International Journal of TESOL Studies (2026)
Vol. 8 (1) 180-215 https://doi.org/10.58304/ijts.260620

Article

A Systematic Review of L2 Listening Assessment: Test Specifications, Validity Evidence, and Reporting Practices

Xuanye Miao
Vahid Aryadoust*
Yanyan Huang
National Institute of Education, Nanyang Technological University, Singapore

Received: 1 March 2026 / Received in revised form: 29 May 2026 / Accepted: 11 June 2026 / Available online: 5 July 2026

Abstract
Over the past few decades, second language (L2) listening assessments have become an integral component of language teaching and learning. However, there have been few systematic reviews examining the validity, reporting practices, and test specifications of existing L2 listening assessments. To gain a comprehensive understanding of L2 listening assessments, we reviewed 126 studies (comprising 160 listening assessments) published between 1997 and 2021. We focused on test specifications, test characteristics, and test validation by adopting established frameworks for test specification alongside the argument-based validation (ABV) framework. The studies were coded using a coding scheme organized into six variable categories, including administrative information, test information, textual features, participant information, validity inferences, and results. The results showed that standardized listening assessments were widely used by L2 listening researchers, and that test specifications across studies were often heterogeneous. In addition, the majority of studies provided limited information on test specifications, thereby limiting the replicability of findings across different contexts. Finally, the distribution of validity evidence across the six ABV inferences was markedly uneven: the generalization inference was the most frequently examined (56.88%), whereas no study provided evidence for the domain description or utilization inferences, and 38.75% of assessments (62 out of 160) presented no validity evidence of any kind. Implications for L2 listening assessment and test development are discussed.

Keywords
Second language listening, systematic review, test specification, validity evidence

1 Introduction

Second language (L2) listening comprehension has been widely recognized as a multidimensional construct involving cognitive, affective, and behavioral processes (Aryadoust, forthcoming; Goh &

*Corresponding author. Email: vahid.aryadoust@nie.edu.sg

Vandergrift, 2022; Halone et al., 1998). Unlike productive skills, listening is an internally processed and latent ability that cannot be directly observed; instead, it must be inferred from performance on tasks that often involve additional modalities such as reading or writing (Buck, 2001; Liu & Aryadoust, 2026). This inferential nature introduces inherent measurement challenges. Specifically, tests of listening ability are particularly vulnerable to construct underrepresentation and construct-irrelevant variance (Aryadoust, forthcoming; Buck, 2001; Liu & Aryadoust, 2026).

Ensuring adequate construct representation and providing sufficient evidence to support score interpretations are therefore central concerns in language testing research. A growing body of scholarship, including narrative reviews, systematic reviews, and meta-analyses, has generated valuable insights into L2 listening constructs, modality effects, cognitive correlates, and research themes in the field (e.g., Aryadoust, forthcoming; Aryadoust & Luo, 2023; Geranpayeh & Taylor, 2008; He & Jiang, 2020; see Section 2 for a detailed review). Collectively, these studies have identified various influential factors in the assessment of L2 listening and highlighted the challenges that test developers face. However, the validity of interpretations and uses of L2 listening test scores remains unclear. In addition, existing reviews have tended to focus on specific constructs or methodological aspects, leaving the transparency of test specifications and the adequacy of validity evidence across studies underexplored.

Addressing this gap carries important theoretical and methodological implications. Test specifications, as conceptualized by Davidson and Lynch (2002) and Owen (2018), articulate how theoretical constructs are operationalized into task features, input characteristics, scoring procedures, and administration conditions, while validity concerns the extent to which score interpretations and uses are supported by evidence and theory. Critically, the two are fundamentally interconnected: test specifications constitute a prerequisite for building a validity argument, as they provide the foundational information needed to evaluate whether a test measures what it claims to measure (Davidson & Lynch, 2002; Kane, 2006). Without clear specifications and validity evidence, it is difficult to evaluate what listening tests actually measure, compare findings across studies, or replicate research in different contexts. Importantly, validity evidence is inherently context-dependent, and therefore the adequacy of score interpretations must be evaluated in relation to specific test purposes, test-taker populations, and use contexts (Kane, 2006; Messick, 1989).

In addition, recent calls for greater transparency and reproducibility in second language research (Marsden et al., 2018) further highlight the importance of detailed reporting of research instruments and procedures. If critical aspects of listening test specifications, including input modality, exposure conditions, item formats, participant characteristics, and administration settings, are not systematically documented, replication becomes difficult and cross-study comparisons are weakened. Moreover, reliability, as a fundamental indicator of measurement consistency, represents an equally critical yet frequently underreported dimension of test quality. Its absence can undermine the interpretability of listening test scores across studies. Insufficient specification reporting may also hinder the accumulation of comparable evidence across studies and impede knowledge development in L2 listening assessment. In this sense, specification transparency is not merely a reporting convention but a prerequisite for cumulative knowledge building in L2 listening assessment research.

To address this gap, the present study provides a systematic review of test specifications and validity evidence of L2 listening tests used in empirical research published between 1997 and 2021. Drawing on established frameworks for operationalizing test specifications (Bejar et al., 2000; Davidson & Lynch, 2002; Marsden et al., 2018; Owen, 2018) and Kane's (1992, 2002, 2006) argument-based validation framework, this review aims to (a) examine how key test specification components are reported in L2 listening studies, (b) analyze the distribution of validity evidence across different inferences, and (c) evaluate the implications of current reporting practices for reproducibility and validity in L2 listening assessment research. This study, therefore, seeks to clarify current measurement and assessment practices and to inform future test development, validation, and reporting standards in L2 listening assessment.

2 Literature Review

2.1 Features of L2 listening assessments

L2 listening assessment is the process of collecting performance data through listening tasks and making inferences about L2 learners' listening ability, strengths and weaknesses, and level of achievement (Aryadoust, 2013; Aryadoust & Luo, 2023; Aryadoust & Zhai, 2026; Buck, 2001; Ockey & Wagner, 2018). Due to the unobservable processes happening in the mind of test takers during listening, the assessment of listening is always carried out via other (construct-irrelevant) language skills, such as reading, writing, or speaking (Buck, 2001; Brunfaut, 2016; Hartshorn & Stephens, 2023; Liu & Aryadoust, 2026; Ockey, 2013; Ockey & Wagner, 2018). This indirect measurement makes listening assessment particularly vulnerable to construct-irrelevant variance and construct underrepresentation, both of which pose serious threats to validity (Buck, 2001; Kane, 2006). Empirical evidence supports this concern: for instance, Cheung (2023) found that vocabulary knowledge accounted for over 66% of the variance in L2 listening scores, suggesting that performance on listening tests is substantially influenced by linguistic knowledge. More recently, Liu and Aryadoust (2026) identified distinct gaze patterns across different listening test formats, indicating that variations in test format may engage different non-listening cognitive processes in addition to listening itself.

The presence of factors unrelated to the listening construct, which can become sources of construct-irrelevant variance in test scores, is a serious threat to the validity of L2 listening assessments (Buck, 2001). Consequently, the design and use of listening tests should be conducted carefully to avoid confounding response processes and test scores with sources of construct-irrelevant variance (Holzknecht et al., 2021; Qiu & Aryadoust, 2024). This concern has been widely discussed in the L2 listening assessment literature, particularly in relation to task design, input modality, and test administration conditions (Liu & Aryadoust, 2026; Liu et al., 2022; Ockey, 2013; Suvorov & He, 2022).

In a systematic review, Aryadoust and Luo (2023) examined how the L2 listening construct has been conceptualized and operationalized across 157 peer-reviewed studies in applied linguistics research. The review revealed that only about one quarter of the studies provided explicit theoretical definitions of the listening construct, and most of these definitions were either very general or limited in scope. Several listening functions, such as discriminative, empathetic, and analytical listening, were found to be largely neglected in construct conceptualization across the literature. The review further identified numerous construct components, including 24 listening subskills, 27 cognitive processes, and 54 listening attributes, showing considerable variation in how listening has been defined and measured. Based on these findings, the authors proposed a multilayered framework to organize these components and argued that clearer construct definitions and more systematic operationalization are needed in future L2 listening research and assessment. These findings underscore the importance of examining not only how listening is assessed, but also how the construct underlying these assessments is defined and operationalized across the research literature.

Buck (2001) summarized the design features and use of listening tests from different perspectives, proposing that the process of test design is complex and that many task characteristics should be considered in test development, including context, item type, input type, accent characteristics, and the environment. Other factors that may be added to this list include the type of test (e.g., standardized versus classroom tests; Buck, 2001), the construct of the test (task-based versus competency-based constructs; see Chapelle et al., 2008), the mode of test delivery (paper-based versus computer-based tests; Ockey, 2013), modality (audio versus audiovisual inputs; Liu & Aryadoust, 2024; Ockey, 2013), type of tasks (e.g., multiple choice versus open-ended items; In'nami & Koizumi, 2009; Liu & Aryadoust, 2026; O'Grady, 2023), test duration and test setting, double exposure (Aryadoust, 2019; Holzknecht, 2019), genre (e.g., monologues versus dialogues; Jung, 2006; Tao & Aryadoust, 2024), and features of auditory stimuli such as accentedness and authenticity (Ockey & French, 2016; Wagner, 2014). These

characteristics collectively inform the theoretical framework adopted in the present study, providing the basis upon which the specifications and validity of L2 listening assessments are systematically examined.

Building on the test characteristics outlined above, several meta-analyses and research syntheses have further examined the effects of specific test characteristics on L2 listening performance. Focusing on input modality, Liu and Aryadoust (2024) found a small positive overall effect of video-based listening tests on L2 listening performance, with the effects moderated by factors such as proficiency level, research design, video features, and test characteristics. Similarly, Suvorov and He (2022) reviewed studies on the use of visuals in L2 listening assessment and concluded that inconsistent findings in the literature are largely attributable to substantial methodological differences across studies, including variations in research design, participant characteristics, test design, and administration procedures. Other meta-analyses have examined the correlates of L2 listening comprehension. In'nami et al. (2022) reported a moderate overall relationship between L2 listening and its components, with linguistic knowledge showing stronger relationships with listening than cognitive and affective factors. Zhang and Zhang (2022) found strong correlations between vocabulary knowledge and L2 listening comprehension, indicating that vocabulary knowledge explains a substantial proportion of variance in listening comprehension. Likewise, Karalık and Merç (2019) showed that linguistic factors had the strongest associations with listening comprehension, followed by metacognitive skills, while working memory showed weaker relationships and affective factors were more strongly related to L2 listening than L1 listening. These findings suggest that L2 listening performance is shaped by a complex interplay of test design characteristics, input modality, and linguistic factors, all of which have direct implications for how listening assessments are specified and validated.

2.2 Specifications of listening tests

Test specifications refer to the articulation of content and process features of a test based on which a test designer can develop a listening test (Davidson & Lynch, 2002; Raymond & Neustel, 2006). The documents that contain the aforementioned information are called test specifications, which are either content-based, process-based, or both (Raymond & Neustel, 2006). The test features spelled out in a test specification document include information about the test format and administration, construct definition, mode and number of test items, local and international requirements, and related test design features (Alderson, 2000). Test specifications also include information about the test takers (e.g., age, gender, level of proficiency) and the target population (e.g., nationality, first and second language) (Alderson, 2000).

Test specifications are not always available to the public, in which case they or parts of them can be "reverse-engineered" based on the available test items and/or published research on the test (Davidson & Lynch, 2002). There are two important advantages to the reverse-engineering of test specifications: (a) it can provide evidence of validity, and (b) it can provide evidence of reproducibility. Since test specifications contain information about test content and process and the population of test takers and related test characteristics, they can be used in the validity argument for the tests (Davidson & Lynch, 2002; Owen, 2018). For example, the information about construct definition and operationalization, target population, and the target domain tasks can constitute appropriate evidence concerning the domain specification and construct representation inferences in validity arguments (Chapelle et al., 2008).

In addition, although the validity argument framework proposed by Kane (2002, 2006) is silent on the reproducibility of validation research, test specifications can be used as necessary information in validation. Recent L2 listening assessment research has stressed the significance of reproducibility in validation (Liu et al., 2022). Liu et al. (2022) presented a framework in which the structural evidence for validity is linked to the reproducibility of research and validation. As the authors argued, a strong validity argument should include evidence supporting the validity of different test forms, while evidence supporting one test version is likely sample-specific and cannot be extrapolated across different samples,

test forms, and test administrations. By gathering information about test specifications, one can establish the degree of reproducibility of the research conducted and test results across different assessment situations and populations. Reproducibility has been presented as an essential feature of second language research by Marsden et al. (2018), highlighting its ever-increasing importance across different fields. We draw on these studies to reverse-engineer features of test specifications to gain an idea about reproducibility as well as the validity of the listening assessments used in the previous research.

