance. These findings may be attributed to the limited linguistic resources of the participants, as the input may have been beyond the listeners' threshold of linguistic knowledge and thereby preventing the executive functions from having much influence on comprehension. If there is insufficient existing knowledge to resolve problems presented by the incoming information, no amount of executive function is going to help. The results also showed that vocabulary size did not moderate the relationship between listening comprehension and executive functioning. Altogether, the findings provide partial support for the core-peripheral model of language proficiency, showing that vocabulary size was most important for listening performance, but that the executive functions may not explain variance in comprehension regardless of how many words are known.

This study is not without its limitations. First, the limited sample size and narrow scope in which the data was collected limit the interpretations of the study. The data was collected from a single location with a homogenous group of participants in Japan. In order to generalize the findings to the broader EFL population, the study would need to be replicated in varied contexts. Also, future studies may consider examining the relationships among the executive functions and listening comprehension for participants with a wider range of proficiency levels. This study looked narrowly at lower-level learners and concluded that the limited linguistic resources prevented the executive functions from sharing variance in listening performance. To provide a more comprehensive view of the relationships among the variables examined in this study, and to further test the core-peripheral language proficiency model, future studies may recruit participants from higher and moderate level proficiencies. Future research may also consider utilizing a listening measure that incorporates a multiple-choice format that does not provide the answer choices before the listening starts. To avoid a possible priming effect, where the vocabulary needed

for the listening is activated before the listening track plays, it is recommended that listening tests provide only the questions prior to listening and reveal the answer options after the listening track has completed. This will give the listeners a goal to listen for, but minimize their lexical activation. This kind of task may be considered more authentic in that it would require listeners to generate new representations of the input as they listen, similar to a realistic listening encounter. Overall, the findings from this study contribute empirical evidence for the relationship between L2 listening comprehension and executive functions, a novel conceptualization of the working memory construct.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Nanyang Technological University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MW and KL conceived the study. MW carried out the experiment, including instrument design, data collection, analysis, and interpretation and took the lead in writing the manuscript. KL provided critical feedback and helped shape the research, including its design, analysis, results interpretation, and manuscript development.

FUNDING

This research was funded by the University of Macau (File no. SRG2018-00138-FAH). This work was also supported by Nanyang Technological University under the scholarship for Ph.D. study. This research was also funded by Educational Testing Service (ETS) under a Committee of Examiners and the Test of English as a Foreign Language Young Students research grant. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the researcher(s).

REFERENCES

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., and Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: an individual differences approach. *Lang. Learn.* 62, 49–78. doi: 10.1111/j.1467-9922.2012.00706.x

Bentler, P. M. (2006). *EQ6 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Bialystok, E. (2015). Bilingualism and the development of executive function: the role of attention. *Child Dev. Perspect.* 9, 117–121. doi: 10.1111/cdep.12116

Brunfaut, T., and Révész, A. (2014). The role of task and listener characteristics in second language listening. *TESOL Q.* 49, 1–28. doi: 10.1002/tesq.168

Buck, G. (2001). *Assessing Listening*. New York, NY: Cambridge University Press.

Cantin, R. H., Gnaedinger, E. K., Gallaway, K. C., Hesson-McInnis, M. S., and Hund, A. M. (2016). Executive functioning predicts reading, mathematics, and theory of mind during the elementary years. *J. Exp. Child Psychol.* 146, 66–78. doi: 10.1016/j.jecp.2016.01.014

Costa, A., and Santesteban, M. (2004). Lexical access in bilingual speech production: evidence from language switching in highly proficient bilinguals and L2 learners. *J. Mem. Lang.* 50, 491–511. doi: 10.1016/j.jml.2004.02.002

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Re. Educ. Res.* 49, 222–251. doi: 10.2307/1169960

Ellefson, M. R., Ng, F. F.-Y., Wang, Q., and Hughes, C. (2017). Efficiency of executive function: a two-generation cross-cultural comparison of samples from Hong Kong and the United Kingdom. *Psychol. Sci.* 28, 555–566. doi: 10.1177/0956797616687812

ETS (2018). *Handbook for the TOEFL Junior tests*. Available online at: http://www.ets.org/s/toefl_junior/pdf/toefl_junior_tests_handbook.pdf (accessed May 24, 2019).

Field, A. (2009). *Discovering Statistics Using SPSS*, 3rd Edn. London: Sage.

Hulstijn, J. H. (2015). *Language Proficiency in Native and Non-native Speakers: Theory and Research*. New York, NY: John Benjamins Publishing Company.

IBM (2016). *IBM SPSS for Windows, Version 24.0*. Armonk, NY: IBM Corp.

Imhof, M. (2010). "What is going on in the mind of the listener? The cognitive psychology of listening," in *Listening and Human Communication in the 21st Century*, ed. A. D. Wolvin (West Sussex: Wiley-Blackwell), 97–126.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* 55, 352–358. doi: 10.1037/h0043688

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling,* (4th Edn.). New York, NY: Guilford Press.

Kroll, J. F., Bobb, S. C., Misra, M., and Guo, T. (2008). Language selection in bilingual speech: evidence for inhibitory processes. *Acta Psychol.* 128, 416–430. doi: 10.1016/j.actpsy.2008.02.001

Lee, K., Bull, R., and Ho, R. M. (2013). Developmental changes in executive functioning. *Child Dev.* 84, 1933–1953. doi: 10.1111/cdev.12096

Linacre, J. M. (2016). *Winsteps® Rasch Measurement Computer Program*. Beaverton, OR: Winsteps.com.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519–530. doi: 10.1093/biomet/57.3.519

McLean, S., Kramer, B., and Begler, D. (2015). The creation and validation of a listening vocabulary levels test. *Lang. Teach. Res.* 19, 1–20. doi: 10.1177/1362168814567889

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Appl. Lang. Learn.* 11, 323–348.

MEXT (2008). *Chugakkou Gakushu Shidou Yoryo Kaisetsu Gaikokugo Hen [Explanatory Comments for the New Study of Course Guideline for Foreign Languages in Junior High Schools]*. Available online at: http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2011/01/05/1234912_010_1.pdf (accessed December 2014).

Miyake, A., and Friedman, N. P. (2012). The nature and organization of individual difference in executive functions: four general conclusions. *Curr. Dir. Psychol. Sci.* 21, 8–14. doi: 10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Morris, N., and Jones, D. M. (1990). Memory updating in working memory: the role of the central executive. *Br. J. Psychol.* 81, 111–121. doi: 10.1111/j.2044-8295.1990.tb02349.x

Muthen, L. K., and Muthen, B. O. (1998-2019). *MPlus User's Guide*, 8th Edn. Los Angeles, CA: Muthen & Muthen.

Rogers, R. A., and Monsell, S. (1995). Costs of predictable switch between simple cognitive tasks. *J. Exp. Psychol. Gen.* 124, 207–231. doi: 10.1037/0096-3445.124.2.207

Rose, S. A., Feldman, J. F., and Jankowski, J. J. (2011). Modeling a cascade of effects: the role of speed and executive functioning in preterm/full-term differences in academic achievement. *Dev. Sci.* 14, 1161–1175. doi: 10.1111/j.1467-7687.2011.01068.x

Schmiedek, F., Hildebrandt, A., Lovden, M., Lindenberger, U., and Wilhelm, O. (2009). Complex span versus updating tasks of working memory: the gap is not that deep. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1089–1096. doi: 10.1037/a0015730

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime Users Guide (Version 2.0)*. Pittsburgh, PA: Psychology Software Tools Inc.

Tsuchihira, T. (2007). L2 working memory capacity and L2 listening test scores of Japanese junior college students. *Bunkyo Gakuin Foreign Lang. Depart. Bunkyo Gakuin Junior Coll.* 7, 159–175.

van Zeeland, H., and Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Appl. Linguist.* 34, 457–479. doi: 10.1093/applin/ams074

Vandergrift, L., and Baker, S. C. (2015). Learner variables in second language listening comprehension: an exploratory path. *Lang. Learn.* 65, 390–416. doi: 10.1111/lang.12105

Vandergrift, L., and Baker, S. C. (2018). Learner variables important for success in L2 listening comprehension in French immersion classrooms. *Canad. Modern Lang. Rev.* 74, 79–100. doi: 10.3138/cmlr.3906

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Work. Pap. in Second Foreign Lang. Assess.* 2, 1–26.

Wallace, M. P. (in press). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Lang. Learn.* 71.

Wolfgramm, C., Suter, N., and Göksel, E. (2016). Examining the role of concentration, vocabulary, and self-concept in listening and reading comprehension. *Int. J. Listen.* 30, 25–46. doi: 10.1080/10904018.2015.1065746

Yntema, D. B. (1963). Keeping track of several things at once. *Hum. Fact.* 5, 7–17. doi: 10.1177/001872086300500102

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 June 2020
doi: 10.3389/fpsyg.2020.01088

Using Corpus Analyses to Help Address the DIF Interpretation: Gender Differences in Standardized Writing Assessment

Zhi Li[1], Michelle Y. Chen[2] and Jayanti Banerjee[2]*

[1] Department of Linguistics, College of Arts and Science, University of Saskatchewan, Saskatoon, SK, Canada, [2] Paragon Testing Enterprises, Vancouver, BC, Canada

Addressing differential item functioning (DIF) provides validity evidence to support the interpretation of test scores across groups. Conventional DIF methods flag DIF items statistically, but often fail to consolidate a substantive interpretation. The lack of interpretability of DIF results is particularly pronounced in writing assessment where the matching of test takers' proficiency levels often relies on external variables and the reported DIF effect is frequently small in magnitude. Using responses to a prompt that showed small gender DIF favoring female test takers, we demonstrate a corpus-based approach that helps address DIF interpretation. To provide linguistic insights into the possible sources of the small DIF effect, this study compared a gender-balanced corpus of 826 writing samples matched by test takers' performance on the reading and listening components of the test. Four groups of linguistic features that correspond to the rating dimensions, and thus partially represent the writing construct were analyzed. They include (1) sentiment and social cognition, (2) cohesion, (3) syntactic features, and (4) lexical features. After initial screening, 123 linguistic features, all of which were correlated with the writing scores, were retained for gender comparison. Among these selected features, female test takers' writing samples scored higher on six of them with small effect sizes in the categories of cohesion and syntactic features. Three of the six features were positively correlated with higher writing scores, while the other three were negative. These results are largely consistent with previous findings of gender differences in language use. Additionally, the small differences in the language features of the writing samples (in terms of the small number of features that differ between genders and the small effect size of the observed differences) are consistent with the previous DIF results, both suggesting that the effect of gender differences on the writing scores is likely to be very small. In sum, the corpus-based findings provide linguistic insights into the gender-related language differences and their potential consequences in a testing context. These findings are meaningful for furthering our understanding of the small gender DIF effect identified through statistical analysis, which lends support to the validity of writing scores.

Keywords: writing assessment, gender differences, corpus analysis, linguistic features, differential item functioning, DIF, validation

OPEN ACCESS

Edited by:
Vahid Aryadoust, Nanyang Technological University, Singapore

Reviewed by:
Clarence Green, Nanyang Technological University, Singapore
Lianzhen He, Zhejiang University, China
Michelle Raquel, The University of Hong Kong, Hong Kong

****Correspondence:***
Michelle Y. Chen
mchen@paragontesting.ca

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

Received: *19 November 2019*
Accepted: *29 April 2020*
Published: *03 June 2020*

Citation:
Li Z, Chen MY and Banerjee J (2020) Using Corpus Analyses to Help Address the DIF Interpretation: Gender Differences in Standardized Writing Assessment. Front. Psychol. 11:1088. doi: 10.3389/fpsyg.2020.01088

INTRODUCTION

The differences in language use between genders have been studied in various fields and are expected to have social consequences (Mulac et al., 2006). In language assessment, for example, if a subgroup of test takers systematically receives lower scores because of a feature of the test (rather than a true difference in language proficiency), they could consistently be denied access to opportunities, such as admission to an English-medium university. Further, assumptions might develop about what the subgroup can and cannot do that are erroneously attributed to their group membership. Since tests and assessments are widely used as a way to evaluate and compare the achievement or proficiency of test takers and since high-stakes decisions, such as graduation or promotion, are made based on test scores, score users need to be confident that the test items function similarly for all test takers regardless of their backgrounds.

In language testing, disparities in performance by subgroups of test takers are viewed from the perspective of fairness and score validity (Kunnan, 2000; Xi, 2010) and are often explored through differential item functioning (DIF) analysis. Gender-related DIF research has been primarily concerned with whether test takers at the same proficiency level might gain higher scores just because of their gender group membership. Nevertheless, only a few studies have investigated gender DIF in standardized writing tests. Most of them reported the existence of DIF effects favoring female test takers. These effects tended to be small and sometimes negligible. While it has been shown that some DIF findings were consistent across different statistical methods (e.g., Welch and Miller, 1995), none of these effects was triangulated through other sources of data such as the writing samples produced by different gender groups. To the best of our knowledge, this is the first study that has examined the linguistic features of test takers' writing samples for the prompts flagged as DIF. A motivation of this study is to address the interpretation and explanation of small gender DIF effects of the writing prompts in standardized language proficiency tests, which have been repeatedly reported in the literature. Evaluating the linguistic features of such writing samples provides unique insights into gender-related language differences and their potential consequences in testing contexts. Doing so may also advance our understanding of DIF results in writing assessment.

This study explores the possibility of using corpus analysis tools to examine gender-related linguistic variations in the writing samples elicited by a timed task on a computer-delivered English proficiency test and evaluates the impact of these differences on writing scores. We first survey the literature on gender differences in language use with a focus on writing. We also review gender DIF studies on writing tests, highlighting potential gaps in the research. Next, we describe our study's research questions, methodology, and results. Finally, we summarize our findings and discuss their implications.

Gender-Related Language Features

Many studies have discussed and reported gender differences in writing. To understand how the two genders communicate, at the macro-level, Gudykunst and Ting-Toomey (1988) proposed a gender-as-culture hypothesis and described four dimensions of inter-cultural styles. They maintained that generally, women may be perceived as being more indirect in expressing their views, more prone to using sophisticated language, more thoughtful with social roles, and more attentive to others' feelings in general interpersonal communication. A later empirical study by Mulac et al. (2006) has supported these hypotheses.

At sentence-level, Mulac and Lundell (1994) studied 40 essays written by undergraduate students at a United States university and found that 9 out of 17 language features differed between gender groups. The features associated with male writers included reference to quantity (e.g., over 30,000), elliptical sentences (e.g., to school), and judgmental adjectives (e.g., distracting); while female writers were found to use more of the following features: uncertain verbs (e.g., seems to be), progressive verbs (e.g., processing), locatives (e.g., upper corner of the frame), reference to emotion (e.g., sad), longer mean sentence length, and sentence-initial adverbials (e.g., rather than . . ., he started . . .).

In addition to sentence-level features, Jones and Myhill (2007) also examined text-level linguistic features of 718 essays written by secondary school students in the United Kingdom. They reported that the gender differences between the two groups were mainly observed in their frequency of using text-level features, rather than sentence-level features. Their study found that the gender groups differed in their use of 18 out of 35 text-level features. Male students used more topical organization, cohesion as in inter-paragraph linkage, and essay ending features. Female students used more paragraphs and repetition of a proper noun. Meanwhile, only 6 out of 24 sentence-level features were divergent between genders, including sentence length and use of finite verbs. Female students wrote shorter sentences, which is different from the findings in Mulac and Lundell (1994), but they used more finite verbs than male students.

Stylistic differences in writing between the gender groups have attracted attention as well. Rubin and Greene (1992) applied an expanded view of both biological and psychological gender to their study of gender differences in writing at a United States university. They coded multiple stylistic features in samples from 88 students on two types of tasks, namely, expressive/reflective writing and argumentative/extensive writing. Their findings indicated that the stylistic differences were less noticeable between the biological genders compared with the differences across the task types. While the similarity in stylistic features between the gender groups may be conditioned by the task characteristics (e.g., level of formality), Rubin and Greene (1992) found that female writers showed higher excitability with more exclamation points, and a lower level of confrontation with greater consideration for opposite views. The psychological gender roles, which were measured by a psychological role orientation scale, were found to have limited effects.

The exploration of possible linguistic features that are gender-specific has also been approached from a computational perspective. For example, Argamon et al. (2003) analyzed 604 documents from the British National Corpus (BNC) for gender-related differences in fiction and non-fiction genres. They

employed machine learning techniques to screen a large number of topic-independent linguistic features (list of function words and list of part-of-speech n-grams), and obtained a set of features that can help identify author gender. Argamon et al. (2003) reported that female works appeared to be more involved since they used first- and second-person pronouns more frequently, while male works contained more informational features with greater numbers of nouns and prepositions and higher type-token ratios.

Overall, these studies suggest that differences in the language used by males and females can be observed in at least four groups of linguistic features, including sentiment (e.g., reference to judgmental adjectives, discussed in Mulac and Lundell, 1994), syntax (e.g., sentence length, discussed in Jones and Myhill, 2007), cohesion (e.g., text-level features, discussed in Jones and Myhill, 2007), and lexical features (e.g., determiners and pronouns, discussed in Argamon et al., 2003). In this study, we focused on these four groups of features to evaluate the effect of gender-specific linguistic differences on test scores. Indeed, syntax, cohesion, and lexical features are commonly used in scoring rubrics to evaluate writing performance (Weigle, 2002). When a writer's tone and his/her task fulfillment are evaluated in a writing task, sentiment-related features may contribute to the overall evaluation of the writing quality as well. Together, these four groups of features partially represent the construct assessed by the writing test of interest. More information about the correspondence between the feature groups and the scoring rubric is provided in section "Materials and Methods."

The salience of gender-related language differences may be influenced by contextual factors (Rubin and Greene, 1992). For example, Leaper and Robnett (2011) found that the settings where language examples were elicited tend to influence the observed magnitude of gender language differences, with research lab setting having more pronounced differences. Likewise, gendered differences in writing performance may be influenced by testing conditions such as time constraints and communication modes as studies have shown that test takers' writing quality in standardized tests may differ from their performance on untimed writing tasks (Riazi, 2016). Nevertheless, studies on gender-related variation in writing have rarely been done in the standardized testing context. When they are, language proficiency, which is another factor that affects the linguistic features produced by test takers, is typically not controlled for. If test takers with the same writing ability have a different probability of receiving the same score on a writing test because of their gender, it will raise concerns about score validity and test fairness. Such concerns are often investigated using DIF methods.

Gender DIF in Writing Tests

In reviewing the writing DIF literature, we observed three emerging issues: (1) studies investigating gender DIF in writing tasks are rare; (2) where they exist, the gender DIF studies identified large numbers of DIF writing prompts with small effect sizes; and (3) there is a paucity of explanations for the gender DIF patterns observed. We elaborate on these three issues in the following paragraphs.

Rarity of Gender DIF Studies in Writing

First, gender DIF has been insufficiently studied on writing tasks compared to other language skills, such as listening and reading, which are often evaluated through multiple-choice items (Zwick et al., 1993). This may be related to the inherent challenges in conducting DIF analyses on writing tests (Welch and Miller, 1995; Chen et al., 2020). One such challenge is the lack of an internal matching variable that could be used in conventional DIF methods to approximate test takers' proficiency levels (e.g., the corrected total score in a test consists of multiple-choice items).

For DIF studies on writing tests, external matching variables have often been used, either in conjunction with writing scores or without them. For example, to investigate DIF on an eighth-grade assessment of writing skill, Welch and Miller (1995) used three matching variables that are created based on scores of different test components of writing skills, namely, multiple-choice questions only, multiple-choice questions and one writing prompt, and multiple-choice questions and two writing prompts. Gender DIF was identified under all three conditions, and the DIF effects appeared weaker when writing prompt(s) was included to create the matching variables. In their study of TOEFL computer-based test (CBT) writing prompts, Breland and Lee (2007) created an English language ability variable by summing up the standardized scores from three multiple-choice question sections, namely, reading, listening, and structure to examinee gender DIF effect.

Similarly, Chen et al. (2016), whose research this study extends, used multiple external matching variables to investigate gender DIF for the Canadian English Language Proficiency Index Program General (CELPIP-General) writing tasks. They matched test takers on their reading and listening scores rather than on their writing scores. This is because the typical small number of writing tasks on a test, two in their case, limits the usefulness of writing scores as an internal matching variable. For example, if one of the writing prompts is investigated for DIF, then an individual's writing proficiency will be solely relied on his/her performance on the other prompt. Also, both the reading and listening scores were highly correlated with the writing scores (r = 0.80 and 0.73, respectively), which enables using them as covariates to account for the effect of different writing proficiency levels.

Prevalence of Small Gender DIF Prompts Favoring Females

Second, in writing DIF studies, it is common for a relatively large number of prompts to be flagged as DIF prompts favoring female test takers but with small effect sizes. Welch and Miller's (1995) study highlighted that under all three matching conditions, gender DIF effects were consistently present in all six writing prompts and female test takers always had a better chance of receiving higher scores. Similar patterns have been reported by Breland and Lee (2007). They found that among the 87 prompts, 86 were flagged with statistically significant uniform DIF effects and 17 with non-uniform DIF effects. All the DIF prompts favored female test takers, although the effect sizes were "negligible." Broer et al. (2005) reported a DIF study on the argument and issue prompts in the Graduate

Record Examination (GRE). They identified DIF in both types of prompts, with females slightly outperforming their male counterparts. Chen et al. (2016) reported a similar pattern: 29 writing prompts showing gender DIF from a pool of 82, 28 of which favored female test takers with small effect sizes. The directionality and magnitude of gender DIF were similar in all these studies, however, the interpretation of these small DIF effects remains unclear and is worth further investigation.

Lack of Methods to Interpret and Explain Gender DIF

Closely tied to our last point, the third issue is a lack of an effective approach to explaining the occurrence of DIF writing prompts. As a statistical inference, the results of a DIF analysis may be influenced by other statistical artifacts (e.g., large sample size leads to statistically significant results without substantive meaning). Therefore, the interpretation of statistical significance will benefit from further evidence to verify the existence of test bias. While the sources of DIF in objective tests are often linked to item features, in performance-based writing tests or other tests that involve human raters (e.g., peer assessment, Aryadoust, 2016), the potential sources of a DIF effect can be multifactorial. A writing DIF effect may be attributed to the features of the writing task, the rubric, the rater(s), the writing samples, and the interactions between these factors. These diverse sources complicate the manifestation of a DIF effect and challenge the identification of its sources.

Commonly used methods for a follow-up analysis of DIF-flagged items include analysis of test content by experts and think-aloud protocols. These methods either assess the features of an item to identify the content-related source of DIF or focus on test takers' use of cognitive skills in their responding process to determine how it could relate to the DIF effect (Pae, 2011). It should be noted that expert judgments may not be as effective as hoped in explaining the sources of DIF items (Ferne and Rupp, 2007). In an age-based DIF study of the listening items on the Certificate in Advanced English (CAE), Geranpayeh and Kunnan (2007) invited five content experts to judge the potential levels of advantage of an item for each of three age groups. Out of 32 items, the expert panel rated seven items as potentially favoring certain age groups. However, three items (out of seven) matched the DIF items identified statistically and only one item was correctly judged regarding DIF directionality. Geranpayeh and Kunnan (2007) concluded that "expert judges could not clearly identify the sources of DIF for the items" (p. 207).

Test takers' think-aloud data are another important tool to explore the causes of DIF tasks (Ercikan et al., 2010). However, the effectiveness of this method is highly dependent on test takers' language proficiency and their ability to verbalize their thinking processes (Alderson, 1990). For example, it may be difficult for highly proficient test takers to realize the automatic processes, such as recognizing a familiar word or phrase.

Besides these methods, an analysis of the linguistic features of writing samples, although rarely used as a follow-up, could be a viable way to investigate the DIF phenomena. Since writing tasks elicit ample linguistic data, the resulting corpus could provide new evidence for validation efforts and studies of fairness (Park, 2014; Xi, 2017). Indeed, it is desirable to take advantage of the advances in corpus linguistics and use corpus-based analysis to evaluate writing DIF. In light of the findings on gender-related language features, the DIF effects identified by analyzing the test scores can be corroborated or refuted with additional evidence from a corpus-based comparative analysis of the essays written by the gender groups matched in the same fashion.

In summary, scholars agree that males and females tend to write differently (Mulac et al., 2006). However, only a handful of DIF studies have investigated gender-related performance differences in writing tests, and no research has examined the extent to which differences in the writing scores can be attributed to gender-related differences in language use. This corpus-based study focuses on an e-mail writing task that demonstrated gender DIF favoring female test takers slightly in Chen et al.'s (2016) writing DIF study and aims to address the following research questions.

Q1. Does the writing of the two gender groups differ in the four groups of construct-related linguistic features, i.e., sentiment and social cognition, cohesion, syntax, and lexical features?
Q2. Do the linguistic differences, if they exist, help explain the divergent scores of males and females on the writing task?

MATERIALS AND METHODS

The Writing Prompt

The writing prompt under investigation is from the CELPIP-General, a general English language proficiency test whose scores are aligned to the Canadian Language Benchmarks (CLB) (Centre for Canadian Language Benchmarks (CCLB), 2012). The test is delivered on a computer. The writing component measures an individual's ability to effectively use written texts to express ideas, influence others, and achieve other communicative functions in social and workplace contexts. The writing test comprises two tasks. The first is an e-mail to a service provider and the second is a response to a survey question asking for an opinion. In the writing test, test takers are supported with a built-in spell checker.

Each writing sample is evaluated by at least two certified raters independently. The scoring rubric assesses four dimensions of the writing construct: coherence and meaning, lexical range, readability and comprehensibility, and task fulfillment (Paragon Testing Enterprises, 2015)[1]. Task- or prompt-level scores are calculated based on the analytic rubric, averaging across raters. The final writing scores are converted and reported on an 11-level scale (M, 3–12), which corresponds to the CLB levels 1–2 (M) and 3–12.

This study focuses on an e-mail task that was flagged as a uniform DIF prompt favoring female test takers with a small effect size (Chen et al., 2016). The selected writing prompt represents one of the common communication functions elicited in the CELPIP-General test, namely, complaining. This prompt asked test takers to write an e-mail of 150–200 words to a restaurant. They were required to describe their recent visit, to

[1]See a brief description of the rating rubric under *Writing Performance Standards* at www.celpip.ca/test-scoring/ (accessed in January, 2019).

complain about the unavailability of menu options to satisfy their dietary restrictions, and lastly, to suggest solutions.

We selected only one DIF writing prompt to rule out prompt effects on the corpus analysis results because different prompts are likely to elicit writing samples of different linguistic features (Weigle, 2002). Although some dimensions of the writing samples are likely to be comparable across prompts (e.g., lexical range), others (e.g., task fulfillment) are probably different depending on communication goals. For example, the writings of the same proficiency level may have different linguistic features depending on whether the communication goal is to effectively complain about a failed service or to offer advice to a friend. Additionally, different prompts tend to have different magnitude, or even direction, of gender DIF effects. Thus, combining writing samples from different prompts may obscure the interpretations of the corpus analysis. Since the prompt was flagged as favoring females with a negligible effect size (R^2 change <0.02), we expected that the impact of gender on linguistic features may not be strong. It was nevertheless chosen as an example because this type of gender DIF has been repeatedly reported in the literature (e.g., Welch and Miller, 1995; Breland and Lee, 2007; Chen et al., 2016). The findings of this study may shed light on the occurrence of such gender DIF in standardized writing assessment.

The Corpus

To remain consistent with the DIF methodology used by Chen et al. (2016), the writing samples were selected based on test takers' reading and listening scores. Similar to the practice in Breland and Lee (2007) and Chen et al. (2020), the reading and listening scores are used to represent overall language proficiency, which is then used to approximate the writing proficiency to overcome the issue of lacking reliable internal matching variables within a writing test. The corpus comprised 826 writing samples (413 female and 413 male test takers), matched by reading and listening component scores.[2] The total number of running words in the corpus is 156,474. On average, the female test takers produced longer pieces. Although the difference was statistically significant, the effect size is small (Cohen's $d_{\text{male-female}} = -0.113$, $p < 0.001$). Besides, most of the linguistic features investigated in this study are normalized, making them less likely to be distorted by the text length.

As **Table 1** indicates, when reading and listening scores are matched, more male test takers are at the lower writing proficiency bands, which is consistent with the DIF result, i.e., compared to male test takers with similar language proficiency levels, females tended to achieve slightly higher writing scores on this prompt. Recall that this prompt is flagged as showing gender DIF favoring female test takers slightly, the differences in the writing scores between gender groups *cannot* be directly interpreted as "true" proficiency differences, rather it might be seen as a result of matching test takers on their English language proficiency (i.e., reading and listening scores in this case).

[2]The top three countries of nationality of these test takers were the Philippines ($N = 192$), India ($N = 135$), and China ($N = 55$).

TABLE 1 | Summary of the CELPIP-General corpus of written samples by three writing proficiency bands.

Gender	Level 4	Levels 5–8	Levels 9–12	Number of samples	Number of words
Male	49	281	83	413	76,855
Female	20	306	87	413	79,619

Selected Linguistic Features and Analytical Tools

Recent development of natural-language-processing (NLP)-based tools has provided new affordance for analyzing writing performance data. In this study, we made use of such tools developed by Kyle and his colleagues as many of their tools have been validated with empirical data by the developers and other researchers.[3] Informed by the writing construct of the CELPIP-General test (see Paragon Testing Enterprises, 2015) as well as the findings on gender-related writing features, we explored four groups of linguistic features—sentiment and social cognition, cohesion, syntactic features, and lexical features in this study. **Table 2** presents how these groups of features could partially represent the scoring dimensions. While these linguistic features are directly or indirectly related to the scoring dimensions, it is worth mentioning that each scoring dimension is more than the sum of the individual linguistic features from the analytical tools. For example, while sentiment and social cognition features may be relevant to the aspects of relevance and tone of the task fulfillment dimension, the same dimension is also concerned with the completeness of responses.

By selecting features that are valued in the rubric, we could tap into the targeted writing construct because the rating rubric is an operationalization of the abstract construct (Weigle, 2002). The four groups of linguistic features are measured by Sentiment Analysis and Cognition Engine (SEANCE) 1.05 (Crossley et al., 2017), Tool for the Automatic Analysis of Cohesion (TAACO; Crossley et al., 2016), Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) 1.0 (Kyle, 2016), and Tool for the Automatic Analysis of Lexical Sophistication (TAALES) 1.4 (Kyle and Crossley, 2015).

Sentiment and social cognition features were assessed by SÉANCE. We selected a range of features including the individual indices from two sentiment dictionaries, the Harvard IV-4 dictionary–based General Inquirer (GI) and National Research Council (NRC) Word-Association Emotion Lexicon (EmoLex), plus 20 composite indices (Crossley et al., 2016). The GI dictionary is chosen for its comprehensiveness in representing both sentiment and social cognition. It is one of the earliest sentiment dictionaries and is still widely used in research. The GI contains 119 word lists representing 16 categories of emotion and social cognitions.[4] Social cognition refers to the cognitive processes related to other people and social situation. EmoLex is a newer list of words annotated for eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). Note that, when

[3]Those tools were obtained from http://www.kristopherkyle.com/tools.html.

[4]See more details at http://www.wjh.harvard.edu/~inquirer/homecat.htm.

TABLE 2 | Summary of analytical tools and relevance of the linguistic features to the CELPIP-General scoring dimensions.

Scoring scale dimensions	Feature groups analyzed	Number of features analyzed	Tool (feature categories)
Task fulfillment: relevance/tone	General Inquirer (GI)–based indices	34	*SEANCE 1.05* (sentiment/ social cognition)
	NRC Word-Association Emotion Lexicon (EmoLex)–based indices	4	
Coherence and meaning: organization	Adjacent lexical/semantic overlaps at sentence level	7	*TAACO* (cohesion features)
Readability and comprehensibility: transitions	Rhetorical connectives	17	
	Repeated words	2	
Readability and comprehensibility: grammar	Clause-based complexity indices	11	*TAASSC 1.4* (syntactic features)
	Noun phrase–based complexity indices	15	
	Usage-based syntactic sophistication indices	12	
	Indices from the L2 Syntactic Complexity Analyzer (L2SCA)	2	
Lexical range: natural use of vocabulary	Frequency of words and n-grams (BNC)	5	*TAALES 1.0* (lexical features)
	Range of words and n-grams (BNC)	7	
	MRC psychological properties of words	3	
	Type token ratios (TTR)	4	*TAACO* (lexical features)

NRC, National Research Council (Canada); BNC, British National Corpus; MRC, Medical Research Council.

judging for sentiment polarity, SEANCE takes the negation markers into account.

The cohesion features were provided by TAACO. We selected the features related to adjacent overlaps of lexical items at sentence level, rhetorical connectives (e.g., basic connectives and the connectives showing rhetorical functions), and occurrence of repeated words. Meanwhile, we excluded paragraph-level adjacent overlap, mainly because the writing samples in the corpus are short, with an average length of 189 words, and many of them are written as a single paragraph.

The syntactic features were captured by TAASSC. We selected various features belonging to the subgroups of the L2 Syntactic Complexity Analyzer (L2SCA) outputs; clause-based complexity indices; noun phrase–based indices; and sophistication indices that focus on verb-argument constructions (VACs, i.e., the units consisting of a verb plus its argument).

To obtain the lexical features, we used both TAALES and TAACO. From TAALES, we selected the features that are calculated with reference to the written corpora such as the written registers in BNC and the Corpus of Contemporary American English (COCA). Also, we utilized word-information-score features based on the Medical Research Council (MRC) Psycholinguistic Database (familiarity, concreteness, imageability, and meaningfulness). Additionally, we paid attention to the type token ratio-based indices from TAACO.

Data Analysis

The selected variables were further screened in the following manner. We removed the indices that demonstrated extremely low variation (SD < 0.005) or contained a large proportion of zeros ($\geq$80%) because they are not widely represented in the corpus data. Then, to identify the linguistic features that contribute to writing scores, we conducted correlation analyses to identify those showing statistically significant correlations with writing performance ($p < 0.05$). Next, we checked redundancy among the indices to reduce the number of similar features that are *not* statistically different from each other. When two or more indices were closely related ($r > 0.90$), we kept the one with the highest correlation with writing performance. After applying these selection criteria, a total of 123 features were retained, including 38 sentiment and social cognition features, 26 cohesion features, 40 syntactic features, and 19 lexical features.