2.3 Argument-based validity

Founded upon Messick's (1989) unitary concept of validity and Toulmin's (2003) argument structure, Kane's argument-based validation (ABV) approach (1992, 2002, 2006, 2013) provides a framework for validating the interpretations and uses of test scores through the evaluation of both theoretical rationales and empirical evidence. This framework consists of two types of arguments: interpretive and validity arguments (Aryadoust & Zakaria, 2026). The interpretive argument specifies the proposed interpretations and uses of test scores and the chain of inferences linking observed performance to score-based decisions. To build the validity argument, the accuracy and plausibility of the claims made in the interpretive argument are then empirically investigated and supported with evidence (Kane, 2006).

2.3.1 Interpretive argument

An interpretive argument is "a network of inferences and supporting assumptions from scores leading to conclusions and decisions" (Tran, 2020, p. 50). It consists of six main validity inferences. The first inference, the domain description inference, characterizes the target language use (TLU) domain by identifying the tasks and contexts in which the test takers are expected to use the language. It provides the foundation for the entire validity argument by establishing the criteria against which the representativeness of test content can be evaluated (Chapelle et al., 2008; Fan & Yan, 2020).

The second inference, called the evaluation or scoring inference, links the observed test performance to an observed score. It makes assumptions about the appropriateness of scoring rubrics or criteria (Kane, 1992, 2013) and whether test takers' performance is accurately observed and appropriately mapped onto score levels or rating categories. The third inference, generalization, which links the observed score to the universe score, assumes that the observed scores will be consistent over numerous test administrations and parallel test versions (Aryadoust, 2013; Aryadoust & Zakaria, 2026). For this inference to be supported, evidence should be presented to show that test takers' scores are reliable and reflect their performance on a variety of similar tasks under similar conditions.

The fourth inference linking the universe scores (the output of the previous step) and the construct of interest is the explanation inference, which is fundamentally the same as traditional construct validity. It posits that the test accurately measures its intended theoretical construct and that the performance of test takers reflects their standing on the continuum of the underlying theoretical construct that the test claims to measure. Like construct validity, this inference assumes that the test includes "the evidence of sufficient construct representation, and lack of construct-irrelevant variance" (Aryadoust, 2013, p. 31).

Next, the extrapolation inference links the universe score to the score in the target language use (TLU) domain. It addresses whether test takers' scores provide sufficient evidence of language competence in the TLU domain (Xi, 2008). Typically, this inference makes assumptions about the representativeness of test takers' scores for their future performance in real-life tasks and contexts (Kane, 2004). Finally, the utilization inference links the target score to the decisions about test takers. The term "utilization" has been adopted in language assessment (e.g., Chapelle et al., 2008) to refer to accurate and appropriate decision-making based on the scores. These six inferences form the core components of an interpretive argument. Notably, they are not a checklist for test validation, and there is no requirement to follow a rigid structure; rather, the flexibility in the development of an interpretive argument allows the researcher

to articulate claims concerning the uses and interpretation of test scores and then provide evidence to support those claims (Kane, 2013).

2.3.2 Validity argument

The validity argument allows for an assessment of the inferences laid out in the interpretive argument (Kane, 2002). Toulmin's (2003) argument structure model was adopted by Kane (2013) for the construction of validity arguments. It consists of six main elements: data, claim, backing, warrant, rebuttal, and qualifier (Toulmin, 2003).

Each inference in the ABV framework starts from the data and leads to a claim (conclusion), which is "a conclusion whose merits we are seeking to establish" (Toulmin, 2003, p. 90). In other words, the claim is the conclusion that we want to draw based on test takers' performance or other evidence. Data refer to "information on which the claim is based" (Toulmin, 2003, p. 90), while warrants are general and hypothetical statements that support the link between data and claims. In addition, backing is a general statement that "legitimize(s) warrants and their application" (Aryadoust, 2013, p. 28), and a rebuttal is a situation or exception in which the warrant is not suitable. Lastly, a qualifier refers to the strength of the relationship between data and claim (Kane, 2013).

The ABV framework has been widely used in validation research on international language tests, such as the Test of English as a Foreign Language (TOEFL; Chapelle et al., 2008), Pearson Test of English (Wang et al., 2012), and the International English Language Testing System (IELTS; Aryadoust, 2013), as well as in local language tests, such as the Locally Created Listening Test (a Vietnamese listening test) (Tran, 2020). It has also been used as a conceptual structure for reviews of second language assessments (Fan & Yan, 2020). Therefore, the ABV framework is adopted in the present review to explore whether the validation of L2 listening tests in prior studies was adequately reported and supported with evidence.

2.4 Research questions

The research questions of the study are as follows:

1. In what types of populations were the listening assessments administered and were the tests used locally or internationally?
2. What test-related specification components of the L2 listening assessments were provided by the researchers in the studies?
3. How were the validity and reliability of the measurement instruments examined? What validity inferences were supported in the language studies?

While research question 3 directly focuses on the validity evidence for the listening assessments, questions 1 and 2 are intended to reverse-engineer the features of test specifications of the assessments and to examine the populations to which the test score interpretations are generalized or extrapolated. Examining questions 1 and 2 also allows us to evaluate the reproducibility and generalizability of the studies, which has been an area of increasing attention in second language research (Liu et al., 2022; Marsden et al., 2018).

3 Methodology

3.1 Dataset

We used the Web of Science (WoS) Core Collection database, a major multidisciplinary citation database, to search keywords comprising "listening", "second language", "test", "assess", "measure"

and "measurable". To avoid missing any relevant studies, the search strategy combined the following keywords: "listening", "second language", and one among "test", "assess", "measure", and "measurable" (see Appendix A for details).

The initial dataset from WoS contained 388 papers, which were then further screened. As shown in Table 1, the dataset was refined based on six inclusion and six exclusion criteria. For example, the first criterion was whether the paper was an empirical study published in a journal, and the second criterion was whether the journal focused on language education and/or language assessment research. Therefore, journals related to biomedicine, economics, and other fields were excluded.

The inclusion and exclusion criteria were developed in accordance with established principles of systematic review methodology (Mayo-Wilson & Grant, 2019; Marsden et al., 2018). The first criterion, requiring studies to be empirical papers published in peer-reviewed journals, reflects standard practice in research synthesis of prioritizing methodologically rigorous evidence over grey literature or non-empirical work (Mayo-Wilson & Grant, 2019, p. 474). The second criterion, restricting the search to journals focused on language education and/or language assessment, was applied to ensure topical relevance and to maintain the focus of the review on L2 listening assessment research. The third criterion, requiring studies to be published in English, was adopted for practical reasons related to coding reliability (Marsden et al., 2018). The fourth criterion regarding accessibility was included to ensure that all selected studies could be fully examined and coded. The fifth and sixth criteria, requiring studies to have used at least one L2 listening comprehension test, were applied to ensure construct relevance, excluding tests that measure other constructs such as vocabulary knowledge, sentence repetition, or L1 listening comprehension, which fall outside the scope of the present review (Buck, 2001; Ockey & Wagner, 2018).

Table 1

Inclusion and Exclusion Criteria

Inclusion Criteria: The paper…	**Exclusion Criteria: The paper…**
1. was an empirical study published in a journal.	1. was a review, a book, a proceeding paper, a book chapter, or a book review.
2. was relevant to language education.	2. was published in journals relating to biomedicine, economics, etc.
3. was published in English.	3. was written in a language other than English such as Afrikaans, French, German, Spanish.
4. was accessible.	4. was unavailable.
5. used at least one listening test.	5.1 did not use any listening test.
	5.2 conducted other tests such as recognition tests, sentence repetition tests, vocabulary tests, judgment tasks, listening span tests, dictation tests, or translation tests, etc.
6. included a test measuring L2 listening comprehension.	6. focused on L1 listening comprehension test.

After the application of the screening procedures based on the foregoing criteria, the search yielded a dataset of 126 papers published in 41 journals. During this process, 29 articles were found to have used more than one listening comprehension test: that is, 24 articles used two listening comprehension tests, 4 articles used three listening tests, and 1 article used five tests, thus totalling 162 listening comprehension tests. Among these, 28 general language proficiency tests were included within the 162 tests, of which 26 clearly stated that the tests had a listening comprehension component, while 2 tests did not specify

the use of a listening subtest and, hence, these test batteries that did not explicitly mention the listening component were excluded. In total, there were 160 listening comprehension tests in the 126 studies.

Overall, the studies were published in 41 journals; the descriptive information of the journals for the top 20 journals is presented in Table 2. The top 10 journals accounted for more than half of the dataset, with the remaining journals each containing fewer than four articles, with an average of 1.58 papers per journal. The journal with the largest number of published articles is Language Testing (n = 12, 9.52%), followed by Language Learning (n = 11, 8.73%), Foreign Language Annals (n = 9, 7.14%), System (n = 9, 7.14%), Modern Language Journal (n = 8, 6.35%), Studies in Second Language Acquisition (n = 7, 5.56%), and Language Assessment Quarterly (n = 6, 4.76%), among others.

Table 2
Journals Publishing L2 Listening Assessment Research

Journal	No. of papers	%
Language Testing	12	9.52%
Language Learning	11	8.73%
Foreign Language Annals	9	7.14%
System	9	7.14%
Modern Language Journal	8	6.35%
Studies in Second Language Acquisition	7	5.56%
Language Assessment Quarterly	6	4.76%
TESOL Quarterly	6	4.76%
Language Teaching Research	5	3.97%
Canadian Modern Language Review	4	3.17%
Computers & Education	3	2.38%
Frontiers in Psychology	3	2.38%
International Journal of Bilingual Education and Bilingualism	3	2.38%
International Journal of Bilingualism	3	2.38%
Language Learning & Technology	3	2.38%
PLOS One	3	2.38%
Reading and Writing	3	2.38%
ReCALL	3	2.38%
Journal of Psycholinguistic Research	2	1.59%
Reading Research Quarterly	2	1.59%

3.2 Coding scheme

A coding scheme comprising 26 variables was designed drawing on Plonsky and Gass (2011), Marsden et al. (2018), and Perez et al. (2013) (see Table 3). The variables were further divided into six groups: administrative information (n = 4), test information (n = 10), textual features of the test (n = 4), participant information (n = 6), validity inferences (n = 2), and results (n = 1). The rationale behind the coding scheme was to enable the reverse-engineering of test specifications and to examine the replicability of the studies and/or the reproducibility of the results. According to Marsden et al. (2018), the presence of reliable and precise information concerning the sample, instruments and their

psychometric features, the population to which the results are generalized, and the study procedures enhances the replicability of the study, as the parameters of the study can be reproduced accurately by future researchers. This information is also essential for test specifications, allowing different test designers to develop test content based on clearly specified criteria (Davidson & Lynch, 2002).

For clarity, we categorized 14 variables into two groups. The test information group included the name of the test, test type, test development, linguistic features, test method, item type, target language, test mode, test duration, and test setting, while the textual features group included four variables: media type, accent, the number of text plays, and material input type. Studies have shown that different test characteristics affect the outcome of the test, such as material input type (Papageorgiou et al., 2012), media type (Holzknecht, 2019), and accents (Shin et al., 2021). Therefore, the types of tests and their characteristics used in the studies were reviewed.

In addition, we coded whether researchers provided details and evidence supporting the validity of the listening assessment(s) used in each study based on the following six validity inferences (domain description, evaluation, generalization, explanation, extrapolation, and utilization; Chapelle et al., 2008; Kane, 2006). The rest of the categories are available in Table 3.