Considering the non-normal distributions of the majority of the linguistic features, we adopted the Mann–Whitney U tests, the non-parametric counterpart of the independent-sample *t*-test, to assess the differences between male and female test takers. Given the relatively large number of linguistic features investigated in this study, we applied Bonferroni adjustment to the significance levels to better control the overall Type I error rate. The alpha values were adjusted to 0.001 for the sentiment and social cognition features (i.e., 0.05/38) and syntactic features (i.e., 0.05/40), 0.002 for cohesion features (i.e., 0.05/26), and 0.003 for lexical features (i.e., 0.05/19).

We chose to compare the individual features between the gender groups, rather than generating latent variables or components via factor analysis (FA) or principal component analysis (PCA) for the following reasons. First, we are primarily interested in pinpointing measurable differences in specific features that would allow us to compare the findings with previous studies in different contexts. An approach evaluating each language features one at a time is suitable to address our first research question. Likewise, we did not choose FA or PCA to aggregate the variables because the resultant factors or components could be difficult to interpret (e.g., the interpretation may be subjective and makes the results less transparent) and may not well represent all the individual linguistic features.

RESULTS

This section presents the linguistic features that are statistically significantly correlated with writing proficiency levels and are distinctive between the two gender groups in the four categories,

i.e., sentiment and social cognition, cohesion, syntactic features, and lexical features.

Sentiment and Social Cognition Features

None of the 38 linguistic features in this category showed statistically significant differences between the gender groups at our pre-set significance level ($\alpha = 0.001$). While female test takers consistently outscored male test takers in most of the features, the effect sizes of gender differences in these features were extremely small (absolute values of Cohen's $d \leq 0.08$).

Cohesion Features

As **Table 3** demonstrates, 2 out of 26 cohesion features displayed statistically significant differences between the gender groups. Of the two, one concerned the use of coordinating conjunctions (e.g., "and," "but," and "or"), and the other was related to the use of pronouns. Nevertheless, the differences were small in magnitude (Cohen's d ranges from −0.11 to −0.14), with the female test takers having higher scores in both features. That is, the female test takers tended to use more coordinating conjunctions and had a higher ratio of pronouns to nouns in their writing samples. Note that these two features were negatively correlated with writing proficiency, indicating higher values on these cohesion indices are associated with lower writing scores.

Syntactic Features

Four out of forty syntactic features were statistically different between the two gender groups. They fall into three categories: noun phrase–based indices [possessives per direct object (no pronoun)], clause-based indices (complex nominals per clause and undefined dependents per clause), and usage-based syntactic sophistication indices (delta P scores). On average, the female test takers outscored their male counterparts in all seven features with small effect sizes (Cohen's d: −0.115 to −0.152).

Table 4 suggests that the female test takers used more sophisticated structures than their male counterparts. The noun phrase–based complexity index that does not count pronouns as part of noun phrases, i.e., possessives per direct object (e.g., "to accommodate my dietary requirements") was higher for the female test takers. This phenomenon is related to the earlier observation that females used more pronouns in general. The same pattern was observed in the two clause-based syntactic indices, namely, complex nominal per clause (e.g., "Even being able to find options within the menu . . .") and undefined dependents per clause (i.e., ungrammatical clauses).

TABLE 3 | Distinctive cohesion features between the two gender groups.

Features	Mann–Whitney U test (male vs. female)		Correlation with writing proficiency level	
	Effect size	*p*	*r*	*p*
Ratio of pronouns to nouns	−0.140	<0.001	−0.307	<0.001
Number of coordinating conjunctions	−0.110	0.002	−0.129	<0.001

TABLE 4 | Distinctive syntactic features between the two gender groups (TAASSC).

Features	Mann–Whitney U test (male vs. female)		Correlation with writing proficiency level	
	Effect size	*p*	*r*	*p*
Possessives per direct object (no pronoun)	−0.129	<0.001	0.143	<0.001
Complex nominals per clause	−0.129	<0.001	0.118	0.001
Delta P scores (verb-construction, SD)	−0.115	0.001	0.112	0.001
Undefined dependents per clause	−0.152	<0.001	−0.324	<0.001

As for syntactic sophistication features that are related to the association strengths of verb argument constructions in reference to COCA written registers, the two gender groups showed some difference in the standard deviations (SD) of delta P scores. Delta P score is a metric of directional strength of association between a verb and a construction with one serving as a cue and another as an outcome or vice versa. A higher value of the SD of the association strengths indicates that females had a larger variation in delta P scores in their use of VACs.

Lexical Features

None of the 19 lexical features was found to diverge between the two gender groups based on the statistical criterion we set (i.e., $\alpha = 0.003$). The absolute values of the effect sizes, as measured by Cohen's d, for these indices were smaller than 0.07.

Summary

The results showed that gender-related writing differences existed in two out of four categories of linguistic features that we explored in this study, namely, cohesion and syntactic features. However, these differences were relatively small, both in terms of the number of statistically significant features and the magnitude of the differences as shown by the effect sizes. **Table 5** shows that out of 123 language features compared across gender groups, only six (about 5%) were significantly different. Of the six significant features, three were positively correlated with higher writing scores and the other three were negative, indicating the impacts of these distinctive features on writing scores are in mixed directions and, when presented in one writing sample, their effects on the writing scores could potentially be canceled out. For example, when a writing sample by a female test taker has higher values on all these six features than the one composed by a male test taker, their writing scores are not necessarily different from each other because of the mixed directions between the language features and writing scores. The overall effect of the gender differences on the writing scores may be attenuated with a balanced distribution of correlation directionalities.

TABLE 5 | Summary of linguistic features studied for the gender groups.

Feature category (*tool*)	Number of features significantly correlated with writing scores	Number of features significantly different between gender groups	Number of gender-distinctive features that is positively correlated with writing scores
Sentiment and social cognition (*SEANCE*)	38	0	0
Cohesion (*TAACO*)	26	2	0
Syntactic (*TAASSC*)	40	4	3
Lexical (*TAALES* and *TAACO*)	19	0	0
Total	123	6	3

DISCUSSION

In this study, we applied a corpus-based analysis to examine an identified gender DIF effect and to investigate its potential linguistic sources. Using test takers' writing samples, we explored the gender DIF prompt through comparisons of multiple language features across male and female test takers, who were matched on listening and reading scores. The results show that, in standardized writing assessment, gender differences in language use are only observed on a small number of linguistic features and the magnitude of such differences is low. When presenting together, the effects of the gender-specific linguistic features on the writing scores are likely to be attenuated because the direction of these effects is mixed (i.e., some features positively affect the score outcome, while others negatively affect the score outcome). Consistent with the previous statistical analysis results (Chen et al., 2016), the findings of this study suggest that this particular writing prompt was not a serious fairness concern. This confirmation serves as an additional piece of evidence relates to test fairness and contributes to a validity argument for the test scores (Kunnan, 2000).

In interpreting the findings, some cautions are worth noting. First, the results of this study reflect the minimal gender DIF effect observed. Indeed, more substantial differences in linguistic features might be observed between the gender groups if the same analyses were to be conducted on a prompt with a large DIF effect. Similar to the results reported in other previous studies, none of the prompts reported by Chen et al. (2016) was associated with a large effect size. This is, of course, to be expected; tasks in a high-stakes context undergo rigorous review and field testing for fairness before they are used operationally. While using DIF prompt with a small effect size may seem as less optimal for studies that aim to explore gender differences, still, this type of study is helpful in addressing the interpretation of statistically flagged DIF items, especially considering the prevalence of writing prompts that were reported slightly favoring female test takers in different exams (e.g., Welch and Miller, 1995; Broer et al., 2005; Breland and Lee, 2007).

Additionally, while using a single prompt helped us focus on the gender-related features in complaint e-mail writing, the generalizability of the findings to other writing tasks may be restricted, as certain distinctive linguistic features may be prompt specific. For example, emotion-laden lexis and words about social cognition may be less important in a neutral inquiry e-mail than in a complaint. Future studies can investigate whether our results apply to other types of writing prompts.

Furthermore, we acknowledge that, with a large number of hypothesis testing, the possibility of observing difference by chance (i.e., the overall Type I error rate) increases. To find a balance between construct representation and number of linguistic features to be investigated, we have focused on those theoretically related to the targeted writing construct and further reduced the number of features for comparison by excluding those not varying across the writing samples or not contributing to the writing scores. Also, we reported the effect sizes to assist the interpretability of the results. We hope this study provides a first step to looking into gender DIF effect through the lens of the linguistic features of writing samples. Based on the findings of the present study, future studies could test more specific research hypotheses or focus on some of the identified indices to better control the overall Type I error rate.

Finally, it is important to be aware that the DIF effect of a writing prompt can be attributed to a number of factors, such as the prompt, the rubric, the raters, the test takers, the test setting, and the interactions of these factors. Previous studies have focused on the features of prompts (Breland et al., 2004) and the effects of raters (Lumley, 2002); the present study has provided a new angle—the linguistic features of writing samples—to seek for explanations of the DIF effect flagged by statistical methods. Future research could look into how other factors and their interactions may lead to a DIF effect in writing tests. Such investigations will extend our understanding of potential sources of DIF, which go beyond the item and test features.

Despite these interpretive considerations, our findings showed that the responses by female and male writers to the same prompt can differ in a limited number of linguistic features. The manifestation of the gender differences, however, is found to be varying across linguistic features. This implies that test developers and users should be aware of the "value statement" brought in by a rating rubric. Depending on which linguistic features are valued in a rating rubric, the scores may be potentially biased against a gender group. For example, if cohesion features are disproportionately privileged in a rating rubric, compared with another scale that is balanced between cohesion and syntactic features, then, this rubric is more likely to induce gender-related DIF effect. Overall, the combination of corpus-based analysis and quantitative DIF methods can be a valuable addition to more traditional approaches to detecting sources of DIF effects. In the following paragraphs, we discuss the findings in relation to the two research questions.

Gender-Related Differences in Writing Features

The first research question concerns gender-related differences in language use on a DIF writing prompt. The results regarding the four categories of language features confirmed some of the previous findings and added new insights into gender differences in writing.

Among the language features explored in this study, some cohesion and syntactic measures showed gender differences. In terms of cohesion, the variations between the gender groups appear in one feature of connectives and one pronoun-related feature. These characteristics, to some extent, echoed the differences found in the previous studies (Rubin and Greene, 1992; Argamon et al., 2003). The e-mails written by the female test takers in this study outscored those written by the male test takers on both cohesion indices, suggesting that the writing of the females was more cohesive through more frequent uses of more coordinating conjunctions and pronouns. Nevertheless, as pointed out by a reviewer, using more coordinating conjunctions and pronouns does not necessarily make the writing samples more coherent. Overly relying on such explicit cohesive devices may add redundancy and make the writing unnatural. Indeed, the negative correlations between the two cohesion indices and writing scores suggest that highly proficient writers are less likely to rely on these features to achieve coherence.

Four syntactic features were found to be different between gender groups with the female test takers outscored their male counterparts on all four features. This trend toward more sophisticated language is somewhat consistent with the general perception of female writing (Gudykunst and Ting-Toomey, 1988). Our findings demonstrate that the female test takers packaged more information at the noun phrase and clause levels with more frequent use of structures like multiple complex nominals per clause. However, Rubin and Greene (1992) noted that writers with a masculine gender role orientation tended to use more complex sentence structures, which contradicts the evidence of this study. The contradiction may be explained by the difference in the writing genres (university academic writing tasks vs. personal e-mails). We also found that the female test takers' e-mails had a larger SD in delta P scores, which suggests that female test takers used structures that showed a larger variability in the strength of association, as measured in delta P scores in reference to all the written corpora in COCA. This may help clarify that the sophistication levels of the language used by all test takers were reflected in their adoption of the more common VAC structures or lexical items employed by native speakers of English (Kyle, 2016).

The present study did not identify gender-related differences in lexical, sentiment, and social cognition features. Although statistically non-significant, the writing of the female test takers showed marginally lower lexical sophistication, with higher MRC familiarity scores as well as higher scores of word frequency and more regular use of trigrams in reference to the BNC written registers.

With regard to sentiment and social cognition features, previous studies suggested that female test takers tended to use more reference to emotion and judgmental adjectives (Mulac and Lundell, 1994) and employ more personal pronouns to refer to themselves, which tends to render their writing more narrative (Rubin and Greene, 1992). Although these differences were not statistically significant in this study, we observed similar patterns showing that female test takers were slightly more likely to use personal pronouns that refer to themselves and use emotion-related words for both negative and positive feelings.

Considering the large number of linguistic features analyzed, the proportion of those that were distinctive is rather small. Some of these features have been confirmed in previous studies (e.g., use of pronouns), while new ones may be considered in future studies on gender-related linguistic features. However, we need to bear in mind that some of the linguistic features identified as distinctive may be more relevant to the writing task (e-mail writing) or environment (writing on a computer and under a time constraint) in this study.

Language Differences and Writing Performance

The second question concerns whether the identified gender-related linguistic features contributed to divergent writing performance between the gender groups. The correlational information between these features and the CELPIP-General writing levels sheds light on the small gender DIF effect observed on the writing scores.

Most of the relationships between the gender-related language features and writing performance are in line with theoretical expectations of the writing construct (see **Table 2**). We hypothesized that the sentiment and social cognition features would contribute to performance on the CELPIP-General writing test with regard to task fulfillment, which includes the relevance of the content, completeness, tone, and length of the text (Paragon Testing Enterprises, 2015). The significant correlations were found in both directions; that is, some sentiment and social cognition features were positively correlated with higher writing scores (e.g., negative sentiment), while others were negative (e.g., positive sentiment and first-person pronouns). Recall that the task was writing a complaint e-mail, the correlations between these features and the writing scores were consistent with our expectation. However, none of them was statistically different across gender. Similarly, although some lexical features were associated with writing scores (e.g., word frequency, trigram), none was significantly divergent between gender groups.

In the two groups of features, cohesion and syntactic features, where gender differences were observed, the two cohesion features, which pertained to the number of coordinating conjunctions and pronouns, were negatively correlated with the writing scores. This pattern of the correlation is consistent with findings of Crossley et al. (2016), where the authors reported that local cohesion (e.g., sentence-level overlaps of verb synonyms) and overall text cohesion (e.g., the pronoun-to-noun ratio and lemma TTR) were negatively correlated with the scores of the essays written on the Scholastic Aptitude Test (SAT) prompts. However, the global cohesion features such as overlaps of certain lexical units (e.g., adverb lemmas, all lemmas, and verb lemmas) among three adjacent paragraphs have been positively associated with writing scores (e.g., Crossley et al., 2016). Also, it has been asserted that features of global cohesion were more predictive of essay quality than local cohesion measures such as the use of connectives (Guo et al., 2013). However, due to the settings of the language proficiency exam, we could not meaningfully report or compare the features measuring cohesion between paragraphs. The large-scale language proficiency test allows test takers limited time to develop their writing responses

(27 min in this case), which leads to short writing samples (mean = 189 words and SD = 38.20). In particular, many test takers submitted a single-paragraph writing sample, which is not uncommon for an e-mail writing task.

Among the four syntactic features that are different between the gender groups, three were positively correlated with writing proficiency and one was negatively correlated with it. Although the configuration of correlations was roughly as expected, compared with other studies, a somewhat unique set of features was associated with this particular writing task. Except for one of the positively correlated features that has been reported in previous studies (i.e., complex nominals per clause, see Lu, 2010), all the others were either not investigated directly or were not found to be related to writing scores. Particularly, the possessive pronoun–related feature is unique to this study; this may have reflected the wide use of possessive pronouns in the e-mail samples. Interestingly, for the delta p scores, a feature based on association strength, its variation exerted more influence on the writing scores than the trait itself.

Overall, female test takers consistently outscored their male counterparts on all the distinctive features identified in the present study. These distinctive language features, however, varied in the magnitude and direction of their correlation coefficients (i.e., from −0.324 to 0.143) with writing performance, suggesting that some of these language features contribute to higher writing scores while others are associated with lower scores. These findings imply that the writing construct of this test, as operationalized through the writing task and the rating rubric is not heavily impacted by the clusters of the linguistic features associated with a gender group. When taken together, they may give an edge to the female test takers, whose texts showed more of these features than those of their male peers. However, given the small to moderate effect sizes of the correlation coefficients, the impact of gender-related differences on the scores was probably minimal. Still, it is worth noting that for female test takers whose profile of linguistic features had more occurrences of the positively perceived features and fewer of the negatively perceived, their advantage in writing scores may be more pronounced.

CONCLUSION

In sum, this study examined the linguistic features of responses to a writing prompt that was flagged showing small gender DIF favoring female test takers—which is a typical finding in the writing DIF studies. Despite the limitations acknowledged at the beginning of section "Discussion," this study demonstrated an additional way for further exploring and understanding the DIF results based on statistical analyses of scores. The finer distinction of dissimilar linguistic features in this corpus-based study provides a good opportunity to examine gender-related differences in greater depth. This approach can be used in other writing tests and hopefully, it will help language testers interpret and explain the DIF effects in large-scale standardized writing tests.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available as the writing samples analyzed in this study were responses to a high-stakes language proficiency test. The samples and the test materials are the properties of the test publisher, Paragon Testing Enterprises.

ETHICS STATEMENT

Ethics review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to share their anonymised responses for research purposes.

AUTHOR CONTRIBUTIONS

ZL and MC designed the study, conducted analysis, and wrote the manuscript. JB contributed to the critical revision of the manuscript and assisted with study concept or design.

REFERENCES

Alderson, J. C. (1990). Testing reading comprehension skills (Part Two): getting students to talk about taking a reading test (A pilot study). *Read. Foreign Lang.* 7, 465–503.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text Interdiscip. J. Stud. Discour.* 23, 321–346. doi: 10.1515/text.2003.014

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Lang. Asses. Q.* 13, 1–24. doi: 10.1080/15434303.2015.1133626

Breland, H., and Lee, Y.-W. (2007). Investigating uniform and non-uniform gender DIF in computer-based ESL writing assessment. *Appl. Meas. Educ.* 20, 377–403. doi: 10.1080/08957340701429652

Breland, H., Lee, Y.-W. W., Najarian, M., and Muraki, E. (2004). *An Analysis of TOEFL CBT Writing Prompt Difficulty and Comparability for Different Gender Groups* (TOEFL Research Reports No. RR-04-05). *ETS Research Report Series*, Vol. 2004. Princeton, NJ: Wiley Online Library, doi: 10.1002/j.2333-8504.2004.tb01932.x

Broer, M., Lee, Y.-W., Rizavi, S., and Powers, D. E. (2005). *Ensuring the fairness of GRE writing prompts: Assessing differential difficulty: ETS Research Report Series*. Princeton, NJ: Wiley Online Library.

Centre for Canadian Language Benchmarks (CCLB) (2012). *Canadian Language Benchmarks: English as a Second Language for Adults*. Available online at: https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf (accessed October 1, 2019).

Chen, M. Y., Lam, W., and Zumbo, B. D. (2016). "Testing for differential item functioning with no internal matching variable and continuous item ratings," in *Poster Presented at the Language Testing Research Colloquium*, Palermo. Available online at: https://www.paragontesting.ca/wp-content/uploads/2018/12/2016-LTRC_Chen-M.-Y.-Lam-W.-Zumbo-B.-D.pdf (accessed July 30, 2019).

Chen, M. Y., Liu, Y., and Zumbo, B. D. (2020). A propensity score method for investigating differential item functioning in performance assessment. *Educ. Psychol. Meas.* 80, 476–498. doi: 10.1177/0013164419878861

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): automatic assessment of local, global, and

text cohesion. *Behav. Res. Methods* 48, 1227–1237. doi: 10.3758/s13428-015-0651-7

Crossley, S. A., Kyle, K., and McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): an automatic tool for sentiment, social cognition, and social-order analysis. *Behav. Res. Methods* 49, 803–821. doi: 10.3758/s13428-016-0743-z

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., and Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educ. Meas. Issues Pract.* 29, 24–35. doi: 10.1111/j.1745-3992.2010.00173.x

Ferne, T., and Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: methodological advances, challenges, and recommendations. *Lang. Assess. Q.* 4, 113–148. doi: 10.1080/15434300701375923

Geranpayeh, A., and Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced english examination. *Lang. Assess. Q.* 4, 190–222. doi: 10.1080/15434300701375758

Gudykunst, W. B., and Ting-Toomey, S. (1988). *Culture and Interpersonal Communication*. Newbury Park, CA: Sage.

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. Writ.* 18, 218–238. doi: 10.1016/j.asw.2013.05.002

Jones, S., and Myhill, D. (2007). Discourses of difference? Examining gender differences in linguistic characteristics of writing. *Can. J. Educ.* 30, 456–482. doi: 10.2307/20466646

Kunnan, A. J. (2000). "Fairness and justice for all," in *Fairness and Validation in Language Assessment*, ed. A. J. Kunnan (Cambridge: Cambridge University Press), 1–13.

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Doctoral dissertation, Georgia State University, Atlanta, GA.

Kyle, K., and Crossley, S. A. (2015). Automatically assessing lexical sophistication: indices, tools, findings, and application. *TESOL Quarterly* 49, 757–786. doi: 10.1002/tesq.194

Leaper, C., and Robnett, R. D. (2011). Women are more likely than men to use tentative language, aren't they? A meta-analysis testing for gender differences and moderators. *Psychol. Women Q.* 35, 129–142. doi: 10.1177/0361684310392728

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *Int. J. Corp. Ling.* 15, 474–496. doi: 10.1075/ijcl.15.4.02lu

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

Mulac, A., Bradac, J., and Gibbons, P. (2006). Empirical support for the gender-as-culture hypothesis: an intercultural analysis of male/female language differences. *Hum. Commun. Res.* 27, 121–152. doi: 10.1111/j.1468-2958.2001.tb00778.x

Mulac, A., and Lundell, T. L. (1994). Effects of gender-linked language differences in adults' written discourse: multivariate tests of language effects. *Lang. Commun.* 14, 299–309. doi: 10.1016/0271-5309(94)90007-8

Pae, T.-I. (2011). Causes of gender DIF on an EFL language test: a multiple-data analysis over nine years. *Lang. Test.* 29, 533–554. doi: 10.1177/0265532211434027

Paragon Testing Enterprises (2015). *CELPIP Study Guide: Reading and Writing*. Vancouver, BC: Paragon Testing Enterprises.

Park, K. (2014). Corpora and language assessment: the state of the art. *Lang. Assess. Q.* 11, 27–44. doi: 10.1080/15434303.2013.872647

Riazi, A. M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: an exploration of textual features. *Assessing Writ.* 28, 15–27. doi: 10.1016/j.asw.2016.02.001

Rubin, D. L., and Greene, K. (1992). Gender-typical style in written language. *Res. Teach. Engl.* 26, 7–40.

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

Welch, C. J., and Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: problems and an example. *J. Educ. Meas.* 32, 163–178. doi: 10.1111/j.1745-3984.1995.tb00461.x

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

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Lang. Test.* 34, 565–577. doi: 10.1177/0265532217720956

Zwick, R., Donoghue, J. R., and Grima, A. (1993). Assessment of differential item functioning for performance tasks. *J. Educ. Meas.* 30, 233–251. doi: 10.1111/j.1745-3984.1993.tb00425.x

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 declared a past collaboration with the authors.

frontiers in Psychology

PERSPECTIVE
published: 19 June 2020
doi: 10.3389/fpsyg.2020.01407

The Input Matters: Assessing Cumulative Language Access in Deaf and Hard of Hearing Individuals and Populations

*Matthew L. Hall**

Department of Communication Sciences and Disorders, Temple University, Philadelphia, PA, United States

Deaf and hard-of-hearing (DHH) children present several challenges to traditional methods of language assessment, and yet language assessment for this population is absolutely essential for optimizing their developmental potential. Whereas assessment often focuses on language outcomes, this Conceptual Analysis argues that assessing cumulative language *input* is critically important both in clinical work with DHH individuals and in research/public health contexts concerned with DHH populations. At the individual level, paying attention to the input (and the person's access to it) is vital for discriminating disorder from delay, and for setting goals and strategies for reaching them. At the population level, understanding relationships between cumulative language input and resulting language outcomes is essential to the broader public health efforts aimed at identifying strategies to improve outcomes in DHH populations and to theoretical efforts to understand the role that language plays in child development. Unfortunately, several factors jointly result in DHH children's input being under-described at both individual and population levels: for example, overly simplistic ways of classifying input, and the lack of tools for assessing input more thoroughly. To address these limitations, this Conceptual Analysis proposes a new way of characterizing a DHH child's cumulative experience with input, and outlines the features that a tool would need to have in order to measure this alternative construct.

Keywords: communication mode, deafness, early intervention, family language planning, language access profile, language assessment, sign language, speech-language pathology

OPEN ACCESS

Edited by:
Thomas Eckes,
Ruhr-Universität Bochum, Germany

Reviewed by:
Amy R. Lederberg,
Georgia State University, United States
Christine Yoshinaga-Itano,
University of Colorado Boulder, United States
Donna Jo Napoli,
Swarthmore College, United States

***Correspondence:**
Matthew L. Hall
matthall.research@gmail.com

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

Received: *17 December 2019*
Accepted: *26 May 2020*
Published: *19 June 2020*

Citation:
Hall ML (2020) The Input Matters: Assessing Cumulative Language Access in Deaf and Hard of Hearing Individuals and Populations. Front. Psychol. 11:1407. doi: 10.3389/fpsyg.2020.01407

INTRODUCTION

Deaf and hard-of-hearing (DHH) children present several challenges to traditional methods of language assessment, and yet language assessment for this population is absolutely essential for optimizing their developmental potential. The Joint Committee on Infant Hearing has been recommending routine and recurring language assessment for DHH children for at least the past 20 years (Joint Committee on Infant Hearing, 2000, 2019; Muse et al., 2013). In DHH populations, language assessment contributes to two important goals that can sometimes seem disconnected from one another: (1) optimizing the outcomes of an individual DHH child, and (2) optimizing the outcomes of the entire population from which the DHH child is sampled.

This latter goal constitutes a public health objective, the achievement of which goes beyond any individual clinician's responsibility. However, because public health data are typically aggregated over large corpora of individual assessment results, and those assessments are usually carried out by clinicians, the two are inextricably linked. This Conceptual Analysis argues that considering a DHH child's language input is vital at both scales.

At the individual level, considering the child's language input provides necessary context for understanding and interpreting assessment results. Language delays are common in DHH children, but for an individual child, these delays can either be unsurprising and (relatively) unimportant, unsurprising but important, or surprising and important (whether positive or negative). Characterizing the child's cumulative experience with language input helps us differentiate these possibilities, and calls attention to recommendations that might otherwise be overlooked.

At the population level, it is essential to identify malleable factors that can optimize a child's developmental potential so that evidence-based recommendations can be presented to future generations. For DHH children, the input that is in their environment throughout infancy and toddlerhood is a malleable factor of major importance; however, we still lack useful information about what kinds of early experiences with input are most likely to maximize language outcomes. A major reason for the absence of such information is the sheer complexity and diversity of DHH children's experiences with linguistic input during the critical language-learning years of infancy and toddlerhood. This poses serious challenges for both clinicians and researchers, as explored below.

In clinical settings, time is precious. Although professional best practices encourage clinicians to take thorough language histories in early intervention contexts (American Speech-Language-Hearing Association [ASHA], 2008), the need to perform a diagnostic assessment may be considered a more urgent priority, especially if a child is participating in a program where assessment outcomes inform the child's continuing eligibility for services, school placement, IFSP/IEP goals, etc. The amount of time devoted to gathering a language history may therefore be very limited, if one happens at all. And because education about the importance of gathering language histories is often provided with respect to multilingual populations (e.g., American Speech-Language-Hearing Association [ASHA], 2010), the clinician may not believe that collecting a thorough, cumulative history from a monolingual family is worth the time. However, even DHH children from monolingual homes have considerably diverse experiences with language input, as the following section on "language exposure" vs. "language access" will explain.

A second significant problem is that even if clinicians are committed to collecting comprehensive data about a DHH child's cumulative experience with linguistic input, they have no empirically tested tools with which to do so. The only formal, research-based tools that are presently available are all developed for multilingual children from hearing families (e.g., LEAP-Q, Marian et al., 2007; BESA, Peña et al., 2018; LEAT, DeAnda et al., 2016, inter al.). Although such tools offer useful frameworks for thinking about input, they would need careful adaptation before becoming suitable for use with DHH populations. But until the need for such tools is more widely appreciated, there is little incentive for them to be developed or used.

In the meantime, clinicians may use informal assessments/interviews, and may even have the opportunity to observe the child's current input in naturalistic settings. But this raises a third problem: using observational language samples to understand the nature of a child's input is only a valid approach when both the input and the child's access to it have remained fairly constant throughout the child's life. For DHH children, neither can be assumed: children's auditory access to spoken language often changes over time, as does their interlocutors' use of and proficiency in various forms of manual communication. Thus, strategies that serve SLPs well when working with hearing children often do not suffice for clinical work with DHH children. Current technology offers no easy solutions: no automated tools for characterizing visual input are available or even on the horizon; nor would their sudden appearance allow for a cumulative history to be obtained. As a result of this constellation of factors, assessment of DHH children's cumulative experience with linguistic input is often limited in clinical contexts, despite the well-established understanding that language input plays a pivotal role in language acquisition.

In the research literature, one strategy has been to rely on recent advances in technology such as Language ENvironment Analysis (LENA) software, which records and to some extent categorizes the auditory input in a child's environment. In DHH populations, this approach is becoming more common as a way of linking a child's language outcomes to their language input (e.g., Aragon and Yoshinaga-Itano, 2012; Van Dam et al., 2012; Wiggin et al., 2012; Suskind et al., 2013, 2016; Ambrose et al., 2014, 2015; Sacks et al., 2014; Vohr et al., 2014). However, it is imperative to understand that LENA systems are inherently limited in the insights that they can offer. First, they provide no data at all about the child's experience before they received and began using a LENA system. Thus, unless a LENA device has been used since birth, there is no way for this information to provide information about a child's cumulative history of linguistic input. Second, LENA systems provide information about auditory input only: they are entirely insensitive to any form of visual communication. Accordingly, LENA systems have no way to differentiate spoken input that is produced without manual accompaniments from spoken input that is accompanied by either signs or cues (which would also not be distinguished from one another). LENA systems would also interpret periods of silence as the absence of input, even if a sign language were being used. Thus, LENA systems are wholly incapable of assessing a DHH child's experience with non-auditory forms of input, which in turn precludes any progress in understanding how difference experiences with such forms of input relate to subsequent language outcomes. Third, a LENA system knows only what *it* hears, which is not the same as what a DHH *child* hears. In order for LENA data to be a valid representation of the child's auditory access, a separate process would need to be implemented that links the LENA recording to datalogging from a child's hearing technology, which is itself only an approximation and

insensitive to the extent of a child's residual hearing. Moreover, even if LENA data were appropriately integrated with datalogging and adjusted for child-specific hearing profiles, the former two problems remain. Thus, although LENA data can illuminate some aspects of the relationship between language input and language outcomes, they cannot document a child's cumulative history of access to various types of input.

A second response to the complexity and diversity of DHH children's cumulative experiences with input has been to rely on the construct of "communication mode" as a proxy for describing DHH children's cumulative experience with language input. Unfortunately, this construct is typically used in ways that are too simplistic to reflect children's actual experiences, and too variable across studies to support meaningful generalization (Hall and Dills, 2020). This Conceptual Analysis argues that clinicians and researchers must reconsider the ways that we assess DHH children's input, adopting methods that recognize its diverse and multidimensional nature throughout the crucial language-learning years of infancy and toddlerhood, and take dose-response functions into consideration.

This manuscript provides only a high-level conceptual overview of what the alternative construct should look like; the primary goal is to underscore the importance of routinely collecting information about DHH children's cumulative experience with language input, not just their language outcomes (or current experience with input), when performing language assessment.

Language Exposure vs. Language Access

Before proceeding, it is necessary to introduce a conceptual distinction that may be new to some readers, particularly those who do not regularly work with or think about DHH populations: namely, the distinction between language *exposure* and language *access*. No child ever learns a language that they are not exposed to. But for DHH populations, language *exposure* (i.e., the presence of input in the child's environment) is not enough. What is necessary is *access*: that is, the child must be able to perceptually receive and cognitively process the signals that are being sent. This distinction is not a new one; Moeller and Tomblin (2015) refer to this distinction as "language input" versus "language experience," and Harris (2013) distinguishes "language input" versus "language uptake." Despite the variations in terminology, the core idea is that for DHH children, it is not enough to simply consider what kinds of linguistic signals are being *sent to* a child. Instead, it is necessary to think about the linguistic signals that that child is *receiving*. Moeller and Tomblin (2015) identify several factors that influence a DHH child's auditory access to spoken input: aided audibility (including appropriate fitting of hearing aids and mapping of cochlear implants), consistent use of hearing technology, and the nature of the linguistic input in the child's environment (quantity and quality). This model can easily be extended to encompass visual forms of communication as well, which is perceptually accessible to DHH children without technology (except in children who also have reduced vision). For the remainder of this paper, unless otherwise noted, the term "input" should be understood as referring to all and only those linguistic signals to which a child has access: whether auditory or visual. Note that it is also possible for a DHH child to have only limited access to any linguistic input; indeed, it is this state of having limited access to input (rather than deafness itself) that creates developmental risk (Hall et al., 2019). More attention is given to this notion of limited access at the population level; first, we consider the importance of assessing cumulative experience with input at the individual level.