Table 3

Variables and Definitions in the Coding Scheme

Variables	Description	References
Administrative information		
Authors	Researchers who conducted the study	Plonsky & Gass (2011); Teimouri et al. (2019)
Title	The title of the article	
Year	The year in which the article was published	
Journal	The journal in which the article was published	
Test information		
Test name	The name of the test used in the study (open coding)	
Test type	1. Standardized test 2. Local test/Classroom test	Perez et al. (2013)
Test development	1. Task-based language assessment (TBLA); 2. Competency-based assessment (CBA); 3. Not stated	Taylor (2013)
Linguistic features	Tests and questionnaires used along with listening tests in studies	Rost (2016); Aryadoust (2019)
Test method	1. While-listening performance (WLP) 2. Post-listening performance (PLP) 3. Not stated	Aryadoust (2012); Aryadoust (2013)
Item type	Response methods 1. Multiple-choice questions (MCQs) 2. Fill-in-the-blank questions (Spoken/Written) 3. T/F 4. Matching 5. Short answer questions (Spoken/Written) 6. Summary (Spoken/Written) 7. Sustained write 8. Mixed: a combination with more than one item type 9. Not stated	

Variables	Description	References
Target language	The language measured in the listening assessment	
Test mode	1. Paper-based assessment 2. Computer-based assessment 3. Others (Oral assessment) 4. Not stated	Öz & Özturan (2018)
Test duration	The length of the input material	
Setting	1. Classroom 2. Quiet room 3. Laboratory 4. A space with computers 5. Not stated	Plonsky & Gass (2011)
Textual feature		
Media type	1. Audio-only 2. Audiovisual (Picture/Video) 3. Not stated	Suvorov (2015)
Accents	The accents of the speaker of the record 1. native speaker 2. nonnative speaker 3. Mixed: a combination with more than one accent 4. Not stated	Shin et al. (2021)
The number of text plays	1. Once 2. Twice 3. More than twice 4. Mixed 5. Not stated	Aryadoust (2019); Holzknecht (2019)
Material input type	1. Monologue 2. Dialogue 3. Sentence 4. Mixed: a combination of monologue and dialogue 5. Not stated	Buck (2001); Papageorgiou et al. (2012)
Participant information		
Sample size	The number of participants that took part in the research	Teimouri et al. (2019); Perez et al. (2013)
L1	The native language of participants	Marsden et al. (2018); Teimouri et al. (2019); Perez et al. (2013)
L2	The second language of participants	
Proficiency level (Ability level)	The second language ability of participants	
Instructional level	The instructional level of the participants	Perez et al. (2013)
Location	The country/region where the study was conducted	Teimouri et al. (2019)
Validity Inference		
Argument-based framework	1. Domain description 2. Evaluation 3. Generalization 4. Explanation 5. Extrapolation 6. Utilization	Chapelle et al. (2008); Fan & Yan (2020)

Variables	Description	References
Data analysis methods of validity inferences	The analysis method used to measure validity inferences	Plonsky & Gass (2011); Marsden et al. (2018)

3.3 Inter-coder reliability

An independent researcher verified the coverage of the dataset, the process of data generation from the WoS, and the coding scheme to ensure inter-coder reliability. The researcher double-checked the coverage of the dataset and the accuracy of the data extraction process. Moreover, the researcher recoded 25% of the studies using the coding scheme presented in Table 3. The selection of 25% of the dataset for double-coding exceeds the proportions adopted in comparable systematic review studies in applied linguistics research, including Plonsky and Gass (2011), who coded 12% of their sample, Fan and Yan (2020), who coded 19.23%, and Marsden et al. (2018), who coded 21%.

For the randomly selected 25% of the studies, the reliability of the double-coding of the 26 variables was assessed using Cohen's Kappa, which ranged from 0.90 to 1.00, providing strong evidence for the inter-coder reliability of the coding scheme and the results (Landis & Koch, 1977).

4 Results

4.1 Research question 1

The first research question of the study focused on the population type and whether the tests used were developed locally or were international tests of second language listening adopted in the studies. Two categories of test types were identified: standardized listening assessments (n = 103, 64.38%) and local assessments (n = 57, 35.63%).

In addition, we extracted information about test takers, including sample size, first language (L1), instructional level, and second language (L2) proficiency level (see Table 4). The first variable in Table 4 is the sample size, which was grouped into three classes for clarity: less than 100 examinees, between 100 and 500 examinees, and more than 500 examinees. The findings showed that half of the listening assessments (n = 80) were conducted in samples between 100 and 500 participants, while the number of test takers in 73 assessments (45.63% of all tests) was less than 100. In addition, 4.38% of assessments were administered to samples that were larger than 500 participants.

The second variable in Table 4 is the L1 of test takers, which is discussed below alongside the distribution of L1 backgrounds. The third variable is the instructional level of test takers. Over half of the assessments (n = 98, 61.25%) were conducted in university contexts, and within this group more than half of the assessments were standardized tests (n = 56, 35.00%). On the other hand, the participants in 26.88% of assessments were from K–12 schools, where standardized assessments (n = 31, 19.38%) were predominantly used. In addition, 19 (11.88%) assessments remained unidentified, as the authors did not provide relevant information.

The last variable in Table 4 is the L2 proficiency level of the participants. Seven proficiency level categories were identified: low, intermediate, high, low-to-intermediate, intermediate-to-high, low-to-high, and not reported. Nearly half of the assessments (n = 76, 47.50%) failed to provide any details regarding the L2 level of the test takers in these studies. In 17.50% of the assessments, the tests were administered to examinees with proficiency levels ranging from low to high. Intermediate-level test takers were represented in 25 assessments (15.63%), among which more than half (11.25%) were local assessments. Similarly, the low proficiency level (6.88%), and from low to intermediate levels (5.63%)

were represented in 11 and 9 assessments, respectively. Six assessments (3.75%) included participants at intermediate to high levels, and five assessments (3.13%) included participants at the high level, which represented the smallest group.

Table 4

Listening Test Takers' Information

Participant information	**Standardized assessment (n = 103)**		**Local assessment (n = 57)**		**Total (n = 160)**	
	# of tests	**%**	**# of tests**	**%**	**# of tests**	**%**
Sample size						
P ≤ 100	39	24.38	34	21.25	73	45.63
100 < P ≤500	58	36.25	22	13.75	80	50.00
500 < P	6	3.75	1	0.63	7	4.38
L1						
Chinese	16	10.00	11	6.88	27	16.88
English	15	9.38	8	5.00	23	14.38
Japanese	16	10.00	0	0.00	16	10.00
Turkish	2	1.25	4	2.50	6	3.75
Korean	3	1.88	1	0.63	4	2.50
Spanish	3	1.88	1	0.63	4	2.50
Dutch	1	0.63	3	1.88	4	2.50
Arabic	0	0.00	3	1.88	3	1.88
Persian	2	1.25	0	0.00	2	1.25
German	1	0.63	1	0.63	2	1.25
Thai	2	1.25	0	0.00	2	1.25
Swedish	2	1.25	0	0.00	2	1.25
French	1	0.63	0	0.00	1	0.63
More than one L1	32	20.00	20	12.50	52	32.50
N/A	7	4.38	5	3.13	12	7.50
Instructional level						
K-12	31	19.38	12	7.50	43	26.88
University	56	35.00	42	26.25	98	61.25
N/A	16	10.00	3	1.88	19	11.88
Second language level						
Low level	7	4.38	4	2.50	11	6.88
Intermediate level	7	4.38	18	11.25	25	15.63
High level	3	1.88	2	1.25	5	3.13
L to I	7	4.38	2	1.25	9	5.63
I to H	5	3.13	1	0.63	6	3.75
L to H	17	10.63	11	6.88	28	17.50
N/A	57	35.63	19	11.88	76	47.50

Note. L to I = low level to intermediate level; I to H = intermediate level to high level; L to H = low level to high level

Finally, in terms of test takers' L1, the studies employed students from diverse first-language backgrounds, with the most frequently studied L1 backgrounds being Chinese, English, and Japanese. Notably, only standardized tests were used to measure the listening skills of students with Japanese

as their L1, although there was very little evidence for the alignment between the listening abilities of the participants and the constructs measured by the standardized tests. The distribution of the L1 backgrounds of participants in the studies may indicate regions where there is greater demand for L2 listening assessment (and language assessment more broadly); the three most researched regions were the US, China, and Japan. The US has the largest English as an L2 population in the world, and China and Japan have large foreign-language education systems with substantial demand for language assessment.

4.2 Research question 2

The summary information pertaining to the specification components of listening comprehension tests, namely test development, test method, item type, target language, test mode, test duration, and test setting, is provided in Table 5. Overall, the proportion of studies failing to report each specification component ranged from 21.25% to 89.38%.

Three categories related to test development were identified: task-based language assessment (TBLA), competency-based assessment (CBA), and not provided. The test development approach of most listening assessments (n = 143, 89.38%) was not reported, regardless of whether the tests were standardized assessments (n = 90, 56.25%) or local assessments (n = 53, 33.13%). Of the remaining assessments, more tests were classified as TBLA (standardized assessment, n = 11, 6.88%; local assessment, n = 3, 1.88%) than CBA, of which two were standardized assessments and one was a local assessment.

The second variable, test method, contains three groups: WLP, PLP, and not provided. As with the first variable, the test method of 110 listening assessments (68.75%) was not mentioned by the authors of the studies. The majority of these unreported cases involved standardized assessments (n = 83, 51.88%), and among the studies that provided information about the test method, WLP tests (n = 28, 17.50%) were slightly more common than PLP tests (n = 22, 13.75%).

In addition, nine item types were identified in the dataset. Multiple-choice questions (MCQ) were the most popular question type across both standardized assessments (n = 35, 21.88%) and local assessments (n = 30, 18.75%), accounting for 40.63% of all listening tests. Twenty-one assessments (13.13%) adopted more than one item type, of which most were standardized listening tests (n = 16, 10.00%). Short answer questions (n = 13, 8.13%) and summary tasks (n = 9, 5.63%) were also commonly used in listening assessments, followed by fill-in-the-blank questions (n = 6, 3.75%), True/False questions (n = 6, 3.75%), and matching questions (n = 5, 3.13%). Only one standardized listening test used sustained writing as the item type. Finally, 34 assessments (21.25%) did not report the item type, most of which (n = 26, 16.25%) were standardized assessments.

Table 5

Test-Related Specification Components of Listening Tests

Test information	Standardized assessment (n = 103)		Local assessment (n = 57)		Total (n = 160)	
	# of tests	%	# of tests	%	# of tests	%
Test development						
TBLA	11	6.88	3	1.88	14	8.75
CBA	2	1.25	1	0.63	3	1.88
Not provided	90	56.25	53	33.13	143	89.38
Test method						
WLP	11	6.88	17	10.63	28	17.50
PLP	9	5.63	13	8.13	22	13.75

Test information	Standardized assessment (n = 103)		Local assessment (n = 57)		Total (n = 160)	
	# of tests	%	# of tests	%	# of tests	%
Not provided	83	51.88	27	16.88	110	68.75
Item type						
MCQ	35	21.88	30	18.75	65	40.63
Short answer	11	6.88	2	1.25	13	8.13
Summary	4	2.50	5	3.13	9	5.63
Fill-in-the-blank	6	3.75	0	0.00	6	3.75
T/F	2	1.25	4	2.50	6	3.75
Matching	2	1.25	3	1.88	5	3.13
Sustained writing	1	0.63	0	0.00	1	0.63
Mixed	16	10.00	5	3.13	21	13.13
Not provided	26	16.25	8	5.00	34	21.25
Target language						
English	73	45.63	34	21.25	107	66.88
Spanish	11	6.88	9	5.63	20	12.50
French	8	5.00	2	1.25	10	6.25
German	3	1.88	5	3.13	8	5.00
Chinese	3	1.88	1	0.63	4	2.50
Japanese	1	0.63	2	1.25	3	1.88
Dutch	2	1.25	0	0.00	2	1.25
Arabic	0	0.00	2	1.25	2	1.25
Korean	1	0.63	0	0.00	1	0.63
Russian	1	0.63	0	0.00	1	0.63
Croatian	0	0.00	1	0.63	1	0.63
Norwegian	0	0.00	1	0.63	1	0.63
Test mode						
Paper-based	28	17.50	12	7.50	40	25.00
Computer-based	21	13.13	16	10.00	37	23.13
Oral-based	7	4.38	2	1.25	9	5.63
Not provided	47	29.38	27	16.88	74	46.25
Test duration						
≤ 0.5 hour	19	11.88	7	4.38	26	16.25
0.5 hour < & <1 hour	22	13.75	6	3.75	28	17.50
≥ 1 hour	4	2.50	0	0.00	4	2.50
Not provided	58	36.25	44	27.50	102	63.75
Test setting						
Classroom	10	6.25	11	6.88	21	13.13
Quiet room	7	4.38	3	1.88	10	6.25
Laboratory	10	6.25	6	3.75	16	10.00
A space with a computer	3	1.88	1	0.63	4	2.50
Not provided	73	45.63	36	22.50	109	68.13

Note. TBLA = task-based language assessment; CBA = competency-based assessment; WLP = while-listening performance; PLP = post-listening performance; MCQ = multiple-choice question; T/F = true or false.

Other factors. Due to space constraints, other components of test specifications are presented in Appendix B. These include variables such as media type, speakers' accents, double exposure, and input type.

4.3 Research question 3

Table 6 presents the distribution of validity evidence across the six validity inferences within the ABV framework for the identified listening comprehension assessments. The results revealed a notably unequal distribution of validity evidence across the six inferences. Specifically, 88 assessments provided evidence related to only one validity inference, while 10 assessments examined two inferences. Notably, no validity inferences were examined in 62 assessments included in the studies.

Accounting for 56.88% of all the tests, the generalization inference (n = 91) of listening comprehension assessments was analyzed most frequently in the L2 listening studies, while the evaluation inference (n = 13) accounted for a much lower portion, representing 8.13% of all the tests. On the other hand, no evidence related to the domain description and utilization inferences of the validity of the tests was reported in any study. The number of tests with evidence related to the explanation and extrapolation inferences was similar, consisting of 3 (1.88%) tests and 1 (0.63%) test, respectively.

Table 7 presents the six inferences across the two listening assessment types. Among all tests grouped under generalization inferences, more than half were standardized assessments (n = 59, 36.88%), while 20% of tests were local assessments (n = 32). A similar trend was found for tests grouped under the evaluation inference, with 11 (6.88%) standardized assessments and only 2 (1.25%) local assessments. The studies that reported evidence for the explanation and extrapolation inferences showed a similar trend, as they consisted exclusively of standardized assessments (n = 3 and 1, respectively). The techniques applied to examine each of the inferences are presented in Appendix C.

Table 6

Representation of Six Validity Inferences in the Listening Comprehension Assessments (n = 160)

Inferences	**Number of tests**	
	# of tests	**%**
Domain description	0	0.00
Evaluation	13	8.13
Generalization	91	56.88
Explanation	3	1.88
Extrapolation	1	0.63
Utilization	0	0.00

Table 7

Representation of Six Validity Inferences across Listening Comprehension Assessments (n = 160)

Inferences	**Standardized assessment**		**Local assessment**	
	# of tests	**%**	**# of tests**	**%**
Domain description (n = 0)	0	0.00	0	0.00
Evaluation (n = 13)	11	6.88	2	1.25
Generalization (n = 91)	59	36.88	32	20.00
Explanation (n = 3)	3	1.88	0	0.00
Extrapolation (n = 1)	1	0.63	0	0.00
Utilization (n = 0)	0	0.00	0	0.00

5 Discussion

This study set out to review the published research relating to L2 listening comprehension assessments published from 1997 to 2021 to examine their test specifications and validity arguments. The findings reveal consistent patterns of incomplete specification reporting and uneven validity evidence across the reviewed studies, with important implications for the reproducibility, construct representation, and validity of L2 listening assessment research. The following discussion addresses these findings in relation to each research question, attending to both the cumulative patterns that emerged from the dataset and the contextual factors that may account for observed variation in reporting and validation practices.

5.1 Research question 1

Research question 1 examined the type of L2 listening tests and the type of populations participating in listening assessment studies. Information about participants (L1, instructional level, and L2 level), particularly the L2 level, was not reported in many studies, thereby limiting the replicability of the studies. Generally, the less information that is provided about the attributes of test takers, the less replicable the study will become (Marsden et al., 2018).

The first notable finding of the study was that a large number of listening tests used in the studies were standardized tests, likely due to the availability of practice test samples or the funding that some of the test developers may offer to researchers. A considerable number of "standardized" listening comprehension assessments were identified in the studies such as IELTS, TOEFL, TOEIC, the Pearson Test of English Academic (PTE Academic), and the Woodcock Language Proficiency Battery (WLPB). As most standardized assessments are developed by groups of experts and may take a long time to create, these tests tend to achieve better reliability and perhaps validity than classroom tests (Van Blerkom, 2017). On the other hand, and in line with Shang et al.'s (2024) findings, fewer researchers used local assessments, some of which were developed based on or modeled after standardized assessments, such as the TOEFL (Yi, 2017) to improve validity. This indicates a significant influence of commercial standardized tests and their features on the field of L2 listening, which has been documented in previous research (e.g., Wagner, 2014). The considerable disadvantage of this dominance is that virtually all of these tests expose language learners to contrived and "polished" audio stimuli that contain no background noise (Fujita, 2022) and/or features of natural speech such as assimilation and rapid speed of delivery (Ockey & Wagner, 2018; Wagner, 2014).

This may also lead to negative washback on L2 listening curricula, wherein teachers may tend to use inauthentic listening materials that do not reflect the realities of spoken English in TLU domains (Wagner, 2014). From a validity perspective, the dominance of standardized tests raises important questions about the domain description inference in the ABV framework (Kane, 2006): if the audio stimuli used in these tests do not adequately represent the characteristics of spoken language in real-life TLU domains, the validity of extrapolating test scores to broader language use contexts becomes questionable (Chapelle et al., 2008; Zhao & Aryadoust, 2025). This further underscores the need for more transparent test specifications that explicitly document the relationship between test content and TLU domain characteristics.

Second, the sample sizes of the studies were mostly below 500 and the largest number of listening tests were administered in universities, with far fewer conducted in K-12 schools, which might be related to the fact that considerably more learners acquire an L2 later in life. This is in line with Aryadoust's (2020) investigation of L1 versus L2 listening research, which found that L2 listening studies were primarily focused on adults, while L1 listening studies were focused on children. We suggest that more L2 studies should be carried out to investigate the listening comprehension of students in basic and/or bilingual education contexts, and the differences between L2 listening across different age levels should be further investigated.

With regard to the participants' L2 level, many researchers explored students from three language levels (low, intermediate, and high), although many tests focused on students from the intermediate level. However, this system of placement appeared to be arbitrary, and no supporting evidence was found. Therefore, listeners from the same level in different studies might not necessarily possess the same level of L2 listening skills. Of note, we found that quite a few researchers decided to group test takers based on one score. Making such decisions requires conducting rigorous standard-setting studies (Papageorgiou, 2010). Admittedly, it may not be pragmatic to expect that researchers conduct such studies prior to their listening research. However, there are two plausible solutions to help optimize the accuracy of the groupings: (a) reporting such arbitrary groupings of students should either be avoided or be backed up by convincing evidence regarding the criteria for grouping; and (b) classification and/or clustering techniques should be applied to group listeners into qualitatively different groups (e.g., Bouveyron et al., 2019).

Further research is required to examine the feasibility of such categorization systems in L2 listening. From a validity perspective, the arbitrary grouping of test takers based on a single score without rigorous standard-setting procedures most directly threatens the decision-making (or utilization) inference in the ABV framework, since the cut scores used to classify listeners into proficiency levels lack a defensible basis. This, in turn, weakens the warrant that the resulting level assignments, and the proficiency-based interpretations and cross-study comparisons that rest on them, are appropriate (Chapelle et al., 2008). Researchers should therefore exercise greater care in documenting their approach to proficiency grouping, as this constitutes an essential element of methodological transparency and directly affects the reproducibility and validity of research findings.

5.2 Research question 2

The findings related to research question 2 reveal a broader pattern of incomplete specification reporting across multiple test characteristics, including test development approach, test methods, item types, test mode, test duration, accent, double exposure, test setting, media, and input types. Each of these characteristics constitutes an essential component of test specifications, and their systematic underreporting has direct implications for the reproducibility of research and the validity of score interpretations.

Test development. Most researchers failed to provide detailed information about whether the tests they used were TBLA or CBA. Despite the lack of explicit descriptions of test design approaches, it appears that many L2 listening tests are TBLA, most of which are standardized listening assessments. On the other hand, several researchers or test developers described the test development of some standardized assessments as TBLA (Stoynoff, 2009; Panahi, 2012). The widespread failure to report test development process has direct implications for the validity of listening assessments. Within the ABV framework, the absence of information about whether a test is constructed from TBLA or CBA perspectives undermines the domain description inference, as it becomes impossible to evaluate whether the test tasks adequately represent the target language use domain (Chapelle et al., 2008). Furthermore, without explicit documentation of test development approaches, the reproducibility of studies is weakened, as future researchers cannot replicate the assessment conditions or evaluate the construct representation of the tests used (Davidson & Lynch, 2002; Marsden et al., 2018).

Test methods. The use of test methods (WLP versus PLP) was not described in most studies, which hampers the reverse-engineering of the test specifications and adversely affects the replication of the studies. This underreporting is particularly consequential from a validity perspective, as the choice between WLP and PLP directly affects the cognitive demands placed on test takers and the construct being measured (Liu & Aryadoust, 2026). The provision of limited information may be attributed to the possibility that researchers have not fully recognized the consequences of adopting WLP versus PLP. In fact, how the use of different test methods influences the outcomes of L2 listening comprehension is still

an underresearched area. The limited available research on this topic found inconsistent results. Koyama et al. (2016) found that students taking WLP tests outperformed those taking PLP tests, which was replicated by Yanagawa and Green (2008), O'Grady (2023), and Aryadoust et al. (2022), although Field (2013) argued that PLP tests possess higher authenticity. These inconsistencies are likely attributable to differences in test specifications across studies, including variations in task design, input characteristics, and participant proficiency levels, which further underscores the importance of transparent specification reporting.

Within the ABV framework, the choice of test method bears directly on the evaluation inference, as it affects whether test takers' observed scores accurately reflect their listening ability or are confounded by the cognitive demands of the response format (Kane, 2006; Buck, 2001). Beyond the evaluation inference, this choice also has implications for the explanation inference (traditionally termed construct validity), since a response format whose demands diverge from the target construct can introduce construct-irrelevant variance into the scores. If, for instance, WLP tests impose additional reading or processing burdens that are unrelated to listening comprehension per se (see Liu & Aryadoust, 2026), then the resulting scores may partly reflect these extraneous demands rather than the intended construct, thereby weakening the warrant that score variation can be attributed to differences in listening ability (Messick, 1989; Kane, 2006). The underreporting of test methods is therefore doubly problematic: it not only obscures how observed scores were generated but also prevents readers from evaluating whether the construct of interest was adequately and exclusively represented, leaving both the evaluation and explanation inferences difficult to scrutinize.

Despite this, more researchers preferred to use WLP tests with the MCQ format, yet the development rationale for most of these tests was not documented. Future research should therefore investigate the differential effects of WLP and PLP formats on construct representation and score validity across different proficiency levels and assessment contexts.

Item types. Both selected-response formats (e.g., Wallace, 2020) and constructed-response formats (e.g., Song, 2008, 2012; Leeser, 2004) were used in the examined research. Nevertheless, the format of some tests was not clearly described, which again affects the replicability of the studies and our understanding of their specifications. Most researchers who reported the item type adopted MCQs in their listening tests, and interestingly both standardized and local tests exhibited a preference for using the MCQ format. This may be attributed to the fact that MCQ items are easier to rate than other test formats like summaries, and they allow reliable scoring of large numbers of test takers in a shorter period compared with summary and constructed-response items (O'Grady, 2023). Some researchers also claimed that MCQs could provide a fairly good elicitation of students' competence that listening tests are intended to measure (Fulcher & Davidson, 2007). However, this is a subject of contention, since test takers can use practical test-taking strategies like guessing and word-matching to achieve higher scores without possessing a high level of L2 listening proficiency (Buck, 2001; Field, 2013; Kho et al., 2023; Liu & Aryadoust, 2026; Qiu & Aryadoust, 2024). In addition, Badger and Yan (2009) indicated that the use of this item type seems to lack interactional authenticity, as it can engage certain cognitive processes that are not typically used in daily life. Some tests adopted other test formats such as summary and matching, which could exert both positive and negative effects on listening performance.

It might be said that there is no one-size-fits-all format that could function well in all circumstances, and thus test developers ought to have a nuanced understanding of the construct and the features of test formats to choose more appropriate methods for measuring listening and maximize authenticity in relation to the TLU domain. From a validity perspective, the choice of item type is not merely a practical decision but a theoretical decision that has direct implications for the explanation inference in the ABV framework. If the selected item format engages cognitive processes that are not representative of real-life listening, it introduces construct-irrelevant variance into test scores, thereby threatening the construct validity of the assessment (Buck, 2001; Liu & Aryadoust, 2026). The widespread use of MCQs without

sufficient supporting evidence, therefore, represents a significant gap in the test specification practices of L2 listening researchers.

Test mode. Most researchers did not provide the details of the test-delivery mode used in the studies, which reduces the replicability of the studies and limits our understanding of test specifications. Both paper-based and computer-based modes have been applied in listening tests, although paper-based delivery was used slightly more frequently in the reviewed studies (see Table 5). This may be attributed to the fact that both researchers and students are more familiar with paper-based testing, as this delivery format has been prevalent since the early days of assessment. Meanwhile, it was found that an increasing number of researchers were interested in exploring the affordances of CBT (computer-based testing), and it is anticipated that CBT might become the dominant test mode in the future, as it is easier to administer and score, and can provide automated scoring and feedback (Chapelle & Douglas, 2006). The use of CBT can also improve authenticity, since it can simulate real-life digital environments, and can also be cost-effective (Öz & Özturan, 2018).