Language Input Matters at the Individual Level

Discriminating Disorder From Delay/Difference

At the group level, language skills in DHH children are often found to be, on average, between 1 and 2 standard deviations below those of test norms (which almost invariably represent monolingual children with typical hearing) or demographically matched hearing controls (Koehlinger et al., 2013; Tobey et al., 2013; Ambrose et al., 2014, 2015; Tomblin et al., 2015; Eisenberg et al., 2016; Geers et al., 2017; Lewis et al., 2017; Hoffman et al., 2018; Lederberg et al., 2019; Antia et al., 2020). Of course, there is considerable heterogeneity at the individual level, and clinicians are charged with supporting one child at a time: assessing their current level of proficiency, making informed inferences about the reasons behind specific areas of weakness or strength, devising individualized interventions, and making recommendations to the child's family and the other allied professionals on the child's team. In the all-too-common event that a child shows language skills that are far behind their chronological age and cognitive potential, one important question is whether this represents a delay/difference or a true language disorder[1]. One way to address that question is to consider growth over time; if the child is making one year's worth of progress in one year's time, then there is little concern about a language disorder. However, this approach requires the passage of time, which is a precious resource in early childhood. Dynamic assessment (Gutiérrez-Clellen and Peña, 2001) is an alternative that is commonly practiced with culturally- and linguistically diverse populations (e.g., Rosemary et al., 1996; Gillam Ronald and Peña Elizabeth, 2004), but has not yet been widely adopted for DHH populations, despite calls to do so (Mann et al., 2014). A third and also-underutilized strategy for discriminating disorder from delay/difference is to consider the child's input. Doing so helps reveal whether the observed outcomes are unsurprising and unimportant, unsurprising but important, or surprising and important.

Unsurprising and unimportant

Hearing children who are successfully acquiring more than one language often appear to score lower on language-specific assessments or to meet language-specific milestones later than

[1]This manuscript considers a language disorder to be a condition that would have compromised a child's ability to acquire a language even in the presence of plentiful and accessible high-quality input.

their monolingual peers (e.g., Hoff et al., 2012). However, it is now widely understood that such putative differences may in fact be epiphenomenal: reflecting weaknesses in a tool's ability to gauge a multilingual child's true developmental state, rather than revealing a meaningful problem (e.g., Pearson et al., 1993). What makes this situation of little concern is evidence that the child's knowledge in the two (or more) languages is complementary and mutually reinforcing, together with evidence that the child is meeting the kinds of milestones that are not language-specific (e.g., increasing MLU, turn-taking, fast mapping, etc.). Such a child is likely to develop age-appropriate command of both languages prior to school entry.

Deaf and hard-of-hearing children may fall into this category if they have had good perceptual access to multiple languages (spoken or signed), such that the primary reason that they fall behind monolingual norms in one language is because they also have knowledge in another language that is not being credited. Unfortunately, this situation is uncommon; language delays in DHH children are much more likely to fall into the next category.

Unsurprising but important

Deaf and hard-of-hearing children often have reduced access to their primary language (spoken or signed) without also having access to another language. If a child has had very little access to input in a given language, then it is unsurprising to find that they are not acquiring that language as a monolingual would. But if this is the case in the child's strongest (or only) language, then the delay becomes highly important, even if its cause is unsurprising. Unlike children with access to multiple languages, it cannot be assumed that children with reduced access to one language will catch up to their typically developing peers, and the delays that they experience are not an epiphenomenon of having knowledge distributed across multiple languages. Instead, these are true delays that have true consequences, such as arriving at kindergarten without the skills needed to succeed (Hall et al., 2019).

In this case, the most straightforward approach to intervention would be to consider how to most effectively increase the child's access to input. This might include attempting to alter the child's perceptual access to the input around them, altering the input around them to be more accessible given the child's perceptual abilities, or both. To determine the most effective course(s) of action, it is important to identify the most significant barriers that have been limiting the child's access to input to date. A child with a late-identified hearing loss may simply need effective amplification. A child who has bilateral cochlear implants but only wears them inconsistently might benefit from parent counseling about strategies for increasing device use. For a child with no auditory nerve whose family refuses to use visual communication, a different kind of counseling is in order. In all cases, it is appropriate to consider what growth rate would be needed in order to achieve age-appropriate skills by school entry, and whether that growth rate is realistically attainable under the current course of action. If the answer is no, it is appropriate to consider whether there may be other courses of action worth pursuing.

Surprising and important (negative)

Although most DHH children experience reduced access to input to some degree, not all will fall into the situation described above, where their cumulative access to input has been so limited that it is an immediate red flag for intervention. A child for whom English has constituted 10% of their input would certainly fall into that category, but not a child whose input has been 90% English. But this raises an important question: at what point should reduced proficiency no longer be attributed to reduced access to input? In other words, how much access to a given language does a child need to have before we are surprised to find that they are not acquiring it?

Although there is surely no hard-and-fast answer to this question, research in hearing multilinguals is beginning to find that when 60% or more of a child's cumulative input has been in Language A, standard scores from monolingual norms can be used without increasing the risk of falsely diagnosing a disorder (Cattani et al., 2014). These findings suggest that if a child scores below the average range despite 60% or more of their input consisting of access to Language A, clinicians are justified in suspecting that something is amiss, and that it is more than can be attributable to reduced access to Language A.

The next question at that point is whether the locus of the problem is within the child's mind or in the child's environment. Here is where dynamic assessment using a range of communication methods is most useful. If the child shows little modifiability across any type of input (signed or spoken), a language disorder may be indicated, and language therapy can be designed accordingly. If, however, the child is responsive to some types of communication, then a language disorder is unlikely and a shift in communication strategy may be warranted.

Surprising and important (positive)

There is another type of surprising and important finding: this time in a positive direction. A child may show surprisingly good command of a language that has constituted only a small proportion of their cumulative experience. This might be revealed through dynamic assessment as described above, but it would ordinarily be missed through static assessment that does not thoroughly characterize the child's cumulative experience with linguistic input. In many cases, only the child's strongest language is assessed; in contrast, input-informed assessment involves evaluating all of the languages that the child is acquiring and considering the observed degree of proficiency in relation to their prevalence in the child's input.

For instance, suppose a child is evaluated in two languages, with a standard score of 80 in Language A and 70 in Language B.[2] A typical outcomes-focused approach to assessment would likely note that the child is doing better in Language A and might recommend prioritizing that language on the assumption that the child has a smaller gap to close there, and that it will accordingly be easier to do so. However, it is not necessarily valid to assume

[2]Results are being presented as standard scores for convenience; it is assumed that a thorough evaluation would not rely entirely on standardized assessments.

that a smaller gap will be easier to close. Suppose we learn that the child's input to date has consisted of 70% Language A and 30% Language B. In this case, the child's proficiency in Language B is both surprising (given how little access there has been to date) and important (in that the child might actually have an easier time closing the gap if more of their input were in Language B). If the child's scores were equivalent in both languages, this would be even more clear – but only if the clinician had information about the child's cumulative experience with linguistic input.

Setting and Tracking Input-Related Goals

In the United States, families of DHH infants and toddlers typically receive early intervention services as part of an Individualized Family Service Plan under Part C of the Individuals with Disabilities Education Act. These plans involve setting specific and measurable goals, most of which are functional in nature (e.g., "Adrian will say which snack he prefers using spoken words."). To facilitate the child's progress toward (or past) these goals, professionals on the IFSP team may include additional goals. For example, where language outcomes are concerned, a speech-language pathologist will likely have identified certain outcome goals that the child is working toward (e.g., more utterances with an MLU > 4, clearer articulation of fricatives, more conversational turns, etc.), and will routinely monitor the child's progress toward those outcomes. However, it can often be valuable to set goals related to the child's *input*. For example, if the child/family struggles with consistent use of hearing technology, a goal might be that the child keep their hearing aids on and working for no less than 8 hours every day. This goal can be tracked using datalogging from the child's hearing devices, but other goals require different approaches to measurement. For example, consider a child who is acquiring more than one spoken language. The SLP may realize or suspect that the family is not providing sufficient input in one language or the other to support the child's acquisition of both, and so might set a goal that for the next 6 months, the child's input consist of nothing less balanced than a 60–40% split, or might recommend creating a family language plan in which the lesser-used language is boosted to at least 3 hours a day. The SLP can then gather information about whether this goal is being met, whether through naturalistic observation, collecting language samples, administering surveys about language use, or conducting conversational interviews with the child's caregivers. The same can be true if, for instance, a family intends their DHH child to become a proficient user of some form of manual communication. If that goal is to be achieved, the child will need to have appreciable amounts of input in that type of communication, and it will need to be tracked across all of the contexts in which the child spends significant amounts of time. Having this information is helpful whether the child is making good progress toward their outcome goals or not. If they are, the family may be informed that their efforts are paying off and be encouraged to maintain their effort. Or, if the family had decided to limit use of their home language in order to support the child's eventual language of education, they might benefit from knowing that their child is doing well enough in the dominant language that they can start using their home language more without compromising the child's success. And if a child is struggling, it is essential to know what the cumulative input has been like in order to determine whether this outcome is unsurprising and unimportant, unsurprising but important, or surprising and important, as discussed above. In the absence of information about the child's cumulative experience with language input, appropriately setting and tracking goals becomes much more difficult.

LANGUAGE INPUT MATTERS AT A POPULATION LEVEL

Despite many cases that would be considered successes at the clinical level (i.e., one child at a time), DHH children as a population remain at serious risk of not developing age-appropriate proficiency in any language by the time they enter school. The lack of true population-based datasets in the United States makes it difficult to know for certain, but large, multi-site/multi-state studies such as CDaCI, OCHL, and NECAP typically report language outcomes in DHH children that are 1–2 standard deviations below their hearing peers, or language quotients below the 80% threshold (Koehlinger et al., 2013; Tobey et al., 2013; Ambrose et al., 2014, 2015; Tomblin et al., 2015; Eisenberg et al., 2016; Geers et al., 2017; Lewis et al., 2017; Hoffman et al., 2018; Yoshinaga-Itano et al., 2018). A separate and more recent study of over 336 DHH children between kindergarten and second grade reported similar outcomes on measures of spoken language, with mean scores again ranging from 1 to more than 2 standard deviations below the normative mean (Lederberg et al., 2019; Antia et al., 2020). These values are commensurate with the findings of a large, longitudinal, population-based study in Australia (LOCHI; see Ching et al., 2010, 2018, for language outcomes at age 3 and 5, respectively). Equally concerning are recent findings from Norway (Wie et al., 2020), where all but two of the deaf children who received early, simultaneous, bilateral cochlear implants were followed from implantation through elementary school. Although not a large n, the data represent virtually the entire population. These were children who had no additional disabilities and received early intervention services focusing on spoken language acquisition, and therefore represent the most optimistic outcomes scenario. The authors reported that although these deaf children appeared to be closing the gap with their hearing peers as they approached school entry, gaps in receptive vocabulary and expressive grammar reappeared and remained present for the duration of the observation period (up to 6 years post-implantation). Outcomes such as these suggest that roughly half of DHH children with bilateral hearing loss[3] -even those without additional diagnoses that might impede language acquisition- are not developing age-appropriate language skills. Clearly, the *status quo* is not allowing DHH children as a group to flourish. Indeed, even those DHH children who

[3]Children with unilateral hearing loss were not included in these studies.

score above the 16th percentile are likely underperforming their true potential.

To those who are accustomed to working with individual children, especially in clinical contexts, it may be tempting to apply the same standards of success to populations. However, to do so is to make a serious mistake. In clinical assessment, it is commonly and correctly understood that although a population may be defined as having a certain expected score on average (e.g., standardized assessments), any individual sampled from the population may deviate from that score to a certain extent without raising suspicion that they may in fact have been sampled from an atypical distribution. The extent of this allowable deviation is commonly termed "the average range," and although conventions vary by discipline and instrument, plus or minus one standard deviation is a common enough criterion that it will suffice to illustrate the present point. It is perfectly reasonable to be fairly unconcerned about an individual who scores an 86 on a standardized assessment where the mean is 100 and standard deviation is 15. However, if the mean of a sample of many individuals is found to be at 86, then that population is evidencing major deviation from expectations. The reason for this seeming double standard is the Central Limit Theorem, according to which the mean of a sample will converge on the mean of the population from which it is drawn as the sample size increases (specifically, in proportion to the square root of the sample size). Therefore, if a sample contains 100 individuals, the "average range" for the mean of that sample is no longer 85 to 115; rather, for a two-tailed test at alpha = 0.05, it would be from a lower bound of 97.06 [i.e., 100–1.96*(15/sqrt(100)] to an upper bound of 102.94 [i.e., 100 + 1.96*(15/sqrt(100)]. This is precisely equivalent to a *z*-test: comparing a sample against a population distribution where the mean and standard deviation are known. Finding that the sample mean falls outside the expected range of variation licenses the inference that the population from where the sample was drawn has a different mean than the reference population; however, this only becomes meaningful if the magnitude of the deviation (i.e., the effect size) is also large. In the case of a sample mean of 86, the mean would be shifted downward by nearly one full standard deviation. Assuming that the sample distribution is normally distributed, this means that roughly 50% of the sample (and, by inference, the population from which it was drawn) would fall below the clinically defined boundaries of the "average range" for individuals. For comparison, only about 16% of individuals in the reference population would be expected to score in that range: a risk ratio of 50/16 = 3.125, which equates to a 212.5% increase in risk relative to the reference population. Unfortunately, such scores are sometimes taken as evidence of success in studies of language outcomes in DHH children (e.g., Wie et al., 2020), rather than evidence that major disparities persist.

The search for ways to better support DHH children continues. As of this writing, the American Centers for Disease Control and Prevention have issued a call for proposals in response to the need for better monitoring of language outcomes and other developmental progress in DHH children after the initial processes of hearing screening, audiological diagnosis, and referral to/enrollment in early intervention. This call draws particular attention to how little is currently known about practices that will optimize DHH children's developmental potential:

> "While collaborative efforts by CDC, states, and other partners have helped lead to the early identification of thousands of children who are D/HH each year, their developmental and language outcomes are often unknown, and these data are not routinely collected by CDC or state EHDI programs. Furthermore, it is currently unclear what actions beyond early identification should be taken by public health to help reduce adverse consequences of hearing loss and ensure that children who are D/HH are ready for success in early childhood" (Centers for Disease Control, 2020).

The call goes on to identify the key role that assessment plays in filling these knowledge gaps:

"The current lack of public health capacity to document and assess the intervention services and associated outcomes of early-identified children who are D/HH at the state and national level makes it challenging to:

- Assess the developmental progress to ensure all children who are D/HH are achieving age-appropriate milestones and are ready for success in early childhood;
- Identify strategies, in addition to those beyond early identification, to help assess and reduce adverse consequences of hearing loss;
- Assess and document the success and impact of EHDI activities across the United States" (Centers for Disease Control, 2020).

In particular, this second goal of identifying strategies to reduce the adverse consequences of hearing loss would be easier if we knew more about DHH children's cumulative experiences with linguistic input. Delayed or incomplete mastery of a first language is one of the most serious adverse outcomes that DHH children face. Although many factors influence language acquisition, the input itself is surely among the most crucial. There may be no guarantee that a child will successfully acquire a language that is present in their input, but if they lack sufficient access to a given language, we can be absolutely sure that they will *not* acquire it.

There has been no shortage of attempts to identify what kinds of early experiences with linguistic input are most likely to yield subsequent language mastery (for recent reviews, see Belzner and Seal, 2009; Fitzpatrick et al., 2016; Erbasi et al., 2017; Demers and Bergeron, 2019). However, these efforts have largely failed to yield consensus, for several reasons. First, there has been disagreement over whether success should be understood as mastery of a *spoken* language, mastery of *at least one* language, or achieving the goals that matter to the child's parents, even if those goals represent less than the child's full potential.[4] The extant research has almost exclusively adopted spoken language acquisition as the barometer of success;

[4] At the individual level, disagreements on success may also stem from the fact that hearing, speech, and language all have different standards of success, but parents may not fully grasp these distinctions.

therefore, very little is known about the factors that support successful acquisition of a sign language by children who are not among the ∼5% born to parents who are already proficient signers. Second, even when looking only at spoken language outcomes, the available results are highly mixed and based on studies of low methodological quality (Fitzpatrick et al., 2016; Demers and Bergeron, 2019). Third, and most relevant to the present argument, the very *construct* that researchers have used in an attempt to answer this question (i.e., "communication mode") is ill-defined. Hall and Dills (2020) point out that in addition to the absence of any uniform operationalization of the term, it typically does not provide any information about what a child's experience was like during infancy and toddlerhood, and it commonly conflates types of input that are very different (e.g., ASL, sign-supported speech, and manually coded English). They identify the desiderata of a better alternative and argue that until such an alternative is available, it will remain impossible to identify the kinds of strategies that the CDC rightly identifies as crucial gaps in knowledge. A high-level conceptual overview of what this new method might look like is provided below (readers interested in a more applied introduction are referred to De Anda and Hall, in prep). However, the primary goal of this section is merely to make the point that if the goal is to identify strategies for improving outcomes, then assessing outcomes alone is insufficient: assessing the input is also necessary. This section further argues that in order to be maximally useful at the population level, measures of input should support bottom-up grouping strategies, and allow exploration of dose-response relationships between language input and language outcomes.

Language Input as an Upstream Determinant of Language Outcomes

At the 2020 Early Hearing Detection and Intervention conference, keynote speaker Dr. Michael Warren (Associate Administrator of the United States Maternal Child and Health Bureau) emphasized the importance of identifying upstream causes of later outcomes. He argued that intervening on upstream factors is a more efficient and more effective approach to public health than attempting to treat problems that arise downstream. Given that language input is necessarily antecedent to language outcomes, efforts aimed at improving language outcomes should pay close attention to language input: particularly to input during infancy and toddlerhood, when the human brain acquires language most readily. However, given the aforementioned limitations of communication mode as a construct, it is worth considering the desiderata of a better measure of language input for DHH children. The following recommendations are drawn from Hall and Dills (2020).

First and foremost, a useful measure of language input should have a clear and consistently applied operational definition. This is a prerequisite for establishing generalizability across studies.

It should capture a child's cumulative experience with linguistic input over a given time window of interest. Ideally, this window would be prior to the point at which outcomes are being evaluated. There is a danger in measuring outcomes as a function of the child's *current* input, since their current situation may be a result of their language proficiency rather than a cause of it. Again, the ultimate goal of population-level outcomes is to identify upstream predictors that can inform recommendations for future generations.

A useful measure of language input should have a way to represent the extent to which a child has had limited access to linguistic input, whether it be because of late identification, delayed availability or inconsistent use of effective hearing technology, delayed onset or infrequent use of visual communication, etc. While many of these reasons may be theoretically preventable, their impact (or lack thereof) on a child's experience is still relevant for understanding that individual child's outcomes, and must be included as part of the construct. Counter-intuitive though it may seem, the necessity of including something like a "limited access" category as part of a child's input can be appreciated by considering two children whose environment consists of nothing but spoken English, of whom one gained excellent auditory access to spoken language at 9 months and the other at 27 months. Without including "limited access" as an input category, both children would appear to have 100% English. Including a "limited access" category reveals that the first child's experience has been 75% English, 25% Limited Access, while the second child has had 75% Limited Access, 25% English. Clearly, the inclusion of this category results in a more faithful representation of their experience.

An existing construct like "hearing age" would likely share variance with a measure of "limited access" for some but crucially not all children. First, "hearing age" measures the time that has elapsed since the onset of auditory access; it does not capture factors that describe the *extent of access* during that time (e.g., appropriateness of fitting/mapping, consistency of device use, and listening environment). Second, "hearing age" would only be a valid proxy for "limited access" among children who did not have access to visual communication prior to the onset of auditory access. For example, consider another hypothetical child whose family began using sign-supported speech as soon as the child referred on their newborn hearing screening, and then switched to spoken English without sign when the child's cochlear implants were activated at 9 months. By 36 months, this child's experience of auditory access to English will be the same as that of the previous child who was also activated at 9 months; however, this child would have 0% Limited Access (and 25% sign-supported speech instead).

Similarly, a construct like "age of acquisition" (more commonly used with respect to sign languages) has comparable limitations: it identifies only the point at which access to a sign language began, but provides no information about how much experience the child then had with signed input. Likewise, it provides no information about the extent to which a child did or did not have auditory access to spoken language prior to (and after) the onset of signing. Thus, the measure of "limited access" would need to be sensitive to all of these considerations.

A useful measure of language input must make distinctions among types of communicative systems that are fundamentally different. For example, cued speech provides phonological information that helps to disambiguate words that look alike while speechreading. Manually coded English systems emphasize morphosyntax by pairing every spoken morpheme with a signed equivalent. Distinct from both of those is a broader

category often called "sign-supported speech." Like the previous two, the utterances in such a system are generated by the grammar of a spoken language (e.g., English). But unlike cued speech, the manual components of this signal have semantic content. And unlike manually coded English, the manual components do not include inflectional or derivational morphemes; often, there are no function words at all. Instead, this type of communication generally involves strings of signs that correspond to selected content words in linear order. This category encompasses practices that include Conceptually Accurate Signed English, simultaneous communication, "total communication" (misnomer though it may be), and baby sign. There may certainly be value in distinguishing among these subtypes of communication; however, distinguishing sign-supported speech from manually coded English and cued speech would be a good first step in the right direction.

Even more importantly, a useful measure of language input would distinguish natural sign languages from the types of communication described in the preceding paragraph. Unlike all of those, utterances produced in a natural sign language are not generated by the grammar of a spoken language. The fact that sign languages have their own grammars seems widely recognized when describing communication options to parents, but it somehow seems to be forgotten when interpreting research that fails to distinguish natural sign languages from other forms of manual communication.

A construct of this nature would be better able to reflect the actual experiences of DHH children than the currently dominant approach of simply identifying a child's "communication mode." Although the examples given above have been purposely simplistic for the sake of convenience, a construct that had the above-described properties would be able to describe more realistic profiles: for example, a child whose input by 36 months has consisted of 40% limited access, 20% English without sign, 15% sign-supported speech, and 5% ASL. Another child might have 40% limited access, and 60% English without sign. Still another might have 10% limited access, 30% Spanish, 30% English, 15% cued Spanish, and 15% cued English. Although it is hopefully now clear how this information is clinically useful at the individual level, such heterogeneity presents challenges to researchers working at a population level, who need either categorical or continuous variables to use as predictors. The constructs described above are perfectly capable of generating continuous values for a predictor variable that focuses on one type of input at a time; however, the argument here is that such an approach might be misleading, in that putative effects of variation in one category may in fact by epiphenomena of changes in another category, since this construct is fundamentally compositional in nature. It is argued that a better approach is to develop a categorical variable whose values represent various combinations of experiences. In this way, a child's complex experience can still be represented with a single categorical value, since that value itself describes a multidimensional experience. A strategy for achieving this is described below.

Top-Down vs. Bottom-Up Grouping

Historically, research on DHH children's experience with linguistic input has involved top-down grouping strategies. That is, a researcher or policy maker makes a set of *a priori* decisions about what groups are relevant to compare, sets criteria for inclusion in those groups, and then proceeds to compare outcomes between/among those groups. Usually, this involves comparing a DHH children who use listening and spoken language exclusively against those who do not (Hall and Dills, 2020). One virtue of this approach is that it covers the entire parameter space, since every child can be characterized as belonging to either one or the other. According to recent data from the National Center for Hearing Assessment and Management [NCHAM] (n.d.) in the United States, this division also results in roughly equal-sized groups: 49% of the 303 families who responded to the survey reporting using listening and spoken language (LSL) exclusively, and 51% did not. However, the 51% reported a diverse set of experiences, including mostly LSL with some signs or cues (17%), roughly equal amount of signed and spoken communication (14%), mostly cued speech (12%), mostly signing with some speech (3%), sign language only (3%), and other (1%). Treating these children as if they all had the same experience with language input precludes the possibility of discovering subsets of children within this group that might have stronger language outcomes than others.

It may be tempting at this point to propose that a better solution might be to simply divide the 51% into smaller groups like those listed above; however, this too has problems, as noted above. Rather than attempting to refine the top-down categories, a better solution may be to abandon them entirely, in favor of bottom-up, data-driven grouping strategies in which DHH children's idiosyncratic and multidimensional experiences are represented as the complex constructs that they truly are. Grouping variables can be discovered through the application of classification algorithms such as hierarchical cluster analysis, latent profile analysis, or related methods. These approaches entail no *a priori* assumptions about what the relevant groups will be; instead, they identify sub-groups of children who have had similar experiences to one another, but different experiences than other sub-groups. A virtue of this approach is that it creates groups that are more internally homogeneous while also reflecting the reality that DHH children's experiences with input are frequently multidimensional. Crucially, this approach can also accommodate information about the extent to which DHH children have lacked access to any form of input. There is of course no guarantee that the resulting profiles will cover the entire parameter space: however, this too turns out to be a virtue, in that it draws attention to areas of the parameter space that are not yet represented in the dataset and therefore potentially worth exploring.

Dose-Response Functions

In healthy adults seeking relief from headache pain, the recommended dosage of aspirin is 300–600 mg every 4–6 hours. If someone takes only 100 mg a day and finds that their headache persists, they are not justified in concluding that aspirin is ineffective at relieving their headache pain. Meanwhile, if someone is taking 600 mg every 4 hours and the headache persists, then they would be justified in concluding that they might benefit from exploring other medications. The same reasoning applies to the relationship between language input and

language outcomes. If Language A has constituted only 10% of a child's input, it would be unsurprising to find that the child has not mastered Language A. But it would also be unjustified that therefore Language A does not benefit the child: rather, it has not been given a reasonable chance to succeed. However, if Language A has constituted upward of 60% of the child's input and the child is not showing age-appropriate language skills, then it does stand to reason that the child -and others like them- may derive greater benefit from other types of input. It would also be important to determine whether the dose-response function is different for these children. For instance, it is possible that some DHH children would respond well to Language A, but only if it constitutes 85% or more of their input. It is also possible that even at this level, DHH children would still struggle to master Language A.

Unfortunately, extant research provides essentially no information about the dose-response relationship for various types of language input. One justifiable reason for this is the multidimensional nature of DHH children's experiences, as described above: there may not be a monotonic relationship between amount of Language A and outcomes in Language A, because different types of input that are Not-A might have different effects. This is the primary justification for treating language input as a categorical rather than continuous predictor, provided that the levels of the categorical variable themselves represent multidimensional values.

More problematic than the absence of this dose-response information is the notion -implicit or explicit- that this information is in fact already known. This notion can surface in many forms. For example, an ASL advocate might promise a hearing family that their child can master ASL even if the primary source of ASL input is parents who are themselves novice learners. Or, an LSL advocate might counsel a family that signing is going to hurt their child's chances of developing spoken language. Empirical evidence exists that is consistent with both of these claims (e.g., Percy-Smith et al., 2010; Allen, 2015; Henner et al., 2016; Geers et al., 2017); however, it is important to recognize that such studies occupy only one individual point somewhere along the broader dose-response function. As such, they cannot appropriately be generalized to other points along the continuum; unfortunately, such overgeneralizations appear to be commonplace. There does not appear to be any research that thoroughly documents the nature of the dose-response function between language input and language outcomes in DHH children. A major reason for this is the historical lack of methods for adequately characterizing language input. Developing and implementing such methods is therefore crucial to the goal of addressing questions such as the priorities identified by the CDC above. If language outcomes are measured but language input is not, how are we ever to know what kinds of input result in the best outcomes?

CONCLUSION

Typically, language assessment focuses on language outcomes. As Moeller and Tomblin (2015) note, this is in part a reflection of theoretical traditions in which variation in linguistic input was thought to play only a minor or peripheral role in language acquisition. It is also a reflection of the tendency, at least in the United States, to treat white, middle class, monolingual children with no disabilities as the default standard to which all other children should be compared. Because such children have largely homogeneous distributions of language input, describing the child's language input was not historically considered essential for understanding language outcomes. More recent work with culturally- and linguistically diverse populations has drawn the field's attention to the importance of these factors, and to the associated drawbacks of relying too much on standardized assessments in clinical practice. Unfortunately, clinical work at the individual level has not always translated these concepts into practice in the most appropriate ways. Meanwhile, work at the population level has little recourse except to rely on the results of standardized tests, and as such is especially dependent on having information about children's experiences with input in order to reach appropriate interpretations.

At the individual level, it would certainly be a mistake to not consider the child's input at all, but it would also be a mistake to summarily dismiss all measures whose norms are derived from typically developing monolinguals. First, DHH children whose only language is English (whether LSL-only or in combination with English-based signing systems[5]) *are* in fact monolinguals: reduced knowledge of English in these children is not compensated by the presence of knowledge in another language. Likewise, it may be unsurprising to find that the mean of a sample of DHH children is likely to be significantly below the expected norm on standardized measures of spoken English, but paying attention to those children's cumulative experience with input can help to discriminate whether this difference is unsurprising and unimportant, unsurprising but important, or perhaps even surprising and important (in a good way or a bad way). A DHH child can be showing progress toward or even achieving their IFSP goals while also still experiencing a significant language delay. Even in children who are showing good progress (e.g., making one year's growth in one year's time), the presence of a language delay can still have serious consequences for the child's cognitive and social-emotional development, school readiness, and academic success. Therefore, intervention plans should look for strategies that are most likely to allow the child to make more than one year's progress in one year's time.

Paying attention to the input can also be a part of setting and tracking individualized goals, especially when there is reason to believe that changes in the child's input would help them achieve their desired outcomes. There has been a lack of good methods for characterizing DHH children's cumulative experience with linguistic input, but new tools are now becoming available that will facilitate these efforts. De Anda and Hall (in prep) provide a practical tutorial in using one such tool; it is hoped that other such tools and trainings will become available as the importance of considering the input becomes more widely appreciated.

[5]Just as learning how to express English in Morse code, Braille, or semaphore does not make someone bilingual, neither does learning how to express English in cued or signed forms.

The present manuscript is offered in part to motivate the development of more resources and tools along these lines.

At the population level, tracking language outcomes without appropriately tracking cumulative language input risks yielding incomplete or even misleading information about upstream strategies that can minimize the adverse consequences of hearing loss. Relying on "communication mode" has now been shown to be deeply flawed, for a number of reasons: there is considerable diversity within children being raised with listening and spoken language (since language *access* is highly variable even within this group) and also within children whose experience includes access to various other forms of communication (e.g., not only variability in auditory access to spoken input, but also variability in the type of manual communication they use, and in the relative distribution of this input over a given period of time). Traditional top-down approaches to creating grouping variables are highly limited in their ability to accurately capture the complex and multidimensional aspects of DHH children's experiences with linguistic input. Instead, bottom-up approaches using various classification algorithms have more potential to reveal insights about strategies that most consistently yield desirable language outcomes. Likewise, bearing in mind dose-response relationships between language input and language outcomes will be necessary in order to avoid prematurely dismissing certain types of communication as ineffective when in reality the dosage may have been too small to have had any appreciable impact. There is of course no guarantee that increasing the "dosage" would necessarily yield more favorable outcomes, and it is understandable that clinicians are reluctant to recommend strategies that remain empirically unproven. However, this also creates a self-fulfilling prophecy: without families who choose to pursue those strategies, crucial data will remain unavailable. This makes it all the more important that when families do pursue lesser-trod paths, public health systems are poised to capture that information in a way that is amenable to investigating natural variation in dose-response relationships, thereby beginning to build more of an evidence base to inform clinical recommendations for future generations. This is only possible if our approach to assessment considers not only the outcomes, but the cumulative input as well.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Allen, T. E. (2015). ASL skills, fingerspelling ability, home communication context and early alphabetic knowledge of preschool-aged deaf children. *Sign Lang. Stud.* 15, 233–265. doi: 10.1353/sls.2015.0006

Ambrose, S., Walker, E., Unflat-Berry, L., Oleson, J., and Moeller, M. P. (2015). Quantity and quality of caregivers' linguistic input to 18-month and 3-year-old children who are hard of hearing. *Ear Hear.* 36, 48S–59S. doi: 10.1097/aud.0000000000000209

Ambrose, S. E., VanDam, M., and Moeller, M. P. (2014). Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear Hear.* 35, 139–147. doi: 10.1097/AUD.0b013e3182a76768

American Speech-Language-Hearing Association [ASHA] (2008). Roles and Responsibilities of Speech-Language Pathologists in Early Intervention: Guidelines [Guidelines]. Available at: http://www.asha.org/policy (accessed May 15, 2020).

American Speech-Language-Hearing Association [ASHA] (2010). *Bilingual Service Delivery*. Available at: https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935225 (accessed May 15, 2020).

Antia, S. D., Lederberg, A. R., Easterbrooks, S., Schick, B., Branum-Martin, L., Connor, C. M., et al. (2020). Language and reading progress of young deaf and hard-of-hearing children. *J. Deaf Stud. Deaf Educ.* 25, 334–350. doi: 10.1093/deafed/enz050

Aragon, M., and Yoshinaga-Itano, C. (2012). Using Language ENvironment Analysis to improve outcomes for children who are deaf or hard of hearing. *Semin. Speech Lang.* 33, 340–353. doi: 10.1055/s-0032-1326918

Belzner, K. A., and Seal, B. C. (2009). Children with cochlear implants: a review of demographics and communication outcomes. *Am. Ann. Deaf* 154, 311–333. doi: 10.1353/aad.0.0102

Cattani, A., Abbot-Smith, K., Farag, R., Krott, A., Arreckx, F., Dennis, I., et al. (2014). How much exposure to english is necessary for a bilingual toddler to perform like a monolingual peer in language tests? *Int. J. Lang. Commun. Disord.* 49, 649–671. doi: 10.1111/1460-6984.12082

Centers for Disease Control (2020). *CDC-RFA-DD20-2005: NCBDDD Outcomes and Developmental Data Assistance Center for EHDI (ODDACE) Programs*. Available at: https://www.grants.gov/web/grants/search-grants.html (accessed March 11, 2020).

Ching, T. Y., Crowe, K., Martin, V., Day, J., Mahler, N., Youn, S., et al. (2010). Language development and everyday functioning of children with hearing loss assessed at 3 years of age. *Int. J. Speech Lang. Pathol.* 12, 124–131. doi: 10.3109/17549500903577022

Ching, T. Y. C., Dillon, H., Leigh, G., and Cupples, L. (2018). Learning from the longitudinal outcomes of children with hearing impairment (LOCHI) study: summary of 5-year findings and implications. *Int. J. Audiol.* 57(Suppl. 2), S105–S111. doi: 10.1080/14992027.2017.1385865

De Anda, S., and Hall, M. L. (in prep) Assessing language input in DHH children: theory and practice.

DeAnda, S., Bosch, L., Poulin-Dubois, D., Zesiger, P., and Friend, M. (2016). The language exposure assessment tool: quantifying language exposure in infants and children. *J Speech Lang. Hear. Res.* 59, 1346–1356. doi: 10.1044/2016_jslhr-l-15-0234

Demers, D., and Bergeron, F. (2019). Effectiveness of rehabilitation approaches proposed to children with severe-to-profound prelinguistic deafness on the development of auditory, speech, and language skills: a systematic review. *J. Speech Lang. Hear. Res.* 62, 4196–4230. doi: 10.1044/2019_JSLHR-H-18-0137

Eisenberg, L. S., Fisher, L. M., Johnson, K. C., Ganguly, D. H., Grace, T., Niparko, J. K., et al. (2016). Sentence recognition in quiet and noise by pediatric cochlear implant users: relationships to spoken language. *Otol. Neurotol.* 37:e75. doi: 10.1097/MAO.0000000000000910

Erbasi, E., Hickson, L., and Scarinci, N. (2017). Communication outcomes of children with hearing loss enrolled in programs implementing different educational approaches: a systematic review. *Speech Lang. Hear.* 20, 102–121. doi: 10.1080/2050571X.2016.1238611

Fitzpatrick, E. M., Hamel, C., Stevens, A., Pratt, M., Moher, D., Doucet, S. P., et al. (2016). Sign language and spoken language for children with hearing loss: a systematic review. *Pediatrics* 137:e20151974. doi: 10.1542/peds.2015-1974

Geers, A. E., Mitchell, C. M., Warner-Czyz, A., Wang, N., Eisenberg, L. S., and CDaCI Investigative Team, (2017). Early sign language exposure and cochlear implantation benefits. *Pediatrics* 140:e20163489. doi: 10.1542/peds.2016-3489

Gillam Ronald, B., and Peña Elizabeth, D. (2004). Dynamic assessment of children from culturally diverse backgrounds. *Perspect. Commun. Disord. Sci. Cult. Linguist. Diverse Populat.* 11, 2–5. doi: 10.1044/cds11.2.2

Gutiérrez-Clellen, V. F., and Peña, E. (2001). Dynamic assessment of diverse children: a tutorial. *Lang. Speech Hear. Serv. Sch.* 32, 212–224. doi: 10.1044/0161-1461

Hall, M. L., and Dills, S. (2020). The limitations of "communication mode" as a construct. *J. Deaf Stud. Deaf Educ.* 9:enaa009.

Hall, M. L., Hall, W. C., and Caselli, N. K. (2019). Deaf children need language, not (just) speech. *First Lang.* 39, 367–395. doi: 10.1177/0142723719834102

Harris, M. (2013). *Language Experience and Early Language Development: From Input to Uptake*. Hove: Psychology Press.

Henner, J., Caldwell-Harris, C. L., Novogrodsky, R., and Hoffmeister, R. (2016). American sign language syntax and analogical reasoning skills are influenced by early acquisition and age of entry to signing schools for the deaf. *Front. Psychol.* 7:1982. doi: 10.3389/fpsyg.2016.01982

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., and Parra, M. (2012). Dual language exposure and early bilingual development. *J. Child Lang.* 39, 1–27. doi: 10.1017/s0305000910000759

Hoffman, M., Tiddens, E., Quittner, A. L., and CDaCI Investigative Team, (2018). Comparisons of visual attention in school-age children with cochlear implants versus hearing peers and normative data. *Hear. Res.* 359, 91–100. doi: 10.1016/j.heares.2018.01.002

Joint Committee on Infant Hearing (2000). Year 2000 position statement: principles and guidelines for early hearing detection and intervention programs. *Am. J. Audiol.* 9, 9–29. doi: 10.1044/1059-0889(2000/005)

Joint Committee on Infant Hearing (2019). Year 2019 position statement: principles and guidelines for early hearing detection and intervention programs. *J. Early Hear. Detect. Intervent.* 4, 1–44. doi: 10.15142/fptk-b748

Koehlinger, K. M., Van Horne Amanda, J., and Owen Moeller, M. P. (2013). Grammatical outcomes of 3- and 6-year-old children who are hard of hearing. *J. Speech Lang. Hear. Res.* 56, 1701–1714. doi: 10.1044/1092-4388(2013/12-0188)

Lederberg, A. R., Branum-Martin, L., Webb, M. Y., Schick, B., Antia, S., Easterbrooks, S. R., et al. (2019). Modality and interrelations among language, reading, spoken phonological awareness, and fingerspelling. *J. Deaf Stud. Deaf Educ.* 24, 408–423. doi: 10.1093/deafed/enz011

Lewis, D. E., Kopun, J., McCreery, R., Brennan, M., Nishi, K., Cordrey, E., et al. (2017). Effect of context and hearing loss on time-gated word recognition in children. *Ear Hear.* 38:e180. doi: 10.1097/AUD.0000000000000395

Mann, W., Peña, E. D., and Morgan, G. (2014). Exploring the use of dynamic language assessment with deaf children, who use American Sign Language: Two case studies. *J. Commun. Disord.* 52, 16–30. doi: 10.1016/j.jcomdis.2014.05.002

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Moeller, M. P., and Tomblin, J. B. (2015). An introduction to the outcomes of children with hearing loss study. *Ear Hear.* 36(Suppl. 1), 4S–13S. doi: 10.1097/AUD.0000000000000210

Muse, C., Harrison, J., Yoshinaga-Itano, C., Grimes, A., Brookhouser, P. E., Epstein, S., et al. (2013). Supplement to the JCIH 2007 position statement: principles and guidelines for early intervention after confirmation that a child is deaf or hard of hearing. *Pediatrics* 131, e1324–e1349. doi: 10.1542/peds.2013-0008

National Center for Hearing Assessment and Management [NCHAM] (n.d.). *Early Intervention for children who are deaf or hard of hearing: Systematic Nationwide Analysis of Program Strengths, Hurdles, Opportunities, and Trends*. Available at: http://www.infanthearing.org/ei-snapshot/docs/ei-snapshot-final-report.pdf (accessed May 15, 2020).

Pearson, B. Z., Fernández, S. C., and Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: comparison to monolingual norms. *Lang. Learn.* 43, 93–120. doi: 10.1111/j.1467-1770.1993.tb00174.x

Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B. A., and Bedore, L. M. (2018). *Bilingual English Spanish Assessment (BESA)*. Baltimore, MD: Brookes.

Percy-Smith, L., Cayé-Thomasen, P., Breinegaard, N., and Jensen, J. H. (2010). Parental mode of communication is essential for speech and language outcomes in cochlear implanted children. *Acta Oto Laryngol.* 130, 708–715. doi: 10.3109/00016480903359939

Rosemary, Q., Brian, G., and Peña Elizabeth, D. (1996). Cultural/linguistic variation in the united states and its implications for assessment and intervention in speech-language pathology. *Lang. Speech Hear. Serv. Sch.* 27, 345–346. doi: 10.1044/0161-1461.2704.345

Sacks, C., Shay, S., Repplinger, L., and Suskind, D. (2014). Pilot testing of a parent-directed intervention (Project ASPIRE) for underserved children who are deaf or hard of hearing. *Child Lang. Teach. Ther.* 30, 91–102. doi: 10.1177/0265659013494873

Suskind, D., Leffel, K., Hernandez, M., Gunderson, E., Sapolich, S., Suskind, E., et al. (2016). A parent-directed language intervention for children of low socio-economic status: a randomized controlled pilot study. *J. Child Lang.* 43, 366–406. doi: 10.1017/s0305000915000033

Suskind, D., Leffel, K., Hernandez, M., Sapolich, S., Suskind, E., Kirkham, E., et al. (2013). An exploratory study of "quantitative linguistic feedback": effect of LENA feedback on adult language production. *Commun. Disord. Q.* 34, 199–209. doi: 10.1177/1525740112473146

Tobey, E. A., Thal, D., Niparko, J. K., Eisenberg, L. S., Quittner, A. L., Wang, N., et al. (2013). Influence of implantation age on school-age language performance in pediatric cochlear implant users. *Int. J. Audiol.* 52, 219–229. doi: 10.3109/14992027.2012.759666

Tomblin, J. B., Harrison, M., Ambrose, S. E., Walker, E. A., Oleson, J. J., and Moeller, M. P. (2015). Language outcomes in young children with mild to severe hearing loss. *Ear Hear.* 36, 76S–91S. doi: 10.1097/AUD.0000000000000219

Van Dam, M., Ambrose, S. E., and Moeller, M. P. (2012). Quantity of parental language in the home environments of hard-of-hearing 2-year-olds. *J. Deaf Stud. Deaf Educ.* 17, 402–420. doi: 10.1093/deafed/ens025

Vohr, B., Topol, D., Watson, V., St Pierre, L., and Tucker, R. (2014). The importance of language in the home for school-age children with permanent hearing loss. *Acta Paediatr.* 103, 62–69. doi: 10.1111/apa.12441

Wie, O. B., Torkildsen, J. V. K., Schauber, S., Busch, T., and Litovsky, R. (2020). Long-term language development in children with early simultaneous bilateral cochlear implants. *Ear Hear.* [Epub ahead of print]. doi: 10.1097/AUD.0000000000000851

Wiggin, M., Gabbard, S., Thompson, N., Goberis, D., and Yoshinaga-Itano, C. (2012). The school to home link: summer preschool and parents. *Semin. Speech Lang.* 33, 290–296. doi: 10.1055/s-0032-1326919

Yoshinaga-Itano, C., Sedey, A. L., Wiggin, M., and Mason, C. A. (2018). Language outcomes improved through early hearing detection and earlier cochlear implantation. *Otol. Neurotol.* 39, 1256–1263. doi: 10.1097/MAO.0000000000001976

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 17 July 2020
doi: 10.3389/fpsyg.2020.01357

The Place of the Bifactor Model in Confirmatory Factor Analysis Investigations Into Construct Dimensionality in Language Testing

Karen J. Dunn[1]* *and Gareth McCray*[2]

[1] Assessment Research Group, British Council, London, United Kingdom, [2] School of Primary, Community and Social Care, Keele University, Keele, United Kingdom

For practical and theoretical purposes, tests of second language (L2) ability commonly aim to measure one overarching trait, general language ability, while simultaneously measuring multiple sub-traits (e.g., reading, grammar, etc.). This tension between measuring uni- and multi-dimensional constructs concurrently can generate vociferous debate about the precise nature of the construct(s) being measured. In L2 testing, this tension is often addressed through the use of a higher-order factor model wherein multidimensional traits representing subskills load on a general ability latent trait. However, an alternative modeling framework that is currently uncommon in language testing, but gaining traction in other disciplines, is the bifactor model. The bifactor model hypothesizes a general factor, onto which all items load, and a series of orthogonal (uncorrelated) skill-specific grouping factors. The model is particularly valuable for evaluating the empirical plausibility of subscales and the practical impact of dimensionality assumptions on test scores. This paper compares a range of CFA model structures with the bifactor model in terms of theoretical implications and practical considerations, framed for the language testing audience. The models are illustrated using primary data from the British Council's Aptis English test. The paper is intended to spearhead the uptake of the bifactor model within the cadre of measurement models used in L2 language testing.

Keywords: language testing, psychometrics, bifactor model, higher-order model, dimensionality, confirmatory factor analysis (CFA)

OPEN ACCESS

Edited by:
Thomas Eckes,
Ruhr University Bochum, Germany

Reviewed by:
Marley Wayne Watkins,
Baylor University, United States
Giorgio Arcara,
San Camillo Hospital (IRCCS), Italy

***Correspondence:**
Karen J. Dunn
karen.dunn@britishcouncil.org

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

Received: *16 December 2019*
Accepted: *22 May 2020*
Published: *17 July 2020*

Citation:
Dunn KJ and McCray G (2020) The Place of the Bifactor Model in Confirmatory Factor Analysis Investigations Into Construct Dimensionality in Language Testing. Front. Psychol. 11:1357. doi: 10.3389/fpsyg.2020.01357

INTRODUCTION

Dimensionality considerations are important for both the development and ongoing validation of tests of second language (L2) ability. For practical and theoretical purposes, language tests are commonly designed to measure one overarching trait, that of general L2 ability, while simultaneously measuring multiple sub-traits (usually L2 reading, listening, speaking, writing). Items are written with the aim of assessing these highly related but conceptually distinct abilities. It is crucial for a strong validity argument that test constructors are able to isolate and examine the similarities and differences between various L2 skill areas. Indeed, the meaningful evidence-based delineation and reporting of scales and possible subscales and their appropriate usage is

an essential aspect in making a construct validity argument for a test (Slocum-Gori and Zumbo, 2010). This has particular ramifications for practical decisions regarding score reporting. Where guidelines are given regarding sub-scores, the requirement for sufficiently high reliability and distinctiveness for all scores is emphasized (e.g., AERA et al., 2014). When reporting a test score on a single scale, the implication is that the test is measuring one unitary skill or trait, and that the scores given reflect the candidate's ability or level on that single trait. Splitting the test into sub-scores and reporting these separately indicates each sub-score should require a sufficiently distinct aspect of ability from the other sub-scales. From a theoretical perspective, an understanding of language tests as straddling both uni- and multi-dimensional structures is now a generally accepted viewpoint within the academic language testing community. Harsch's (2014) "state of play" summary on dimensionality in L2 language testing emphasizes that "language proficiency can be conceptualized as both unitary *and* divisible, depending on the level of abstraction and the purpose of the assessment and score reporting" (Harsch, 2014, p. 153). Nonetheless, achieving a balance between these concurrent theorizations can generate sometimes vociferous debate about the precise nature of, and relationship between, the construct(s) measured.

This paper will compare the kinds of insights various confirmatory factor analysis (CFA) models can offer into the underlying dimensional structure of sets of test items designed to tap into different but highly related knowledge domains. Four model structures are described and discussed: the unidimensional model, the correlated traits model, the higher- (or second-) order model, and the bifactor model. The first three are frequently used in analysis of L2 language tests, while the fourth, the bifactor model, is less commonly employed in this field for analysis. The ultimate aim of the paper is to gauge what added value the bifactor model can bring to the assessment of dimensionality and, thus, to place its usefulness in the language test researcher's CFA toolkit. Two illustrative studies are presented, which employ language testing datasets, plus a brief literature review on the background to the dimensionality debates surrounding each area addressed. The first illustrative study examines the evidence for the divisibility and, thus, the appropriateness of sub-score reporting, of a grammar and vocabulary test component with 50 items, 25 intended to measure grammar and 25 intended to measure vocabulary. The second illustrative study examines the evidence for multidimensionality in data representing the traditional four skills (L2 listening, reading, speaking, and writing) comprising an overall measure of general second language (L2) ability. The abovementioned psychometric models will be fitted to the data and then interpreted. It is important to note that while this paper does fit a battery of models, which address common debates in L2, the focus here is primarily on demonstrating the modeling and inferential process, particularly regarding the bifactor model, rather than generalizing theory from the substantive interpretations of the results. Furthermore, note that this paper will illustrate why using only an assessment of model fit statistics to choose the most appropriate form for score reporting is a limited and inappropriate analytical strategy. The theoretical, statistical, and practical differences between the four models will be discussed, and recommendations for usage in the language testing context will be provided.

LITERATURE REVIEW

Dimensionality of Large-Scale Language Tests

Over the past decade and beyond, work exploring the dimensionality of a range of large-scale language tests has supported the interpretation of multi-skill tests as comprising a series of strongly related, yet distinct, dimensions. A large body of such studies have employed CFA techniques to show either the correlated factor or higher-order factor model to be the most appropriate model to represent the underlying measurement qualities of the test in question (Shin, 2005; Stricker et al., 2005; Sawaki et al., 2009; In'nami and Koizumi, 2012; Sawaki and Sinharay, 2013, 2017; In'nami et al., 2016; Kim and Crossley, 2019). Usually, a CFA study involves proposing various theoretically informed structures for the relationships between sets of items purported to be measuring different dimensions. Statistical models are then fitted to collected data, which operationalize these theoretical structures, and evidence is gathered on which of the models best describes the data and, thus, which of the structures is most likely closest to that under which the data was generated. The tests in these analyses included: TOEIC,® which was found to be best represented by a correlated factor model for reading and listening (In'nami and Koizumi, 2012); the TEAP test, represented by a higher-order model for the four skills (In'nami et al., 2016); the TOEFL iBT® meanwhile has been subject to a large number of studies with a higher-order model being favored in some projects (Stricker and Rock, 2008; Sawaki et al., 2009) and the correlated four-factor model in others (Sawaki and Sinharay, 2013, 2017).

Several of the studies explored the use of a bifactor model as a possible representation of a multidimensional structure hypothesized to underlie a test (Sawaki et al., 2009; Sawaki and Sinharay, 2017). This modeling framework is currently uncommon in language testing, but gaining traction in other disciplines (Reise, 2012). The bifactor model incorporates a general factor, onto which all items load directly, plus a series of orthogonal (i.e., specified as uncorrelated) factors each loading on a sub-set of items (Reise, 2012). Where the higher-order and correlated factor models account for commonalities within and across each of the subscales, the bifactor model explicitly models the general commonality between all items in the test and the residual variance for each skill area beyond that of general L2 proficiency, with equal weight (see below for further details). It is important to note, however, that statistically, the higher-order model has been shown to be nested within the bifactor model (Yung et al., 1999; Rijmen, 2010; Markon, 2019). The subordinate factors in a higher-order model mediate the relationship with the more general factor, but the higher-order factor can be expressed in terms of their direct relationship with the observed variables following mathematical transformation. The two models are not, therefore, as far removed from one another as it may first appear, however, employing the more flexible bifactor model

has implications for interpretation of multidimensionality in language tests as shall be explored in this paper.

The aim of the majority of the language testing studies cited above exploring the dimensionality of large-scale language tests is to justify score reporting practices, which break down an overall score into a number of sub-scores for each skill area. As observed by Sawaki and Sinharay (2017), "conceptual distinctness among section scores does not necessarily guarantee their psychometric distinctness from one another" (Sawaki and Sinharay, 2017, p. 530–531). The importance of these studies is to provide empirical backing for theoretical assumptions about the underlying structure of L2 language tests. This is of particular interest to test developers, since stakeholders often expect detailed feedback and are perhaps not overly concerned with its justification from a measurement perspective. It is, therefore, unsurprising to find that few of the studies explored alternative groupings of the sub-scales in the tests. In other words, because of the influence of stakeholders, scores are often reported in a traditional way, e.g., an overall score with reading, listening, reading, and writing sub-scores, whether or not the subscales are shown to be psychometrically distinct. There are, however, a couple of exceptions to this rule. One example is Kim and Crossley's (2019) investigation of the latent structure of the Examination for the Certificate of Competency in English (ECCE) across test sections, addressing ability in reading, listening, writing, speaking, and lexico-grammatical ability (Kim and Crossley, 2019). These researchers identified a three-factor solution, with one factor representing reading, listening, and lexico-grammar, and the additional two factors representing writing and speaking abilities, respectively. In addition, they found this structure to hold across age and gender sub-groups of the data. Another example of an alternative factor structure was presented by Stricker et al. (2005), whose modeling of the LanguEdge test showed speaking to load on one factor while reading, listening, and writing all jointly loaded on a second. From a measurement perspective, when an alternative structure is indicated for a test, there will be an implication for score reporting. However, stakeholder expectations may be resistant to, or ultimately prohibit, changes in this regard owing to the use of language sub-scores in decision-making processes. This point is emphasized by Sawaki and Sinharay (2017), who focused in their study on the degree to which section scores can offer value-added information to stakeholders. In addition to investigating the overall factor structure of the TOEFL iBT®, these researchers explored the extent to which section sub-scores are reliable and, importantly, distinct, from other sub-scores. These researchers employed a classical test theory-based sub-score analysis (Haberman, 2008). The current paper, meanwhile, discusses how the bifactor model provides a tool to explore such considerations within the CFA framework.

Dimensionality of Grammar and Vocabulary Tests

Considerations regarding test dimensionality are also pertinent in addressing the distinction between grammar and vocabulary knowledge. While superficially these two aspects of language may seem different, separating the constructs is not as clear-cut is it may first appear. From an analytic perspective, and with much dependent on how the constructs are operationalized, the likelihood is that candidates scoring highly on vocabulary items would have a strong tendency to score highly on grammar items. However, this does not necessarily mean that they are indivisible constructs or that it is desirable to treat them on a unidimensional scale in all cases; indeed, researchers have been mixed in their recommendations on conceptualizing grammar and vocabulary on a uni- or bi-dimensional scale. Taking examples from studies that aim to describe the components of reading ability, it can be seen that Purpura (1999), for instance, drew on both vocabulary and grammar measures to form a single "lexico-grammatical ability" factor, while Shiotsu and Weir (2007) and Shiotsu (2010) maintained a distinction.

Test developers need to be sensitive to the manner in which the grammar and vocabulary constructs are operationalized in any given test before assuming a united or divided treatment of these language knowledge areas. This notion was demonstrated by Alderson and Kremmel (2013), who warned about the need to be cognizant of the "slipperiness of the slope" between grammar and vocabulary knowledge, and for the test constructor be able to define and defend their decisions to report the constructs separately (Alderson and Kremmel, 2013, p. 550). In addition, it should be recognized that grammar and vocabulary may well be activated differently within each language domain. For example, while readers can rely on linguistic information in the text via bottom–up processes, the "online" nature of listening means that learners tend to draw more on top–down processes (Lund, 1991; Park, 2004). In practical terms, this means that the listener will perhaps compensate for lack of specific vocabulary knowledge by drawing on other, more general or metacognitive, areas of knowledge, but this is less common in reading (van Zeeland and Schmitt, 2013, p. 461). Consequentially, when considering the theoretical arguments for the dimensionality of grammar and vocabulary, one should carefully consider the specific operationalization of the constructs and not make overgeneralizations that grammar and vocabulary are always *or* never distinct.

Data from measures underpinned by such closely related constructs, and which tap into such tightly interrelated knowledge domains, very often result in item responses that are consistent with both unidimensional and multidimensional interpretations (Reise et al., 2010). Described as a "dimensionality quagmire" by researchers working in clinical psychology settings (Reise et al., 2018), a similar state of affairs is equally applicable to the language testing context. The choice of measurement model is, nonetheless, crucial for both score reporting and assessing score reliability (Brunner et al., 2012). It then becomes the job of the researcher to take into account information from a range of sources when considering the dimensionality of a test, of which statistical evaluation is just one aspect.

Current Aims and Research Question

The aim of the current paper is to illustrate, in some detail, the usefulness of a range of factor analytic models in answering questions about test dimensionality. While most of the models

will be familiar to L2 language test researchers working with CFA, the intention here is to encourage the integration of the bifactor model into the cadre of models already employed by academics and test developers in this field. Two key points are emphasized in this paper:

1. CFA models are to be viewed as tools used to gather evidence, rather than truth-makers. *Inferences on dimensionality should not be solely based on statistical fit any more than they should be purely based on expert judgment of item content.*
2. All tests, and indeed subscales of tests, with more than one distinct item are multidimensional to some extent. *It is the job of the researcher and the test constructor to investigate the tenability of assuming unidimensionality and reporting a single score for these scales and subscales.*

To elaborate on these points, the current paper will describe the way in which information from various models offered within the CFA framework can be used to complement theoretical understandings and practical requirements. The comparative nature of this paper aims to provide a framework for researchers to evaluate what the bifactor model might bring to their assessment of L2 language test dimensionality in addition to the oft-used models. Particular focus is given therefore to interpretation and applicability of the bifactor model which, as a less commonly used model, is more vulnerable to misinterpretation, particularly of the meaning of the trait-specific factors (DeMars, 2013). The following research question is addressed via two worked examples, using data from two variants of the British Council's Aptis test:

RQ: What insights useful to both test and theory developers can the: (i) unidimensional; (ii) correlated factors; (iii) higher-order; and (iv) bifactor model provide about the dimensionality and score interpretation of the underlying construct(s) when applied to L2 language test data?

MODEL DESCRIPTIONS

Each of the models employed in the worked examples below are introduced in the following four sub-sections.

Unidimensional Model

Unidimensionality is a key assumption within almost all scoring models in both classical as well as item response test theory (Gustafsson and Åberg-Bengtsson, 2010). The unidimensional model hypothesizes a single factor to explain the variance across all observed variables (i.e., the variance in test scores across all items), with no differentiation between sub-groups of items. This model is illustrated in **Figures 1A**, **2A** below. A series of estimated loadings indicate the strength of the relationship between the single factor and each of the observed variables. An error term (omitted in the figures in this paper) is also estimated against each observed variable, since the latent factor is not assumed to provide a perfect explanation of the observed variance. Standardized loadings can be directly compared, and smaller loadings on the general factor will be associated with a higher degree of error and, thus, the response to an item providing less information about a test-taker's trait score.

The unidimensional model is the most commonly applied (or assumed) model in psychometrics, and it is particularly valuable as it can be used to model items measuring various aspects of a construct on the same scale and report a single score to represent the ability of the test taker. A key question to answer when using this model is: "Is this test unidimensional?" Or, in other words, can a large proportion of variance in observed test scores be explained with reference to the same underlying construct? When modeling language test data, this factor is often hypothesized as general L2 ability.

Correlated Factors Model

The correlated factors model (e.g., Brown, 2015) includes two or more latent variables, which are allowed to correlate (see **Figures 1B**, **2B** for illustrations). Observed variables are grouped by shared features and act as indicators for a factor hypothesized to reflect this commonality. This explicitly models the multidimensionality of a test. The correlated factors model does not incorporate any general or underlying factor, however, the correlations between each of the latent variables indicate shared variation across all pairs of latent variables in the model. A series of loadings indicate the strength of the relationship between the observed variables and their associated factor. Again, error terms are estimated against each observed variable. Note that each observed variable in the model is assumed to be only associated with a single factor.

The correlated factors model is often used as a point of comparison with the unidimensional model described above. A language testing researcher might want to ask: "Is this test multidimensional?" or perhaps he or she will have more specific questions regarding whether a particular group of items constitutes a subscale.

Higher-Order Model

The higher-order model (Thurstone, 1944) incorporates at least one superordinate (higher-order) factor and a series of subordinate factors upon which specified sub-group of items load (see **Figure 2D** for an illustration). This second- or higher-order factor explicitly models the shared variance between subordinate factors, meaning that these first-order grouping factors are conditionally independent of one another, and each one mediates the relationship between the overarching, or superordinate, factor and the observed variables.

The higher-order model estimates two sets of loadings: those showing the relationships between the observed variables and the relevant grouping, or subordinate, factor, plus those showing the relationship between the higher-order factor and each of the subordinate factors. Error terms against each of the observed variables show that the model is not hypothesized to perfectly explain the variance of the observed variables, and error terms on the factors (termed *disturbances* in CFA literature) indicate that this higher-order factor does not explain all the variance of each of the subordinate factors.

Higher-order models are often used for theory testing (Brown, 2015), and they enable the researcher to explore theoretical

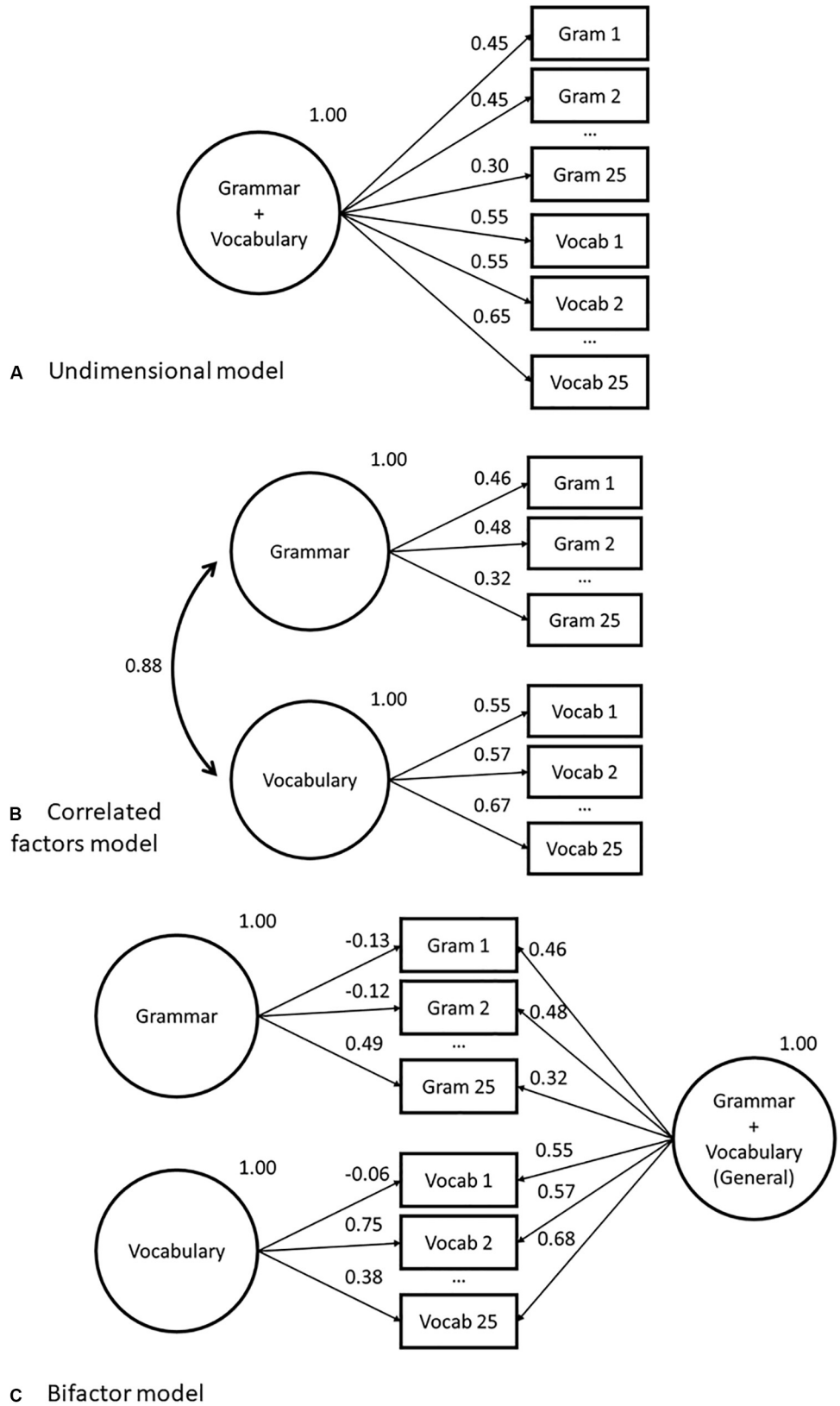


FIGURE 1 | Abbreviated factor loading diagrams for confirmatory factor analysis (CFA) models fit to grammar and vocabulary dataset in study 1. Models as follows: **(A)** Unidimensional; **(B)** Correlated factors; **(C)** Bifactor.

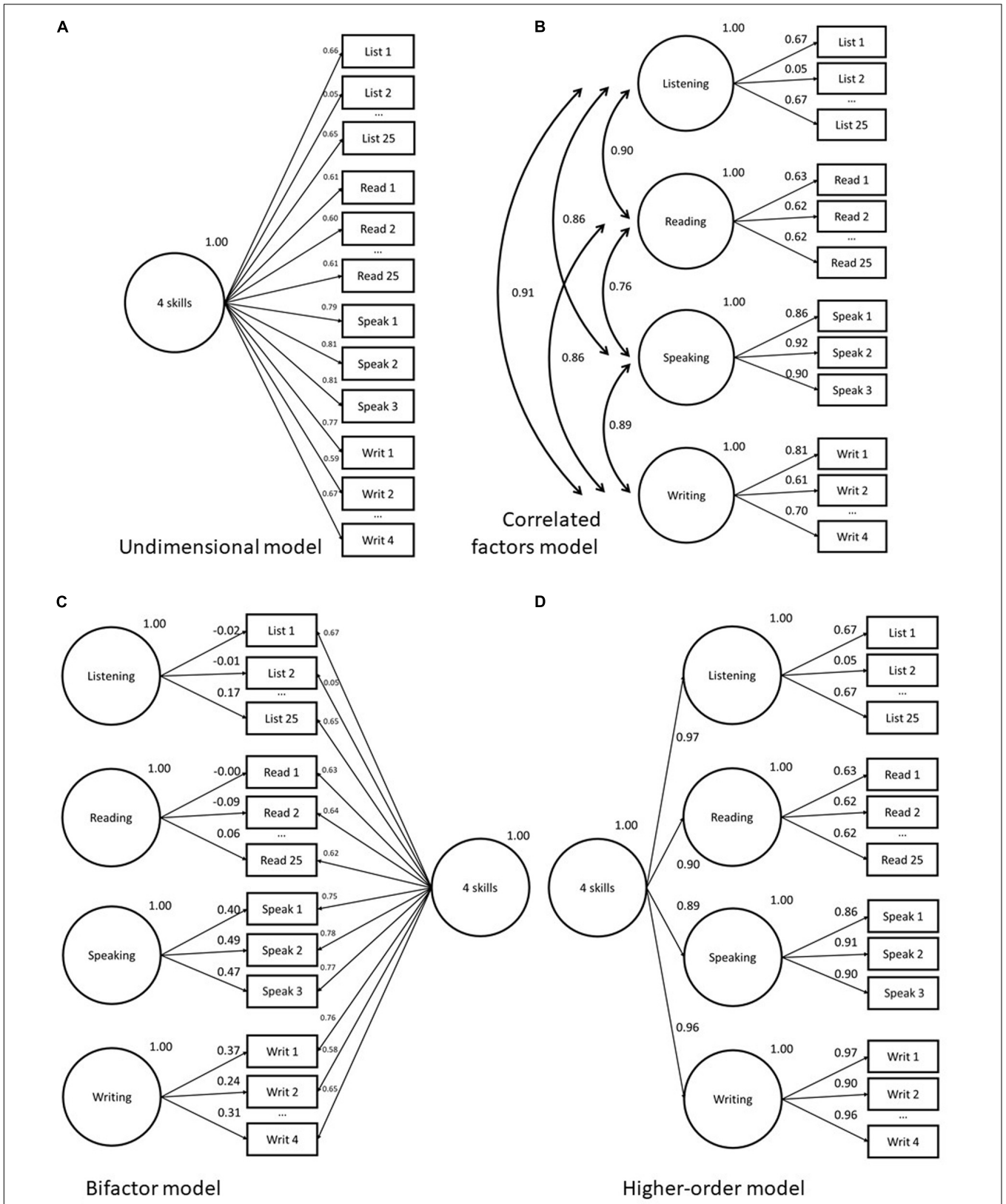


FIGURE 2 | Abbreviated factor loading diagrams for CFA models fit to four-skill dataset in study 2. Models as follows: **(A)** Unidimensional; **(B)** Correlated factors; **(C)** Bifactor; **(D)** Higher-order.

understandings of the relationship between a series of sub-tests as distinct from one another, but also united by a common factor, which attempts to explain the scores in the higher-order factor. The researcher might ask: "Can I justify reporting this multi-skill test as an overall scale?" This is a highly relevant question in language testing, where the researcher or test developer may seek empirical justification for the reporting of an overall score in addition to sub-scores for each language domain incorporated in the test. If the loadings between the higher-order and subordinate factors are satisfactorily high, it can be concluded that there is enough commonality between the sub-skills to justify this reporting both sub-scores and an overall score.