More future studies should investigate the validation of CBT and its different types (e.g., mobile-administered tests, laptop-based tests and importantly AI-generated tests), as it is likely to become the mainstream model of assessment in L2 listening. The shift from paper-based to computer-based delivery also has important implications for test specifications and validity. Changes in test delivery mode can affect the representation of the construct being measured, as different modes may elicit different cognitive processes from test takers (Ockey, 2013). Within the ABV framework, this bears most directly on the explanation and generalization inferences. With respect to the explanation inference, a delivery mode that engages processes extraneous to listening comprehension (e.g., navigation, screen-reading, or device-specific demands) can introduce construct-irrelevant variance, as shown by eye tracking research (e.g., Holzknecht et al., 2021; Kho et al., 2023; Liu & Aryadoust, 2026; Qiu & Aryadoust, 2024). This can weaken the warrant that score variation reflects differences in listening ability rather than the affordances of the mode itself (Kane, 2006). With respect to the generalization inference, if test scores obtained under different delivery modes are not equivalent, then comparing results across studies or generalizing findings to different testing contexts becomes questionable (Kane, 2006; Chapelle et al., 2008). Researchers should therefore document the test delivery mode as a standard component of test specifications and provide validity evidence to support score comparability across modes.

Test duration. Most researchers did not provide details of the listening test duration, a reporting gap that is more consequential than it might initially appear. Test duration is not only a logistical parameter but importantly a specification variable with direct implications for construct validity. Prolonged testing can induce fatigue and additional cognitive load, which may introduce construct-irrelevant variance into test scores by reducing test takers' performance for reasons unrelated to their listening ability (Buck, 2001; Ockey & Wagner, 2018). Many researchers controlled L2 listening assessments by creating tests lasting less than an hour, as longer tests can cause fatigue and affect test takers' concentration. Therefore, most standardized listening tests last less than an hour, such as approximately 40 minutes for IELTS, 45 minutes for TOEIC, and about 50 minutes for the TOEFL iBT.

However, the rationale for these specific durations is rarely documented, which limits researchers' ability to evaluate whether test length is appropriate for the intended construct and target population. Within the ABV framework, the absence of duration-related specifications weakens the explanation inference, for the same reason set out above in connection with response format and delivery mode: a source of variance unrelated to listening ability undermines the warrant that score differences are attributable to the construct (Chapelle et al., 2008). What distinguishes duration from those cases, however, is the temporal character of the threat. Whereas response format and delivery mode introduce extraneous demands that are relatively stable across the test, fatigue accumulates dynamically as the test unfolds, so that the very meaning of a score may shift between its early and late portions and the construct may be represented unevenly across the testing occasion. This dynamic quality makes

the omission of duration information especially difficult to diagnose after the fact, since neither the magnitude nor the onset of any fatigue effect can be reconstructed from an unreported test length. Future research should therefore investigate the relationship between test duration, cognitive load, and score validity, and researchers should document test duration as a standard component of their test specification reporting.

Accent. The dominant accent type of speakers was standard native-speaker accents (see Aryadoust, 2024, for a corpus-based test review), which is reminiscent of the dominance of native-speakerism in L2 learning (Alptekin, 2002). However, it has also been emphasized that more language accent varieties should be included in L2 listening assessments to improve the authenticity of the target language use (He & Jiang, 2020; Harding, 2012; Taylor & Geranpayeh, 2011). For instance, some international English standardized listening assessments (e.g., IELTS, TOEFL, and TOEIC) have started to adopt both inner- and outer-circle accents, which nevertheless remain limited when compared with English L1 accents. The results also showed that some researchers adopted both native and nonnative speaker accents in separate parts of one listening test (e.g., Kang et al., 2020; Major et al., 2005). However, Elliott and Wilson (2013) contended that the use of non-native speakers might impair the validity of assessments, since listeners with shared L1 backgrounds might have a higher chance of performing better (Major et al., 2002; Harding, 2012). Overall, these studies show that when listening to unfamiliar accents, students may fail to comprehend parts of the input, mostly due to difficulties in segmentation and word recognition (Buck, 2001). Therefore, advantaging a shared-L1 group over another group by using familiar accents could result in bias and unfairness as well as the emergence of sources of construct-irrelevant variance (Harding, 2012).

Within the ABV framework, accent selection occupies an unusual position in that it implicates the chain of inferences at both ends simultaneously. At one end stands the domain description inference: the range of accents represented in a test should reflect the linguistic diversity of the TLU domain (see Aryadoust, forthcoming), so that a test confined to standard native-speaker accents underrepresents a domain in which learners must increasingly comprehend inner-circle, outer-circle, and L2-accented speech (Aryadoust, 2024; Chapelle et al., 2008). At the other end stands the extrapolation inference, the warrant that test performance predicts real-world listening: a narrow accent sample may license accurate inferences about comprehension of that accent while systematically failing to extrapolate to the accent varieties learners actually encounter beyond the test.

What makes accent distinctive among the specification variables considered here is that the same selection decision can also generate construct-irrelevant variance of a differential kind, as noted above, advantaging listeners who share an L1 with the speaker, so that the threat is not merely one of incomplete domain coverage but of systematically unequal coverage across examinee subgroups, which becomes a fairness concern that cuts across the inferential chain. Future research should therefore develop principled frameworks for accent selection that are grounded in TLU domain analyses, and researchers should document their accent selection decisions as an explicit component of test specifications.

Double exposure. Most researchers (n = 33) adopted once-heard texts (e.g., Medina et al., 2020; Wallace & Lee, 2020), which might be attributed to the fact that once-hearing is more common in real-life communication and has higher authenticity (Taylor & Geranpayeh, 2011). However, with technological development, listeners are exposed to digital environments in which they can listen to or watch the input multiple times if they fail to understand the meaning (Geranpayeh & Taylor, 2008). Overall, whether to use repeated input is still an ongoing debate in the listening assessment field (e.g., Aryadoust, 2019; Holzknecht, 2019), as listening to the input multiple times has been shown to facilitate L2 listeners' comprehension regardless of proficiency levels (Sakai, 2009), while it could have varying effects on authenticity under certain circumstances (Aryadoust, 2019). Therefore, to include repetition, test developers must first determine how it would affect the validity and authenticity of listening assessments, taking into account the characteristics of the target language use (TLU) domain.

Setting. Most researchers did not describe the assessment setting in the studies. This omission is significant because test setting constitutes a contextual variable that can systematically affect test takers' performance and the validity of score interpretations. Among the studies that reported the test setting, classrooms were the most popular place for researchers to conduct assessments. Since students are familiar with classroom environments, this assessment setting may help reduce test anxiety. However, familiarity with the setting does not eliminate the potential impact of environmental conditions on performance. In addition, some researchers highlighted that they conducted listening assessments in a quiet room (e.g., Lei et al., 2020), underscoring the importance of test environments, as unfavorable environments with high levels of noise or inappropriate lighting or temperature may negatively affect L2 listening comprehension (Chen & Ou, 2021; Wong et al., 2017).

Within the ABV framework, test setting is directly relevant to the generalization inference: if assessment conditions vary significantly across studies, the extent to which observed scores can be generalized across different testing contexts becomes questionable (Kane, 2006). Researchers should therefore document test setting as a standard component of test specification reporting, and future studies should investigate the differential effects of various testing environments on listening performance and score comparability.

Media. Most L2 listening studies solely used audio input, which has been an indispensable component of listening tests. Even though many researchers argued that audiovisual input tests are more authentic than audio-only tests (e.g., Suvorov, 2015; Liu & Aryadoust, 2024), the use of audiovisual input could also introduce construct-irrelevant variance (He & Jiang, 2020), while audio-only input tests are associated with fewer confounding factors. In addition, studies indicate that listeners receiving audiovisual input outperformed listeners who received audio-only input (e.g., Suvorov, 2008; see also Liu & Aryadoust, 2024, for a meta-analysis), which suggests a test method effect. The studies that adopted audiovisual input mainly used local listening assessments and mostly investigated the effects of multimedia listening, whereas most standardized assessments used audio-only input. Likely due to the influence of standardized tests, which was previously discussed, researchers mostly prefer to use audio-only input to measure L2 listening comprehension.

Input types. Finally, a considerable number of researchers (n = 38) preferred to administer more than one input type in an L2 listening comprehension assessment (most of them adopted both monologues and dialogues). This is attributed to the fact that both monologic and dialogic discourse types are frequent in the TLU domains and are important in listening assessments. For example, academic L2 listening assessments are normally intended to measure test takers' listening ability to understand lectures and academic group work, which requires comprehension of both monologic and dialogic discourse (e.g., Wallace, 2020; Vafaee & Suzuki, 2020; Matthews & Cheng, 2015). Therefore, researchers and teachers should consider the differences between different input types and choose the type(s) that are relevant to the TLU domains. On the other hand, the difficulty of the two discourse types remains a controversial topic. Some researchers argued that monologic discourse is easier to understand (Read, 2002), while other researchers suggested that dialogues are easier (Brindley & Slatyer, 2002; Papageorgiou et al., 2012). In addition, while research shows that the internal and discourse structure of monologues and dialogues varies depending on the context and aims of communication (Goh & Burns, 2012), there is little research on the effect of input structure on listening performance. It is important to investigate the effect of these structures on listening comprehension and test performance in future research.

5.3 Research question 3

The findings related to research question 3 reveal an unbalanced distribution of validity evidence across the six ABV inferences in the 160 listening assessments examined. Moreover, no researcher provided evidence supporting all six validity inferences, and some researchers provided no evidence at all, which raises concerns about the validity of the tests used.

The unbalanced distribution of validity evidence across the six inferences warrants deeper examination beyond mere categorization. Several factors may account for researchers' failure to provide comprehensive validation evidence. First, practical constraints play a significant role: collecting validity evidence for all six inferences requires substantial resources, time, and access to test takers, raters, and TLU domain experts, which may not always be feasible, particularly in small-scale or classroom-based research contexts (Fan & Yan, 2020; Aryadoust, 2020). This is especially evident in assessments of learning and assessments for learning in classrooms, where teachers are primarily concerned with evaluating the extent to which students have achieved the learning objectives specified in the curriculum (Aryadoust, forthcoming). Although curricula generally reflect aspects of a broader TLU domain, teachers are often less concerned with extrapolating test performance to real-world language use or investigating the potential consequences of test use. Instead, their primary focus is on making accurate judgments about students' attainment of intended learning outcomes.

Second, there is an epistemological dimension to this issue: many researchers who use listening tests as instruments in their studies may not perceive test validation as their primary responsibility, viewing the tests as established measurement tools rather than objects of validation inquiry in their own right (Im et al., 2019). Third, the demands of the ABV framework itself may be a contributing factor. As Kane (2013) acknowledged, the framework is inherently flexible and does not require researchers to address all inferences in every validation study; rather, it encourages researchers to focus on the inferences most relevant to their specific interpretive and validity arguments. However, this flexibility may inadvertently permit researchers to selectively address only the most accessible inferences, such as generalization, while neglecting others that are more challenging to investigate empirically.

It is also important to recognize that validity is inherently context-dependent, and the ABV framework, while comprehensive, should not be applied as a universal checklist without consideration of the specific contextual factors that shape assessment practices (Kane, 2013; Weir, 2005). The present findings reveal a stark contrast between standardized and local assessments in terms of the validity evidence provided: researchers using standardized tests were more likely to report validity evidence than those using local or classroom-based assessments. This disparity reflects the different contextual realities that researchers face. Standardized tests, developed by large testing organizations with dedicated resources, typically come with purported pre-existing validity evidence that researchers can consider. In contrast, local and classroom-based assessments are often developed with limited resources and within tight time constraints, making it impractical to conduct comprehensive validation studies for every assessment use (Aryadoust, 2020). Furthermore, the purpose of the assessment matters: a high-stakes standardized test used for admission decisions demands a more rigorous validity argument than a classroom test used for formative purposes. Future researchers should therefore consider the contextual factors, including the stakes of the assessment, the resources available, and the intended uses of test scores, when determining the scope and depth of validation evidence required.

The evaluation and generalization inferences were examined most frequently in the studies. The provision of evidence supporting these two inferences is comparatively less challenging for researchers since the former deals with issues of scoring consistency, rater reliability, and the functioning of rating scales at the task level, and the latter with reliability, which can explain the high number of studies examining the generalization inference, particularly in standardized assessments. Researchers who adopted standardized assessments appeared to attach more importance to validation, which may explain why researchers using standardized tests are more likely to provide evidence supporting the validity of tests.