It is important to note that in this model, there is no direct relationship hypothesized between the more general (higher-order) factor and the observed variables. The observed variables act as indicators of the subordinate factors, and therefore, the commonality modeled by the higher-order factor is between the scales already established for each sub-group. This mediating role for the subordinate factors means that the higher-order factor, therefore, represents a "distilled" estimate of general ability rather than a more direct estimate, which accounts for commonalities between all observed variables as per the unidimensional model. This distance is termed by Markon (2019) and others as a "level of abstraction," with the higher-order model the choice of the researcher for whom the subordinate factors are "theoretically salient" (Markon, 2019, p. 53). In practice, this means that there may be commonalities between items across different subscales that are not captured by the higher-order model. If, say, individual items across reading, listening, and writing factors depended, in part, on a particular aspect of knowledge (for example, the "past-perfect tense"), the higher-order model may not see those items load as high on the general factor, after distillation, as they would on a unidimensional model, or indeed the bifactor model, described below.

Bifactor Model

The bifactor model (Holzinger and Swineford, 1937), also described as a nested-factor (NF) model (Gustafsson and Åberg-Bengtsson, 2010; Brunner et al., 2012), or a hierarchical model (Markon, 2019), incorporates a general factor, which loads directly onto all of the observed variables in the model and, in addition to this, grouping factors, which load onto sub-groups of the same set of observed variables (see **Figures 1C**, **2C** for an illustration. One of the defining features of the bifactor model is that the grouping factors in the model are hypothesized to be orthogonal (uncorrelated) with the general factor. Grouping factors, themselves, can be either correlated or uncorrelated (Reise et al., 2018); however, the focus in this paper is on bifactor models with uncorrelated grouping factors, as providing a more readily interpretable solution. Additionally, unlike the CFA model structures presented above, the bifactor model does not offer a "simple structure" solution in which each observed variable only loads onto a single factor (Gustafsson and Åberg-Bengtsson, 2010). Observed variables, by design, in this model load onto more than one factor, meaning that the variance explanation is split between (at least) two latencies. Each observed variable in the bifactor model is an indicator of both the general factor and one grouping factor. This means that each observed variable has two loading estimates in the model; the first will show its relationship with the general factor and the second with its allocated grouping factor.

While the interpretation of the loadings on the general factor can be understood as per the single factor in the unidimensional model, it is important to note that the estimates for the grouping factors in the bifactor model are *not* analogous to the subordinate factor loadings in the higher-order model or the skill-specific factors in the correlated factors model. The grouping factors in the bifactor model give an estimate of the shared variance between sub-groups of items once the common variance between all observed variables captured by the general factor has been partitioned out. This can be thought of as the relationship between residuals[1]. With respect to scoring considerations, DeMars (2013) described how constructing a group factor score for the bifactor model can be achieved by algebraically combining the loading on the grouping factor and the general factor. Statistical packages do not commonly provide this score by default, as sub-score generation is virtually never the reason a bifactor model is fit. In language testing, a bifactor model of a four-skill test would include the general factor as representative of overall L2 ability, and the grouping factors as representative of a shared aspect of knowledge within each skill area that is not captured by the information about overall L2 ability. This point is discussed in more detail in the second worked example below.

A key distinguishing feature between the bifactor model and the higher-order model is that the general factor is hypothesized to load directly on each of the observed variables. This grants the general factor greater theoretical salience than the grouping factors, the reverse scenario to that of the higher-order model, which foregrounds the skill-specific factors (Markon, 2019). With respect to the accepted understanding of general L2 language proficiency as both "unitary and divisible" (Harsch, 2014), this distinction in emphasis between the two models are not necessarily at odds with one another. As noted earlier, the higher-order model is nested within the bifactor model. This is illustrated by studies that show the possibilities for expressing the direct relationship between the superordinate factor in a higher-order model and the observed variables via a process known as the Schmid–Leiman transformation (Schmid and Leiman, 1957; Yung et al., 1999). The resulting estimates are structurally equivalent to those of a bifactor model subject to certain constraints (Brunner et al., 2012; Markon, 2019). These two models should perhaps not, therefore, be viewed as competing structures, but rather different means of accounting for the multidimensionality in language tests, the estimates from which are useful in different ways, as explored below.

An important question that the bifactor model can help the researcher to answer is: "Is this test unidimensional enough to be reported on a single scale, and relatedly, does it make sense to also report domain sub-scores?" In some respects, the bifactor model fleshes out the insight gained from the unidimensional

[1]Note that in the bifactor model, it is not the case that a unidimensional model is fitted, and then group factors are fitted to the residuals. Rather, the general and specific factors are fitted at the same time, and thus, the specific factors are only analogous to residuals.

model in cases where the researcher knows that there are likely to be dependencies between sub-groups of items within the test. Researchers in other disciplines suggest that this factor structure can, in fact, lead to greater conceptual clarity than alternative CFA model structures (e.g., Chen et al., 2012) and are particularly valuable for evaluating the plausibility of subscales (Reise et al., 2010, 2018).

In summary, each of the four models described above acknowledge some interrelatedness between all items. This is an important requirement when modeling tests that assesses L2 knowledge. Echoing the comparative description of these four-factor analytic models from Brunner et al. (2012), the unidimensional model focuses exclusively on general language ability, and the correlated factors model on specific abilities, while the higher-order and the bifactor models both "consider the ability hierarchy in its entirely, containing a mix of general and specific constructs" (Brunner et al., 2012, p. 813). However, in terms of model estimates (without transformations), the language test researcher will note that there is a different division in terms of the manner the overall commonality between items is addressed. The unidimensional and bifactor models directly model shared variance between observed responses, while the correlated factors and higher-order model mediate this relationship by the inclusion of grouping factors at the individual sub-skills or first-order level. These varying structures accord the researcher different insights into the measurement properties of a test. This is demonstrated in the following sections using two examples of the application of CFA models to the kind of data typically analyzed in language testing. An interpretation of the findings is given, which considers the utility of each model fitted to address subtly different sets of questions about the underlying factor structures of the data and the practical ramifications for test constructors of the inferences drawn from the models.

ILLUSTRATIVE STUDY 1 – APTIS GENERAL GRAMMAR AND VOCABULARY

The first illustrative study examines the insights that can be gathered from fitting three models to explain the score variance seen in a selection of grammar and vocabulary items: (i) unidimensional; (ii) correlated factors; and (iii) bifactor model. Note that the higher-order model was not fitted in this study as a model with only two first-order factors (i.e., grammar and vocabulary) loading onto the higher-order factor is not statistically identifiable (Brown, 2015). The models fitted are illustrated in **Figure 1**.

Study 1 Dataset

The Grammar and Vocabulary component of the Aptis General test variant (O'Sullivan and Dunlea, 2015) was delivered to a large global population (N = 17,227) between April 2018 and June 2019. Representation in the dataset was from more than 60 different countries. Ability levels ranged from pre-A1 to above B2 on the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) according to their Aptis score designation. Each test taker completed 25 grammar items and 25 vocabulary items, scored dichotomously. All items were multiple choice, but the vocabulary items had pooled response options (10 options per each block of five items). A description of the test, including sample questions, are available in the Aptis Candidate Guide (British Council, 2019a).

Study 1 Method

The three CFA models, unidimensional, correlated factors, and bifactor, were fit to these data using the latent variable modeling software Mplus (Muthén and Muthén, 2017) using robust maximum likelihood (MLR) estimation, analogous to marginal maximum likelihood (MML) estimation in an Item Response Theory framework. Response outcomes were modeled at the individual item level; for each candidate (random missingness aside), this was a series of 50 responses, 25 from each test section.

In terms of evaluating each model, for completeness, the model chi-square values and associated p-values are reported; however, note that these should not be relied upon for model acceptance or rejection as they are acutely sensitive to sample size (Schermelleh-Engel et al., 2003; Vandenberg, 2006), and the models described have been fit to a large sample. With respect to global fit, i.e., how well the data fit the predictions of the model, root mean squared error of approximation (RMSEA), and comparative fit index (CFI) are reported. As recommended by Hu and Bentler (1999) we report both an absolute index of fit, the RMSEA, where a model is compared against a perfectly fitting model, and a relative index of fit, the CFI, where a model is compared against a baseline, or null, model. For acceptable fit, the RMSEA should be below 0.06, while the CFI should be greater than 0.95 (Hu and Bentler, 1999). Given that the performance of fit indices can vary according to aspects of the model, it is generally recommended to take into account a range of fit statistics when considering model appropriateness (Brown, 2015). Note, as MLR estimation method does not furnish RMSEA and CFI values, when modeling categorical variables, we report statistics that come from fitting the same model on the same data using the weighted least squares mean and variance adjusted (WSLMV) estimation method in Mplus with both models having probit link functions.

Given that the models are not all nested, some comparative fit measures, i.e., measures that allow us to compare the ability of non-nested models to explain the score variance seen, were also extracted. Specifically, AIC, BIC, and sample-size-adjusted BIC are reported. As a reflection of the practical differences between models, two sets of metrics have been generated: mean absolute error (MAE) of the factor loadings for various latent traits against the unidimensional model. The MAE is a measure of the size of an average difference between a parameter on the unidimensional model and a corresponding parameter from another model fitted to the same data[2]. A low MAE indicates a high similarity between sets of parameter values, and a high MAE indicates the converse. To aid in interpretability, factor scores have been rescaled from

[2] $MAE = \frac{\sum |x_i - y_i|}{n}$, where x is a parameter for the unidimensional model, y is a corresponding parameter from the model in question, and n is the sum of comparisons.

having a mean of 0 and an SD of 1, to having a mean of 50 and an SD of 10.

It should be noted that there are also various well-documented statistical methods that can be employed to evaluate the usefulness of several of the models described, for example, the commonly calculated "omega" and the "omega hierarchical." These statistics provide a model-based approach to assess scale and subscale reliability and are relevant to the bifactor model, and the higher-order model following transformations (Gustafsson and Åberg-Bengtsson, 2010; Brunner et al., 2012; Reise et al., 2013; Rodriguez et al., 2016a,b; Reise et al., 2018). There are also statistics that directly address the question of sub-score utility (Rodriguez et al., 2016a,b). A full review of these methods is beyond the scope of the current paper, but the reader is encouraged to explore the literature.

Study 1 Results

The fit measures and average loadings for the three models are shown in **Table 1**. None of the models have non-statistically significant chi-square *p*-values, but this is unsurprising given the large sample size. All models have acceptable levels of fit on the absolute index, RMSEA, but none of the models have acceptable levels of fit on the relative index of fit, the CFI, though at 0.896, the bifactor model is very close to the suggested threshold of 0.9 for a reasonable model. All three indices of comparative fit indicate that the bifactor model is the best fit by some margin, followed by the correlated factors model, then, the unidimensional model.

With respect to loading estimates, **Table 1** shows that for the unidimensional model, the average loading on the single general factor is 0.54. For the correlated factors model, the average loading for grammar is 0.47, the average loading for vocab is 0.64, and the correlation between the two traits is 0.88. For the bifactor model, the mean loading on the general factor is 0.50, and then the mean loadings for the grouping factors are much lower at 0.21 and 0.25 for grammar and vocab, respectively. In terms of loadings, there is an MAE of 0.03 and 0.01 difference between the grammar and vocabulary loadings on the unidimensional versus the correlated factors model, respectively. There is a 0.08 MAE between the unidimensional model loading and the general factor of the bifactor model. In terms of scores, when put onto a mean = 50 and *SD* = 10 scale, there is a 0.90 and 1.18 average difference in the scores that would be given by the unidimensional model compared to the correlated factors model, for grammar and vocabulary, respectively. There is an average difference of 4.11 between the scaled scores from the unidimensional model and those from the general factor of the bifactor model. Given the fact that the scale is set to have an SD of 10, these differences are minimal and likely to be no greater than error.

Figures 1A–C give traditional factor loading diagrams for the three models used in this section. Note that they are abbreviated, in that all not all observed variables are displayed. **Supplementary Table A1** in the appendix provides the full list of item loadings. Conditional formatting has been applied to the table to help with interpretation, where lighter cells are lower and darker cells are higher values. We can see there is little difference between the unidimensional and correlated factors model loadings. There are a series of items from Vocab 6 to Vocab 15 that have higher loadings (> 0.7) than the rest of the items on the scale. We do see some differences, however, between the aforementioned two models and the bifactor model loadings. In the bifactor model, we do not see an overall uniform loading on the grouping factors (i.e., all items loading approximately similarly indicating shared variance). Rather, we see several items which load lower on the general factor but load higher on the grouping factor, for example, Vocab 2, 4, 21, and 24 all have loadings above 0.5 on the grouping factor.

Study 1 Interpretation

The first aspect to note about the estimated loadings on the unidimensional model is that they are all positive, and the majority – across both grammar and vocabulary items – are larger than 0.32, the rule of thumb value to consider loadings as statistically meaningful (Tabachnick and Fidell, 2007). For standardized solutions, this value indicates that the latent factor explains more than 10% of the variance in the item, a figure obtained by squaring the loading [see Brown (2015, p. 52) for an accessible explanation of the relationship between indicator variables and latent factors]. The fact that none of the loadings are negative indicates that all the items are measuring the latent trait in the same way, i.e., a positive response indicates higher ability. This is very much to be expected on such a rigorously designed assessment tool, and any negative loadings would be a serious cause for concern and item removal. Vocabulary items tend to load higher on the general factor than the grammar items. However, as noted above, the vocabulary items have pooled response options, which means that they will have at least some dependencies between items because of this. This may be the cause of the higher loadings, and it would be advisable to inspect correlations in the error terms further.

The correlated factor model, meanwhile, is indicated by the lower AIC and BIC to provide a more accurate description of the data. However, the high positive correlation between the two factors ($r = 0.88$) implies a poor discriminant validity between factors (Brown, 2015, p. 28), in combination with the similarities in magnitude of the loadings on the correlated two-factor model compared with the loadings in the unidimensional model mean. It is not fully clear, from a substantive perspective, what additional insight is gained from splitting the construct into two factors. Given that we would be expecting some difference purely as a result of random error, the small MAE values showing little difference in factor loading or score over and above the unidimensional model, bear out the suggestion that there seems little to be gained from reporting this particular set of items with separate grammar and vocabulary scores. While from a statistical perspective better comparative fit indices indicate a better representation of the data, when we cast it in practical terms, the improvement is minimal. It is here that the loading estimates from the bifactor model can provide additional insight.

Recall that the loadings on the skill-specific grouping factors in the bifactor model are not interpreted in the same way as those in the correlated factors model. In the bifactor model, these estimates indicate the degree of shared variance between groups of items after accounting for the general factor. In the case of the

TABLE 1 | Grammar and vocabulary fit measures and average loadings.

	Model descriptive statistics		
Fit measure	**Unidimensional**	**Correlated factors**	**Bifactor**
Chisq	70,533 (1,175), $p < 0.001$	66,999 (1,174), $p < 0.001$	42,169 (1.125), $p < 0.001$
RMSEA	0.059	0.057	0.046
CFI	0.824	0.833	0.896
AIC	868,283	865,840	848,767
BIC	869,058	866,623	850,034
ADJ BIC	868,740	866,302	849,558
Mean (SD) of factor loadings			
General	0.54(0.17)	*NA*[a]	0.50(0.19)
Grammar	*NA*[a]	0.47(0.13)	0.21(0.25)
Vocab	*NA*[a]	0.64(0.16)	0.25(0.27)
Factor correlations			
Gra-Voc	*NA*[b]	0.88	NA[b]

Mean absolute error (MAE)[c] of parameters from those on the unidimensional model

	Gra	Voc	General
Factor loadings	0.03	0.01	0.08
Factor scores[d]	0.90	1.18	4.11

[a]Not provided as the factor was not measured by the model. [b]Not provided as no correlation between factors was measured in the model. [c]MAE = $\frac{\sum|x_i - y_i|}{n}$, where x is a parameter for the unidimensional model, y is a corresponding parameter from the model in question, and n is the sum of comparisons. [d]Scores set to mean = 50, SD = 10.

current concerns, we can discern whether there is any systematic association between grammar items, or between vocabulary items, *once the more general construct has been taken into account.* A set of strong, relatively uniform, loadings on the grouping factors would indicate a dependence between items that is not picked up on by the more general factor, i.e., something unique to vocabulary knowledge over and above the lexico-grammatical knowledge accounted for by the general factor. However, rather than consistent strong loadings for the grouping factors, the estimates show a minimal number of items with high loadings on the grouping factor, either for vocabulary or for grammar. There are four grammar items and four vocabulary items, which have a loading higher than 0.5, which indicates that 25% or more of the original observed variance is explained by the grouping factor. As a diagnostic, this indicates a deviation from the general construct among these few items. Examining the content of these two sets of items, respectively, would be recommended in real-life test construction or evaluation to identify if there are any characteristics that make them distinct from the rest of the grammar or vocabulary items. Overall, however, the mean loadings on the grouping factors are 0.21 (grammar) and 0.25 (vocabulary), which indicates that, collectively, around 5% of the observed variance in each group of items can be explained with reference to a skill-specific grouping factor, once the general lexico-grammatical ability has been taken into account.

As a researcher armed with this information, the question here is whether these discrepancies contribute something distinct enough at the point of use to merit reporting on two separate scales. Clearly, the multidimensionality route provides the statistically better fitting solution, however, is this enough to require or allow a meaningful division of the scores? There is some evidence of multidimensionality from item fit statistics. However, it was demonstrated that item fit statistics should not be the only criterion used to guide decisions, as they can be sensitive to non-construct relevant variance. It would, therefore, be acceptable to conclude that there is no compelling evidence that these items require to be reported on separate scales. Indeed, in the case of the Aptis test, the grammar and vocabulary are reported as a single score (O'Sullivan and Dunlea, 2015). This reporting structure is supported empirically in a study, which marries bifactor analysis with other methodologies to generate a battery of evidence on which to base dimensionality considerations (McCray and Dunn, in press). Treating this set of grammar and vocabulary items on the same scale can be viewed as reflecting both insights about the underlying constructs the two sets of items are designed to measure, as well as the onward consequences and application of the score. A point to note is that this decision regarding the reporting structure of this Aptis test component does not necessarily generalize all grammar and vocabulary items as operationalized in other testing scenarios.

ILLUSTRATIVE STUDY 2 – APTIS FOR TEENS FOUR SKILLS

In this section, we turn our attention to a commonly specified theoretical model in language testing, the four-skill model (e.g., Stricker and Rock, 2008; Sawaki et al., 2009; Sawaki and Sinharay, 2013, 2017; In'nami et al., 2016). The four-skill model posits that the receptive skills reading and listening, along with the productive skills, speaking and writing, are fundamental, divisible, and separately scorable abilities as part of the construct

of general L2 ability. Here, we fit four different models: (i) unidimensional, (ii) correlated factors, (iii) higher-order, and (iv) bifactor model, and compare the inferences about underlying dimensionality, we can make from each. The models fitted are illustrated in **Figure 2**.

Study 2 Dataset

The current illustrative example utilizes data from the Aptis for Teens test. This is a variant within the British Council's Aptis suite of tests designed for the use of learners of English aged between 13 and 17 years. Further information about this test is available in the Aptis for Teens Candidate Guide (British Council, 2019b).

The scoring system for the test components is different for receptive and productive skills. Listening and reading each comprise a series of four testlets, which address a candidate's ability to interpret an input text common to each testlet (written or aural, as relevant). Each item is scored dichotomously, though the independence assumption is violated to some extent by the testlet format. Speaking and writing, meanwhile, require the test taker to respond to a series of four tasks each, which are submitted for marking by a human rater who apportions a score to a maximum of between 4 and 7 points depending on the task. Only three tasks from the speaking component are included in the modeling exercise, as this component allocates the first task randomly from a pool of items, leading to a large degree of structural missingness. It would not be possible to retain the response data from this task for analysis without introducing inconsistencies into the analysis.

Score data analyzed in the current study is taken from a sample of 1,432 15-year-old students from the Madrid region of Spain who sat the test in 2017 as part of a wider British Council project (Shepherd and Ainsworth, 2017)[3]. Full involvement and approval of Madrid Ministry of Education was obtained prior to conducting the original study. Individual participation in the study was contingent on receiving written parental consent, with conditions agreed with the Madrid government.

Study 2 Method

The methodology for this illustrative study follows that of study 1 (see above), with the addition that the higher-order model is also fit to this dataset. With each of the four-skill factors as first-order factors, the higher-order model is identified and, therefore, a statistically viable alternative to consider. The Mplus code used for the analysis is available in **Supplementary Data Sheet 1**. All four models are illustrated in **Figure 2**.

Study 2 Results

Figures 2A–D give traditional factor loading diagrams for the four models used in this section. Note that they are abbreviated, in that all not all observed variables are displayed. **Supplementary Table A2** in the appendix provides the full list of item loadings. For all models, the chi-square *p*-values are statically significant, as can be seen in **Table 2**. Again, however, this is no particular cause for concern. All models have good levels of fit on RMSEA and CFI. In terms of statistical measures of comparative fit, the best fit is achieved by the bifactor model, followed by the correlated factors, then the higher-order model, with the comparatively worst fit yielded by the unidimensional model (though, as noted, still within the acceptable thresholds on key indicators).

The average loading of items on the unidimensional model is 0.62, providing a mean explanation of 38% of the variance of the observed variables. This relatively high loading indicates that meaningful measurement of the construct is taking place (Tabachnick and Fidell, 2007). For the correlated factors model, the average loading of the speaking (0.89) and writing (0.74) are higher than that of the reading (0.67) and listening (0.58) items. This is likely a consequence of the polytomous nature of the response options for the productive skills, generating fewer but strong correlations with the general factor rather than the many weaker, yet still informative, correlations seen in the binary items used in the receptive skills. The average loadings for the subordinate factors in the higher-order model are virtually the same as those in the correlated factors model, unsurprisingly. All four first-order factors load very highly (0.89–0.97) on the general L2 ability factor. In the bifactor model, the average loading on the general factor is very close to that of the unidimensional model (0.62), however, the mean loadings for the grouping factors from 0.27 (listening) to 0.45 (speaking) indicate that there may be persuasive evidence of multidimensionality for some of the grouping factors. In terms of loading MAE scores, there is a large difference between the loadings on the unidimensional model and those on the speaking factor (0.09) of the correlated factors and second-order models. Regarding the MAE values for score, there is a sizeable difference between the scores of the unidimensional model versus those on the speaking (7.2) and, to a lesser extent, writing (3.9) components of the correlated factors model.

Study 2 Interpretation

As mentioned above, the fits statistics indicate that all the models presented offer a reasonable explanation of the data (CFI > 0.95; RMSEA < 0.06). In effect, any confirmatory question we ask about the dimensionality of the test as modeled by any of these models we could justify statistically. In this situation, the value of the different models lies in the information they give us about the comparative ways of handling the dimensionality of the test. For example, based on model fit statistics alone, if we were to ask the question "Is this test unidimensional enough to treat on a single scale?" We would cite RMSEA 0.045 and CFI 0.952 and answer "yes." However, as we saw above, selecting purely based on comparative fit statistics is likely unwise. The difference between speaking and writing compared with general L2 ability scores from the unidimensional model highlighted by the MAE value indicates a practical need to report scores on more than one dimension. The general L2 ability score is not a suitable proxy for the writing and, to a lesser extent, speaking skills measured by Aptis.

The latent variables in the correlated factors model all correlate strongly, the highest being between writing and listening

[3]Not all cases from the original study were available for item-level analysis, so this represents a sub-set of the full project dataset. Exclusions were related to a technical aspect of version allocation and were not contingent on candidate-related factors.

TABLE 2 | Four-skills model fit and parameter comparisons.

	Model descriptive statistics			
Fit measure	**Unidimensional**	**Correlated factors**	**Bifactor**	**Higher order**
Chisq	5627 (1430), $p < 0.001$	4255 (1424), $p < 0.001$	2454 (1375), $p < 0.001$	4273 (1426), $p < 0.001$
RMSEA	0.045	0.037	0.023	0.037
CFI	0.952	0.968	0.988	0.967
AIC	89592	88384	87315	88446
BIC	90329	89152	88612	89204
ADJ BIC	89884	88687	87992	88746
Mean (SD) of factor loadings on observed variables				**Loadings on higher-order factor**
General	0.62 (0.15)	*NA*[a]	0.62 (0.15)	0.93 (0.42)	*NA*[b]
Listening	*NA*[a]	0.58 (0.19)	0.13 (0.11)	0.58 (0.19)	0.97
Reading	*NA*[a]	0.67 (0.14)	0.13 (0.27)	0.67 (0.14)	0.90
Speaking	*NA*[a]	0.89 (0.02)	0.45 (0.04)	0.89 (0.02)	0.89
Writing	*NA*[a]	0.74 (0.10)	0.27 (0.08)	0.75 (0.10)	0.96
Factor Correlations				
LI-RE	*NA*[c]	0.90 (0.01)	*NA*[c]	*NA*[c]
LI-SP	*NA*[c]	0.86 (0.01)	*NA*[c]	*NA*[c]
LI-WR	*NA*[c]	0.91 (0.01)	*NA*[c]	*NA*[c]
RE-SP	*NA*[c]	0.76 (0.01)	*NA*[c]	*NA*[c]
RE-WR	*NA*[c]	0.86 (0.01)	*NA*[c]	*NA*[c]
SP-WR	*NA*[c]	0.89 (0.01)	*NA*[c]	*NA*[c]

Mean Absolute Error (MAE)[d] of parameterscompared toequivalent(s) in unidimensional model

	Correlated factors				Bifactor	Higher order				
	Li	**Re**	**Sp**	**Wr**	**Gen**	**Gen**	**Li**	**Re**	**Sp**	**Wr**
Factor loadings	0.02	0.03	0.09	0.03	0.02	*NA*[f]	0.01	0.03	0.09	0.04
Factor scores[e]	0.9	1.7	7.2	3.9	1.24	1.02	*NA*[g]	*NA*[g]	*NA*[g]	*NA*[g]

[a]Not provided as the factor was not measured by the model. [b]No loading of general factor on itself. [c]Not provided as no correlation between factors measured in the model. [d]$MAE = \frac{\sum |x_i - y_i|}{n}$ where x is a parameter for the unidimensional model, y is a corresponding parameter from the model in question, and n is the sum of comparisons. [e]Scores set to mean = 50, SD = 10. [f]Not provided as items do not load directly onto the general factor. [g]Not provided as the results of the higher order factor are analogous to those on the unidimensional model, not those of the lower order factors.

($r = 0.91$), and the lowest between speaking and reading ($r = 0.76$). The comparative fit of the higher-order model is not favorable to the correlated factors model, with a marginally lower score on each of the information criterion. Given the good global fit statistics, however (RMSEA 0.037; CFI 0.967), again, in a confirmatory factor analytic approach, this higher-order model would be accepted as a solid way of understanding the factor structure. It is clear from the strong positive loadings of the higher-order factor on the four subordinate factors that this factor is a good summary of the four skill areas, with very little associated error.

At this point, if the researcher is interested in investigating the nature of the multidimensionality further, he or she may wish to model the data using a bifactor model. Looking to the loadings for the skill-specific grouping factors in the bifactor model for these data, a number of points can be observed. The first is that all three speaking tasks have loadings of greater than 0.32 on the grouping factor, with a mean of 0.45. This shows that the grouping factor is explaining more than 20% of the observed variance across these task responses, which is suggestive of a systematic deviation from the variance explained by the general factor. While the mean loadings on the grouping factor for writing are not as high at 0.27, it is still markedly higher than the mean loadings for reading and listening at 0.13. In some respects, it is not unexpected to see this pattern, given the role of individual difference in explaining performance in tests of the productive skill areas (see, e.g., Kim and Crossley, 2019). The other grouping factor with several items loading higher than 0.32 is reading. In this case, however, this pattern is only observed for items associated with "task 2" in the reading component. This indicates that individual item responses associated with this particular task have a strong dependency distinct from the explanation provided by the general factor[4]. In this respect, the

[4]The scoring approach taken for this particular testlet was, in fact, revised since the recording of the scores in the current dataset, owing to issues with dependence and the representation of construct (Spiby and Dunn, 2018).

bifactor model has highlighted a source of systematic construct irrelevant variance.

This brief illustration has shown the dual usefulness of the bifactor model in estimating the magnitude of dependencies between sub-groups of items beyond the general ability hypothesized. The additional variance on the speaking and writing tasks may validly be attributable to a feature of test performance that is distinct from the overall L2 ability represented by the general factor. Meanwhile, in the case of the reading test, we see an example of what Reise et al. (2010) highlight as *nuisance* dimensions – "factors arising because of content parcels that potentially interfere with the measurement of the main target construct" (Reise et al., 2010, p. 5). In cases where the skill-specific grouping factors indicate substantial degree of shared variance over and above the explanation provided by the general factor, it rests on the researcher to bring their knowledge of the item content and the context of testing to the interpretation, and ultimately whether this is viewed to be worth accounting for separately.

DISCUSSION

The two illustrative examples presented above show how dimensionality might be assessed in language testing-specific contexts using CFA models. In the first example, looking at grammar and vocabulary items, evidence of multidimensionality was indicated by model fit statistics. However, it was shown in various ways, e.g., similar loading and factor scores between uni- and multidimensional models, that the practical ramifications of ignoring that multidimensionality would be small. Furthermore, the dominance of a small number of items on the bifactor skill-specific grouping traits led to the conclusion that there would be little to be gained from splitting and reporting separate scores for grammar and vocabulary in this case. The second illustrative study examined the dimensionality of the four-skill model. Again, evidence from item fit indicated multidimensionality. However, in this example, the need to report sub-scores for the productive skills (i.e., speaking and writing) was indicated by the fit of the correlated factor model, the high loadings on some bifactor grouping traits, and the large differences between the MAE values on factor loading and score for some subscales. In both cases, the bifactor model, alongside the more traditionally fitted models greatly aided the evidence-gathering process. This highlights a central methodological point, that rather than viewing the bifactor model as providing an opposing latent structure to test against, it should be understood as providing the researcher with the capacity to investigate the assumption of a combination of general and skill-specific abilities more thoroughly. This is consistent with the first of our standpoints stated at the outset of the studies, encouraging a move away from the approach, which asks, "which CFA model fits the data best?" toward understanding each CFA model as a tool, offering related, but distinct, insights to the researcher. In fact, the nesting relationship between the bifactor model and higher-order model already alluded to in this paper, shows the bifactor to be the less restricted of the two models. The bifactor model is, therefore, often able to more flexibly account for variance in the data than the higher-order model, and is thus more likely to yield favorable fit statistics when modeling real-world data with potentially unaccounted for complexities (for a more detailed explanation, see Yang et al., 2017).

Recalling the second standpoint posited at the beginning of the paper, it has been advanced here that the researcher is best placed to initiate their investigation from an understanding that tests are not either unidimensional *or* multidimensional, but that all tests with more than one item are multidimensional to some extent (Gustafsson and Åberg-Bengtsson, 2010; Reise et al., 2010). Echoing Reise et al.'s statement that, "when a scale is subjected to "confirmatory" factor analyses, the conclusion is, almost without exception, that the data are multidimensional" (Reise et al., 2010, p. 16), we found evidence of multidimensionality from the comparison of fit statistics between the unidimensional and multidimensional models in both illustrative studies. The four models presented and discussed in this paper, rather than being viewed as competitors in providing the *best* explanation of a dataset, via model selection of minimal AIC/BIC or some other criterion, can be seen as tools to be employed in exploring and understanding the latent structure of a test. We would suggest, again in line with Reise et al. (2010), that some method of assessing the practical impact of multidimensionality be undertaken. In practice, this means answering dimensionality questions by scrutinizing the nature and relative size of loading estimates rather than solely through comparisons of model fit. As illustrated in the studies described above, this could take the form of looking at the differences in loadings in scores between the uni- and correlated factors model, or, equally, by examining the size and distribution of the loadings on the grouping factors of the bifactor model. Relatively uniform loadings on the grouping factor indicate score variance common to all subscale items that is untapped by the general factor. The magnitude of the loadings is reflective of the extent to which reporting a separate score for that factor is important. Non-uniform loadings indicate correlations between specific items, which should be investigated further. This level of detail enables the researcher to pick up on nuances that are not so easily discernible from higher-order model estimates[5].

To elaborate further using the example of modeling four-skill data, employing the higher-order model in a CFA framework has often been a natural step to take, since this factor structure provides an intuitive reflection of the current theoretical conceptions of language tests (Stricker and Rock, 2008; Sawaki et al., 2009; Harsch, 2014). In order to understand the closeness of the relationship between the sub-skills and the overarching factor in this model, the researcher will look to the disturbance estimates (the error associated with the first-order factors) against each of the subordinate factors. In fact, these disturbance estimates in the higher-order factor model are analogous to the skill-specific

[5]Unless transformed using the Schmid–Leiman transformation referenced above. Regardless, the higher-order model imposes a greater degree of constraint on the relationship between the general factor and the observed variables (Markon, 2019).

grouping factors in the bifactor model (Reise et al., 2010, p. 5). From the single disturbance estimate for each skill area, we would have been able to discern slightly larger overall disturbance estimates for writing and, in particular, speaking, compared to the other skill areas. This could lead to the same conclusion regarding a distinct underpinning of the speaking, and also writing components, perhaps due to individual differences. However, from a single disturbance estimate, it would not be possible to identify clusters of items within an individual skill area that might be the driving source of additional variance. This was seen in the four-skill bifactor model above, where reading items grouped in a single testlet displayed an interdependence distinct from the general factor. This extra information about individual observed variables highlights the added value from the bifactor model.

Broadening this out to a consideration of how the bifactor model can be understood as complimentary to the higher-order model, it is useful to bring in considerations of the similarities between the two models. As observed by Markon, "These two paradigms differ in how levels of abstraction are modeled: In one, superordinate factors are at a greater level of abstraction because they influence subordinate factors; in the other, superordinate factors are at a greater level of abstraction because they influence a greater breadth of observed variables" (Markon, 2019, p. 53). This thinking is also presented by Gustafsson and Åberg-Bengtsson (2010). These researchers suggested that it is a misconception to distinguish between the two models based on the differing "distance from reality" of the general factors, i.e., whether they load directly on the observed variables. They highlight the fact that both models share two types of factors, exerting broad and narrow influence, respectively, with the key difference between models lying on whether a simple or complex structure is retained, rather than any fundamental distinction in theoretical underpinnings. While on the face of it, the bifactor model is more complex as a latent structure than the higher-order model, the interpretation of the variance explanation becomes much more straightforward, owing to the clear separation between general and grouping factors. The bifactor model can, therefore, be recognized as a powerful means of assessing multidimensionality assumptions, in a manner that is consistent with current theoretical understandings of the latent structure of language tests.

When making a decision about how scores are to be reported, it should equally be recognized that the statistical evidence is only one consideration. Researchers should look to both explore the observed properties of the responses, as well as to establish a structure that suits the data in the light of the uses and interpretations that the test score report will need to fulfill. Often, the expectation of score users, whether rightly or wrongly, may trump the measurement considerations. For example, although we found evidence in illustrative study 2 that there is no prohibition against reporting the receptive skills on the same scale, doing so would represent such a break with conventions that is unlikely to be implemented in a large-scale test in practice. We agree with Rijmen (2010) who comments that good statistical practice should balance the modeling and empirical fit considerations with substantive theory. However, we would argue that in language testing, stakeholder expectations also need serious consideration.

CONCLUSION

This paper illustrated how the bifactor model can be used alongside other traditionally employed psychometric models to assess the underlying dimensional structures of the construct(s) measured by a test. Fundamentally, the bifactor model lets the researcher look in detail at what variance is common in a subscale that is *not* explained by a general factor. An examination of the patterns and magnitudes of the loadings not explained by a general factor is tremendously valuable for assessing the weight of evidence for uni- or multidimensionality and also for diagnosing problematic groups of items. Through the illustrative examples, each of which came to substantively different conclusions about the dimensionality of the test, it is hoped that a template for the usage of the bifactor model in language testing research has been provided, and recommendations have been given on how to approach inference from the model. We have argued for, and hope to see, a more multifaceted approach to dimensionality assessment through CFA in the future that not only takes account statistical model fit and theoretical presuppositions but also considers the practical impact of score/subscore reporting and stakeholder expectations of what will be reported in the final analysis.

DATA AVAILABILITY STATEMENT

Commercial confidentiality means the data are not publicly available. However, access to Aptis test data is available for research purposes to British Council Assessment Research Grant holders. Further information at: https://www.britishcouncil.org/exam/aptis/research/assessment-advisory-board/awards/assessent-grants.

AUTHOR CONTRIBUTIONS

KD and GM made substantial contributions to the conception of the project, drafting and critical redrafting of the manuscript. KD ran the statistical models. Both authors worked in detail on individual aspects of the analysis and interpretation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01357/full#supplementary-material

REFERENCES

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Alderson, J. C., and Kremmel, B. (2013). Re-examining the content validation of a grammar test: the (im)possibility of distinguishing vocabulary and structural knowledge. *Lang. Test.* 30, 535–556. doi: 10.1177/0265532213489568

British Council (2019a). *Aptis Candidate Guide*. Available online at: https://www.britishcouncil.org/sites/default/files/aptis_candidate_guide-web_0.pdf (accessed May, 2020).

British Council (2019b). *Aptis for Teens Candidate Guide*. Available online at: https://www.britishcouncil.org/sites/default/files/aptis_for_teens_candidate_guide_0.pdf (accessed May, 2020).

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York, NY: The Guildford Press.

Brunner, M., Nagy, G., and Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *J. Pers.* 80, 796–846. doi: 10.1111/j.1467-6494.2011.00749.x

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., and Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: a comparison of the bifactor model to other approaches. *J. Personal.* 80, 219–251. doi: 10.1111/j.1467-6494.2011.00739.x

Council of Europe (2001). *Common European Framework of Reference for Languages*. Cambridge, MA: Cambridge University Press.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *Int. J. Test.* 13, 354–378. doi: 10.1080/15305058.2013.799067

Gustafsson, J. E., and Åberg-Bengtsson, L. (2010). "Unidimensionality and the interpretability of psychological instruments," in *Measuring Psychological Constructs*, ed. S. Embretson (Washington DC: American Psychological Association), 97–121. doi: 10.1037/12074-005

Haberman, S. J. (2008). When can subscores have value? *J. Educ. Behav. Stat.* 33, 204–229. doi: 10.3102/1076998607302636

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

Holzinger, K. J., and Swineford, F. (1937). The bi-factor method. *Psychometrika* 2, 41–54. doi: 10.1007/bf02287965

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

In'nami, Y., and Koizumi, R. (2012). Factor structure of the revised TOEIC® test: a multiple-sample analysis. *Lang. Test.* 29, 131–152. doi: 10.1177/0265532211413444

In'nami, Y., Koizumi, R., and Nakamura, K. (2016). Factor structure of the test of english for academic purposes (TEAP®) test in relation to the TOEFL iBT® test. *Lang. Test. Asia* 6:3. doi: 10.1186/s40468-016-0025-9

Kim, M., and Crossley, S. A. (2019). *LatentStructure of the ECCE: Discovering Relationships Among Reading, Listening, Writing, Speaking, and Lexico-Grammatical Ability (2018-01 ed.)*. Available online at: https://michiganassessment.org/wp-content/uploads/2020/02/20.02.pdf.Res_.LatentStructureoftheECCE-DiscoveringRelationshipsamongReadingListeningWritingSpeakingandLexico-GrammaticalAbility.pdf

Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *Modern Lang. J.* 75, 196–204. doi: 10.2307/328827

Markon, K. E. (2019). Bifactor and hierarchical models: specification, inference, and interpretation. *Annu. Rev. Clin. Psychol.* 15, 51–69. doi: 10.1146/annurev-clinpsy-050718-095522

McCray, G., and Dunn, K. J. (in press). *Validity and Usage of the Aptis Grammar and Vocabulary (Core) Component: Dimensionality Investigations*. London: British Council.

Muthén, L. K., and Muthén, B. O. (2017). *Mplus User's Guide*, 8th Edn. Los Angeles, CA: Muthén & Muthén.

O'Sullivan, B., and Dunlea, J. (2015). *Aptis General Technical Manual Version 1.0*. Available online at: https://www.britishcouncil.org/sites/default/files/aptis_general_technical_manual_v-1.0.pdf (accessed May, 2020).

Park, G.-P. (2004). Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Lang. Ann.* 37, 448–458. doi: 10.1111/j.1944-9720.2004.tb02702.x

Purpura, J. (1999). *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach*. Cambridge, MA: Cambridge University Press.

Reise, S. P. (2012). Invited paper: the rediscovery of bifactor measurement models. *Multivariate Behav. Res.* 47, 667–696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., and Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *J. Pers. Assess.* 95, 129–140. doi: 10.1080/00223891.2012.725437

Reise, S. P., Bonifay, W. E., and Haviland, M. G. (2018). "Bifactor modelling and the evaluation of scale scores," in *The Wiley Handbook of Psychometric Testing: A Mulitidiciplinary Reference on Surevey, Scale and Test Development*, Vol. 2, eds P. Irwing, T. Booth, and D. J. Hughes (Hoboken, NJ: John Wiley & Sons, inc), 677–708.

Reise, S. P., Moore, T. M., and Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J. Pers. Assess.* 92, 544–559. doi: 10.1080/00223891.2010.496477

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and asecond-order multidimensional IRT model. *J. Educ. Meas.* 47, 361–372. doi: 10.1111/j.1745-3984.2010.00118.x

Rodriguez, A., Reise, S. P., and Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *J. Pers. Assess.* 98, 223–237. doi: 10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., and Haviland, M. G. (2016b). Evaluating bifactor models: calculating and interpreting statistical indices. *Psychol. Methods* 21, 137–150. doi: 10.1037/met0000045

Sawaki, Y., and Sinharay, S. (2013). Investigating the value of section scores for the TOEFL iBT ® test. *ETS Res. Rep. Ser.* 2013, i–113. doi: 10.1002/j.2333-8504.2013.tb02342.x

Sawaki, Y., and Sinharay, S. (2017). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Lang. Test.* 35, 529–556. doi: 10.1177/0265532217716731

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

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodnessof-fit measures. *Methods Psychol. Res. Online* 8, 23–74.

Schmid, J., and Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika* 22, 53–61. doi: 10.1007/bf02289209

Shepherd, E., and Ainsworth, V. (2017). *English Impact: An evaluation of English Language Capability*. Madrid: British Council.

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Lang. Test.* 22, 31–57. doi: 10.1191/0265532205lt296oa

Shiotsu, T. (2010). *Components of L2 Reading: Linguistic and Processing Factors in the Reading Test Performances of Japanese EFL Learners*. Cambridge, MA: Cambridge University Press and Cambridge ESOL.

Shiotsu, T., and Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Lang. Test.* 24, 99–128. doi: 10.1177/0265532207071513

Slocum-Gori, S. L., and Zumbo, B. D. (2010). Assessing the unidimensionality of psychological scales: using multiple criteria from factor analysis. *Soc. Indicators Res.* 102, 443–461. doi: 10.1007/s11205-010-9682-8

Spiby, R., and Dunn, K. (2018). "Cohesive scoring devices: towards fair assessment of sentence-sequencing reading tasks," in *Paper presented at the 15th European Association for Language Testing and Assessment Conference*, Bochum.

Stricker, L. J., and Rock, D. A. (2008). *Factor Structure of the TOEFL® Internet-Based Test Across Subgroups (Vol. No. TOEFL-iBT-07)*. Princeton, NJ: ETS.

Stricker, L. J., Rock, D. A., and Lee, Y.-W. (2005). *Factor structure of the LanguEdge test Across Language Groups*. Princeton, NJ: ETS.

Tabachnick, B. G., and Fidell, L. S. (2007). *Using Multivariate Statistics*, 5th Edn. New York, NY: Allyn & Bacon.

Thurstone, L. (1944). Second-order factors. *Psychometrika* 9, 71–100. doi: 10.1007/bf02288715

van Zeeland, H., and Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Appl. Linguist.* 34, 457–479. doi: 10.1093/applin/ams074

Vandenberg, R. J. (2006). Statistical and methodological myths and urban legends. *Organ. Res. Methods* 9, 194–201. doi: 10.1177/1094428105285506

Yang, R., Spirtes, P., Scheines, R., Reise, S. P., and Mansoff, M. (2017). Finding pure sub-models for improved differentiation of bi-factor and second-order models. *Struct. Equ. Model.* 24, 402–413. doi: 10.1080/10705511.2016.1261351

Yung, Y.-F., Thissen, D., and McLeod, L. D. (1999). On the relationship between the higher order factor model and the hierarchical factor model. *Psychometrika* 64, 113–128. doi: 10.1007/bf02294531

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.

frontiers in Psychology

REVIEW
published: 03 September 2020
doi: 10.3389/fpsyg.2020.02123

Assessing Second Language Listening Over the Past Twenty Years: A Review Within the Socio-Cognitive Framework

Lianzhen He and Ziyun Jiang*

Institute of Applied Linguistics, Zhejiang University, Hangzhou, China

The assessment of second language (L2) listening has received much attention. To understand the state-of-the-art research on L2 listening assessment, a total of 87 studies published in 14 peer-reviewed journals and two research report series between 2001 and 2020 were reviewed, using the socio-cognitive framework for developing and validating listening tests proposed by Weir (2005). Thirteen research themes were identified in relation to the six components of the framework, including test-taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity. Context validity was the most investigated component, covering three research themes, that is, task setting, linguistic demands (input and output), and speakers. Based on a detailed analysis of the 13 research themes, recommendations for future research in L2 listening assessment were given.

Keywords: second language listening assessment, socio-cognitive framework, listening comprehension, research theme, validity

OPEN ACCESS

Edited by:
Vahid Aryadoust,
National Institute of Education,
Nanyang Technological University,
Singapore

Reviewed by:
Franz Holzknecht,
University of Innsbruck, Austria
Andrew Wolvin,
University of Maryland, United States

****Correspondence:***
Lianzhen He
hlz@zju.edu.cn

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

Received: *14 May 2020*
Accepted: *30 July 2020*
Published: *03 September 2020*

Citation:
He L and Jiang Z (2020) Assessing Second Language Listening Over the Past Twenty Years: A Review Within the Socio-Cognitive Framework. Front. Psychol. 11:2123. doi: 10.3389/fpsyg.2020.02123

INTRODUCTION

Listening is the most frequently used mode of human communication, and "more than forty-five percent of our total communication time is spent in listening" (Feyten, 1991, p. 174). As one of the crucial components of successful human communication (Field, 2008; Rost, 2011), listening lies at "the heart of language learning" (Vandergrift, 2007, p. 191) and facilitates second language (L2) learning (Buck, 2018; Ockey and Wagner, 2018). As a multidimensional construct, listening consists of affective, behavioral, and cognitive processes (Halone et al., 1998; Worthington and Bodie, 2017). Assessing such a complex construct is challenging (Brindley, 1998; Buck, 2001, 2017; Vandergrift, 2007; Wagner, 2013b) and has become a focus of listening scholarship due to its significant role in education, politics, and society (Weir, 2013), with considerable efforts made to provide measures that are valid indicators of listening (Bodie and Worthington, 2017). Compared with listening in a first language (L1), L2 listening has more comprehension barriers which require L2 listeners to perform additional processes (Flowerdew and Miller, 2005).

Over the past 20 years, the field of L2 listening assessment has witnessed important development, and the importance of authenticity has been particularly underscored (Elliott and Wilson, 2013; Ockey and Wagner, 2018). An authentic assessment requires that the way test takers interact with the task corresponds to their use of language in the real-life communication contexts

(Bachman and Palmer, 1996; Buck, 2001). As pointed out by Weir (2005, p. 98), "to test listening we must understand the processing that takes place in real-life situations and attempt to see that communication in our tests is anchored in the real world as far as possible." The growing interest in authenticity has spurred research on the innovation of L2 listening assessment practices. For instance, large-scale standardized tests like the Test of English as a Foreign Language Internet-based Test (TOEFL iBT) were driven to embrace a wider view of listening (Weir and Vidakovic, 2013) and incorporate integrated tasks that involve listening and other skills (i.e., reading, speaking, and writing). Meanwhile, advances in computer technology have not only improved the quality of acoustic input in L2 listening assessment (Geranpayeh and Taylor, 2013) but also caused a surge of interest in the development and application of video-based listening (e.g., Wagner, 2010b), cognitive diagnostic assessment (e.g., Lee and Sawaki, 2009), computerized dynamic assessment (e.g., Poehner et al., 2015), and computerized adaptive testing (e.g., He and Min, 2017). These advances are evidenced by the increasing number of research articles published in peer-reviewed journals and research report series.

A handful of reviews on L2 listening assessment research have been conducted over the past two decades. Some discussed recent development and challenges in the field (e.g., Wagner, 2013b), and others focused on a specific theme of L2 listening assessment (e.g., Taylor and Geranpayeh, 2011). Taylor and Geranpayeh (2011) reviewed approaches to assessing listening for academic purposes. Drawing on the socio-cognitive framework (Weir, 2005), they focused on how to define and operationalize the construct of academic listening proficiency. These reviews provide helpful insights into the complex factors and challenges involved in L2 listening assessment. However, a comprehensive understanding of the state-of-the-art research in the field is still lacking, and it is unclear what research themes are important.

This study aims to give a comprehensive review of research on L2 listening assessment in journal articles and research reports published between 2001 and 2020 to facilitate the understanding of the state-of-the-art research in the field and to try to point out avenues for future research. As an influential theory of developing and validating language tests, the socio-cognitive framework (Weir, 2005; Geranpayeh and Taylor, 2013) was used to categorize research themes to make the review more coherent.

THE SOCIO-COGNITIVE FRAMEWORK

The socio-cognitive framework (Weir, 2005) views the ability to be tested as the mental processes of test takers and conceives the use of language as a social rather than a purely linguistic phenomenon (Taylor, 2013). In relation to four macro skills of reading, listening, speaking, and writing, the framework has been widely used in a variety of contexts, especially in test development and validation projects. A typical example is its application in the validation of University of Cambridge ESOL Examinations (Shaw and Weir, 2007; Khalifa and Weir, 2009; Taylor, 2012; Geranpayeh and Taylor, 2013). Although the framework has been criticized for separating out many types of validity, which is a departure from Messick's (1989) unitary theory of validity (Knoch and Chapelle, 2018), it presents a unified approach to conceptualizing and assembling different types of validity evidence in a comprehensive and coherent way (Taylor, 2013). In addition, it provides a transparent and plausible system for researchers and helps to analyze the key features of L2 listening assessment (Taylor and Geranpayeh, 2011, 2013). Therefore, it is considered suitable for the review of research on L2 listening assessment.

The framework contains six key components, namely test-taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity (Weir, 2005). The first component is test-taker characteristics, which is divided into three types – physical/physiological characteristics, psychological characteristics, and experiential characteristics. Test-taker characteristics should be considered "at every stage of test development and continuously throughout live administrations of a test" (Taylor and Geranpayeh, 2013, p. 323). It is necessary that test developers attempt to design tests to elicit test-takers' best performance through understanding test-taker characteristics and promoting feelings of comfort in test takers (Bachman and Palmer, 1996).

Related to test-takers' cognitive or mental processing activated by the test task, the second component is cognitive validity, which addresses the extent to which test tasks require test takers to engage in cognitive processes that resemble those employed in a real-life listening situation (Field, 2013). Given that L2 listening involves a complex mechanism, the importance of understanding cognitive processes in L2 listening assessment has been underscored (Weir, 2005; Field, 2013). Drawing upon Cutler and Clifton (1999) model of L1 listening, Field (2013) presented a five-level processing model of L2 listening including input decoding, lexical search, parsing, meaning construction, and discourse construction, which can be divided into lower-level processing (i.e., input decoding, lexical search, and parsing) and higher-level processing (i.e., meaning construction and discourse construction).

The third component, context validity, concerns the contextual parameters of the test task, including linguistic content parameters and sociocultural contexts (Taylor, 2013), and is related to the extent to which test tasks are "representative of the larger universe of which the test is assumed to be a sample" (Weir, 2005, p. 19). Context validity is affected by multiple aspects, including task setting, administration, linguistic demands (task input and output), and speakers. These aspects are important to the development of tasks that are representative of the target language use (TLU) domain and the target language proficiency levels (Elliott and Wilson, 2013).

As the fourth component, scoring validity is related to the reliability of test scores and all aspects of the scoring process (Weir, 2005; Geranpayeh, 2013). The parameters of scoring validity include test difficulty, item bias, internal consistency, error of measurement, and grading and awarding. Developing valid items in terms of cognitive and contextual parameters

matters little if student responses are not reported consistently (Taylor and Geranpayeh, 2013), so examination boards must devote considerable efforts to all aspects of scoring validity (Geranpayeh, 2013).

The fifth component, consequential validity, is concerned with test washback and impact and is closely related to fairness and ethics (Taylor, 2005; Hawkey, 2013). Test washback refers to the effect of tests on teaching and learning, and test impact is related to wider influences of tests in terms of educational systems and society in general (Hawkey, 2006, 2013). When tests are misused or abused, they can be viewed as unethical and unfair (Shohamy, 1997) and entail detrimental consequences for stakeholders (Bachman and Palmer, 2010). Therefore, it is important for test developers to consider the intended and unintended influences of tests (Bachman and Palmer, 2010).

The last component is criterion-related validity, including three aspects – comparison with different forms of the same test, cross-test comparability, and comparability with external standards and frameworks. Criterion-related validity is important because there would be no basis for meaningful score interpretation if different forms of a test are not comparable or tests which measure the same ability yield results that are not comparable to each other (Lim and Khalifa, 2013). In addition, it is necessary that the relationship between tests and external realities is consistently appropriate (Lim and Khalifa, 2013) because external standards and frameworks situate tests within larger contexts, which enhances the transparency and meaning of test results (Lim and Khalifa, 2013; Papageorgiou et al., 2019).

MATERIALS AND METHODS

Given the time and space limit, 14 peer-reviewed journals were targeted due to their relevance to the present study and the quality of the articles published in those journals. In addition, Educational Testing Service (ETS) and the International English Language Testing System (IELTS) research report series were included to provide a comprehensive picture of L2 listening assessment research. The two research report series were chosen because they include rigorous studies conducted by leading researchers from all over the world.

The articles and research reports were retrieved online *via* keyword search. Variations of the following terms were used in the search: *listening assessment*, *listening test*, and *listening task*. Two selection criteria were used in our examination of the titles and/or abstracts of the studies: (1) the study involved L2 test takers and focused on L2 listening assessment, or it investigated the assessment of multiple skills with specific discussion on L2 listening assessment and (2) the study was an empirical study or a systematic review. A total of 89 studies – 79 journal articles and 10 research reports – were initially retrieved. After careful reading of all the studies, two research reports were excluded because they had the same research design and used the same data with two journal articles included in the current study, resulting in a final dataset of 87 studies. **Table 1** presents the number of studies included in the dataset for the current study.

Table 2 presents a coding scheme based on the socio-cognitive framework (Weir, 2005; Geranpayeh and Taylor, 2013). The coding was done manually. First, the two authors read each study carefully and coded it independently. Some studies were coded into more than one category since they investigated multiple components of the socio-cognitive framework. The initial intercoder agreement was high, reaching 89.66%. Incongruence between the coding results was discussed between the authors, and another expert in the field was invited if the incongruence remained unresolved. For instance, the authors

TABLE 1 | Number of articles taken from the 14 journals and two research report series.

Journal/research report series	Number of selected articles	%
Language Testing	20	22.99
Language Assessment Quarterly	14	16.09
System	11	12.64
TESOL Quarterly	7	8.05
Applied Linguistics	5	5.75
IELTS Research Report Series	4	4.60
Language Learning	4	4.60
ETS Research Report Series	4	4.60
Journal of English for Academic Purposes	3	3.45
Modern Language Journal	3	3.45
Studies in Second Language Acquisition	3	3.45
Computer Assisted Language Learning	2	2.30
Foreign Language Annals	2	2.30
Frontiers in Psychology	2	2.30
Language Learning and Technology	2	2.30
Journal of Educational Research	1	1.15
Total	87	100

TABLE 2 | The coding scheme based on the socio-cognitive framework.

Components		Research themes
Test-taker characteristics	1	Physical/physiological characteristics
	2	Psychological characteristics
	3	Experiential characteristics
Cognitive validity	4	Cognitive processes
	5	Task setting
	6	Setting: administration
Context validity	7	Linguistic demands (task input and output)
	8	Speakers
	9	Test difficulty
	10	Item bias
Scoring validity	11	Internal consistency
	12	Error of measurement
	13	Grading and awarding
Consequential validity	14	Washback on individuals in classroom/workplace
	15	Impact on institution and society
Criterion-related validity	16	Comparison with different forms of the same test
	17	Cross test comparability
	18	Comparability with external standards and frameworks

disagreed on the coding of Wei and Low (2017), a study on test-takers' score change pattern and increase rate. After discussion with the expert, the authors agreed that this study should be coded into comparison with different forms of the same test under criterion-related validity.

RESULTS

Five out of the 18 research themes in the coding scheme were not addressed in our dataset, that is, administration, test difficulty, error of measurement, impact on institution and society, and comparison with different forms of the same test. Therefore, only 13 research themes were identified, as is shown in **Table 3**. Among the six components, context validity was the most investigated (N = 57, 65.52%), followed by test-taker characteristics (N = 21, 24.14%), cognitive validity (N = 12, 13.79%), scoring validity (N = 8, 9.2%), criterion-related validity (N = 4, 4.6%), and consequential validity (N = 1, 1.15%). And among the 13 research themes identified, task setting (N = 34, 39.08%) was the most investigated, followed by linguistic demands (task input and output; N = 14, 16.09%) and cognitive processes (N = 12, 13.79%). The 13 research themes will be discussed in detail in the following sections.

Test-Taker Characteristics

Physical/Physiological Characteristics

Physical/physiological characteristics cover obvious biological features shared by test takers like gender and age, short-term ailments like a heavy cold, and long-term disabilities such as dyslexia (O'Sullivan, 2000; Weir, 2005; Elliott, 2013). A common approach to investigating physical/physiological characteristics is differential item functioning (DIF) analysis, which is used to detect the variation of responses across different subgroups of test takers. DIF exists when the probability of answering one item correctly differs for subgroups of test takers with comparable ability (Min and He, 2020). Geranpayeh and Kunnan (2007) conducted bias analyses of listening test items of the Certificate in Advanced English examination in terms of age. In their study, test takers were divided into three age groups (i.e., 17 and younger, 18–22, and 23 and older). Although they reported that no age group was clearly disadvantaged, it was observed that the 17 and younger group performed worse than the other two groups. One possible reason was that the test topics were less attractive to younger test takers.

Similarly, researchers investigated whether DIF existed across gender subgroups in listening tests, and gender-based DIF was detected (Park, 2008; Aryadoust et al., 2011). Conducting DIF analysis of the Michigan English Language Assessment Battery (MELAB) listening test, Aryadoust et al. (2011) observed that males with lower listening proficiency were likely to score higher on some items than females and males with higher listening proficiency. Apart from exploring test-takers' responses, recent studies probed into the gender effect in test preparation and test-taking processes. For instance, Chou (2019) investigated whether gender predicted self-efficacy in test preparation for the listening section of the University Entrance Examination test in Taiwan and reported that gender was not associated with self-efficacy, test anxiety, and strategy use. Moreover, Aryadoust et al. (2020) conducted a neuroimaging study and employed functional near-infrared spectroscopy (fNIRS) to uncover the test-takers' neurocognitive mechanisms involved in listening tests. They observed differences in neural substrates across genders, although differences in the test scores of males and females were not statistically significant.

In addition to age and gender, research interest in dyslexia has emerged. Dyslexia is one of the most common learning difficulties test takers have and is categorized into physical/physiological characteristics together with other long-term illnesses or disabilities such as speech defects (O'Sullivan, 2000; Weir, 2005; Elliott, 2013). Dyslexic learners are characterized by the "underlying weakness in the areas of working memory, executive functioning, and processing speed" (Kormos et al., 2019, p. 835). In Kormos et al.'s (2019) study, the listening test performance of young dyslexic and non-dyslexic learners was compared, and dyslexic test takers performed worse than their non-dyslexic peers. In some countries, there is a legal requirement that test takers with specific learning difficulties such as dyslexia should be accommodated (Weir, 2005). However, it is controversial as to what special arrangements should be offered to test takers to make tests assess abilities rather than disabilities, ensuring fair tests for every test taker without compromising test validity is challenging to test developers (Kosak-Babuder et al., 2019).

TABLE 3 | Summary of research themes based on the socio-cognitive framework.

Components		Research themes	Number of articles (%)
Test-taker characteristics	1	Physical/physiological characteristics	7 (8.05)
	2	Psychological characteristics	13 (14.94)
	3	Experiential characteristics	1 (1.15)
Cognitive validity	4	Cognitive processes	12 (13.79)
Context validity	5	Task setting	34 (39.08)
	6	Linguistic demands (task input and output)	14 (16.09)
	7	Speakers	9 (10.34)
Scoring validity	8	Item bias	5 (5.75)
	9	Internal consistency	2 (2.3)
	10	Grading and awarding	1 (1.15)
Consequential validity	11	Washback on individuals in classroom/workplace	1 (1.15)
Criterion-related validity	12	Comparison with different forms of the same test	1 (1.15)
	13	Comparability with external standards and frameworks	3 (3.45)

Fifteen studies (17.24%) were coded into multiple research themes, with 14 (16.09%) coded into two themes and one (1.15%) into three themes.

Psychological Characteristics

Psychological characteristics include cognitive characteristics such as memory and affective characteristics like motivation (Elliott, 2013). Four psychological characteristics have received much research attention, including working memory, metacognition, motivation, and anxiety. Working memory is the ability to "keep track of ongoing mental processes and moment-to-moment changes in the immediate environment" (Logie, 2011, p. 240) and is essential for complex cognitive activities (Olive, 2004). Brunfaut and Revesz (2015) investigated the correlation between test-takers' performance on working memory tasks and 11 listening tasks of Pearson Test of English Academic (PTE Academic). Results showed that test-takers' listening scores were positively correlated with their working memory capacity, and listening tasks assessing local comprehension (i.e., listening for specific details) put higher demands on working memory than those assessing global comprehension (i.e., listening for main ideas).

Metacognition refers to learners' ability to control their thoughts and regulate their own learning (Vandergrift and Goh, 2012), which plays an important role in learning to listen (Vandergrift and Goh, 2012). Researchers have investigated test-takers' use of metacognitive strategies, such as planning for, monitoring, and evaluating listening. More specifically, Wang and Treffers-Daller (2017) used Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006) to measure the effect of metacognition on the listening scores of College English Test Band 4 (CET 4). A significant positive correlation between test-takers' listening scores and metacognitive awareness was reported, although it was relatively low ($r = 0.19$), compared with test-takers' vocabulary size ($r = 0.44$) and general language proficiency ($r = 0.36$).

Closely related to metacognition, motivation is a continuum consisting of amotivation, extrinsic motivation, and intrinsic motivation in self-determination theory (Deci and Ryan, 1985, 1995). Drawing on this theory, Vandergrift (2005) provided empirical evidence for the interplay between motivation and metacognition and for their effect on listening scores. In his study, a greater use of metacognitive strategies was related to a higher level of motivation. Moreover, test-takers' listening scores were correlated negatively with amotivation, while a high level of motivation did not appear to be a reliable predictor of L2 listening proficiency. Another study on motivation was conducted by Xu (2017), who used expectancy-value theory (Wigfield and Eccles, 2000) to conceptualize test-taking motivation. He observed the mediating effect of metacognition on the relationship between motivation and the listening scores of CET 4. The findings revealed that the effect of motivation on listening scores was pronounced, and increased listening metacognitive awareness improved test-takers' listening performance when their motivation level was stable.

Anxiety is another important psychological characteristic explored in our dataset. Foreign language listening anxiety has received some attention, which is the type of anxiety experienced by learners in the listening context, and consists of communication apprehension, test anxiety, and fear of negative evaluation (Horwitz et al., 1986). The negative effect of foreign language listening anxiety was observed by Zhang (2013), who investigated the causal relations between foreign language listening anxiety and IELTS listening test scores and found that anxiety negatively affected test-takers' performance on the IELTS listening test. This negative effect was also observed by Brunfaut and Revesz (2015) who reported that less anxious test takers performed better on the listening section of PTE Academic. Instead of focusing on foreign language listening anxiety, In'nami (2006) explored the relationship between test-takers' test anxiety and performance in familiar listening tasks (i.e., multiple choice questions and open-ended questions) and found that test anxiety did not influence test performance, suggesting that test anxiety can be independent of the other two components of foreign language listening anxiety (i.e., communication apprehension and fear of negative evaluation).

Experiential Characteristics

Experiential characteristics concern test-takers' experience in preparing and taking tests and their familiarity with the test, including test-takers' educational and cultural background (Elliott, 2013). The effect of test-takers' preparation on their IELTS listening test scores was investigated by Winke and Lim (2017), who explored the effects of listening test preparation on listening scores, test-taking strategies, and anxiety. Three types of instruction were given in their study, that is, explicit preparation (i.e., test-taking-strategies instruction and practice tests), implicit preparation (i.e., vocabulary instruction and practice tests), and conversation classes plus a practice test. They found that all of the three types of instruction helped test takers perform better in listening tests, while there were no differential effects on scores, strategy use, or anxiety levels among the three types. They concluded that concise test preparation (i.e., one simple practice test) helped test takers perform better, and extensive test preparation lasting months or years might not be necessary.

Cognitive Validity

It is common that listening is assessed as a composite of several subskills (Worthington, 2017). Listening subskills reflect core cognitive processes measured in L2 listening tests, and researchers have not reached consensus on what subskills make up L2 listening. A popular approach to investigating listening subskills is the use of cognitive diagnosis models. Listening subskills were found to be different in terms of various grain sizes (Sawaki et al., 2009), and the contribution of a particular listening subskill was not consistent across items (Yi, 2017), indicating the vague definition of L2 listening subskills (Aryadoust, 2020). To address this gap, Aryadoust (2020) used the document co-citation analysis to give a systematic review of research on comprehension subskills. An integrative framework of comprehension subskills was provided, which included a total of 18 L2 comprehension subskills.

In addition to listening subskills, items targeting different levels of listening comprehension, such as local (i.e., explicit and factual) and global (i.e., inferential) comprehension, have been investigated. For instance, Becker (2016) examined the

extent to which the two types of items differentiated between test takers with different proficiency levels. Since items targeting different levels of listening comprehension were able to distinguish different proficiency groups, and items targeting local comprehension were easier than those targeting global comprehension for all groups, Becker provided empirical evidence for the hierarchy of cognitive processes and the relative difficulty of items targeting different cognitive processes.