Beyond validity, reliability constitutes another important component of measurement quality in L2 listening assessment, and its relationship with validity deserves more explicit attention. Reliability refers to the consistency and stability of test scores across different testing occasions, raters, and forms (Shang et al., 2024). Within the ABV framework, reliability is most directly addressed through the generalization

inference, which concerns the extent to which observed scores can be generalized across tasks, raters, and test versions (Kane, 2006). The present findings show that the generalization inference was the most frequently examined in the dataset (n = 91, 56.88%), suggesting that researchers do attend to reliability to some extent. One possible explanation is that reliability indices are relatively straightforward to compute and report.

However, the analyses reveal that most of this evidence was limited to internal consistency measures such as Cronbach's alpha, with relatively few studies examining inter-rater reliability, test-retest reliability, or reliability of different test forms (where applicable). This is a notable gap, as listening assessment often involves subjective scoring decisions, particularly in constructed-response formats such as summaries and short answers, where rater consistency is a critical concern (Buck, 2001). Furthermore, some scholars have argued that reliability and validity are not independent properties of a test; insufficient reliability undermines the validity of score interpretations, as inconsistent scores cannot support meaningful interpretations and uses (e.g., Chapelle et al., 2008; Messick, 1989; Kane, 2006).

With regard to the explanation and extrapolation inferences, very few studies provided evidence for these inferences, and all of these studies involved standardized tests. As noted earlier, explanation is analogous to construct validity, and extrapolation is related to criterion-related validity (Aryadoust, 2013). While standardized tests are often created based on explicit construct definitions, classroom tests are often created based on a set of criteria. While it is not surprising that the explanation inference is not examined in classroom-based studies, what is surprising is the frequent use of standardized tests in studies without corresponding construct-related validation evidence. It is also concerning that there is no criterion-based evidence to support the classroom-based assessments.

No researcher provided in-depth evidence of domain description and utilization in either standardized assessments or local assessments. As the starting and end points of test development and the validation framework, the significance of these two inferences is evident (Chapelle, 2012). Im et al. (2019) indicated that domain description is "the most important aspect for test design to identify language knowledge, skills, and abilities" (p. 19), and the utilization inference involves the intended uses of test scores.

5.4 Limitations and future research

The findings of the current study can also be interpreted in a different way. Although the ABV framework sets high standards and expects researchers to provide evidence for all the ABV inferences (interpretations and uses of test scores), in reality, researchers often focus only on a subset of inferences (Fan & Yan, 2020; Aryadoust, 2020). Many language assessment studies provide evidence of reliability and psychometric validity, whereas fewer provide evidence of construct validity and fewer still explore other validity inferences (Aryadoust, 2020). The demands of the ABV framework may be difficult to meet in some contexts, which may explain why certain inferences were left unexamined.

For example, in a study on composite scores, Kane and Case (2004) evaluated score quality primarily through reliability and a narrowly defined, content-based validity coefficient tied to the weighting of the two tests, rather than through the full chain of inferences the argument-based approach prescribes. This is striking because, by this point, Kane had already articulated the mature framework, endorsing Messick's integrated definition of validity and conceptualizing validation as an argument spanning evaluation, generalization, extrapolation, explanation, and decision-making inferences (Kane, 2001). In fairness, a composite of this kind, validated against a content-defined target rather than an external criterion, approximates what Kane (2001) calls an "observable attribute," an interpretation he argues can legitimately be supported by a *modest* argument resting mainly on evaluation and generalization; on that reading, the reliance on reliability and a content-based coefficient is less a lapse than the framework operating as intended for a low-theory interpretation. Even so, Kane and Case's (2004) tendency to treat validity as a property of scores (rather than a property of interpretations and uses of scores) illustrates

how easily even the framework's principal architect reverts to older language in technical practice, and how context-dependent its standards are in application.

It is therefore worth revisiting and interrogating the demands of the ABV framework as a standard validation method. Recent studies by Aryadoust (2026) and Aryadoust and Zakaria (2026) have discussed the limitations of the ABV framework in classroom-based and small-scale assessment contexts and, more fundamentally, its emphasis on justification rather than truth-seeking. These studies argue that the version of the ABV framework that is adapted in language assessment, i.e., the assessment use argument (AUA, Bachman & Palmer, 2010), primarily focuses on building a persuasive argument to justify score interpretations and uses, rather than on uncovering the underlying truth about what a test actually measures. In this sense, validation becomes an argumentative process centered on plausibility rather than veracity, which may lead to situations where sufficient supporting evidence is accumulated to defend score interpretations even when the underlying construct representation or measurement assumptions are neither fully accurate nor complete. In short, one limitation of the ABV framework is that it prioritizes argument-based justification and coherence of evidence over the discovery of the true nature of the construct and measurement processes.

To address these limitations, future researchers can apply the socio-cognitive validity framework of Weir (2005), which conceptualizes test validity as arising from the interaction between the cognitive processes of test takers, the characteristics of test tasks and input, and the contextual conditions under which assessment occurs. Unlike the AUA framework, which focuses primarily on the justification of score interpretations and uses through a chain of inferences, Weir's framework places stronger emphasis on the internal cognitive validity of assessment tasks, providing a complementary perspective for examining whether test tasks actually elicit the intended language abilities and processes. This framework presents a clear program for examining the external and internal factors that affect an assessment and provides a more comprehensive approach to validation by explicitly considering context validity, cognitive validity, scoring validity, consequential validity, and criterion-related validity.

However, it should be noted that Weir's (2005) framework and the ABV/AUA framework are not mutually exclusive; rather, they can be applied in a complementary manner to strengthen the validity argument for L2 listening assessments from multiple theoretical perspectives. This framework can therefore be applied to place greater emphasis on test design, task characteristics, and test-taker cognitive processes, rather than focusing primarily on the justification of score interpretations and uses. A notable example of the application of Weir's (2005) framework in the context of listening assessment is presented in several chapters of the edited volume by Geranpayeh and Taylor (2013).

Another limitation of the present study is that the reviewed studies were conducted largely before the widespread use of generative artificial intelligence (GenAI) in language assessment research and practice. GenAI is becoming increasingly relevant in language testing as it offers many affordances for test development, item generation, task design, scoring, and validation processes (Aryadoust & Jia, 2026; Aryadoust & Zhai, 2026; Zheng & Wang, 2026). Recent work has shown that GenAI can be used across multiple stages of the assessment cycle, including test specification development, item and task generation, simulation of test-taker responses, automated scoring, and the collection of validity evidence (Aryadoust & Jia, 2026). From the perspective of test specifications, GenAI can assist test developers in systematically generating tasks that conform to predefined specifications and in documenting test characteristics more transparently. From a validation perspective, GenAI can also be used to generate simulated data, examine task characteristics, and support validation research through large-scale automated analyses.

Finally, since the present study focused on L2 listening assessment research conducted prior to the widespread adoption of GenAI, future research should investigate how GenAI can be integrated into L2 listening test development, specification design, and validation to improve transparency, reproducibility, and the overall quality of listening assessments (Aryadoust & Jia, 2026; Aryadoust & Wong, 2026;

Aryadoust & Zhai, 2026). GenAI-assisted test development has the potential to strengthen validity. For example, by generating tasks that conform to predefined specifications, it can support the domain description inference; and by facilitating systematic documentation of test characteristics and scoring procedures, it can enhance the scoring inference. However, the integration of GenAI also introduces new challenges, particularly in relation to construct representation and the potential for construct-irrelevant variance, which future research should address.

6 Conclusion

To our knowledge, the present study is the first systematic review to simultaneously examine test specification transparency and the distribution of validity evidence across the full ABV inferential chain in L2 listening assessment research. The distinctive feature of the study is the synthesis of L2 listening assessment research from various perspectives, thus providing an in-depth multidimensional analysis of the L2 listening field. The findings indicate a growing stream of research in L2 listening assessment in the past few years, which nevertheless reveals consistent limitations in terms of reproducibility, clarity of test specifications, and validity evidence.

This study aimed to provide a comprehensive and systematic review of L2 listening comprehension assessment by examining L2 listening assessment research published in the past 24 years indexed in WoS. Three research questions addressing the attributes of test takers, the characteristics of tests used in the reviewed studies, and the validation of tests were investigated. The ABV framework was adopted as the analytical framework for evaluating the validation of tests (Chapelle et al., 2008; Kane, 2013). The major findings are summarized as follows.

For research question one, both standardized and local listening assessments were found in studies, but standardized tests were predominantly adopted. In addition, the sample size, instructional level, L2 level, and L1 of examinees were found to be heterogeneous. The overwhelming majority of the researchers using standardized tests seemed to have a high level of confidence in the suitability of the tests for their target samples, and therefore did not take steps to examine or support the validity of the tests for their specific contexts.

Research question two also concerns the specifics of L2 listening comprehension assessment in studies. Among the assessments for which test development and method were reported, TBLA and WLP tests were the most commonly identified test development approaches and test methods, respectively; MCQs and paper-based modes were the most frequently reported item format and test mode overall. In addition, English was the most commonly assessed L2 in listening assessment; most tests were under an hour long; and the most popular setting to conduct tests was classrooms. Regarding the textual features of L2 listening assessments, audio-only input (media type), native speaker input (speakers' accent), single exposure (exposure times), and mixed monologue and dialogue inputs (material types) were frequently used. However, a significant finding was that most researchers failed to explicitly provide details about the tests they adopted, which affects the validity and replicability of tests. Overall, the data show that the majority of studies lacked sufficient information on test specifications, rendering full replication in other contexts extremely difficult.

Finally, question three explored the validation of the L2 listening tests by using the ABV framework as a criterion. An unbalanced distribution of evidence pertaining to each inference across standardized and local assessments was observed. The generalization inference was the most frequently examined (n = 91, 56.88%), with evidence including internal consistency and interrater reliability; the evaluation inference was examined in a much smaller proportion of assessments (n = 13, 8.13%); a few researchers investigated the explanation and extrapolation inferences of tests, and no study provided evidence for the domain description and utilization inferences of the tests.

The findings of this study collectively indicate that, while some researchers have attended to test validation by providing supporting evidence for select inferences in the ABV framework, the majority of

studies in the reviewed corpus did not provide sufficient validation evidence to support the interpretation and use of their listening assessment scores. This represents a gap in the field, as the absence of transparent test specifications and comprehensive validity evidence not only limits the reproducibility of individual studies but also constrains the cumulative development of knowledge in L2 listening assessment research. It is hoped that the present review will serve as a resource for researchers and practitioners in identifying the areas of test specification and validation that require greater attention, and in making more informed decisions when developing, selecting, and reporting on listening assessments in future research.

Appendix A Search Codes of the Web of Science

TOPIC: (listening) AND TOPIC: ("second language")

Refined by: TOPIC: (test OR assess OR measu*) AND DOCUMENT TYPES: (ARTICLE) AND [excluding] SOURCE TITLES: (DIAGNOSTICA OR EDUCACION XX1 OR JOURNAL OF EDUCATIONAL PSYCHOLOGY OR EDUCATIONAL TECHNOLOGY SOCIETY OR EUROPEAN PSYCHOLOGIST OR JOURNAL OF EXPERIMENTAL PSYCHOLOGY HUMAN PERCEPTION AND PERFORMANCE OR EXPERT SYSTEMS WITH APPLICATIONS OR JOURNAL OF NEUROLINGUISTICS OR IEEE SENSORS JOURNAL OR IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS OR INTERACTIVE LEARNING ENVIRONMENTS OR INTERNATIONAL JOURNAL OF AUDIOLOGY OR NEUROREPORT OR INTERNATIONAL JOURNAL OF LANGUAGE COMMUNICATION DISORDERS OR PERCEPTUAL AND MOTOR SKILLS OR INTERNATIONAL JOURNAL OF PSYCHOPHYSIOLOGY OR INTERNATIONAL JOURNAL OF SPEECH LANGUAGE AND THE LAW OR SCANDINAVIAN JOURNAL OF PSYCHOLOGY OR JAPANESE JOURNAL OF EDUCATIONAL PSYCHOLOGY OR AGING CLINICAL AND EXPERIMENTAL RESEARCH OR JOURNAL OF CHILD LANGUAGE OR JOURNAL OF COGNITIVE PSYCHOLOGY OR ANNALS OF DYSLEXIA OR APPLIED ACOUSTICS OR APPLIED ECONOMICS OR ASSESSING YOUNG LEARNERS OF ENGLISH GLOBAL AND LOCAL PERSPECTIVES OR JOURNAL OF NURSING EDUCATION OR AUGMENTATIVE AND ALTERNATIVE COMMUNICATION OR JOURNAL OF PSYCHOPHYSIOLOGY OR JOURNAL OF SPEECH LANGUAGE AND HEARING RESEARCH OR BRAIN IMAGING AND BEHAVIOR OR JOURNAL OF UNIVERSAL COMPUTER SCIENCE OR JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA OR BRITISH JOURNAL OF EDUCATIONAL TECHNOLOGY OR JOURNAL OF THE AMERICAN ACADEMY OF AUDIOLOGY OR CEREBRAL CORTEX OR MULTI USER VIRTUAL ENVIRONMENTS FOR THE CLASSROOM PRACTICAL APPROACHES TO TEACHING IN VIRTUAL WORLDS OR NATURAL LANGUAGE ENGINEERING OR CLINICAL LINGUISTICS PHONETICS OR NEURAL PLASTICITY OR CONTEMPORARY EDUCATIONAL PSYCHOLOGY OR NEUROIMAGE OR BRAIN RESEARCH OR CRITICAL REFLECTIONS ON DATA IN SECOND LANGUAGE ACQUISITION OR NEUROPSYCHOLOGIA OR DEVELOPMENTAL SCIENCE OR NEUROSCIENCE OR EAR AND HEARING)

Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH.