A variety of methods were used to probe into test-takers' cognitive processes, such as stimulated recall protocols, questionnaires, content analysis, and advanced technology. One typical example is Field (2009), who investigated the cognitive validity of a lecture-based note-taking task in the IELTS listening test by comparing the cognitive processes of participants under test and non-test conditions. Evidence in the verbal report revealed that cognitive processes under the two conditions were incongruent. More precisely, participants adopted test-wise strategies under test conditions. Also, the processing of many participants was superficial under test conditions as they focused on lexical matches instead of the overall meaning. Carrell (2007) focused on test-takers' note-taking behavior on academic lecture tasks consisting of multiple-choice questions. A significant correlation between content words in the notes and listening scores was observed and test takers tended to write down content words following the linear order of the lectures instead of using abbreviations, symbols, or paraphrasing. Carrell's study contributed to the understanding of the content and quality of test-takers' notes in L2 listening assessment.

Instead of focusing on tasks that only require listening, Rukthong and Brunfaut (2019) explored the cognitive processes involved in an integrated task (i.e., a listening-to-summarize task). With an increasing popularity, integrated tasks require test takers to complete tasks employing at least two language skills (Rukthong and Brunfaut, 2019) and have been acclaimed for authenticity (Wagner, 2013b) as well as positive washback (Taylor and Geranpayeh, 2011). Based on data collected from a stimulated recall protocol and perception questionnaire, they found that test takers relied on various listening processes, including both higher-level and lower-level processing. The cognitive processes of listening play a crucial role in completing integrated tasks which involve listening.

Advanced technology has been employed in the investigation of cognitive processes, including eye-tracking technology (Suvorov, 2015; Holzknecht et al., 2020) and neuroimaging (Aryadoust et al., 2020). Test-takers' eye movement during the listening test can be recorded by eye-tracking technology to understand their oculomotor engagement with test items, such as the stems and options of multiple-choice questions. For instance, Suvorov (2015) recorded test-takers' eye movement during the video-based listening test including context and content videos, and no significant difference was observed in test-takers' oculomotor engagement with content and context videos. More recently, Holzknecht et al. (2020) observed that test takers paid significantly less attention to later options when answering listening items from the Aptis Test using eye-tracking technology. Aryadoust et al. (2020) investigated brain activation patterns under test conditions using functional magnetic resonance imaging (fMRI). Among the main techniques of understanding how different parts of the brain are engaged in psychological and behavioral functions (Burunat and Brattico, 2017), fMRI has been used by neuroscientists and physicians and was first applied to L2 listening assessment by Aryadoust et al. They introduced the notion of neurocognitive validity, which means that a listening test should engage the neurocognitive processes which are required in real-life contexts. The use of advanced technology has provided deeper insights into cognitive processes, which may have implications for test development and validation.

Context Validity

Task Setting

Task setting is the most investigated research theme in our dataset, which is not surprising due to the important role of task characteristics in L2 listening assessment. A wide range of task setting parameters have been investigated, and the complexity of interactions between these parameters was observed (Brindley and Slatyer, 2002; Brunfaut and Revesz, 2015). Four aspects of task setting received much attention, that is, task purpose and rubric, response method, modality/channel of presentation, and time constraints.

Five studies in our dataset have explored task purpose and rubric. Researchers have investigated listening tasks that are developed for assessing translanguage and those for assessing pragmatic competence. Specifically, Baker and Hope (2019) developed a translanguaged French/English listening task for university professors. In their study, text types were chosen from the TLU domain, including short telephone messages, an introduction and biography of a guest speaker, and a departmental meeting. Also, listening scripts were developed based on the recordings of authentic departmental meeting to incorporate authentic syntactic and discourse functions into the task. In addition to translanguaged listening tasks, pragmatic listening tasks were developed to assess test-takers' ability to comprehend speakers' intentions (Taguchi, 2005, 2007, 2008a,b). Taguchi (2005) incorporated dialogues with the interactive characteristics of spoken English, such as discourse markers, interjections, or hesitation markers, and Taguchi (2008b) gleaned linguistic features from the synthesis of a literature review, survey, and field notes, tapping into different types of implied meaning.

Second, researchers had much interest in response methods, with a particular focus on multiple-choice questions, open-ended questions, partial dictation, and note-taking tasks. A given response method only tests part of the listening construct, and over-reliance on a single response method may lead to construct under-representation (Elliott and Wilson, 2013). Therefore, it is generally desirable to use various response methods in listening assessment (Khalifa and Weir, 2009). For example, 11 different response methods are employed in the listening section of PTE Academic, which are designed to assess a wide range of listening skills (Wei and Zheng, 2017).

As a mainstay of listening assessment, multiple-choice questions provide retrieval cues which facilitate recall of information from the listening input (Chung, 2002).

The prevalence of multiple-choice questions could be attributed to practical benefits such as grading and editing (Elliott and Wilson, 2013). Many issues related to multiple-choice questions have been investigated, including the effect of item preview (Chang and Read, 2006; Yanagawa and Green, 2008; Koyama et al., 2016), the mode of presenting items (Chang and Read, 2013), the language of questions (Filipi, 2012), the number of options (Lee and Winke, 2013), and response order (Holzknecht et al., 2020).

Different from multiple-choice questions, open-ended questions, partial dictation, and note-taking tasks are constructed response formats, which require test takers to formulate their own answers with words or phrases and can effectively evaluate test-takers' listening and their ability to reconstruct what they have heard (Cheng H., 2004). Researchers compared open-ended questions with multiple-choice questions and found that test takers performed better on multiple-choice questions (Chung, 2002; Cheng H., 2004; In'nami and Koizumi, 2009). Targeting partial dictation tasks, Cai (2013) investigated the difficulty and internal consistency of phrasal and single-word partial dictation tasks and found that the two types of partial dictation tasks were comparable. In terms of note-taking tasks, the outline format and blank format of note-taking tasks were explored in Song (2012), who found that note quality indices, especially the number of topical ideas and the organization of notes, were good indicators of listening proficiency, and the outline format was a more reliable measure of L2 academic listening than the blank format.

Third, 14 studies explored modality/channel of presentation, with a particular focus on the use of visual input, such as images and videos. Although the use of visual input is an important aspect of promoting authenticity, whether to use visuals in listening assessment remains open for discussion (Kellerman, 1992; Gruba, 1997; Buck, 2001; Taylor and Geranpayeh, 2011; Wagner and Ockey, 2018). Allowing test takers to employ visual input in understanding the aural input tends to bring about construct-irrelevant variance. Traditionally, L2 listening assessment is "typically concerned with mastery of the language itself, not that of pancultural, *ad-hoc*, gesture-based communication" (Batty, 2015, p. 17). However, trying to separate the effect of visuals from audio elements is unproductive (Gruba, 1997). Most real-life listening involves visual input which aids in comprehension, and various channels are employed by listeners to construct the meaning of what they are hearing (Gruba, 2004, 2006) and videos have become an important part of the listening construct due to the technological advances.

Research on the role of videos in L2 listening tests produced mixed results. Non-verbal information in videos was found to improve test scores (Ginther, 2002; Jones and Plass, 2002; Sueyoshi and Hardison, 2005; Wagner, 2010b, 2013a; Dahl and Ludvigsen, 2014). However, the score difference was not pronounced (Coniam, 2001; Cubilo and Winke, 2013; Batty, 2015; Suvorov, 2015). Using the Rasch model, Batty (2015) found that the difference in item difficulty of video-based and audio-only tasks was small. Test takers varied in their attitudes toward videos, some interacting extensively with videos and preferring video-based tasks to audio-only tasks (Sueyoshi and Hardison, 2005; Ockey, 2007; Wagner, 2007, 2008, 2010a; Cubilo and Winke, 2013), while others reporting that visuals were distracting (Coniam, 2001).

Lastly, as an important aspect of context validity, time constraints have been explored. In L2 listening teaching and assessment practices, the input is sometimes repeated to make the information more comprehensible. However, second hearings are often not possible in the TLU domain, and once-heard texts have greater authenticity (Taylor and Geranpayeh, 2011). Elkhafaifi (2005) found that the repeated exposure to the listening passage improved test-takers' performance, concurring with findings of other studies (Brindley and Slatyer, 2002; Sakai, 2009; Holzknecht et al., 2020). Sakai (2009) divided test takers into two listening proficiency groups according to their pretest scores and explored the interactional effect between repetition and proficiency levels. Their performance on the free written recall tasks in the first and second hearing conditions was compared. Results showed that the repetition of listening passages led to more precise comprehension and was effective for both proficiency groups.

Linguistic Demands (Task Input and Output)

In terms of linguistic demands, the type of input texts (i.e., monologic/dialogic texts and scripted/unscripted texts) has received much research interest. For instance, Read (2002) found that a monologue was significantly easier than a dialogue of the same content. Papageorgiou et al. (2012) examined the difference between monologic and dialogic texts through statistical and content analyses. They found that monologues, compared with dialogues, were more structured and contained additional explicit statements, and the relative difficulty of monologic and dialogic texts varied across items. Apart from monologues and dialogues, unplanned informal conversations and formal written language have been compared. The inclusion of unscripted texts is considered to be more authentic (Wagner, 2013b) and more challenging (Read, 2002; Wagner and Toth, 2014), probably because test takers are more familiar with scripted texts than unscripted texts (Read, 2002) and the spoken input learners hear often consists of textbook texts which lack the characteristics of the unplanned discourse mode (Wagner and Toth, 2014).

Another line of research focused on the role of lexical and grammatical resources in L2 listening tests. The relative importance of lexical and syntactic knowledge in L2 listening test was investigated. It was found that both lexical and syntactic resources played an important role in successful L2 listening, and the role of lexical resources was more important than that of syntactic resources (Cai, 2020; Vafaee and Suzuki, 2020). Furthermore, empirical evidence showed that vocabulary knowledge is a strong predictor of L2 listening performance (Andringa et al., 2012; Matthews and Cheng, 2015; Wang and Treffers-Daller, 2017). Staehr (2009) investigated the depth and breadth of test-takers' vocabulary knowledge and their listening performance and found that a lexical coverage of 98% was needed in the listening test. In van Zeeland and Schmitt (2013) study, most L2 participants understood everyday narrative texts with a lexical coverage of 90–95%. More recently, researchers

have explored the effect of aural vocabulary knowledge (Cheng and Matthews, 2018; Matthews, 2018; Li, 2019), which refers to the knowledge of words mediated through the aural modality (Matthews, 2018). A significant positive correlational relationship between test-takers' aural vocabulary size and listening scores was found (Matthews, 2018; Li, 2019).

In addition, the lexical complexity of listening passages has garnered much research attention. Brunfaut and Revesz (2015) found that the lexical complexity of listening input was significantly correlated with item difficulty. They reported that listening passages including low-frequency phrases were significantly more difficult. However, Paribakht and Webb (2016) did not find any correlation between the lexical coverage of academic words in listening passages and test-takers' listening performance. One possible reason was that other factors such as test-takers' strategy use and content knowledge will impact the outcomes.

Speakers

With the diversity of accents that English speakers are exposed to in the TLU domain for which many listening tests are designed (Taylor and Geranpayeh, 2011), L2 listening assessment has been argued to reveal the changing demographics in English speaking contexts (Ockey and French, 2014) by incorporating accented speech. For example, inner and outer circle English accents have been used in high-stakes listening tests, including the TOEFL iBT, Test of English for International Communication (TOEIC), and IELTS (Kang et al., 2019). However, concerns about the inclusion of non-standard accents have been raised. According to the interlanguage speech intelligibility benefit (Bent and Bradlow, 2003), also called a shared-L1 advantage phenomenon, test takers who share the same L1 with the speakers of listening passages can understand listening materials more easily. If the inclusion of non-standard accents results in a subgroup of test takers being advantaged, using non-standard accents may introduce construct-irrelevant variance (Elliott and Wilson, 2013) and have detrimental effects on test fairness.

Empirical evidence provided partial support for a shared-L1 advantage phenomenon (Major et al., 2002; Harding, 2012; Dai and Roever, 2019; Kang, et al., 2019). Major et al. (2002) found that Spanish-L1 test takers scored higher when listening to Spanish-accented speech, but Chinese-L1 test takers performed worse when listening to Chinese-accented speech. However, Harding (2012) observed that Chinese-L1 test takers were advantaged on Chinese-accented items, while the facilitative effect of L1 accents was not clearly observed in the group of Japanese-L1 test takers. Dai and Roever (2019) divided Chinese-L1 adolescent test-takers into four groups, each of which took one accented version of the same English listening test. Results showed that the Chinese-accented group scored highest, followed by the Spanish, Australian, and Vietnamese-accented groups. Additionally, the beneficial shared-L1 effect was strongest for gap completion items, indicating the highly complex interplay between the effect of accents and task types. Kang et al. (2019) found that Indian-L1 and South African-L1 test takers benefited from their own accent, but they did not observe the shared-L1 effect on test scores because test takers performed significantly better when listening to standard American or British English.

In addition, the effect of accent strength and familiarity has been investigated (Matsuura et al., 2014; Ockey and French, 2014). Ockey and French (2014) developed a strength of accent scale based on salience and comprehensibility and a survey assessing test-takers' familiarity with accents. They found that listening scores decreased as strength of accent increased and familiarity with accents was an advantage for test takers. Likewise, Matsuura et al. (2014) found that L2 listeners performed worse when listening to nonnative English speech, and less familiar accent was more difficult than a more familiar one.

Another line of research focused on the intelligibility of accents (Kang et al., 2018a,b, 2020). Intelligibility refers to the extent to which the speakers' intended utterance is understood by listeners, which is generally measured by transcription tasks (Kang et al., 2018a,b). Kang et al. (2018b) examined the relationship between the phonetic/phonological features of speakers and intelligibility, which helps test developers to select speakers with different English accents for listening input. More recently, Kang et al. (2020) examined the relationship between test-takers' proficiency levels and comprehension of different accents. They found that test-taker's proficiency levels affected their comprehension of accented speech, and the performance of intermediate-level test takers, whose TOEIC scores were between 305 and 400 (i.e., 61–80th percentile), was more sensitive to speech with different accents than the beginner and advanced groups.

Scoring Validity

Item Bias

One important aspect of scoring validity is that test results are free from bias (Weir, 2005). A test may be considered biased when there is systematically differential performance among subgroups of test takers with the same ability (Geranpayeh, 2013). Four studies in the dataset examined if test results biased toward a subgroup of test takers in terms of their L1 background (Harding, 2012), gender (Park, 2008; Aryadoust et al., 2011), and age (Geranpayeh and Kunnan, 2007). In addition, Batty (2015) conducted differential distractor functioning (DDF) analysis, similar to DIF analysis, to examine if test takers interacted with a particular distractor in video-based and audio-only multiple-choice questions. Batty found that one item revealed significant DDF, and it was difficult to explain the sources of DDF. Although research on item bias provides information about potential sources of bias and contributed to a better understanding of score-based decisions (Min and He, 2020), it is challenging to identify the reasons for items exhibiting significant DIF (Geranpayeh and Kunnan, 2007; Batty, 2015).

Internal Consistency

As a key parameter of scoring validity, internal consistency contains many aspects, including internal consistency coefficients, composite reliability, marker reliability, G-theory, and Item Response Theory (IRT)-based reliability (Geranpayeh, 2013; Geranpayeh and Taylor, 2013). IRT or Rasch models have been widely used to investigate internal consistency.

For instance, IRT analyses were conducted to estimate the internal consistency for the listening scores across different groups of test takers and across different items (Pardo-Ballester, 2010).

Widely used in L2 listening assessment, testlets refer to sets of items that are based on the same input (Eckes, 2014). Testlets tap into higher-level skills and make item writing and test administration more efficient; however, items nested within testlets might violate one of the assumptions of IRT models, that is, the local independence assumption (Eckes, 2014). This assumption is maintained if a person's response to an item does not affect the probability of the person's response to another item (Eckes, 2014). As testlets may have negative influence on the precision of ability estimates and test reliability, Eckes (2014) examined the testlet effect of the listening section of the Test of German as a Foreign Language (TestDaF) and observed small or moderate testlet effects. Eckes compared different approaches of analyzing testlet-based tests, including the use of independent-items models, the polytomous-items model, and the testlet response theory (TRT; Wainer et al., 2007) model. Eckes found that treating testlet items as independent items (i.e., the use of independent-items models) or as a single polytomous superitem (i.e., using the polytomous IRT model) led to the inaccurate estimation of test reliability and test-takers' ability.

Grading and Awarding

Listening tests often consist of multiple components targeting different communication goals (Choi and Papageorgiou, 2020). Scores on each component of the listening test, also called listening subscores, may provide added value over the total score. To examine the justifiability of reporting subscores at the individual and school levels, Choi and Papageorgiou (2020) explored the reliability and distinctiveness of listening and reading subscores of the TOEFL Primary test. Four listening subscores based on different communication goals were targeted, that is, Monologue, Dialogue, Narrative, and Academic subscores. They found that the individual-level subscores lacked psychometric added value, while the school-level subscores provided fine-grained information about the strengths and weaknesses of test takers from different schools, indicating that it is necessary to consider in score reporting what is reported and who is the intended user.

Consequential Validity

One study in our dataset explored consequential validity, focusing on washback (Nguyen and Gu, 2020). The researchers investigated the washback of the TOEIC listening and reading tests, which were used as an exit requirement, on teaching in Vietnam. Moreover, to understand the mechanism of washback, they explored three types of factors in washback – test factors, personal factors, and context factors. They found that teachers tended to tailor their teaching content and methods to the demands of the test by focusing on the tested skills while devoting less time to communicative activities. In relation to the mechanism of washback, test and personal factors played a significant role and influenced teachers' tendency to teach to the test and their use of communicative activities. In comparison, context factors were not closely related to the perceived washback. They argued that washback of the TOEIC in the Vietnamese context had not been fully understood and follow-up studies were needed to elucidate the reasons why these factors were correlated with washback.

Criterion-Related Validity

Comparison With Different Forms of the Same Test

As the only study on the comparability of test forms, Wei and Low (2017) examined the longitudinal score change pattern of 19,855 repeaters – test takers who took the test six times in 68 administrations over a period of 4 years – by analyzing the scores of the monthly administered TOEIC listening and reading tests. The starting month and the spacing of the six test-taking months varied across the repeaters. Linear growth modeling results showed that the repeaters' scores were stable over time (i.e., months) as their monthly score increases were small (i.e., a 1.6 score point increase per month), suggesting a high reliability of test scores across forms and across administrations. They also found that test scores varied much more between test takers than they varied overtime within test takers, and test-takers' background variables, especially gender, educational levels, and test-taking experience, had impacts on their listening score growth patterns and increase rate.

Comparability With External Standards and Frameworks

Three studies have explored the comparability between listening tests and criteria measures, including academic lecture tasks (Sawaki and Nissan, 2009), final grades in degree courses (Breeze and Miller, 2011), and local tests (Wagner, 2016). Since TOEFL iBT can be interpreted as a measure of academic listening ability (Sawaki and Nissan, 2009), it is important to gather empirical evidence about the relationship between TOEFL iBT listening test and an appropriate criterion measure of academic listening. Sawaki and Nissan (2009) investigated the relationship between test-takers' performance on TOEFL iBT listening test and academic lecture tasks that L2 English speakers encounter in their daily academic life. The researchers found that the listening test scores and the results of the criterion measure were positively correlated, indicating that they measured a similar academic listening construct.

Scores on large-scale L2 proficiency tests like TOEFL iBT and IELTS are used for many purposes, such as admission, placement, and exit. Breeze and Miller (2011) investigated the predictive validity of IELTS listening test as an entry requirement for admission to degree courses taught partly in English in a Spanish university. They found that test-takers' listening test scores were correlated with their final grades in programs in Humanities, Law, and Medicine, which justified the use of IELTS listening test for admission to academic programs. To be noted, IELTS listening test scores only accounted for a small part of academic success, which was not surprising given that aspects other than listening ability may determine students' academic success.

Research on the comparability with external standards and frameworks not only justifies the use of L2 listening tests

but also helps score users to make better decisions. Specifically, Wagner (2016) investigated the use of TOEFL iBT speaking and listening tests for international teaching assistants (ITAs) screening purposes. Three criteria measures of ITAs' language proficiency and teaching competence were included in his study, that is, the SPEAK test assessing ITA's oral proficiency, the TEACH test that measured ITAs' mastery of the curriculum, and undergraduate students' evaluations of their ITAs' language proficiency and teaching competence. TOEFL iBT listening test scores had significant correlations with the criteria measures. More importantly, TOEFL iBT listening test scores predicted ITAs' teaching competence better than TOEFL iBT speaking test scores, as the listening test scores accounted for an additional 15.3% of the variance of students' assessment of ITAs' teaching competence, whereas the speaking test scores accounted for only 5.9%. Wagner concluded that listening played an important part in teaching competence and TOEFL iBT listening scores should be used for ITA screening purposes.

Summing Up

As is shown above, 87 studies in our dataset were conducted to explore L2 listening assessment from a wide range of perspectives, tapping into 13 research themes in relation to the six components of the socio-cognitive framework. The vast majority of the studies explored test-taker characteristics, cognitive validity, context validity, and scoring validity, accounting for 94.25%. As important variables influencing listening test scores, a variety of test-taker characteristics were investigated. Research on cognitive validity examined items targeting different listening subskills and levels of listening process. Various research methods were used to uncover the complex cognitive processes, with innovative technology used to investigate test-takers' eye movement and brain activation patterns. In terms of context validity, task setting, linguistic demands (task input and output), and speakers have received considerable attention. Three parameters (i.e., item bias, internal consistency, and grading and awarding) influencing the scoring validity of L2 listening assessment were explored. In comparison, there is a small number of studies focusing on consequential validity and criterion-related validity, with only one study addressing the issue of test washback, and three studies exploring criterion-related validity. While helping to deepen our understanding of listening assessment from different perspectives, this review also brings to light many questions that need to be answered and a large amount of work that needs to be done.

DISCUSSION AND RECOMMENDATIONS FOR FUTURE RESEARCH

Findings of the present study suggest that more research efforts are needed in the field of L2 listening assessment. Recommendations for future research are discussed below from two perspectives, one on the four components which have been extensively investigated and the other on the two components which did not receive much attention (i.e., consequential validity and criterion-related validity).

Although research on physical/physiological characteristics underscores the importance of understanding test-takers' special needs, it is challenging to accommodate test takers with special needs, since it is not clear how test fairness and validity are affected by providing special arrangements for a particular group of test takers. In relation to experiential characteristics, the effect of test preparation was explored, indicating that test-takers' familiarity with the test format and preparation for listening tests are important variables influencing test performance. Future research should consider the role of test-takers' listening proficiency in test preparation. Moreover, with young learners constituting a large proportion of language learners, more studies are needed to explore the physical/physiological, psychological, and experiential characteristics of young test takers.

Studies on cognitive validity revealed that L2 listening is a complicated and dynamic cognitive operation. Moreover, research on L2 listening subskills and levels of comprehension indicates that it is challenging for test developers to operationalize the construct of L2 listening systematically. Recent years have witnessed an increasing use of advanced technology, such as eye-tracking technology and neuroimaging, which has brought about important development in the field. For instance, the notion of cognitive validity has been expanded, as researchers probed into the neurocognitive mechanism of test takers (Aryadoust et al., 2020). However, it is still difficult to understand test-takers' cognitive processes due to the highly overlapping and synergistic nature of comprehension (Alderson, 1990). For instance, test takers may simultaneously use higher-level and lower-level processing to comprehend the input (Brindley, 1998), and it is challenging to distinguish different levels of processing. Therefore, the authors think that research on cognitive processes is an important area where new perspectives are still unfolding and more research is needed to elucidate the relationship between cognitive processes and listening performance.

It is not surprising that a high proportion of studies investigated context validity since test developers should design tasks and adjust task characteristics that can retain key features of language use contexts and the way test tasks are designed and controlled has a direct effect on test authenticity (Bachman, 1990). Despite the abundance of research on context validity, the authors think that more efforts should be made to increase task authenticity and to avoid construct under-representation and construct irrelevance. As discussed previously, the use of visuals in listening assessment improves task authenticity as real-life listening usually involves visual input, but it may introduce construct-irrelevant variance if the test is designed to assess test-taker' mastery of the language itself. Similarly, whether to incorporate varieties of accents remains open for discussion. The use of diverse accents in L2 listening tests resembles the real-life context which requires multidialectal listening ability, but certain test takers may be advantaged due to the shared-L1 effect, which raises concerns about test fairness.

Therefore, more research is needed to elucidate the shared-L1 effect and justify the use of non-standard accents in listening assessment.

In relation to scoring validity, the theme of grading and awarding warrants more research endeavors. With descriptive and interpretable score reporting required for improving instructional designs and guiding students' learning (Alderson, 2005; Jang, 2008), more meaningful descriptors should be attached to listening scores. Future studies can consider providing richer and more detailed feedback of listening assessment for test users and convert test scores to plausible statements about test-takers' listening ability (Taylor and Geranpayeh, 2013). Also, more research is needed to explore the utility of feedback for L2 listening test users, including learners, teachers, and institutions.

The following are some recommendations for future research on the two components which did not receive much attention in our dataset, i.e., consequential validity and criterion-related validity. Consequential validity is one of the key areas for future research, and themes of test use, consequences, test fairness, and ethics warrant more research efforts, given that test washback and impact have become major areas of study in the field of language testing (Alderson, 2004). As Shohamy (2007, p. 117) pointed out, "the quality of tests is not judged merely by their psychometric traits but rather in relation to their impact, ethicality, fairness, values, and consequences."

There is a scarcity of research on the washback and impact of listening tests in our dataset, probably due to the complex mechanism of washback and impact in different social and educational contexts (Alderson and Wall, 1993; Hawkey, 2013). Washback and impact are affected by simply changing test methods and educational contexts (Cheng, 1997; Alderson, 2004) and may be independent of the original intentions of the test developers (Cheng et al., 2004). Therefore, the investigation of test washback and impact is time-consuming and complicated by a wide range of variables influencing learning and teaching, which requires a long-term and relatively complicated research program (Alderson and Wall, 1993; Nguyen and Gu, 2020). Furthermore, the study of washback and impact in the field of L2 listening assessment is more challenging due to the complexity of listening construct (Hawkey, 2013).

More research efforts are needed to explain the mechanism of washback and impact of L2 listening tests with education innovation and change in various contexts. The study of test washback and impact should be situated within the micro contexts (e.g., the school setting) and macro contexts (e.g., the sociocultural environment where the test is used; Cheng L., 2004). Considering the rapid change in educational policy and the needs of stakeholders, a better understanding of how the washback and impact of L2 listening tests occur is needed. In addition, with the increasingly widespread use of high-stakes tests that have important consequences for individuals and institutions (Bailey, 1999; Alderson, 2004; Green, 2013), future research should investigate the washback and impact of high-stakes listening tests.

In addition to consequential validity, criterion-related validity is also important with the development of language proficiency scales, such as the Common European Framework of Reference for languages: Learning, Teaching, Assessment (CEFR) and the recently released China's Standards of English Language Ability (CSE). One of the aims of these proficiency scales is to promote communication between researchers and practitioners in the fields of language learning, teaching, and assessment (Council of Europe, 2001; National Education Examinations Authority, 2018). Although aligning tests to proficiency scales is conducive to bridging the gap between learning and assessment, the procedure of alignment is complex (Harsch and Rupp, 2011). Thus, future research is needed to provide evidence for the validity of using these proficiency guidelines for listening assessment.

CONCLUSION

In the present study, a review of research on L2 listening assessment was conducted using Weir's (2005) socio-cognitive framework. With a total of 87 studies collected, 13 research themes were identified in relation to the six components of the framework and analyzed. Recommendations for future research in the field were discussed from the perspectives of the four components that were extensively investigated and the other two components which did not receive much attention in our dataset, that is, consequential validity and criterion-related validity. While trying to give a comprehensive review of relevant research, the authors are fully aware of the limitations of the present study. For one thing, only studies from 14 peer-reviewed journals and two research report series were reviewed, and research on L2 listening assessment published in other journals, research report series, conference proceedings, or book series were not included due to time and space limit. For another, studies written in languages other than English were not included as a result of resource and space constraints. Despite the limitations mentioned above, this study provides valuable insights into various factors that can influence test-takers' performance in L2 listening assessment and sheds light on the state-of-the-art research in L2 listening assessment.

AUTHOR CONTRIBUTIONS

LH designed the study, coded the data, and drafted the manuscript. ZJ collected the data, coded the data, and drafted the manuscript together with LH. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to thank Dr. Dajian Chen and all the reviewers for their comments and suggestions on the previous drafts of this article.

REFERENCES

Alderson, J. C. (1990). Testing reading comprehension skills (Part Two). *Read. Foreign Lang.* 7, 465–503.

Alderson, J. C. (2004). "Foreword" in *Washback in language testing: Research contexts and methods*. eds. L. Cheng, Y. Y. Watanabe and A. Curtis (New Jersey: Lawrence Erlbaum Associates Publishers), 12–17.

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

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

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., and Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: an individual differences approach. *Lang. Learn.* 62(Suppl. 2), 49–78. doi: 10.1111/j.1467-9922.2012.00706.x

Aryadoust, V. (2020). A review of comprehension subskills: a scientometrics perspective. *System* 88:102180. doi: 10.1016/j.system.2019.102180

Aryadoust, V., Goh, C. C. M., and Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Lang. Assess. Q.* 8, 361–385. doi: 10.1080/15434303.2011.628632

Aryadoust, V., Ng, L. Y., Foo, S., and Esposito, G. (2020). A neurocognitive investigation of test methods and gender effects in listening assessment. *Comput. Assist. Lang. Learn.* 1–21. doi: 10.1080/09588221.2020.1744667

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

Bachman, L. F., and Palmer, A. S. (1996). *Language testing in practice*. 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*. Princeton, NJ: Educational Testing Service.

Baker, B., and Hope, A. (2019). Incorporating translanguaging in language assessment: the case of a test for university professors. *Lang. Assess. Q.* 16, 408–425. doi: 10.1080/15434303.2019.1671392

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

Becker, A. (2016). L2 students' performance on listening comprehension items targeting local and global information. *J. Engl. Acad. Purp.* 24, 1–13. doi: 10.1016/j.jeap.2016.07.004

Bent, T., and Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *J. Acoust. Soc. Am.* 114, 1600–1610. doi: 10.1121/1.1603234

Bodie, G. D., and Worthington, D. L. (2017). "Measuring listening" in *The sourcebook of listening research: Methodology and measures*. eds. D. L. Worthington and G. D. Bodie (New York, NY: Wiley Blackwell), 21–44.

Breeze, R., and Miller, P. (2011). *Predictive validity of the IELTS listening test as an indicator of student coping ability in Spain. Vol. 12*. IELTS Research Report.

Brindley, G. (1998). Assessing listening abilities. *Annu. Rev. Appl. Linguist.* 18, 171–191. doi: 10.1017/S0267190500003536

Brindley, G., and Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Lang. Test.* 19, 369–394. doi: 10.1191/0265532202lt236oa

Brunfaut, T., and Revesz, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Q.* 49, 141–168. doi: 10.1002/tesq.168

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

Buck, G. (2018). "Preface" in *Assessing L2 listening: Moving towards authenticity*. eds. G. J. Ockey and E. Wagner (Amsterdam: John Benjamins), 11–16.

Burunat, I., and Brattico, E. (2017). "Functional magnetic resonance imaging (fMRI)" in *The sourcebook of listening research: Methodology and measures*. eds. D. L. Worthington and G. D. Bodie (New York, NY: Wiley Blackwell), 290–298.

Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: evidence from confirmatory factor analysis. *Lang. Test.* 30, 177–199. doi: 10.1177/0265532212456833

Cai, H. (2020). Relating lexical and syntactic knowledge to academic English listening: the importance of construct representation. *Front. Psychol.* 11:494. doi: 10.3389/fpsyg.2020.00494

Carrell, P. (2007). *Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks (TOEFL Monograph Series No. MS-35)*. Princeton, NJ: Educational Testing Service.

Chang, A. C., and Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Q.* 40, 375–397. doi: 10.2307/40264527

Chang, A. C., and Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System* 41, 575–586. doi: 10.1016/j.system.2013.06.001

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Lang. Educ.* 11, 38–54. doi: 10.1080/09500789708666717

Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Lang. Ann.* 37, 544–553. doi: 10.1111/j.1944-9720.2004.tb02421.x

Cheng, L. (2004). "The washback effect of a public examination change on teachers' perceptions toward their classroom teaching" in *Washback in language testing: Research contexts and methods*. eds. L. Cheng, Y. Y. Watanabe and A. Curtis (New Jersey: Lawrence Erlbaum Associates Publishers), 147–170.

Cheng, J., and Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Lang. Test.* 35, 3–25. doi: 10.1177/0265532216676851

Cheng, L., Watanabe, Y., and Curtis, A. (eds.) (2004). *Washback in language testing: Research contexts and methods*. New Jersey: Lawrence Erlbaum Associates Publishers.

Choi, I., and Papageorgiou, S. (2020). Evaluating subscore uses across multiple levels: a case of reading and listening subscores for young EFL learners. *Lang. Test.* 37, 254–279. doi: 10.1177/0265532219879654

Chou, M. (2019). Predicting self-efficacy in test preparation: gender, value, anxiety, test performance, and strategies. *J. Educ. Res.* 112, 61–71. doi: 10.1080/00220671.2018.1437530

Chung, J. (2002). The effects of using two advance organizers with texts for the teaching of listening in English. *Foreign Lang. Ann.* 35, 231–240. doi: 10.1111/j.1944-9720.2002.tb03157.x

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: a case study. *System* 29, 1–14. doi: 10.1016/S0346-251X(00)00057-9

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

Cubilo, J., and Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: considering the impacts of visual-cue interpretation and note-taking. *Lang. Assess. Q.* 10, 371–397. doi: 10.1080/15434303.2013.824972

Cutler, A., and Clifton, C. (1999). "Comprehending spoken language: a blueprint of the listener" in *The neurocognition of language*. eds. C. M. Brown and P. Hagoort (New York, NY: Oxford University Press), 123–166.