Appendix B The results pertinent to the seven factors of assessments

For the variable of target language, 12 languages were examined in the dataset. Notably, English was the most frequently assessed second language among all tests (n = 107, 66.88%). This was followed by Spanish (n = 20, 12.50%), French (n = 10, 6.25%), German (n = 8, 5.00%), Chinese (n = 4, 2.50%), and

Japanese (n = 3, 1.88%). Dutch and Arabic were each investigated in two studies, while Korean, Russian, Croatian, and Norwegian were tested in one study each.

The next variable is test mode, which was divided into four categories: paper-based assessment, computer-based assessment, oral-based assessment, and not stated. Some listening tests were conducted using face-to-face oral tests, such as interviews, which belong to neither paper-based assessment nor computer-based assessment; hence, these tests are grouped into other modes (oral-based assessment). Nearly half of researchers did not provide the mode of the listening tests used in their studies (n = 74, 46.25%), although more than half of the tests were identified as standardized assessments (n = 47, 29.38%). Among the studies that did report these details, researchers preferred paper-based assessments and computer-based assessments which accounted for 25.00% and 23.13%, respectively, of all listening assessments. Only nine studies used oral-based assessments.

Test duration is the sixth variable. A notable result was that the test duration of more than half of assessments (n = 102, 63.75%) was not stated by the researchers, of which standardized assessments accounted for 36.25% of studies and local assessments accounted for 27.50%. The duration of 17.50% of tests lasted from 30 minutes to one hour, while 16.25% of tests finished in 30 minutes or less. Only four standardized listening tests lasted for more than one hour. Therefore, a test duration less than one hour was preferred by most researchers.

The last variable is setting, which refers to the examination environment. Of note, the examination setting was not reported in over half of the studies (n = 109, 68.13%). Most researchers chose to conduct tests in a classroom (n = 21, 13.13%) or a laboratory (n = 16, 10.00%), while some researchers conducted listening tests in a quiet room (n = 10, 6.25%). Due to the needs of the examinations, some tests were conducted in a place with computers (n = 4, 2.50%).

Table B1 provides details of the textual features of the listening comprehension assessments. Four variables were identified, including the media type of the listening material, the accent of the speakers, the number of text plays, and the material input type.

The first variable is the media type. Two media types, audio-only input and audiovisual input, were used in the assessments. The audio-only media type (n = 75) was the most popular input type in listening assessments, accounting for 46.88% of all tests. Audiovisual input (n = 18, 11.25%) was used less frequently than audio-only input, of which most (9.38%) tests were local assessments. The media type of nearly half of listening assessments (n = 67, 41.88%) was not clearly stated by the authors of the studies, although most of the tests were identified as standardized assessments (n = 53, 33.13%) and only 8.75% of tests were local assessments.

For the second variable of accent, four groups were identified: native speaker, nonnative speaker, mixed accents, and not reported. Notably, the accent of the input material of most tests (n = 126, 78.75%) was unknown, of which more than half (n = 87, 54.38%) were standardized assessments. Of the studies that specified the accent, 16.88% of listening tests (n = 27) used recordings of native speakers, among which 16 (10.00%) were local assessments, and the recordings of 2.50% of tests were adopted from nonnative speakers. The audio of three tests was provided by native and nonnative speakers, accounting for 1.88% of listening tests.

The next variable is the number of text plays. Like for the previous variable, most studies (70.00%) did not report this information, of which more than half were standardized assessments (n = 80, 50.00%). The listening material of 33 tests (20.63%) was played once during the test, of which 11.25% were local assessments. Occupying the next position, 6.25% of assessments adopted double play. Three (1.88%) listening assessments used a mixed play mode within one test, that is, part of the materials was played once while the other materials were played twice. Two listening tests (1.25%) were conducted with the materials played more than two times.

The last variable is material input type, which was categorized into five groups: monologue, dialogue, sentence, mixed input types, and not stated. The material input type of 50 assessments (31.25%)

was unclear, of which 25.00% were standardized assessments. The use of mixed input types (n = 38) was the most popular choice, accounting for 23.75% of assessments. Thirty listening comprehension assessments used monologue as the input type, accounting for 18.75% of all tests. Sentences and dialogues stood in third and fourth place, accounting for 13.75% and 12.50% of assessments, respectively.

Table B1

Textual Features of Listening Comprehension Assessments

Test information	Standardized assessment (n = 103)		Local assessment (n = 57)		Total (n = 160)	
	# of tests	%	# of tests	%	# of tests	%
Media type						
Audio-only	47	29.38	28	17.50	75	46.88
Audiovisual	3	1.88	15	9.38	18	11.25
N/A	53	33.13	14	8.75	67	41.88
Accent						
Native speaker	11	6.88	16	10.00	27	16.88
Nonnative speaker	3	1.88	1	0.63	4	2.50
Mixed	2	1.25	1	0.63	3	1.88
N/A	87	54.38	39	24.38	126	78.75
Number of text plays						
Once	15	9.38	18	11.25	33	20.63
Twice	7	4.38	3	1.88	10	6.25
More than twice	0	0.00	2	1.25	2	1.25
Mixed	1	0.63	2	1.25	3	1.88
N/A	80	50.00	32	20.00	112	70.00
Material input type						
Monologue	15	9.38	15	9.38	30	18.75
Dialogue	11	6.88	9	5.63	20	12.50
Sentence	13	8.13	9	5.63	22	13.75
Mixed	24	15.00	14	8.75	38	23.75
N/A	40	25.00	10	6.25	50	31.25

Appendix C The techniques and methods used to investigate the validity inferences of L2 listening tests

Table C1

Analysis Methods for Evaluation Inference (n = 13)

Methods	Standardized assessment		Local assessment		Total	
	# of tests	%	# of tests	%	# of tests	%
Rasch	5	38.46	2	15.38	7	53.85
ANCOVA	3	23.08	0	0.00	3	23.08
Item analysis	1	7.69	0	0.00	1	7.69
Descriptive statistics	1	7.69	0	0.00	1	7.69
Pearson	1	7.69	0	0.00	1	7.69

Table C2

The Two Types of Generalization Inference (n = 91)

Generalization	Standardized assessment		Local assessment	
	# of tests	%	# of tests	%
Internal reliability (n=81)	53	65.43	28	34.57
Interrater reliability (n=16)	10	62.50	6	37.50

Note. The total exceeds 91 because some assessments provided evidence for both internal reliability and interrater reliability.

Table C3

Analysis Methods for the Generalization Inference of Internal Reliability of Listening Comprehension Tests (n = 81)

Methods	Standardized assessment		Local assessment		Total	
	# of tests	%	# of tests	%	# of tests	%
Cronbach's alpha	39	48.15	25	30.86	64	79.01
Spearman-Brown split-half	6	7.41	0	0.00	6	7.41
Pearson	3	3.70	1	1.23	4	4.94
Rasch	2	2.47	0	0.00	2	2.47
Cohen's kappa	1	1.23	1	1.23	2	2.47
Unclear	2	2.47	1	1.23	3	3.70

Table C4

Analysis Methods for the Generalization Inference of Interrater Reliability (n = 16)

Methods	Standardized assessment		Local assessment		Total	
	# of tests	%	# of tests	%	# of tests	%
Descriptive statistics	3	18.75	1	6.25	4	25.00
Cohen's kappa	2	12.50	2	12.50	4	25.00
Pearson	2	12.50	1	6.25	3	18.75
Cronbach's alpha	1	6.25	1	6.25	2	12.50
Spearman rank-order	1	6.25	0	0.00	1	6.25
Hoyt reliability measure	1	6.25	0	0.00	1	6.25
Unclear	0	0.00	1	6.25	1	6.25

References

Alderson, C. (2000). *Assessing reading.* Cambridge University Press.

Alptekin, C. (2002). Towards intercultural communicative competence in ELT. *ELT Journal*, *56*(1), 57–64. https://doi.org/10.1093/elt/56.1.57

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the International English Language Testing System (IELTS) listening module. *International Journal of Listening, 26*(1), 40–60. https://doi.org/10.1080/10904018.2012.639649

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Cambridge Scholars Publishing.

Aryadoust, V. (2019). An integrated cognitive theory of comprehension. *International Journal of Listening, 33*(2), 71–100. https://doi.org/10.1080/10904018.2017.1397519

Aryadoust, V. (2020). A review of comprehension subskills: A scientometrics perspective. *System*, *88*, Article 102180. https://doi.org/10.1016/j.system.2019.102180

Aryadoust, V. (2024). Topic and accent coverage in a commercialized L2 listening test: Implications for test-takers' identity. *Applied Linguistics, 45*(5), 765–785. https://doi.org/10.1093/applin/amad062

Aryadoust, V. (2026). Justifying the score or informing the stakeholder? Transparency challenges in language testing. *Language Testing*. Advance online publication. https://doi.org/10.1177/02655322261444781

Aryadoust, V. (forthcoming). *Assessing listening in the age of generative artificial intelligence.* Routledge.

Aryadoust, V., Foo, S., & Ng, L. Y. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Language Testing*, *39*(1), 56–89. https://doi.org/10.1177/02655322211026876

Aryadoust, V., & Jia, Y. (2026). Generative artificial intelligence in listening assessment. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave encyclopaedia of computer-assisted language learning* (pp. 1–10). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-51447-0_202-1

Aryadoust, V., & Luo, L. (2023). The typology of second language listening constructs: A systematic review. *Language Testing, 40*(2), 375–409. https://doi.org/10.1177/02655322221126604

Aryadoust, V., & Wong, J. (2026). How to train your dragon: Evaluating prompting and fine-tuning for GPT-based item generation in L2 listening assessment. Computers and Education: *Artificial Intelligence*, Article 100623. Advance online publication. https://doi.org/10.1016/j.caeai.2026.100623

Aryadoust, V., & Zakaria, A. (2026). A critical assessment of the argument-based validity framework. In H. K. Tan, H. Y. Tay, K. L. Chue, & A. Yeo (Eds.), *Effective assessment in schools: Beyond basics and best practices* (Vol. 2). Routledge. Advance online publication.

Aryadoust, V., & Zhai, J. (2026). Researching Listening: Mixed methods research. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (2nd ed., pp. 1–10). Wiley. https://doi.org/10.1002/9781405198431.wbeal20016

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

Badger, R., & Yan, X. (2009). The use of tactics and strategies by Chinese students in the Listening component of IELTS. In P. Thompson (Ed.), *International English Language Testing System (IELTS) research reports* (Vol. 9, pp. 67–98). British Council and IELTS Australia.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). TOEFL 2000 listening framework. *Educational Testing Service*. https://www.ets.org/Media/Research/pdf/RM-00-07-Bejar.pdf

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science, with applications in R*. Cambridge University Press.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369–394. https://doi.org/10.1191/0265532202lt236oa

Brunfaut, T. (2016). Assessing listening. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 97–112). De Gruyter Mouton.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19–27. https://doi.org/10.1177/0265532211417211

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

Chen, Q., & Ou, D. (2021). The effects of classroom reverberation time and traffic noise on English listening comprehension of Chinese university students. *Applied Acoustics, 179,* Article 108082. https://doi.org/10.1016/j.apacoust.2021.108082

Cheung, S. W. L. (2023). Exploring the impact of lexical threshold on listening comprehension. *International Journal of TESOL Studies*, *5*(4), 37–54. https://doi.org/10.58304/ijts.20230403

Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Elliott, M., & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 152–241). Cambridge University Press.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, *11*, Article 330. https://doi.org/10.3389/fpsyg.2020.00330

Field, J. E. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge University Press.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment.* Routledge.

Fujita, R. (2022). The role of speech-in-noise in Japanese EFL learners' listening comprehension process and their use of contextual information. *International Journal of Listening, 36*(2), 118–137. https://doi.org/10.1080/10904018.2021.1963252

Geranpayeh, A., & Taylor, L. B. (Eds.). (2013). *Examining listening: Research and practice in assessing second language listening*. Cambridge University Press.

Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. *Cambridge ESOL: Research Notes*, *32*, 2–5. https://www.cambridgeenglish.org/Images/23151-research-notes-32.pdf

Goh, C. C. M., & Burns, A. (2012). *Teaching speaking: A holistic approach.* Cambridge University Press.

Goh, C. C. M., & Vandergrift, L. (2022). *Teaching and learning second language listening: Metacognition in action* (2nd ed.). Routledge. https://doi.org/10.4324/9780429287749

Halone, K. K., Cunconan, T. M., Coakley, C. G., & Wolvin, A. D. (1998). Toward the establishment of general dimensions underlying the listening process. *International Journal of Listening*, *12*(1), 12–28. https://doi.org/10.1080/10904018.1998.10499016

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*(2), 163–180. https://doi.org/10.1177/0265532211421161

Hartshorn, K. J., & Stephens, C. (2023). The effects of transcript use on advanced ESL listening comprehension. *International Journal of TESOL Studies, 5*(4), 55–72. https://doi.org/10.58304/ijts.20230404

He, L., & Jiang, Z. (2020). Assessing second language listening over the past twenty years: A review within the socio-cognitive framework. Frontiers in Psychology, 11, Article 2123. https://doi.org/10.3389/fpsyg.2020.02123

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2021). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing, 38*(1), 41–61. https://doi.org/10.1177/0265532220917316

Holzknecht, F. (2019). *Double play in listening assessment* [Doctoral dissertation, Lancaster University]. ProQuest Dissertations and Theses Global.

Im, G. H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia*, *9*(1), 1–26. https://doi.org/10.1186/s40468-019-0089-4

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi.org/10.1177/0265532208101006

In'nami, Y., Koizumi, R., Jeon, E. H., & Arai, Y. (2022). L2 listening and its correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 235–284). John Benjamins.

Jung, E. H. (2006). Misunderstanding of academic monologues by nonnative speakers of English. *Journal of Pragmatics, 38*(11), 1928–1942. https://doi.org/10.1016/j.pragma.2005.05.001

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342. https://www.jstor.org/stable/1435453

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31–41. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135–170. https://doi.org/10.1207/s15366359mea0203_1

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger Series on Higher Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*(3), 221–240. https://doi.org/10.1207/s15324818ame1703_1

Kang, O., Moran, M., Ahn, H., & Park, S. (2020). Proficiency as a mediating variable of intelligibility for different varieties of accents. *Studies in Second Language Acquisition, 42*(2), 471–487. https://doi.org/10.1017/S0272263119000536

Karalık, T., & Merç, A. (2019). Correlates of listening comprehension in L1 and L2: A meta-analysis. *Eurasian Journal of Applied Linguistics, 5*(3), 353–383. https://doi.org/10.32601/ejal.651387

Kho, S. Q. E., Aryadoust, V., & Foo, S. (2023). An eye-tracking investigation of the keyword-matching strategy in listening assessment. *Education and Information Technologies, 28*(4), 3739–3763. https://doi.org/10.1007/s10639-022-11322-y

Koyama, D., Sun, A., & Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning & Technology*, *20*(1), 148–165. https://doi.org/10.64152/10125/44450

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Leeser, M. J. (2004). The effects of topic familiarity, mode, and pausing on second language learners' comprehension and focus on form. *Studies in Second Language Acquisition*, *26*(4), 587–615. https://doi.org/10.1017/S0272263104040033

Lei, M., Miyoshi, T., Dan, I., & Sato, H. (2020). Using a data-driven approach to estimate second-language proficiency from brain activation: A functional near-infrared spectroscopy study. *Frontiers in Neuroscience*, *14,* Article 694. https://doi.org/10.3389/fnins.2020.00694

Liu, T., & Aryadoust, V. (2024). Does modality matter? A meta-analysis of the effect of video input in L2 listening assessment. *System*, *120*, Article 103191. https://doi.org/10.1016/j.system.2023.103191

Liu, T., & Aryadoust, V. (2026). An eye-tracking study of the impact of item presentation and format on cognitive processing in L2 listening assessment. *Studies in Second Language Acquisition, 1–28.* Advance online publication. https://doi.org/10.1017/S0272263126101648

Liu, T., Aryadoust, V., & Foo, S. (2022). Examining the factor structure and its replicability across multiple listening test forms: Validity evidence for the Michigan English Test. *Language Testing, 39*(1), 142–171. https://doi.org/10.1177/02655322211018139

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173–190. https://doi.org/10.2307/3588329

Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language Learning, 55*(1), 37–69. https://doi.org/10.1111/j.0023-8333.2005.00289.x

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System, 52,* 1–13. https://doi.org/10.1016/j.system.2015.04.015

Mayo-Wilson, E., & Grant, S. (2019). Transparent reporting: Registrations, protocols, and final reports. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 471–483). Russell Sage Foundation.

Medina, A., Socarrás, G., & Krishnamurti, S. (2020). L2 Spanish listening comprehension: The role of speech rate, utterance length, and L2 oral proficiency. *The Modern Language Journal, 104*(2), 439–456. https://doi.org/10.1111/modl.12639

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

O'Grady, S. (2023). Adapting multiple-choice comprehension question formats in a test of second language listening comprehension. *Language Teaching Research, 27*(6), 1431–1455. https://doi.org/10.1177/1362168820985367

Ockey, G. J. (2013). Assessment of listening. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 212–218). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0048

Ockey, G. J., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity*. John Benjamins.

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics, 37*(5), 693–715. https://doi.org/10.1093/applin/amu060

Owen, N. (2018). Test specifications. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–6). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0354

Öz, H., & Özturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language & Linguistics Studies*, *14*(1), 67–85. https://dergipark.org.tr/en/pub/jlls/issue/43213/527697

Panahi, A. (2012). Binding task-based language teaching and task-based language testing: A survey into EFL teachers and learners' views of task-based approach. *English Language Teaching*, *5*(2), 148–159. http://doi.org/10.5539/elt.v5n2p148

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing, 27*(2), 261–282. http://doi.org/10.1177/0265532209349472

Papageorgiou, S., Stevens, R., & Goodwin, S. (2012). The relative difficulty of dialogic and monologic input in a second-language listening comprehension test. *Language Assessment Quarterly, 9*(4), 375–397. https://doi.org/10.1080/15434303.2012.721425

Perez, M. M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720–739. https://doi.org/10.1016/j.system.2013.07.013

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*(2), 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Qiu, Y., & Aryadoust, V. (2024). The predictive value of gaze behavior and mouse-clicking in testing listening proficiency: A sensor technology study. *System, 126*, Article 103440. https://doi.org/10.1016/j.system.2024.103440

Raymond, M. R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 181–223). Lawrence Erlbaum Associates Publishers.

Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes, 1*(2), 105–119. https://doi.org/10.1016/S1475-1585(02)00018-8

Rost, M. (2016). *Teaching and researching listening (3rd ed.)*. Routledge.

Sakai, H. (2009). Effect of repetition of exposure and proficiency level in L2 listening tests. *TESOL Quarterly, 43*(2), 360–372. https://doi.org/10.1002/j.1545-7249.2009.tb00179.x

Shang, Y., Aryadoust, V., & Hou, Z. (2024). A meta-analysis of the reliability of second language listening tests (1991–2022). *Brain Sciences, 14*(8), 746. https://doi.org/10.3390/brainsci14080746

Shin, S. Y., Lee, S., & Lidster, R. (2021). Examining the effects of different English speech varieties on an L2 academic listening comprehension test at the item level. *Language Testing*, *38*(4), 580–601. https://doi.org/10.1177/0265532220985432

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. Language Testing, 25(4), 435–464. https://doi.org/10.1177/0265532208094272

Song, M.-Y. (2012). Note-taking quality and performance on an L2 academic listening test. *Language Testing*, *29*(1), 67–89. https://doi.org/10.1177/0265532211415379

Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching, 42*(1), 1–40. https://doi.org/10.1017/S0261444808005399

Suvorov, R. S. (2008). *Context visuals in L2 listening tests: The effectiveness of photographs and video vs. audio-only format*. Iowa State University.

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. *Language Testing*, *32*(4), 463–483. https://doi.org/10.1177/0265532214562099

Suvorov, R., & He, S. (2022). Visuals in the assessment and testing of second language listening: A methodological synthesis. *International Journal of Listening, 36*(2), 80–99. https://doi.org/10.1080/10904018.2021.1941028

Taguchi, N. (2011). The effect of L2 proficiency and study-abroad experience on pragmatic comprehension. *Language Learning*, *61*(3), 904–939. https://doi.org/10.1111/j.1467-9922.2011.00633.x

Tao, X., & Aryadoust, V. (2024). A multidimensional analysis of a high-stakes English listening test: A corpus-based approach. *Education Sciences, 14*(2), 137. https://doi.org/10.3390/educsci14020137

Taylor, L. (2013). *Testing reading through summary: Investigating summary completion tasks for assessing reading comprehension ability.* Cambridge University Press.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes, 10*(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*(2), 363–387. https://doi.org/10.1017/S0272263118000311

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Tran, D. (2020). *Argument-based validation of a high-stakes listening test in Vietnam* [Doctoral dissertation, Victoria University of Wellington]. Research Archive. https://doi.org/10.26686/wgtn.17151692

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition, 42*(2), 383–410. https://doi.org/10.1017/S0272263119000676

Van Blerkom, M. L. (2017). *Measurement and statistics for teachers*. Routledge. https://doi.org/10.4324/9781315464770

Wagner, E. (2014). Using unscripted spoken texts in the teaching of second language listening. *TESOL Journal, 5*(2), 288–311. https://doi.org/10.1002/tesj.120

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning, 72, 5–44*. https://doi.org/10.1111/lang.12424

Wallace, M. P., & Lee, K. (2020). Examining second language listening, vocabulary, and executive functioning. *Frontiers in Psychology, 11,* Article 1122. https://doi.org/10.3389/fpsyg.2020.01122

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson test of English academic: Building an assessment use argument. *Language Testing*, *29*(4), 603–619. https://doi.org/10.1177/0265532212448619

Weir, C. J. (2005). Language testing and validation. Palgrave Macmillan.

Wong, S. W., Tsui, J. K., Chow, B. W. Y., Leung, V. W., Mok, P., & Chung, K. K. H. (2017). Perception of native English reduced forms in adverse environments by Chinese undergraduate students. *Journal of Psycholinguistic Research, 46*(5), 1149–1165. https://doi.org/10.1007/s10936-017-9486-y

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 2316–2335). Springer.

Yanagawa, K., & 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*(1), 107–122. https://doi.org/10.1016/j.system.2007.12.003

Yi, Y.-S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing*, *34*(3), 337–355. https://doi.org/10.1177/0265532216646141

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research, 26*(4), 696–725. https://doi.org/10.1177/1362168820913998

Zhao, Y., & Aryadoust, V. (2025). An automatized semantic analysis of two large-scale listening tests: A corpus-based study. *Language Testing, 42*(3), 312–343. https://doi.org/10.1177/02655322241288598

Zheng, C., & Wang, Y. (2026). Generative AI in EFL contexts: Q-methodological analysis of cognitive dissonance patterns and self-regulation strategies among Chinese tertiary learners. *Acta Psychologica, 264*, Article 106531. https://doi.org/10.1016/j.actpsy.2026.106531

Xuanye Miao received her MA in applied linguistics from the National Institute of Education of Nanyang Technological University. She currently works as an independent researcher and English teacher. Her research interests include systematic review, language assessment, and generative AI in secondary school English teaching.

Vahid Aryadoust is an Associate Professor of language assessment at the National Institute of Education, Nanyang Technological University. His research focuses on generative AI in language assessment, computational and quantitative methods, and the application of sensor technologies such as eye tracking and neuroimaging in language assessment. He has published widely in leading journals and has received several awards for his research and teaching.

Yanyan Huang is a Doctor of Education (EdD) candidate in English Language Education at the National Institute of Education, Nanyang Technological University, Singapore. Her research interests include young learner language assessment, listening construct validation, metacognitive instruction, extensive listening, and second language acquisition in educational contexts.