Dahl, T. I., and Ludvigsen, S. (2014). How I see what you're saying: the role of gestures in native and foreign language listening comprehension. *Mod. Lang. J.* 98, 813–833. doi: 10.1111/modl.12124

Dai, D. W., and Roever, C. (2019). Including L2-English varieties in listening tests for adolescent ESL learners: L1 effects and learner perceptions. *Lang. Assess. Q.* 16, 64–86. doi: 10.1080/15434303.2019.1601198

Deci, E. L., and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, E. L., and Ryan, R. M. (1995). "Human autonomy: the basis for true self-esteem" in *Efficacy, agency and self-esteem*. ed. M. H. Kernis (New York: Plenum), 31–48.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: a testlet response theory modeling approach. *Lang. Test.* 31, 39–61. doi: 10.1177/0265532213492969

Elkhafaifi, H. (2005). The effect of prelistening activities on listening comprehension in Arabic learners. *Foreign Lang. Ann.* 38, 505–513. doi: 10.1111/j.1944-9720.2005.tb02517.x

Elliott, M. (2013). "Test-taker characteristics" in *Examining listening: Research and practice in assessing second language listening. Vol. 35*. eds. A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 36–76.

Elliott, M., and Wilson, J. (2013). "Context validity" in *Examining listening: Research and practice in assessing second language listening. Vol. 35*. eds. A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 152–241.

Feyten, C. M. (1991). The power of listening ability: an overlooked dimension in language acquisition. *Mod. Lang. J.* 75, 173–180. doi: 10.1111/j.1540-4781.1991.tb05348.x

Field, J. E. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press.

Field, J. E. (2009). *The cognitive validity of the lecture-based question in the IELTS listening paper*. *Vol. 9*. IELTS Research Report.

Field, J. E. (2013). "Cognitive validity" in *Examining listening: Research and practice in assessing second language*. eds. A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 77–151.

Filipi, A. (2012). Do questions written in the target language make foreign language listening comprehension tests more difficult? *Lang. Test.* 29, 511–532. doi: 10.1177/0265532212441329

Flowerdew, J., and Miller, L. (2005). *Second language listening: Theory and practice*. New York: Cambridge University Press.

Geranpayeh, A. (2013). "Scoring validity" in *Examining listening: Research and practice in assessing second language listening. Vol. 35*. eds. A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 242–272.

Geranpayeh, A., and Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Lang. Assess. Q.* 4, 190–222. doi: 10.1080/15434300701375758

Geranpayeh, A., and Taylor, L. (eds.) (2013). *Examining listening: Research and practice in assessing second language listening. Vol. 35*. Cambridge: Cambridge University Press.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Lang. Test.* 19, 133–167. doi: 10.1191/0265532202lt225oa

Green, A. (2013). Washback in language assessment. *Int. J. English Stud.* 13, 39–51. doi: 10.6018/ijes.13.2.185891

Gruba, P. (1997). The role of video media in listening assessment. *System* 25, 335–345. doi: 10.1016/S0346-251X(97)00026-2

Gruba, P. (2004). Understanding digitized second language videotext. *Comput. Assist. Lang. Learn.* 17, 51–82. doi: 10.1076/call.17.1.51.29710

Gruba, P. (2006). Playing the videotext: a media literacy perspective on video mediated L2 listening. *Lang. Learn. Technol.* 10, 77–92. https://www.lltjournal.org/item/2549

Halone, K. K., Cunconan, T. M., Coakley, C. G., and Wolvin, A. D. (1998). Toward the establishment of general dimensions underlying the listening process. *Int. J. List.* 12, 12–28. doi: 10.1080/10904018.1998.10499016

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: a DIF perspective. *Lang. Test.* 29, 163–180. doi: 10.1177/0265532211421161

Harsch, C., and Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: a test-centered approach. *Lang. Assess. Q.* 8, 1–33. doi: 10.1080/15434303.2010.535575

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.

Hawkey, R. (2013). "Consequential validity" in *Examining listening: Research and practice in assessing second language listening. Vol. 35*. eds. A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 273–302.

He, L., and Min, S. (2017). Development and validation of a computer adaptive EFL test. *Lang. Assess. Q.* 14, 160–176. doi: 10.1080/15434303.2016.1162793

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., et al. (2020). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Lang. Test.* 1–21. doi: 10.1177/0265532220917316

Horwitz, E. K., Horwitz, M. B., and Cope, J. (1986). Foreign language classroom anxiety. *Mod. Lang. J.* 70, 125–132. doi: 10.1111/j.1540-4781.1986.tb05256.x

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System* 34, 317–340. doi: 10.1016/j.system.2006.04.005

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: 10.1177/0265532208101006

Jang, E. E. (2008). "A framework for cognitive diagnostic assessment" in *Towards adaptive CALL: Natural language processing for diagnostic language assessment*. eds. C. A. Chapelle, Y. R. Chung and J. Xu (Ames, IA: Iowa State University), 117–131.

Jones, L. C., and Plass, J. L. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations. *Mod. Lang. J.* 86, 546–561. doi: 10.1111/1540-4781.00160

Kang, O., Moran, M., Ahn, H., and Park, S. (2020). Proficiency as a mediating variable of intelligibility for different varieties of accents. *Stud. Second Lang. Acquis.* 42, 471–487. doi: 10.1017/S0272263119000536

Kang, O., Thomson, R., and Moran, M. (2018a). Empirical approaches to measuring intelligibility of different varieties of English in predicting listener comprehension of tests. *Lang. Learn.* 68, 115–146. doi: 10.1111/lang.12270

Kang, O., Thomson, R., and Moran, M. (2018b). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Appl. Linguis.* 41, 453–480. doi: 10.1093/applin/amy053

Kang, O., Thomson, R., and Moran, M. (2019). The effects of international accents and shared first language on listening comprehension tests. *TESOL Q.* 53, 56–81. doi: 10.1002/tesq.463

Kellerman, S. (1992). I see what you mean: the role of kinesic behaviour in listening and implications for foreign and second language learning. *Appl. Linguis.* 13, 239–258. doi: 10.1093/applin/13.3.239

Khalifa, H., and Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Knoch, U., and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Lang. Test.* 35, 477–499. doi: 10.1177/0265532217710049

Kormos, J., Kosak-Babuder, M., and Pizorn, K. (2019). The role of low-level first language skills in second language reading, reading-while-listening and listening performance: a study of young dyslexic and non-dyslexic language learners. *Appl. Linguis.* 40, 834–858. doi: 10.1093/applin/amy028

Kosak-Babuder, M., Kormos, J., Ratajczak, M., and Pizorn, K. (2019). The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Lang. Test.* 36, 51–75. doi: 10.1177/0265532218756946

Koyama, D., Sun, A., and Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Lang. Learn. Technol.* 20, 148–165. https://www.lltjournal.org/item/2936

Lee, Y., and Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Lang. Assess. Q.* 6, 239–263. doi: 10.1080/15434300903079562

Lee, H., and Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Lang. Test.* 30, 99–123. doi: 10.1177/0265532212451235

Li, C. (2019). Using a listening vocabulary levels test to explore the effect of vocabulary knowledge on GEPT listening comprehension performance. *Lang. Assess. Q.* 16, 328–344. doi: 10.1080/15434303.2019.1648474

Lim, G. S., and Khalifa, H. (2013). "Criterion-related validity" in *Examining listening: Research and practice in assessing second language listening. Vol. 35*. eds. A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 303–321.

Logie, R. H. (2011). The functional organization and capacity limits of working memory. *Curr. Dir. Psychol. Sci.* 20, 240–245. doi: 10.1177/0963721411415340

Major, R., Fitzmaurice, S., Bunta, F., and Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: implications for ESL assessment. *TESOL Q.* 36, 173–190. doi: 10.2307/3588329

Matsuura, H., Chiba, R., Mahoney, S., and Rilling, S. (2014). Accent and speech rate effects in English as a lingua franca. *System* 46, 143–150. doi: 10.1016/j.system.2014.07.015

Matthews, J. (2018). Vocabulary for listening: emerging evidence for high and mid-frequency vocabulary knowledge. *System* 72, 23–36. doi: 10.1016/j.system.2017.10.005

Matthews, J., and Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System* 52, 1–13. doi: 10.1016/j.system.2015.04.015

Messick, S. (1989). "Validity" in *Educational measurement*. ed. R. L. Linn (New York: American Council on Education and Macmillan), 13–103.

Min, S., and He, L. (2020). Test fairness: examining differential functioning of the reading comprehension section of the GSEEE in China. *Stud. Educ. Eval.* 64:100811. doi: 10.1016/j.stueduc.2019.100811

National Education Examinations Authority (2018). *China's standards of English language ability*. Beijing, China: Higher Education Press & Shanghai Foreign Language Education Press.

Nguyen, H., and Gu, Y. (2020). Impact of TOEIC listening and reading as a university exit test in Vietnam. *Lang. Assess. Q.* 17, 147–167. doi: 10.1080/15434303.2020.1722672

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing*. Unpublished PhD Dissertation. University of Reading.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Lang. Test.* 24, 517–537. doi: 10.1177/0265532207080771

Ockey, G. J., and French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Appl. Linguis.* 37, 693–715. doi: 10.1093/applin/amu060

Ockey, G. J., and Wagner, E. (eds.) (2018). *Assessing L2 listening: Moving towards authenticity*. Amsterdam: John Benjamins.

Olive, T. (2004). Working memory in writing: empirical evidence from the dual-task technique. *Eur. Psychol.* 9, 32–42. doi: 10.1027/1016-9040.9.1.32

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. Q.* 9, 375–397. doi: 10.1080/15434303.2012.721425

Papageorgiou, S., Wu, S., Hsieh, C. N., Tannenbaum, R. J., and Cheng, M. (2019). Report No.: RR-19-44. Mapping the TOEFL iBT® Test Scores to China's Standards of English Language Ability: Implications for Score Interpretation and Use (ETS Research). Princeton, NJ: Educational Testing Service.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: test usefulness evaluation. *Lang. Assess. Q.* 7, 137–159. doi: 10.1080/15434301003664188

Paribakht, T. S., and Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *J. Engl. Acad. Purp.* 21, 121–132. doi: 10.1016/j.jeap.2015.05.009

Park, G. (2008). Differential item functioning on an English listening test across gender. *TESOL Q.* 42, 115–123. doi: 10.1002/j.1545-7249.2008.tb00212.x

Poehner, M. E., Zhang, J., and Lu, X. (2015). Computerized dynamic assessment (C-DA): diagnosing L2 development according to learner responsiveness to mediation. *Lang. Test.* 32, 337–357. doi: 10.1177/0265532214560390

Read, J. (2002). The use of interactive input in EAP listening assessment. *J. Engl. Acad. Purp.* 1, 105–119. doi: 10.1016/S1475-1585(02)00018-8

Rost, M. (2011). *Teaching and researching listening. 2nd Edn*. Harlow, UK: Pearson.

Rukthong, A., and Brunfaut, T. (2019). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Lang. Test.* 37, 31–53. doi: 10.1177/0265532219871470

Sakai, H. (2009). Effect of repetition of exposure and proficiency level in L2 listening tests. *TESOL Q.* 43, 360–371. doi: 10.1002/j.1545-7249.2009.tb00179.x

Sawaki, Y., Kim, H., and Gentile, C. (2009). Q-matrix construction: defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Lang. Assess. Q.* 6, 190–209. doi: 10.1080/15434300902801917

Sawaki, Y., and Nissan, S. (2009). Report No.: RR-09-02. Criterion-related validity of the TOELF iBT listening section (ETS Research). Princeton, NJ: Educational Testing Service.

Shaw, S., and Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Shohamy, E. (1997). Testing methods, testing consequences: are they ethical? Are they fair? *Lang. Test.* 14, 340–349. doi: 10.1177/026553229701400310

Shohamy, E. (2007). Language tests as language policy tools. *Assess. Educ.* 14, 117–130. doi: 10.1080/09695940701272948

Song, M. (2012). Note-taking quality and performance on an L2 academic listening test. *Lang. Test.* 29, 67–89. doi: 10.1177/0265532211415379

Staehr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Stud. Second Lang. Acquis.* 31, 577–607. doi: 10.1017/S0272263109990039

Sueyoshi, A., and Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Lang. Learn.* 55, 661–699. doi: 10.1111/j.0023-8333.2005.00320.x

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: a comparison of context videos and content videos. *Lang. Test.* 32, 463–483. doi: 10.1177/0265532214562099

Taguchi, N. (2005). Comprehending implied meaning in English as a foreign language. *Mod. Lang. J.* 89, 543–562. doi: 10.1111/j.1540-4781.2005.00329.x

Taguchi, N. (2007). Development of speed and accuracy in pragmatic comprehension in English as a foreign language. *TESOL Q.* 41, 313–338. doi: 10.1002/j.1545-7249.2007.tb00061.x

Taguchi, N. (2008a). Cognition, language contact, and the development of pragmatic comprehension in a study-abroad context. *Lang. Learn.* 58, 33–71. doi: 10.1111/j.1467-9922.2007.00434.x

Taguchi, N. (2008b). Pragmatic comprehension in Japanese as a foreign language. *Mod. Lang. J.* 92, 558–576. doi: 10.1111/j.1540-4781.2008.00787.x

Taylor, L. (2005). Washback and impact. *ELT J.* 59, 154–155. doi: 10.1093/eltj/cci030

Taylor, L. (ed.) (2012). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Taylor, L. (2013). "Introduction" in *Examining listening: Research and practice in assessing second language listening. Vol. 35*. eds. A. Geranpayeh and L. Taylor, (Cambridge: Cambridge University Press), 1–35.

Taylor, L., and Geranpayeh, A. (2011). Assessing listening for academic purposes: defining and operationalizing the test construct. *J. Engl. Acad. Purp.* 10, 89–101. doi: 10.1016/j.jeap.2011.03.002

Taylor, L., and Geranpayeh (Eds.) (2013). "Conclusions and recommendations" in *Examining listening: Research and practice in assessing second language listening. Vol. 35*. (Cambridge: Cambridge University Press), 322–341.

Vafaee, P., and Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Stud. Second Lang. Acquis.* 42, 383–410. doi: 10.1017/S0272263119000676

van Zeeland, H., and Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Appl. Linguis.* 34, 457–479. doi: 10.1093/applin/ams074

Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Appl. Linguis.* 26, 70–89. doi: 10.1093/applin/amh039

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Lang. Teach.* 40, 191–210. doi: 10.1017/S0261444807004338

Vandergrift, L., and Goh, C. C. (2012). *Teaching and learning second language listening: Metacognition in action*. New York: Routledge.

Vandergrift, L., Goh, C. C., Mareschal, C. J., and Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: development and validation. *Lang. Learn.* 56, 431–462. doi: 10.1111/j.1467-9922.2006.00373.x

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Lang. Learn. Technol.* 11, 67–86.

Wagner, E. (2008). Video listening tests: what are they measuring? *Lang. Assess. Q.* 5, 218–243. doi: 10.1080/15434300802213015

Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System* 38, 280–291. doi: 10.1016/j.system.2010.01.003

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Lang. Test.* 27, 493–513. doi: 10.1177/0265532209355668

Wagner, E. (2013a). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Lang. Assess. Q.* 10, 178–195. doi: 10.1080/15434303.2013.769552

Wagner, E. (2013b). "Assessing listening" in *The companion to language assessment*. ed. A. J. Kunnan (Hoboken, NJ: John Wiley), 47–63.

Wagner, E. (2016). Report No.: RR-16-18. A study of the use of the TOEFL iBT test speaking and listening subscores for international teaching assistant screening (ETS Research). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12104

Wagner, E., and Ockey, G. J. (2018). "An overview of the use of audio-visual texts on L2 listening tests" in *Assessing L2 listening: Moving towards authenticity*. eds. G. J. Ockey and E. Wagner (Amsterdam: John Benjamins), 130–144.

Wagner, E., and Toth, P. D. (2014). Teaching and testing L2 Spanish listening using scripted vs. unscripted texts. *Foreign Lang. Ann.* 47, 404–422. doi: 10.1111/flan.12091

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wang, Y., and Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: the contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System* 65, 139–150. doi: 10.1016/j.system.2016.12.013

Wei, Y., and Low, A. (2017). Report No.: RR-17-54. Monitoring score change patterns to support TOEIC listening and reading test quality (ETS Research). Princeton, NJ: Educational Testing Service.

Wei, W., and Zheng, Y. (2017). An investigation of integrative and independent listening test tasks in a computerized academic English test. *Comput. Assist. Lang. Learn.* 30, 864–883. doi: 10.1080/09588221.2017.1373131

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. (2013). "An overview of the influences on English language testing in the United Kingdom 1913-2012" in *Measured constructs: A history of Cambridge English language examinations 1913-2012. Vol. 37*. eds. C. J. Weir, I. Vidaković and E. D. Galaczi (Cambridge: Cambridge University Press), 1–102.

Weir, C., and Vidakovic, I. (2013). "The measurement of listening ability 1913-2012" in *Measured constructs: A history of Cambridge English language examinations 1913-2012. Vol. 37*. eds. C. J. Weir, I. Vidakovic and E. D. Galaczi (Cambridge: Cambridge University Press), 347–444.

Wigfield, A., and Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemp. Educ. Psychol.* 25, 68–81. doi: 10.1006/ceps.1999.1015

Winke, P., and Lim, H. (2017). The effects of test preparation on second-language listening test performance. *Lang. Assess. Q.* 14, 380–397. doi: 10.1080/15434303.2017.1399396

Worthington, D. L. (2017). "Modeling and measuring cognitive components of listening" in *The sourcebook of listening research: Methodology and measures*. eds. D. L. Worthington and G. D. Bodie (New York, NY: Wiley Blackwell), 70–96.

Worthington, D. L., and Bodie, G. D. (eds.) (2017). *The sourcebook of listening research: Methodology and measures*. New York, NY: Wiley Blackwell.

Xu, J. (2017). The mediating effect of listening metacognitive awareness between test-taking motivation and listening test score: an expectancy-value theory approach. *Front. Psychol.* 8:2201. doi: 10.3389/fpsyg.2017.02201

Yanagawa, K., and Green, A. (2008). To show or not to show: the effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System* 36, 107–122. doi: 10.1016/j.system.2007.12.003

Yi, Y. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: an application of cognitive diagnostic models. *Lang. Test.* 34, 337–355. doi: 10.1177/0265532216646141

Zhang, X. (2013). Foreign language listening anxiety and listening performance: conceptualizations and causal relationships. *System* 41, 164–177. doi: 10.1016/j.system.2013.01.004

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.

frontiers in Psychology

SYSTEMATIC REVIEW
published: 04 September 2020
doi: 10.3389/fpsyg.2020.01941

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

Vahid Aryadoust[1]*, *Azrifah Zakaria*[1], *Mei Hui Lim*[2] 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

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

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

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 aspects[1]. 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).

[1]We 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 logistically 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 publication[2]. 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 articles[3], 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 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-citations[4]. DCA was conducted twice times—once for each dataset obtained from Scopus, as previously discussed. We further investigated the duration of burstness (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.

[2]We 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.

[3]In 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 presents study.

[4]CiteSpace, 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 ($\sum$). 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

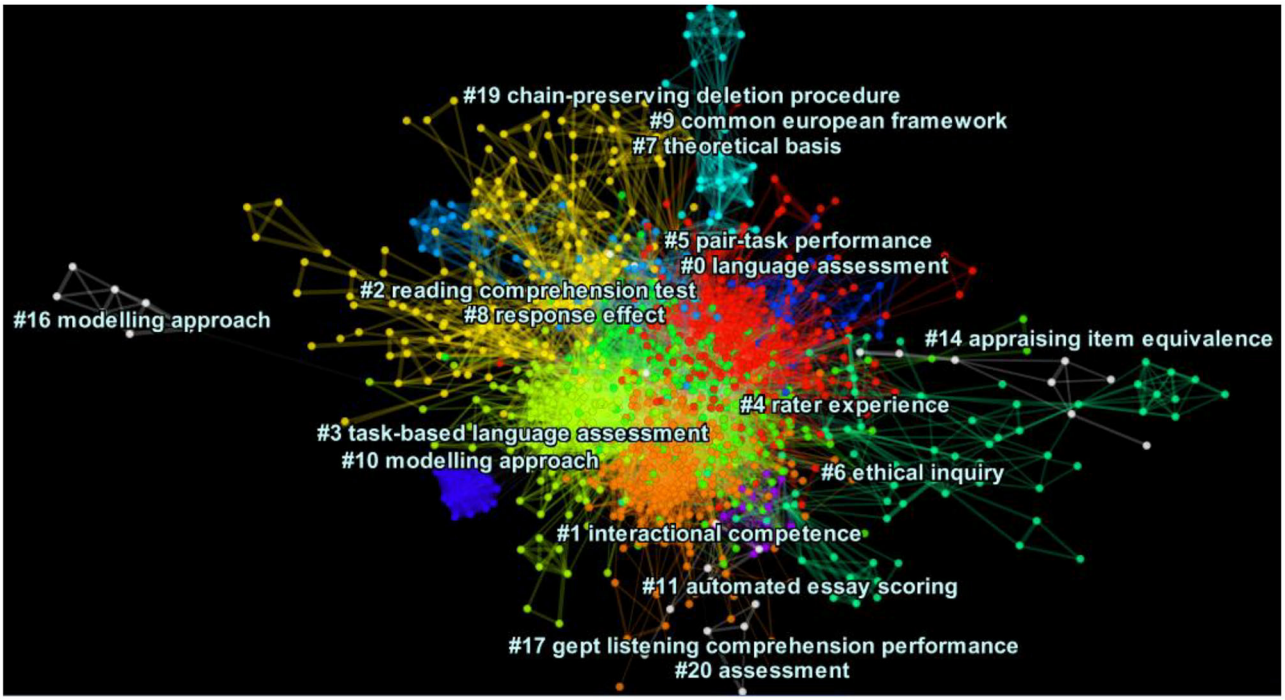


FIGURE 1 | The cluster view of network in the core journals dataset (modularity Q = 0.541, average silhouette score = 0.71), generated using CiteSpace, Version 5.6.R3.

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

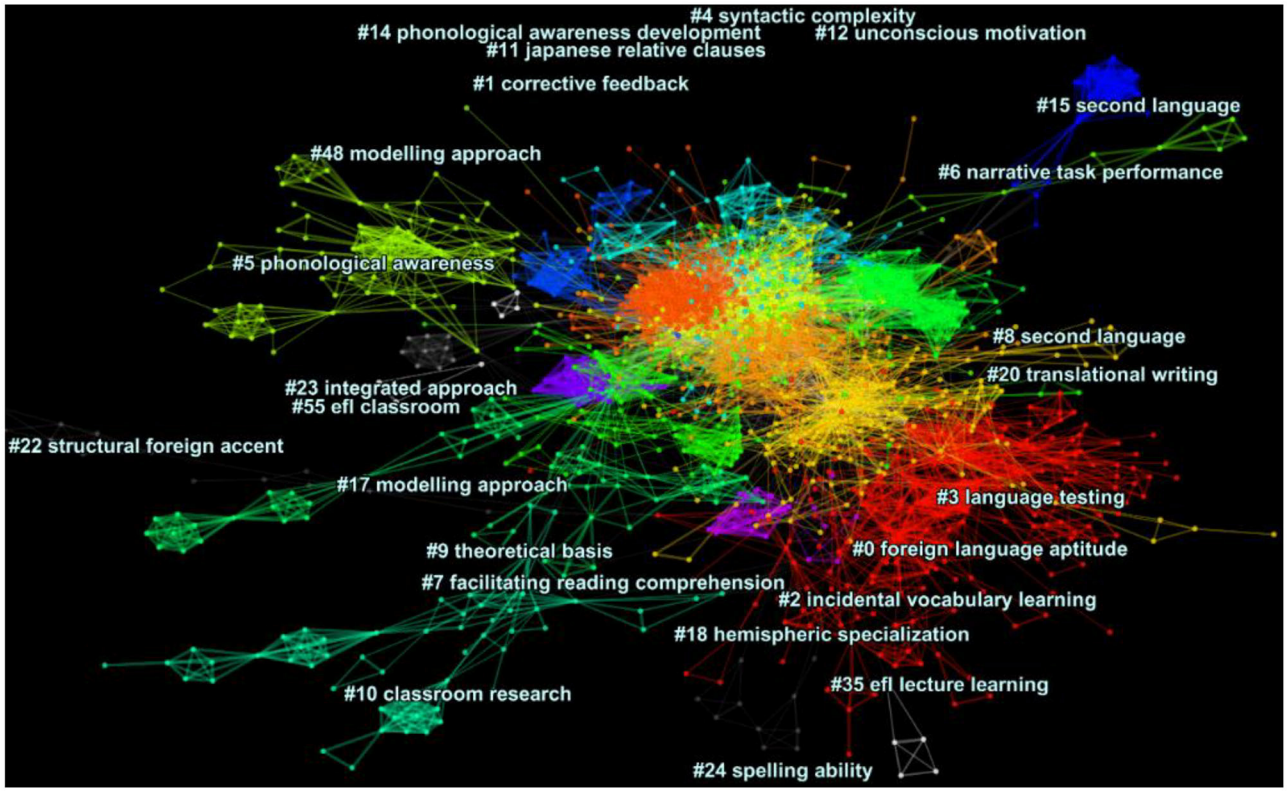


FIGURE 2 | The cluster view of network in the general journals dataset (modularity Q = 0.6493, average silhouette score = 0.787), generated using CiteSpace, Version 5.6.R3.

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	Cloze/ Likert/ multiple choice	Essays and writing	Oral/interview	Total
1	10	32	21	63
4	17	17	9	43
0	20	5	13	38
5	4	0	11	15
2	8	4	2	14
3	2	3	3	8
Total	61	61	59	181

Reliability

Cluster ID	Reported reliability	Did not report reliability	Total
1	49	36	85
0	30	29	59
4	26	4	30
3	8	18	26
2	13	8	21
5	10	9	19

Generalization

Cluster ID	Reported generalizability evidence	Did not report generalizability evidence	Total
1	6	79	85
0	1	58	59
4	6	24	30
3	0	26	26
2	1	20	21
5	3	16	19

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
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; Lallmamode 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
Levelt (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

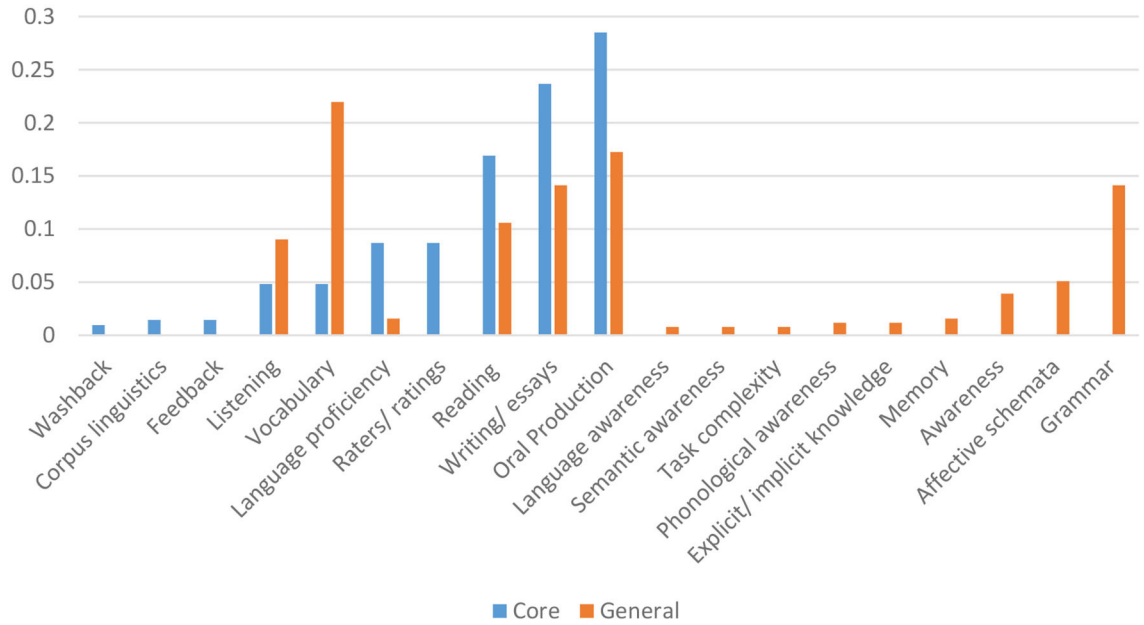


FIGURE 3 | Comparison of domain specifications in the core and general journals.

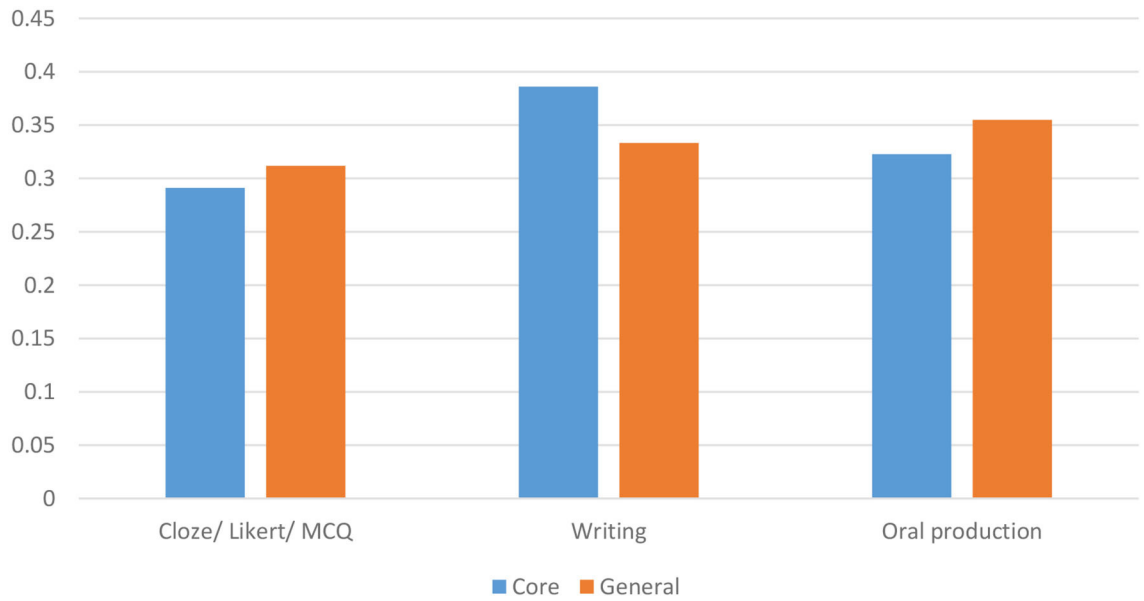


FIGURE 4 | Comparison of construct operationalization in the core and general journals.

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

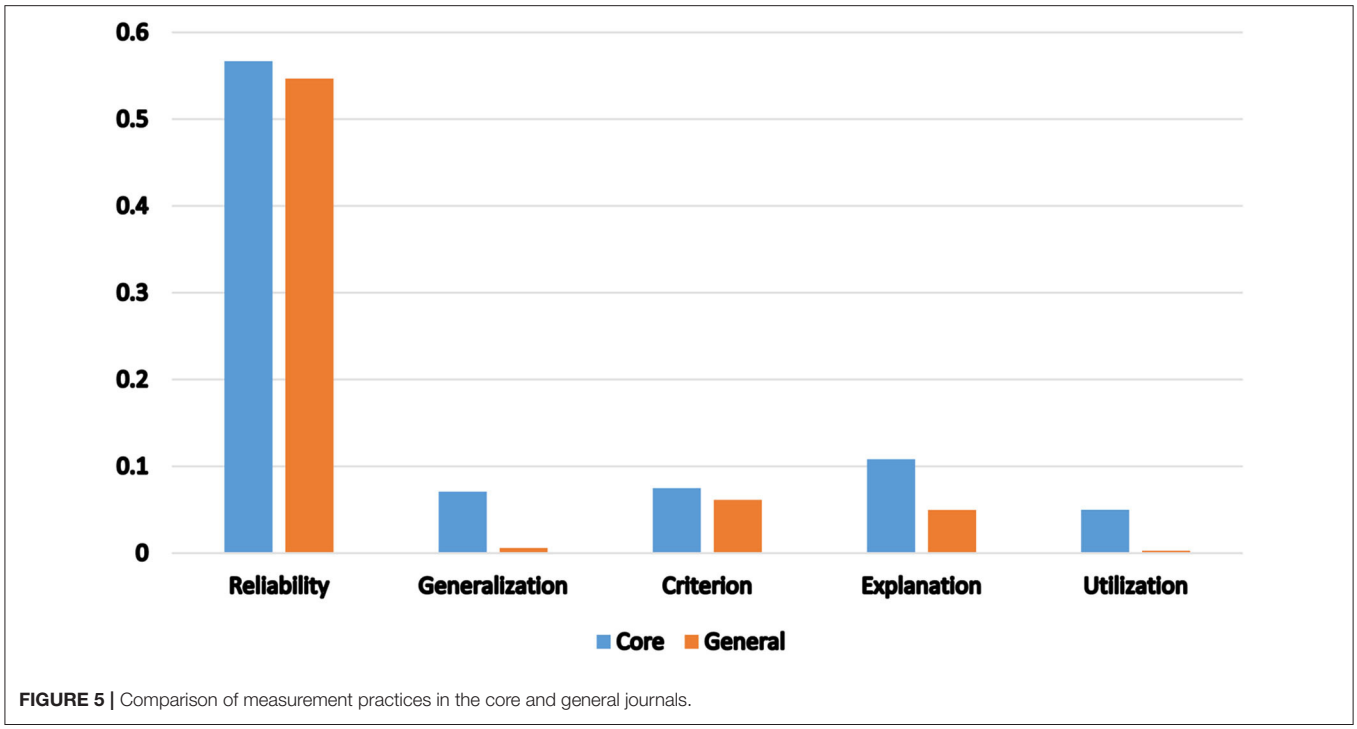


FIGURE 5 | Comparison of measurement practices in the core and general journals.

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 generals. 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'nami 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., Ortuño, 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/Research/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 model 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. Linguis.* 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.* Availaable 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.datak.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: 10.1177/026553228500200204

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: 10.1057/978-1-137-59900-1_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., Ünaldi, 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: 10.1177/0265532208101006

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.

Lallmamode, 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

Levelt, 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 Roever, 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.asw.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.asw.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.asw.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: Paedagogike 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," i*n 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., andde 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, drawling, 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 of 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership