Linguistic and Philosophical Investigations
ISSN: 1841-2394, e-ISSN: 2471-0881
Vol 25 (1) 2026
pp. 177–199

Eye Tracking in L2 Listening Assessment: A Systematic Review of Tasks, Cognitive Processes, and Cognitive Validity Evidence

[1] Zoulikha DJOUADI *, [2] Khadidja Samira ZITOUNI

[1] University Centre of Barika, Algeria, SEPRADIS Laboratory
zoulikha.djouadi@cu-barika.dz
[2] University Centre of Barika, Algeria, Sodidlem Laboratory
khadidja.zitouni@cu-barika.dz

ABSTRACT: Eye tracking (ET) has become an increasingly important methodological tool for investigating how learners process listening tasks in second language (L2) assessment contexts. Despite a growing body of empirical work, research in this area remains fragmented, limiting evaluation of the cognitive validity of L2 listening assessments and principled test design. Accordingly, the current systematic review seeks to combine eye tracking research to outline task characteristics, identify the investigated cognitive processes, and evaluate the relationships between eye movements and cognitive validity claims. Guided by Weir's socio-cognitive framework, this study systematically reviews 22 empirical studies employing eye tracking in L2 listening assessment. The review shows that most studies used academically oriented listening tasks drawn from high-stakes standardized tests (e.g., IELTS, TOEIC, Aptis, TOEFL), predominantly employing multiple-choice formats in laboratory-based, computer-delivered contexts. Eye-tracking metrics such as fixation counts, dwell time, proportions of looking time, and scanpaths were used to operationalize cognitive processes including visual attention allocation, bottom-up decoding, lexical access, inferencing, and visual–auditory integration. The synthesis suggests that eye tracking provides response-process evidence supporting cognitive validity by verifying construct-relevant processing, identifying construct-irrelevant variance, and revealing test method effects not visible in test scores. This review provides a clear overview of the field's current state, such as research focus and the use of eye movements as indicators in supporting cognitive validity in L2 listening assessment.

Keywords: eye tracking; L2 listening; cognitive processes; cognitive validity; systematic review

Received: 30 11 2025 **Revised:** 13 03 2026 **Accepted:** 02 05 2026

Introduction

Listening has often been regarded as the least explored among the four skills because of its transient and complex nature (Brown & Abeywickrama ,2019; Vandergrift ,2011). Traditional product-oriented approaches rely on comprehension questions to gauge listening, yet they often disregard the cognitive processes underlying L2 listening. As a result, such approaches provide limited insight into how learners actually comprehend and interact with spoken input. More recently, there has been a shift toward process-oriented approaches that seek to reveal the cognitive processes involved in listening comprehension, particularly in instructed second language acquisition contexts where audio-visual input serves as a primary source of linguistic exposure (Nguyen & Abbott, 2017; Suvorov, 2022).

Among these approaches, eye-tracking (ET) has emerged as a valuable tool to investigate learners' visual attention allocation. It allows researchers to examine how learners process visual or audiovisual material (Winke et al. 2013; Tóthová & Rusek 2025). Eye-tracking data derived from gaze behaviors provide insights into how learners interact with test materials, such as questions, answer options, and visual aids, while interpreting spoken language. These gaze patterns have been used to clarify higher-order cognitive

mechanisms, including comprehension monitoring, inference-making, and test-taking strategies (Suvorov, 2015a; Winke & Lim ,2014; Zhai & Aryadoust, 2022). During the past two decades, ET has gained momentum in second language (L2) research and has yielded important insights into cognitive processes underpinning reading, speaking, and listening (Conklin & Pellicer-Sánchez, 2016).

Despite the growth of the research body in second language acquisition, eye tracking studies in L2 contexts have largely been reviewed at a broad level. They studied eye tracking from a learning- and processing-oriented perspective, rather than from an assessment and validation perspective. These reviews typically synthesize ET studies that investigate cognitive mechanisms underlying reading, speaking, or listening development, without explicitly situating ET as a source of validity evidence for language assessment.

To date, reviews of ET research in L2 contexts have appeared mainly in three forms: narrative (Abdel Latif ,2019; Godfroid 2020; Godfroid & Hui ,2025), systematic (Hu & Aryadoust 2024; Godfroid et al., 2025), or scientometric forms (Aryadoust & Ang, 2019). However, they have largely overlooked L2 listening assessment. To our knowledge, no systematic review has examined the application of eye tracking in this subfield. This gap leaves unresolved questions about the cognitive processes examined, the eye-tracking measures employed, and the methodological standards adopted across studies.

A lack of synthesis in this area is problematic. Without it, neither researchers nor test designers can critically evaluate how, when, and under what conditions eye-tracking (ET) data meaningfully contribute to the cognitive validity of listening tests. It also hinders efforts to establish systematic links between eye-tracking metrics and the listening processes they are intended to capture, particularly across different task types and assessment formats. As a result, test developers face limited guidance on how ET evidence should inform task design and validation. Meanwhile, validation arguments remain vulnerable to construct underrepresentation by failing to capture key listening processes or construct-irrelevant variance, when eye-movement patterns reflect task artifacts rather than listening ability.

To address this gap, the present study conducts a systematic review of eye-tracking research in L2 listening assessment. The review follows explicit inclusion criteria and transparent procedures to ensure replicability and principled synthesis across heterogeneous studies. Specifically, it has three key objectives: to pinpoint the cognitive processes examined and the eye-tracking measures employed; to investigate how eye-tracking data are used to demonstrate cognitive validity by confirming that test-takers engage in relevant processes and by detecting sources of irrelevant influence; to examine task features and research methods across studies in order to support sound test design and validation. In addition to these goals, the review advances the field by consolidating existing research and provide implications for future validation and assessment work.

Based on socio-cognitive validity framework proposed by Weir which argue that cognitive validity stands as evidence that test tasks engage the mental processes they are intended to measure, this review elucidates how eye-tracking evidence can be linked to task demands and cognitive processing. Cognitive validity is prioritized because eye-tracking provides direct response-process evidence, making it uniquely suited to evaluating whether listening assessments elicit construct-relevant cognition rather than task-related to cognitive distortions. Therefore, this review aims to clarify the role of ET in strengthening the cognitive validity of L2 listening assessment. In doing so, it supports more informed interpretation of eye-tracking data in listening research and charts new directions for future studies.

Literature review

1.The Nature of L2 listening: Cognitive foundations

Second language (L2) listening can be understood as a cognitive process that occurs in real time. Listeners construct meaning from disappearing spoken input under limited time (Goh, 2000). L2 listening comprehension is a cognitively demanding activity because it unfolds in real time and listeners must interpret the speech before it fades due to time constraints (Imhof, 2010). Unlike reading, spoken language unfolds in real time and cannot be readily revisited. As result, listeners unlike readers must continuously process, interpret, and integrate information as it becomes available (Field, 2008; Buck ,2001; Vandergrift & Goh, 2012). At the initial stage, comprehension depends on perceptual and attentional processes. These processes make listeners orient to the speech signal, discriminate relevant acoustic features, and segment

continuous input into meaningful units (Mattys et al., 2012). It is important to note that these processes are strongly affected by external factors such as background noise, unfamiliar accents, listeners themselves and divided attention. Higher level comprehension can be influenced by disruptions that occur at early level (Pratiwi, 2019; Van Engen & McLaughlin, 2018).

From a cognitive perspective, listeners engage in bottom-up processing which involves decoding sounds into words and sentences. Listeners through bottom-up processes decode the acoustic signal by segmenting the speech stream. They also identify phonological patterns, accessing lexical representations, and parsing syntactic structures. L2 listeners face many challenges such as, limited vocabulary knowledge or slower decoding speed. Such issues are sources for increased cognitive load, leaving fewer mental resources available for building meaning (Scharenborg & van Os, 2019). In parallel, top-down processing depends on context and cues related to prior knowledge to interpret the message (Richards, 2008). Effective listening depends on the dynamic interaction between these processes rather than on either mechanism in isolation.

Beyond decoding, listeners rely on higher-order cognitive processes. As a result, they can construct coherent discourse representations. These processes include inferencing, integrating incoming information with prior knowledge, and maintaining global coherence across extended spoken texts such as lectures or narratives. At the discourse level, listeners engage in multiple processes such as monitoring topic development, interpreting structural cues and relating local details to the overall meaning (Flowerdew & Miller, 2010). Metacognitive regulation reinforces comprehension. During metacognitive processes, listeners perform monitoring, evaluation and compensatory strategies. To cope with mental overload, listeners redirect attention, focus on key information, or tolerate ambiguity (Goh, 2018; In'nami & Koizumi, 2021).

In recent years, multimodal and integrated tasks become dominant. In these tasks, auditory input is accompanied by visual information such as facial expressions, gestures, or written prompts. Visual cues can facilitate comprehension if they are relevant. However, irrelevant cues may increase cognitive demands. Together, perceptual processing, decoding, higher-level integration, metacognitive regulation, and multimodal coordination constitute the cognitive architecture of L2 listening comprehension and provide a principled framework for examining how real-time measures capture listening processes as they unfold (Plakans & Park, 2024).

Working memory and attention are central to this interaction. These mechanisms assist listeners to maintain, manage, and interpret spoken input in real time. Automaticity is limited in a second language and requires simultaneous decoding, interpretation, and monitoring. Consequently, working memory resources are rapidly taxed (Field, 2008). Predictive processing further facilitates listening. The use of linguistic and situational cues allows listeners to anticipate upcoming content. The anticipation reduces processing load and strengthens comprehension continuity (Brothers et al., 2019).

Anderson and Lynch (1988) argue that the same characteristics that define listening namely its transience, incrementality, and internal nature also render its cognitive processes difficult to observe directly. Momentary lapses in attention or failures in early perceptual processing can have cascading effects on comprehension, yet these processes leave no visible trace in final performance outcomes. As a result, traditional product-based measures such as test scores or retrospective self-reports cannot provide direct and complete evidence of how listening unfolds in real time (Wagner, 2013).

Recent reviews within the socio-cognitive framework demonstrate that L2 listening involves dynamic, overlapping cognitive processes that cannot be fully captured through product-based measures alone, thereby motivating the adoption of process-sensitive and process-tracing methodologies (He & Jiang, 2020). The cognitive perspective is essential for this review because it relates eye tracking metrics to mental processes the test takers engaged in during taking listening test.

2. Listening assessment in second language contexts

Approaches of listening assessment in second language (L2) contexts has been influenced by several distinct paradigms. Each paradigm is based on contrasting views of listening ability definition and measurement (Buck, 2001). Early assessment practices used a discrete-point paradigm. listening assessment relied on isolated test elements, such as phonemes, vocabulary, and grammatical forms (Lado, 1961). In contrast, integrative approaches attempted to assess listening as a holistic skill. They aim to evaluate test takers on processing long input and complete tasks such as cloze or information transfer. Thus, engaging multiple language components simultaneously (Oller, 1979). More recently, communicative paradigms have foregrounded authenticity and real-world relevance. listening assessment is therefore seen as the ability to understand meaning in rich and purposeful communicative contexts (Fulcher & Davidson, 2007; Aryadoust, 2017).

In these paradigms, a variety of listening task types are commonly employed. Academic lectures and talks are implemented in higher-education to evaluate the comprehension of extended monologic

discourse, whereas dialogues and conversations are designed to represent interactional listening demands characteristic of daily communication (Taylor & Geranpayeh, 2011). Listening input may be presented in audio-only formats or accompanied by visual information in video-based tasks, with each mode imposing distinct processing demands and implications for construct representation (Gruba, 1998).

Assessment practices also vary in terms of response formats. Multiple-choice questions are widely employed because of their practicality and ease of scoring (Chang & Read, 2013). Whereas gap-fill and table-completion tasks are aimed at assessing information extraction and selective listening (Field, 2024). Regarding short-answer and summarization tasks, they require test takers to engage in deeper processing and synthesis of meaning and produce written or oral form (Rukthong, 2020).

Listening assessments are further differentiated according to their testing contexts and delivery conditions. Tests may take place in laboratory-based or classroom-based environment. In addition, tests may be delivered through institutionally controlled computer-delivered tests, web-based or remotely administered formats (Chapelle & Douglas, 2006). Finally, tasks may require responses during listening or after listening, with while-listening formats placing greater concurrent processing demands on test takers and post-listening formats allowing separation of comprehension and response (Wagner, 2013)

3. Eye tracking as a process-tracing method in SLA

Eye tracking has become an increasingly prominent method in second language acquisition (SLA) research because it documents attention distribution during task performance. As a process-tracing tool, eye tracking records observable indicators of visual attention (Godfroid & Hui, 2025). As eye-tracking technology has become more accessible, it has increasingly been adopted in language assessment research, including L2 listening, because it allows researchers to track attentional allocation and infer cognitive processes during task performance in real time (Suvorov & Irgin, 2026).

To understand the role of eye tracking in documenting these processes, researchers rely on several eye metrics. The most widely used eye tracking measures are fixation-based metrics. They represent periods of relative gaze stability on a visual element; visits or dwell time, which capture the cumulative duration of attention allocated to specific areas of interest; and scanpaths, which reflect the sequential patterns of gaze movement across visual stimuli. These measures present indirect access to ongoing cognitive activity by revealing learners ' distributional attention while processing task-related information (Conklin et al., 2019).

The interpretation of eye-tracking data is based on the eye–mind assumption, which posits a systematic relationship between perceptual orientation and cognitive processing. It claims that where individuals look and what they are processing cognitively. It should be noted that this relationship is probabilistic rather than deterministic because gaze does not always correspond directly to conscious processing, particularly in complex, multimodal tasks. As a result, eye-tracking evidence must be interpreted cautiously and within a strong theoretical framework (Holmqvist et al., 2011).

One of the main strengths of eye tracking is its temporal sensitivity. Eye trackers help researchers to observe moment-to-moment changes in attention as cognitive processes unfold in real time. This feature is especially valuable in SLA contexts, where language processing is transient and incremental. In addition, eye tracking enables non-reactive data collection, as gaze behavior can be recorded without interrupting task performance or relying on introspective reports (Roberts & Siyanova-Chanturia, 2013).

Despite these advantages, the method has notable methodological limitations. Gaze patterns may be influenced by visual salience effects rather than linguistic relevance. They may also be shaped by interface design or layout features and do not indicate underlying language processing. Consequently, eye tracking captures visual attention but cannot, on its own, provide a complete account of cognitive activity (Rayner, 2009). For this reason, the triangulation of eye-tracking data with complementary sources such as performance outcomes, verbal reports, or other process measures is essential. Properly situated, eye tracking functions not as a score predictor but as a theoretically informed source of response-process evidence that supports inferences about cognition in SLA research (Godfroid, 2020).

4.Operationalizing listening processes through eye-tracking metrics

In L2 listening research, eye-tracking metrics have been increasingly used to make unobservable cognitive processes observable through relating patterns of visual attention to stages of comprehension. Measures based on fixations, including fixation count, mean fixation duration, and total dwell time, are taken as evidence of attentional allocation and levels of processing effort. Longer or more frequent fixations on task-relevant areas (e.g., questions or response options) are typically taken to reflect increased cognitive load, difficulties in comprehension, or heightened processing demands during concurrent listening and reading. In contrast, shorter fixations may suggest more automatic processing or greater task familiarity (Conklin et al., 2019; Cullipher & VandenPlas, 2018).

Besides fixations, researchers used visit-based metrics including number of visits and revisits to specific areas of interest to infer listeners active integration and rechecking processes. When listeners

revisit response options or task prompts, this may indicate listeners' attempts to integrate auditory input with visual information. They also indicate the effort to resolve uncertainty, or confirm interpretations as listening progresses. Repeated visits are especially meaningful in tasks where listeners must coordinate information across different modes or postpone decision making (Kwon & Yu, 2024).

From a more global perspective, scanpaths and gaze transition patterns can give glimpse on strategic behavior and metacognitive monitoring during listening. The order and direction of eye movements across task elements can identify listeners behavior. They can reveal whether listeners adopt linear path or rely on options. They may show repeated alternation between questions and options, or shift attention in response to comprehension problems. Researchers view these dynamic patterns as signs of strategic deployment., self-monitoring, and flexible attention adjustment (Holmqvist et al., 2011; Suvorov,2015).

Despite these promising applications, the field lacks a clear framework that links eye movements with cognitive processes. Across the literature, the same metrics are frequently interpreted differently. Simultaneously, similar processes are captured using divergent measures, consequently, this limits comparability and diminishes the strength of cumulative interpretation (Godfroid et al., 2025). As a result, inferences about cognition may vary as a function of methodological choice rather than underlying processing differences. Moreover, there is a lack of standardized conventions regarding the choice of metric, definition of area-of-interest, and the timing relationship between visual data and auditory input (Hessels et al., 2016). The acknowledgement of these constraints is essential for combining findings from the current literature and for developing more systematic, theory-based links between eye-tracking metrics and listening processes.

5. Eye tracking as response-process evidence for validity

Eye tracking (ET) contributes to validity research in L2 listening assessment. This non-invasive technique visualizes response-process evidence that reveals how test takers engage cognitively with listening tasks as they unfold in real time (Cao & Ma, 2025). In particular, ET makes it possible to distinguish between construct-relevant processing, such as sustained attention to task-relevant information during auditory comprehension. It also can distinguish construct-irrelevant processing, including excessive visual search, premature focus on response options, or reliance on keyword-matching strategies that bypass meaning construction (Schmidt & Pastorino-Campos, 2024). These distinctions are difficult, if not impossible, to identify through test scores alone. Although score-based evidence indicates test takers' performance, it offers limited information about the cognitive processes underpin performance (Pellegrino et al., 2001; Finn et al., 2014). Process-based evidence obtained from ET complements traditional outcome measures. Eye tracking makes qualitative differences visible in processing even when test takers achieve similar scores. Research has repeatedly shown that test takers with similar scores may exhibit markedly different gaze patterns, indicating divergent strategies, attentional allocation, or levels of cognitive load (Godfroid ,2020).

In addition to tracking listeners' cognitive processes, ET plays a critical role in examining test method effects. It shows that response format, time pressure, or interface design shape listening behavior. By tracing the visual attention across task formats, ET helps determine whether observed performance reflects listening ability or the test design effect. This capability is particularly relevant for issues of equity and fairness, as certain task designs may advantage test takers who are more adept at visual scanning or strategic test-taking rather than listening comprehension per se (Suvorov, 2024).

Cognitive validity, understood as the extent to which test tasks elicit the mental processes they are intended to measure, remains insufficiently specified in existing research. It should be noted that ET alone is not sufficient to validate the test quality. Instead, it enhances the validity arguments by theoretically presenting empirically grounded evidence. This evidence supports or challenges assumptions about the cognitive processes elicited by specific tasks (Bax & Chan, 2019; Zhang, 2023). In line with contemporary validity theory, such evidence contributes to evaluating the plausibility of inferences linking performance to score interpretation and use (Kane, 2013). However, ET cannot replace other forms of validity evidence, including content analysis, psychometric evaluation, and theoretical justification. Properly integrated, ET serves as a powerful methodological bridge between cognitive theory and assessment practice, directly informing the evaluation of cognitive validity in L2 listening (Gruba & Suvorov, 2019).

Research gap

Systematic reviews have become increasingly important in applied linguistics and language assessment because they integrate methodological rigor, transparency, and replicability in synthesizing fragmented empirical findings (Chalmers et al., 2023). Unlike narrative reviews, systematic reviews employ explicit protocols in identifying and selecting related studies. This method contributes to reducing bias and enable to accumulating knowledge. Systematic reviews are particularly valuable for language assessment

research. Studies in this path of research differ substantially in theoretical orientation, task design, and analytic methods.

Eye tracking has gained considerable prominence over the past decade as a process-tracing methodology in second language research, offering a means of examining cognitive processes that cannot be captured through outcome-based measures alone (Godfroid, 2020). It is effective for observing listening comprehension because listening involves quick and real-time cognitive processes unfold rapidly and concurrently. Despite the growing use of eye tracking in L2 listening assessment, the literature remains fragmented. Existing studies vary widely in listening activities, assessment formats, input modalities (audio-only versus audiovisual), response formats, timing of responses, analytical approaches, and theoretical assumptions. Consequently, findings are scattered across isolated studies. The fragmented studies make it difficult to draw standardized conclusions about the use of eye tracking to investigate listening-related cognitive processes or the contribution of such evidence to validity arguments in L2 listening assessment.

A major gap in the literature appears to be the absence of a systematic synthesis that integrates findings across eye-tracking studies in L2 listening assessment. Although prior research has yielded insights into specific tasks or experimental conditions, relatively few reviews have attempted to assemble the range of listening activities and assessment formats used in L2 listening assessment. Task design features including the type of the input, response format, and response timing play a significant role in shaping the cognitive load. Further, they affect attentional processes, and strategies employed by test takers (Buck, 2001; Field, 2008; Kho et al., 2022). Without a systematic account of these design variables, it remains unclear how task characteristics shape eye-movement patterns or whether findings are comparable across studies.

A related gap concerns the treatment of cognitive processes in eye-tracking research on L2 listening assessment. Listening comprehension has been regarded as a multidimensional construct that involves processes such as bottom-up decoding, top-down inferencing, integration of information over time, and metacognitive monitoring (Rost, 2011; Vandergrift & Goh, 2012). However, it is unclear which of these processes have been empirically targeted in existing studies or how consistently they have been operationalized. Although eye tracking provides a range of metrics (e.g., fixation duration, dwell time, scanpaths), there has been no systematic examination of how these measures have been selected to represent specific cognitive processes, limiting interpretability and cumulative knowledge development.

Questions about the role of eye-tracking in supporting validity arguments in L2 listening assessment remain unresolved. Modern validity frameworks stress the importance of response-process evidence to evaluate whether test tasks trigger the cognitive processes predicted by theory (Bachman & Palmer, 1996; Kane, 2013; Weir, 2005). This is particularly important in listening tests because visually mediated tasks may trigger construct-irrelevant processes that encourage test takers may depend on reading or strategic visual search (Field, 2008; Kho et al., 2022). Although eye tracking has the potential to illuminate such processes, it remains unclear whether existing studies have systematically integrated eye-movement data into principled validity arguments or have primarily used eye tracking as an exploratory descriptive tool.

All these gaps call for a systematic, theory-informed, and methodologically critical review of eye-tracking research in L2 listening assessment to clarify its contribution to cognitive validity research and to guide future empirical and theoretical developments. To obtain these aims, the study is guided by the following research questions:

RQ1: What types of listening activities and assessment formats do L2 listening eye-tracking studies employ, and in which contexts are these studies conducted?

RQ2: Which cognitive processes and linguistic processing stages are investigated, and how are they quantified through eye-tracking metrics?

RQ3: How do eye-tracking measures help establish the cognitive validity of L2 listening assessments? By addressing these questions through a systematic and transparent synthesis of existing research, the present study endeavors to clarify the existing body of eye-tracking research in L2 listening assessment. Thus, it provides a principled foundation for future methodological and theoretical developments in the field.

Mthod

This study followed the PRISMA 2020 guidelines (Page et al., 2021) for conducting a systematic literature review. The review consisted of four successive stages: (1) database selection and search strategy development, (2) definition of inclusion and exclusion criteria, (3) study identification and screening, and (4) coding and analysis of the eligible studies.

2.1 Database and search terms

Comprehensive coverage was ensured through Scopus and ERIC. They were selected as the primary databases for this study due to their broad coverage of research in applied linguistics, language assessment, and educational technology, particularly in relation to eye-tracking research in L2 listening**.** To enhance retrieval, the database search was supplemented with grey literature**,** including testing organizations' research reports (e.g., *IELTS Research Reports*) and relevant book chapters. The grey literature was searched through institutional repositories, then evaluated according to the same eligibility of peer-reviewed studies to ensure methodological consistency. The search targeted studies published between January 2014 and January 2025**,** reflecting the systematic emergence of eye-tracking applications in L2 listening assessment following earlier developments in L2 reading and SLA research (Godfroid, 2020).

2.3. Study selection

Search phrases were designed around four core constructs: eye-tracking**,** listening**,** second language learning**,** and assessment**,** and were adapted to the syntax of each database. In Scopus, searches were restricted to the *TITLE-ABS-KEY* fields, whereas ERIC searches covered titles and abstracts. The search sets were logically combined using the AND operator. The complete database-specific search strings are reported in Appendix A (Database Search Strategy).

2.2 Inclusion and exclusion criteria

Selected studies were screened based on predefined inclusion and exclusion criteria (see Table 1). Firstly, only English-language publications were considered to maintain consistency in methodological consistency and facilitate comparability across the selected studies. Theses, dissertations, or abstracts are excluded because of their variability in reporting formats and accessibility which may limit reliable comparison. Secondly, studies focus on L2 listening assessment or test-related tasks were included. Thirdly, only empirical studies used eye-tracking methods and reported quantitative eye-movement data were involved. The review questions require quantitative eye- movements metrics which allow cross-studies systematic comparison. Finally, non-empirical studies, studies unrelated to listening, or those lacking eye-tracking data were excluded.

Table1. Inclusion and exclusion criteria for study selection

Inclusion criteria	Exclusion criteria
Published in English in peer-reviewed journals, research reports, or book chapters	Not written in English or published as theses, dissertations, or abstracts.
Participants are L2/EFL learners or mixed L1–L2 groups	Participants are exclusively L1 listeners or non-language populations
	Qualitative, theoretical, or review studies
Quantitative empirical studies (including mixed methods)	Focus on non-listening skills or non-assessment contexts
Focus on L2 listening assessment or test-related listening tasks	No eye-tracking methodology
Use of eye-tracking methodology	No quantitative eye-tracking data
Reporting quantitative eye-movement data (e.g., fixations, dwell time, scanpaths)	

2.3 Study selection

In this study, we reviewed relevant English-language literature published between January 2014 and January 2025. A total of 90 records were identified through database searches. 61 records were retrieved from Scopus ,22 records from ERIC, and 7 from grey literature sources. After removing 26 duplicate records**,** 64 records remained for title and abstract screening. Based on predefined inclusion and exclusion criteria, 32 records were excluded. Full-text assessment was conducted on the remaining 32 reports**.** 10 of them were excluded for not meeting the eligibility criteria. Consequently, 22 studies were retained and included in the final systematic review. The list of studies included in the final synthesis (N = 22) is provided in Appendix D. Figure 1 presents an overview of the study selection process.

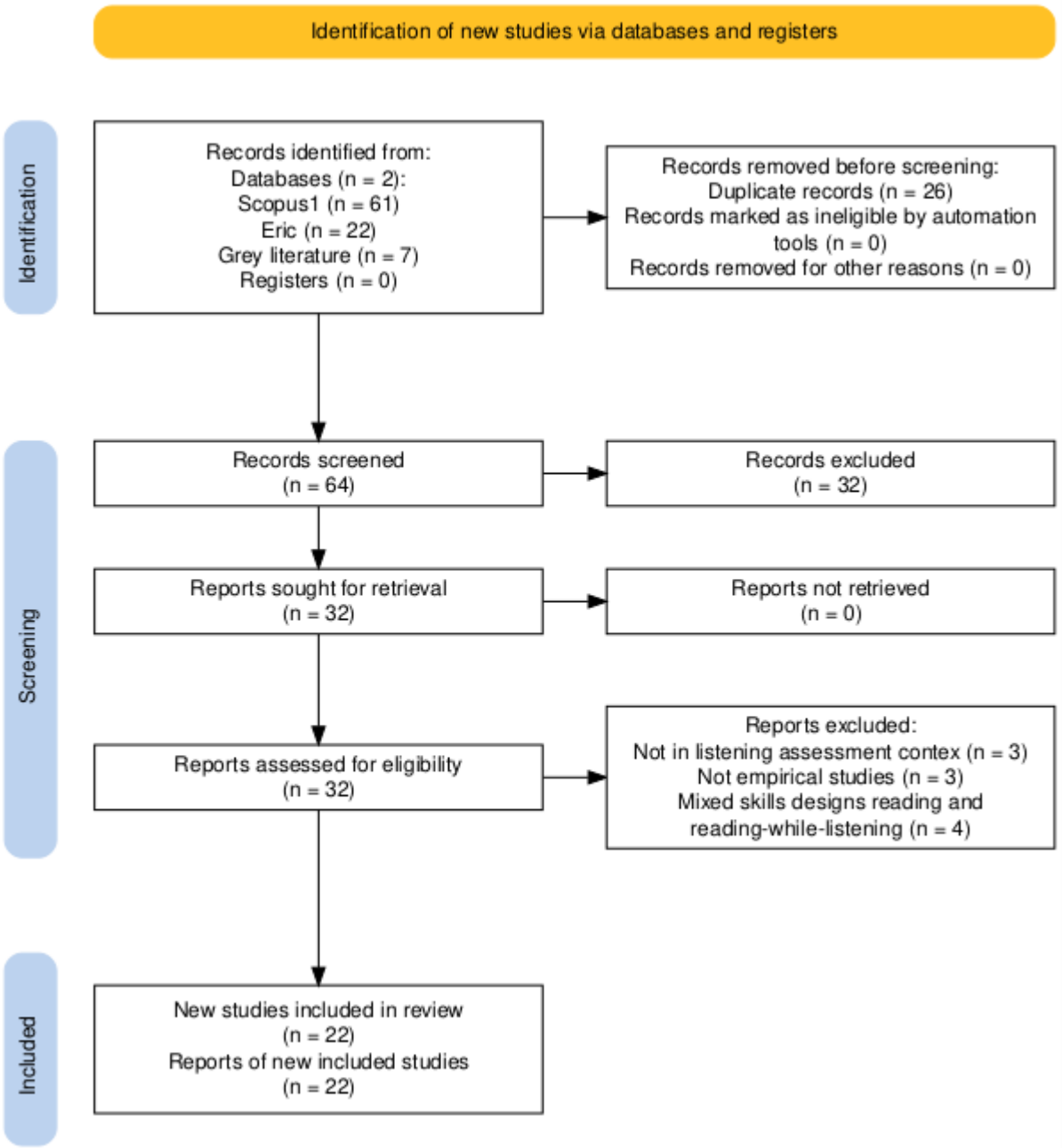

Figure1. Studies selection process for this systematic review

2.4. Coding procedure

We followed and extended the coding framework of prior systematic reviews of L2 listening research, particularly the typology of second language listening constructs proposed by Aryadoust and Luo (2022). We adopted this approach due to the absence of a established framework synthesizing the use of eye-tracking research in L2 listening assessment. (1) bibliographic information (authors, year, publication venue);
(2) study context and location;
(3) participant characteristics (language background, proficiency, age);
(4) listening activity and assessment format;
(5) eye-tracking technology and experimental setup;
(6) targeted cognitive processes or processing stages;
(7) eye-tracking metrics employed; and
(8) the role of eye-tracking evidence in supporting validity arguments.

At the initial stage, both authors worked collaboratively to code a subset of studies to develop and refine the coding criteria. Face to face discussions were used to resolve coding disagreements. Consensus was reached to ensure a shared understanding of the coding framework and its application. After calibration, one author conducted the coding of the remaining studies based on the finalized coding rules. Any

ambiguities or uncertainties encountered during coding were discussed by the two authors until agreement was achieved. Formal inter-rater reliability coefficients were not calculated because coding consistency was established through iterative calibration and consensus procedures.

4.Result

4.1. Types of listening tasks, assessment formats, and study Contexts

With respect to listening task types, 11 studies employed academic listening activities, most frequently academic lectures, presented in either audio-only or video-based formats. Of these, 4 studies used lecture-based tasks were commonly drawn from or modeled on standardized language tests, including CAEL, Aptis, TOEFL Primary, TOEFL iBT, IELTS, MET, EIKEN, and VALT. In addition to lecture-based tasks, 11 studies included dialogic listening activities that use every day conversational dialogues and short conversational videos. Some studies incorporated audiovisual input that involve context and content videos featuring gestures, facial expressions, and other non-verbal cues. A smaller number of studies employed alternative task types, such as listen-to-summarize activities.

Regarding assessment formats, multiple-choice questions (MCQs) were used in the majority of studies. These varied in structure, including three-, four-, and five-option formats, and in some cases were designed to target explicit or implicit comprehension. In several studies, MCQs were combined with other response formats, including gap-fill tasks, matching, table completion, and short-answer questions, particularly in studies using IELTS-based or researcher-developed instruments. A limited number of studies employed open-ended response formats, most notably listen-to-summarize tasks. Several studies implemented while-listening performance (WLP) and post-listening performance (PLP) formats.

Most eye-tracking studies were conducted in laboratory settings. The majority used computer-based or computer-delivered testing environments using electronic platforms such as Tobii Studio, Moodle, and Gorilla Experiment Builder reported across studies. A smaller number of studies were conducted in web-based or online environments. No studies were conducted in classroom-based or purely instructional contexts.

Table 2. Types of listening activities and assessment formats employed in L2 listening eye-tracking studies

Study ID(s)	Listening Task Type	Assessment Format	Context
Winke & Lim (2014); Batty (2021)	Academic lectures (CAEL); everyday and academic listening	MCQs, matching, gap-fill	Laboratory; web-based
Suvorov (2015a)	Short conversational videos (gestures, objects, facial expressions)	One four-option MCQ per video (explicit & implicit)	Laboratory; Moodle
Suvorov (2015b); Holzknecht et al. (2017)	Aptis Listening (A1–B2); standardized test listening	25 four-option MCQs	Laboratory; Tobii Studio
Suvorov (2018)	TOEFL Primary (Steps 1 & 2); child-focused listening	Three-option MCQs	Laboratory; simulated test
Holzknecht et al. (2021); Kho et al. (2022)	Academic lectures (video-based)	15 MCQs	Online; Gorilla Experiment Builder
Zhai & Aryadoust (2022); Hui et al. (2022)	Dialogues and academic lectures (CSAT; NELT)	15 five-option MCQs	Laboratory; computer-based
Aryadoust et al. (2022)	Academic lectures (Economics; History)	MCQs, gap-fill, matching	Laboratory; computer-based
Low & Aryadoust (2023)	EIKEN Pre-1	Listen-to-summarize task	Laboratory; Tobii Studio
Kim (2023)	TOEFL iBT Listening	MCQs	Laboratory

Nishikawa et al. (2024); Kim (2024); Kwon & Yu (2024)	VALT & MET; dialogic vs. monologic listening; context vs. content videos	MCQs	Laboratory; computer-based
Kwon (2024)	MET (listening, grammar, reading)	Mixed item types	Laboratory; Moodle
Qiu & Aryadoust (2024)	IELTS practice tests	Table completion, gap-fill, MCQs	Laboratory; computer-based
Aryadoust, Foo, & Ng (2022); Domínguez-Lucio & Aryadoust (2022); Suvorov (2024); Batty (2025)	Audio-only listening (dialogues, radio monologues, academic lectures); video-based conversational listening	MCQs (explicit & implicit), gap-fill, short-answer, WLP & PLP formats	Laboratory; computer-delivered

4.2. Eye-tracking measures of cognitive and linguistic processing

The included studies were synthesized at the level of cognitive processes and linguistic processing stages, focusing on how these constructs were operationalized through eye-tracking metrics rather than on task formats or test scores.

At early stages of processing, studies investigated visual attention allocation and orienting, quantified through fixation count, fixation rate, time to first fixation, and proportion of looking time. These indicators capture how test takers initially distribute attention across task-relevant stimuli. At the bottom-up linguistic processing stage, the measures of average and total fixation duration were used to infer lexical access, decoding effort, and processing intensity. Higher-level integrative and discourse-level processes, including inferencing and meaning construction, were operationalized through visit-based measures (visit count, visit duration). Furthermore, alternating gaze patterns were examined which reflect re-inspection and information integration over time. These measures are used as interpretive indicators rather than direct measures of cognitive processes.

Metacognitive regulation and comprehension monitoring were examined using re-visits, scanpaths, and gaze switching. By contrast, strategic processing (e.g., keyword matching) was inferred from fixation patterns on written response options. Finally, cognitive load and method-driven processing differences, central to arguments about test equity and scoring validity in Weir's framework, were examined through fixation rate, dwell time, and normalized gaze metrics. They often reveal discrepancies between observed processes and test scores.

Table3. Mapping cognitive and linguistic processing stages to eye-tracking metrics in L2 listening

Cognitive processes	**Linguistic processing stage**	**Eye tracking metrics used**	**Studies**
Visual attention allocation	Early perceptual and attentional processing	Fixation count; fixation rate time to first fixation (TTFF); proportion of looking time (PLT)	Winke & Lim (2014); Holzknecht et al. (2017); Hui et al. (2022); Kwon (2024); Qiu & Aryadoust (2024)
Visual search and orienting	Early task engagement	Fixation count; scanpaths; gaze transitions	Holzknecht et al. (2021); Aryadoust et al. (2023); Low & Aryadoust (2023)
Lexical access and decoding	Bottom-up linguistic processing	Average fixation duration ; total fixation duration	Kho et al. (2022); Zhai & Aryadoust (2022); Nishikawa et al. (2024); Winke & Lim (2014)

Processing effort	Bottom-up and integrative processing	Total fixation duration; fixation density; normalized fixation metrics	Kho et al. (2022); Zhai & Aryadoust (2022); Domínguez Lucio & Aryadoust (2022)
Multimodal (audio–visual) integration	Integrative processing	Dwell time; PLT on visual cues	Suvorov (2015a, 2015b); Batty (2021); Kim (2023); Kim (2024); Batty (2025)
Inferencing and meaning construction	Higher-order discourse-level processing	Visit count; visit duration; alternating gaze patterns	Suvorov (2015b); Batty (2021); Batty (2025)
Pragmatic and affective interpretation	Discourse-pragmatic processing	Dwell time on faces; alternating gaze patterns	Batty (2021); Batty (2025); Kim (2024)
Comprehension monitoring	Metacognitive regulation	Re-visits; scanpaths; gaze switching	Suvorov (2018); Low & Aryadoust (2023)
Test-taking strategies (e.g., keyword matching)	Strategic processing	Fixation duration on written items; visit frequency; scanpaths	Winke & Lim (2014); Kho et al. (2022); Suvorov (2018)
Decision-making and response evaluation	Post-listening evaluative processing	Visit count; visit duration on response options	Holzknecht et al. (2021); Domínguez Lucio & Aryadoust (2022)
Cognitive load	Cross-stage processing demand	Fixation rate; dwell time; normalized fixation and visit metrics	Kho et al. (2022); Zhai & Aryadoust (2022); Domínguez Lucio & Aryadoust (2022)
Hidden process inequivalence (test equity)	Method-driven processing differences	Fixation count; visit count; PLT	Domínguez Lucio & Aryadoust (2022); Qiu & Aryadoust (2024)

4.3. Eye-tracking and cognitive validity in L2 listening assessment

To synthesize findings across studies, themes were derived by coded results based on conceptual similarity and their validity relevance. As summarized in Table 4, the reviewed studies collectively show that eye-tracking frequently provides response-process evidence that indicates construct alignment, diagnoses construct-irrelevant variance. In addition, eye tracking differentiates test methods and item types, and clarifies modality, proficiency, and equity effects. Even when score differences are minimal or absent, gaze data reveal meaningful cognitive distinctions, thereby strengthening the cognitive validity argument for L2 listening assessments.

Table4. Synthesis of eye-tracking evidence supporting cognitive validity in L2 listening assessment

Cognitive validity theme	**What eye tracking demonstrates?**	**Key studies**
Construct-relevant listening processes	Test takers' gaze patterns align with intended listening processes (e.g., meaning construction, discourse processing), even when score differences are absent	Suvorov (2015a, 2015b); Holzknecht et al. (2017); Batty (2021); Kim (2023); Kim (2024); Batty (2025)

Detection of construct-irrelevant variance	Scores are influenced by non-listening behaviors such as keyword matching, test-wiseness, response order, or interaction behaviors	Suvorov (2018); Holzknecht et al. (2021); Kho et al. (2022); Aryadoust & Foo (2023); Qiu & Aryadoust (2024)
Test method effects and cognitive load	Different listening formats (while-listening vs. post-listening; audio-only vs. video-based) elicit different cognitive processes and levels of cognitive load	Zhai & Aryadoust (2022); Aryadoust et al. (2022); Kwon (2024)
Role of visuals and authenticity	Visual input guides attention and supports authenticity without necessarily improving comprehension or altering core listening cognition	Suvorov (2015a, 2015b); Kim (2023); Kim (2024); Kwon (2024); Kwon & Yu (2024)
Item-type sensitivity	Different item types (explicit vs. implicit) direct attention to different visual cues, activating different cognitive resources	Batty (2025); Kwon & Yu (2024)
Proficiency- and group-specific processing	Lower-proficiency or L2 listeners rely more on written input and keyword matching than L1 listeners	Aryadoust & Foo (2023); Nishikawa et al. (2024); Qiu & Aryadoust (2024)
Strategy–performance alignment	Gaze behavior predicts performance more reliably than self-reports, revealing actual strategy use	Low & Aryadoust (2023); Suvorov (2018)
Equity and comparability of test forms	Comparable cognitive and neurocognitive engagement across equated test forms, despite minor gaze differences	Domínguez Lucio & Aryadoust (2022); Winke & Lim (2014)
Methodological contribution to validation	Eye-tracking overcomes limitations of verbal reports by providing non-reactive response-process evidence	Suvorov (2024); Winke & Lim (2014)

5.Discussion

5.1. Task types and construct representation

As shown in Table 2, the majority of reviewed studies relied on academically oriented listening tasks, most commonly monologic lectures drawn from or modeled on standardized tests such as CAEL, IELTS, TOEFL iBT, Aptis, MET, EIKEN, and VALT. From a cognitive validity perspective, lecture-based tasks afford sustained attention, macro-level meaning construction, and information integration across extended input. As a result, they make them well suited to operationalizing higher-level academic listening constructs (Buck, 2001; Field, 2008). At the same time, their dominance reflects a narrow functional framing of listening that prioritizes academic comprehension over interactional, pragmatic, or socially situated listening abilities.

Dialogic listening tasks and conversational exchanges particularly those associated with TOEIC-, CSAT-, or NELT-type materials were present but comparatively underrepresented. This imbalance poses a risk of incomplete representation of the target construct. Dialogic listening imposes cognitive demands which involve turn-taking, rapid inferencing, perspective-taking, and pragmatic interpretation (Floyd,

2010). Similarly, a number of studies used audiovisual or multimodal materials that involve gestures, facial expressions, and contextual cues. However, the use of these elements was designed for assessment purposes and with limited focus on simulating real life communication (Suvorov, 2015; Batty, 2021). As a result, audiovisual input was often treated as supplementary information rather than to elicit interactive meaning-making processes. This perspective reinforces a conception of listening as an individual, receptive process rather than a co-constructed activity (Rost, 2011). The choice of such tasks facilitates standardization and comparability; nevertheless, they do not capture cognitive processes involved in real-world listening.

5.1.2. Assessment formats and cognitive validity implications

Multiple-choice questions (MCQs) as response format were the overwhelmingly dominant across the reviewed studies. Their prevalence is unsurprising because of their psychometric advantages namely scoring reliability, efficiency, and scalability (Alderson, Clapham, & Wall, 1995; Buck, 2001) as well as their methodological compatibility with eye-tracking research. MCQs provide clearly delineated visual regions of interest, enabling fine-grained analysis of fixation duration, dwell time, and option-level attention (Holzknecht et al., 2021).

However, from a cognitive validity standpoint, MCQ-dominated designs introduce identifiable risks. Recognition-based formats may place less emphasis on integrative listening processes such as inferencing, prediction, and discourse-level coherence building. These processes are pivotal to authentic listening (Field, 2008; Polat, 2020). By contrast, they may promote construct-irrelevant strategies such as keyword matching, option elimination, or visual scanning of distractors processes (Chang & Read, 2013). To detect these strategies, eye tracking is the optimal choice because it provides detailed evidence of response processes. Nevertheless, its data is not indicative of listening ability itself (Kho et al., 2022). Because task formats affect measurement, many studies emphasize fixation-based metrics associated with decision-making at the response stage. In contrast, gaze transitions or scanpaths reflect continuous comprehension and multimodal integration (Tóthová & Rusek, 2025). Only a limited number of studies used open-ended or integrative response formats, such as summarization, short-answer responses, or matching activities. Although such formats are difficult to score and control experimentally, they are more effective in supporting cognitive validity. They require learners to construct meaning, synthesize information, and individualized interpretation (Rukthong, 2020). Their infrequent use indicates that current eye-tracking research continues to rely on psychometrically convenient formats at the cost of cognitively richer task designs.

5.1.3. Study contexts and ecological validity

With respect to research context, Table 2 indicates a strong dominance of laboratory-based, computer-mediated environments. Most studies were conducted under highly controlled conditions using platforms such as Tobii Studio, Moodle, or Gorilla Experiment Builder. These virtual environments allow for precise calibration, synchronized presentation stimulus, and reliable high-quality gaze data. All these features are significant because they play a critical role for ensuring methodological rigor in eye-tracking research (Godfroid, 2020).

Nevertheless, this emphasis on experimental control threats the ecological validity. The lack of classroom or instructional contexts reduces the generalizability of the results to real life learning environments. Even when studies used online or hybrid delivery modes, they typically framed tasks as formal assessments rather than pedagogically embedded activities. Most studies focus on listening behavior in assessment conditions. As a result, the research on how cognitive processing unfolds during interactive or instructional listening remain underexplored (Aryadoust et al., 2022).

5.2. Cognitive Processes and Linguistic Processing Stages

5.2.1. Predominance of attentional processes

Research on L2 listening cognition has drawn on a wide range of methodological approaches, including think-aloud protocols, stimulated recall, questionnaires, interviews, eye tracking, and neuroimaging (Field, 2008; Goh, 2000; Aryadoust et al., 2022). Introspective and self-report methods help us understand strategic behavior but they are inherently indirect, retrospective, and influenced by memory and awareness. Unlike self-report methods, eye tracking and other technology-mediated methods provide real-time indicators of attentional allocation, processing effort, and cognitive load during task performance. These methods show that listening involves visual and multimodal elements, particularly in assessment contexts where auditory input is accompanied by written prompts or visual stimuli (Aryadoust & Luo, 2022).

Visual attention allocation was the most frequently examined process, typically captured using fixation-based metrics such as fixation count, fixation duration, dwell time, time to first fixation (TTFF), and proportion of looking time (PLT). These measures are well established as indicators of attentional engagement and processing effort (Just & Carpenter, 1980; Rayner, 1998) and offer clear analytic

affordances in test-based listening tasks with well-defined visual regions of interest mainly measures. While such metrics offer valuable insights into attentional allocation and processing effort, their dominance reflects a theoretical narrowing of the listening construct. In complex listening tasks, visual attention does not necessarily equate to comprehension or meaning construction. Fixation behavior may be influenced by interface design, visual salience, task layout, or the reading demands imposed by response formats, rather than by auditory processing per se (Holmqvist et al., 2011; Negi & Mitra, 2020). Strong interpretations of fixation behavior may exaggerate the eye–mind assumption by treating gaze as a transparent proxy for cognitive processing, without adequately considering alternative explanations such as interface-driven gaze behavior, visual search strategies, or reading load imposed by test formats (Godfroid, 2020). As a result, attentional measures may represent test takers navigation behavior interfaces more than auditory comprehension processes.

5.2.2. Bottom-up linguistic processing and lexical access

A limited number of studies analyzed bottom -up linguistic mechanisms with main focus on lexical access and decoding processes. The inference of these processes was primarily based on average and total fixation duration referring on the assumption that longer fixations signal increased processing difficulty or lexical retrieval effort (Holsanova, 2014; Mulder et al., 2024). Such metrics were most often applied when listeners interacted with written prompts, captions, or response options during listening.

Although these studies contribute to understanding early-stage processing, their interpretive power remains limited. Fixation duration alone cannot reliably differentiate between lexical difficulty, strategic hesitation, uncertainty, or format-induced processing demands. Without triangulation or explicit modeling of linguistic complexity, bottom-up processing is inferred rather than directly substantiated, which constrains explanatory depth and raises questions about construct interpretability (Godfroid, 2020).

5.2.3. Integrative and multimodal processing

Multimodal integration (the coordination of auditory input with visual information) was examined in a limited but growing body of research. The use of metrics such as dwell time, PLT on visual cues, and gaze alternations between auditory- and visually relevant regions to capture audiovisual integration (Suvorov, 2015a, 2015b; Batty, 2021; Kim, 2024). The incorporation of scanpaths and gaze transitions represented a methodological advance that enable researchers to analyze temporal sequencing and cross-modal coordination rather than static attention snapshots (Godfroid, 2020).

Despite these developments, multimodal processing remains under-theorized and under-measured relative to its fundamental role in real-world listening. Few studies explicitly linked gaze patterns to inferencing, discourse construction, or meaning integration over time, even though these processes are fundamental to authentic listening comprehension (Field, 2008; Rost, 2011). Consequently, much of the existing evidence indicates where attention is allocated rather than how meaning emerges through coordinated processing across modalities.

5.2.4. Higher-order cognition, metacognition, and strategy use

The assessment of higher-order cognitive processes, such as inferencing, comprehension monitoring, and strategic regulation, were predominantly captured through visit-based metrics, re-visits, scanpaths, and gaze switching. The use of such measures reflects theoretical models that characterize listening as an active, recursive, and strategic activity (Bachman & Palmer, 1996; Weir, 2005). Nevertheless, they were applied inconsistently and commonly confined to a secondary analysis. Strategy use, for example, was frequently inferred from disproportionate attention to written response options interpreted as evidence of keyword matching rather than from explicit modeling of listening strategies or triangulated evidence (Kho et al., 2022; Suvorov, 2018). This pattern reinforces long-standing concerns that metacognitive processes in L2 listening are difficult to operationalize and are often inferred indirectly through behavioral proxies, increasing the risk of construct underrepresentation (Aryadoust, 2022).

Taken together, these findings suggest that current eye-tracking research captures only a partial view of higher-order cognition in listening assessment. The predominance of indirect behavioral proxies limits the extent to which gaze data alone can substantiate claims about inferencing, discourse integration, or metacognitive regulation. From a validity perspective, this underscores the need for more integrated interpretive frameworks that situate gaze behavior within broader models of cognitive processing, thereby strengthening the inferential link between observed attentional patterns and theoretically defined higher-order constructs.

5.2.5. Cognitive load, test equity, and validity considerations

A number of studies applied eye-tracking metrics to investigate cognitive load and method-related processing variations to highlight their significance for test equity. Measures such as fixation rate, normalized fixation duration, and visit density were used to identify mismatches between observed processing patterns and test scores (Domínguez Lucio & Aryadoust, 2022; Qiu & Aryadoust, 2024). The findings provide response-process evidence in line with Weir's (2005) framework of cognitive validity. It

demonstrates how eye tracking can reveal construct-irrelevant variance and processing inequivalences that remain invisible in score-based analyses.

At the same time, these studies emphasize the limits of eye tracking as validity evidence. Based on Kane's (2013) argument-based approach, eye-tracking data can support but not independently establish validity claims. Without clear theoretical constraints on metric–process mapping, interpretations risk theoretical overreach, particularly when alternative explanations for gaze behavior are not explicitly addressed.

Generally, these results suggest that the contribution of eye tracking to cognitive validity lies in its integration within broader interpretive frameworks rather than in isolated metric–outcome associations. More principled use of gaze data conceptually anchored and triangulated with complementary evidence strengthens its role as a theory-driven tool for identifying cognitive load effects and equity-related threats in L2 listening assessment.

5.3.Eye-tracking and cognitive validity in L2 listening assessment

The synthesis of the reviewed studies indicates that eye tracking makes a substantive contribution to cognitive validity arguments in L2 listening assessment by providing fine-grained response-process evidence. Across studies, gaze data illustrate how test takers engage with listening tasks in real time, allowing researchers to examine whether observed cognitive processes plausibly align with intended construct definitions. In this respect, eye tracking complements traditional score-based validation by illuminating underlying processing mechanisms rather than treating test scores as the sole indicators of listening ability.

Within Kane's (2013) argument-based validity framework, eye tracking primarily supports interpretation and use inferences. The reviewed studies show that gaze behavior can indicate engagement in construct-relevant processes such as meaning construction, inferencing, and audiovisual integration, while also revealing method effects and construct-irrelevant strategies (e.g., keyword matching or excessive visual search) that may distort score interpretation. However, eye-tracking evidence provides only indirect support for extrapolation and cannot, on its own, justify high-stakes decision inferences, as gaze behavior in testing contexts does not confirm transfer to real-world listening or consequential outcomes.

A central contribution of eye tracking lies in strengthening the plausibility of construct-relevant processing while simultaneously identifying construct-irrelevant variance. Multiple studies reported gaze patterns consistent with theoretically motivated listening processes, even when test scores were insensitive to experimental conditions or group differences (e.g., Suvorov, 2015a, 2015b; Holzknecht et al., 2017; Batty, 2021; Kim, 2023, 2024). At the same time, eye tracking exposed sources of construct contamination—such as test-wiseness strategies and disproportionate attention to written response options—that remain largely invisible in outcome-based analyses. From a cognitive validity perspective, such evidence is corroborative rather than confirmatory, strengthening construct interpretation without independently verifying construct representation.

The reviewed studies also demonstrate that eye tracking is particularly informative for understanding test method effects and cognitive load. Comparisons across task formats (e.g., while-listening vs. post-listening; audio-only vs. video-based tasks) consistently revealed distinct patterns of attentional allocation and processing effort, often in the absence of corresponding score differences. These findings suggest that method effects operate primarily at the level of cognitive processing, underscoring the importance of evaluating how task design shapes response processes, not merely performance outcomes.

Finally, emerging work indicates that eye tracking can contribute to equity and comparability arguments by examining whether different test forms or groups engage similar cognitive processes. Although existing evidence suggests broadly comparable processing patterns across forms (Domínguez Lucio & Aryadoust, 2022; Winke & Lim, 2014), the limited scope of such studies constrains generalization.

To sum up, the evidence suggests that eye tracking strengthens cognitive validity arguments by revealing how listening tasks are processed in real time. Its value lies not in functioning as a standalone validation tool, but in complementing psychometric and qualitative approaches to clarify construct plausibility, diagnose construct-irrelevant variance, and evaluate method effects in L2 listening assessment.

6.Limitations and future research

Despite the contributions of the current systematic review, it is a subject to several limitations that should be acknowledged when interpreting the findings. First, the majority of the reviewed studies relied on academic lecture-based tasks adapted from standardized tests. This dominance facilitated conceptual consistency and facilitated comparison; however, it narrowed the functional scope of listening under investigation. Dialogic, interactional, and socially situated listening tasks are largely underrepresented. As

a result, the synthesized evidence may not capture the cognitive processes that are fundamental in real-world listening, such as turn-taking, pragmatic inference, and co-constructed meaning. Second, ecological validity is constrained because of the dominance of multiple-choice response formats and laboratory-based, computer-mediated settings. The use of such designs emphasizes attentional allocation and decision-making processes related to response selection, but underrepresent integrative meaning construction and discourse-level comprehension. Consequently, the predominance of test-based task designs tends to illustrate listening behavior rather than the cognitive demands inherent in authentic listening situations.

Third, the reviewed literature shows a strong reliance on fixation-based attentional metrics. Although these measures offer valuable insights into processing effort and visual engagement, their prevalence limits the interpretive depth of cognitive validity claims (Godfroid, Winke, & Conklin, 2020). Higher-order processes including inferencing, metacognitive regulation, and discourse integration are often inferred indirectly and inconsistently. Thus, this inference increases the risk of underrepresenting the target construct.

Future research should overcome these limitations by extending eye-tracking investigations to include interactive, dialogic, and classroom-embedded listening tasks, as well as more open-ended and integrative response formats. Methodologically, greater emphasis on temporal, process-oriented analyses is required. Multi-method approaches that triangulate gaze data with complementary evidence would strengthen the validity of inferences about underlying cognitive processes. Such advances would enhance the contribution of eye tracking to theory-driven cognitive validity arguments in L2 listening assessment.

Conclusions

This study aims to analyze and summarize eye-tracking research in L2 listening assessment. It seeks to address three central questions regarding task characteristics, cognitive processes, and the employed eye movements measures associated with validity claim. The synthesis indicates that standardized, academically oriented listening tasks using laboratory-based formats and fixation-based metrics have dominated the reviewed studies. It also shows that eye tracking provides evidence about response processes relevant to evaluating construct alignment. Eye tracking data identifies construct-irrelevant influences, and differentiates test methods, proficiency levels, and task features.

The findings strengthen the existing knowledge. It suggests that eye tracking should not viewed as a direct method to test listening ability itself. Instead, it can be used as a tool for examining the underlying processes. Thus, it explains how test outcomes are achieved.

The review offers several implications. For test developers, it stresses the importance of task design features that elicit construct-relevant processing. For researchers, it emphasizes the need to triangulate eye tracking with other approaches. In addition, it highlights the inclusion of different types of listening situations. For language assessment, it promotes the integration of process evidence into validation frameworks with numerical score statistics.

Appendix A. Database Search Strategy

TITLE-ABS-KEY(
"eye tracking" OR "eye-tracking" OR eyetracking OR "gaze behavior*" OR "visual attention" OR "visual cue*") AND
TITLE-ABS-KEY(
"L2 listening" OR "second language listening" OR "listening assessment*" OR "listening test*" OR "video-based listening")

1	Winke et al.	*The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation* (IELTS Research Report)	2014
2	Suvorov	*Interacting with visuals in L2 listening tests: An eye-tracking study* (AR-A/2015/1, British Council)	2015
3	Suvorov	*The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos*	2015
4	Holzknecht et al.	*Looking into listening: Using eye tracking to establish the cognitive validity of the Aptis Listening Test* (AR-G/2017/3, British Council)	2017
5	Suvorov	*Investigating test-taking strategies during the completion of computer-delivered items from the Michigan English Test (MET): Evidence from eye tracking and cued retrospective reporting*	2018
6	Batty	*An eye-tracking study of attention to visual cues in L2 listening tests*	2021
7	Holzknecht et al.	*The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test*	2021
8	Aryadoust et al.	*What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments?*	2022
9	Domínguez-Lucio & Aryadoust	*Neurocognitive evidence for test equity in an academic listening assessment*	2022
10	Hui et al.	*Reading aloud listening test items to young learners: Attention, item understanding, and test performance*	2022
11	Kho et al.	*An eye-tracking investigation of the keyword-matching strategy in listening assessment*	2022
12	Zhai & Aryadoust	*The metacognitive and neurocognitive signatures of test methods in academic listening*	2022
13	Aryadoust et al.	*An eye-tracking investigation of visual search strategies and test performance of L1 and L2 listening test takers*	2023
14	Low & Aryadoust	*Investigating test-taking strategies in listening assessment: A comparative study of eye tracking and self-report questionnaires*	2023
15	Kim	*Test takers' interaction with context videos in a video-based listening test: A conceptual replication and extension of Suvorov (2015)*	2023
16	Nishikawa et al.	*The impact of test takers' proficiency on their listen-to-summarize task performance*	2024
17	Kim	*Second language listeners' emotion and eye gaze: A web-based eye-tracking study*	2024
18	Kwon	*A comparative study on audio-only and video-based listening tests: The impact of visual input*	2024
19	Kwon & Yu	*The effect of viewing visual cues in a listening comprehension test on second language learners' test-taking process and performance: An eye-tracking study*	2024
20	Qiu & Aryadoust	*The predictive value of gaze behavior and mouse clicking in testing listening proficiency: A sensor technology study*	2024

21	Suvorov	*The use of eye tracking in validating L2 listening assessments* (in *Language Test Validation in a Digital Age*)	2024
22	Batty	*Attention to visual cues in explicit and implicit item types on video L2 listening tests: An eye-tracking study*	2025

("eye tracking" OR "eye-tracking" OR eyetracking OR "gaze behavior*" OR "visual attention" OR "visual cues")
AND
("L2 listening" OR "second language listening" OR "listening assessment" OR "listening test" OR "video-based listening")
Appendix B: Included Studies (N = 22).

AI use disclosure
The authors used ChatGPT (GPT-4) for language editing and limited text refinement. All content was reviewed and validated by the authors, who take full responsibility for the manuscript in accordance with Taylor & Francis AI policy.

References

[1] Abdel Latif, M. M. M. (2019). Eye-tracking in recent L2 learner process research: A review of areas, issues, and methodological approaches. System, 83, 25–35. https://doi.org/10.1016/j.system.2019.02.008

[2] Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge University Press.

[3] Anderson, A., & Lynch, T. (1988). Listening. Oxford University Press.

[4] Aryadoust, V. (2017). Communicative testing of listening. In J. I. Liontas (Ed.), The TESOL encyclopedia of English language teaching. John Wiley & Sons. https://doi.org/10.1002/9781118784235.eelt0617

[5] Aryadoust, V., & Ang, B. H. (2019). Exploring the frontiers of eye tracking research in language studies: A novel co-citation scientometric review. Computer Assisted Language Learning, 34(7), 898–933. https://doi.org/10.1080/09588221.2019.1647251

[6] Aryadoust, V. (2022). The known and unknown about the nature and assessment of L2 listening. International Journal of Listening, 36(2), 69–79. https://doi.org/10.1080/10904018.2022.2042951

[7] Aryadoust, V., Foo, S. W. L., & 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

[8] Aryadoust, V., & Luo, L. (2022). The typology of second language listening constructs: A systematic review. Language Testing, 39(4), 537–566. https://doi.org/10.1177/02655322221126604

[9] Aryadoust, V., & Foo, S. W. L. (2023). An eye tracking investigation of visual search strategies and test performance of L1 and L2 listening test takers. Research and Practice in Technology Enhanced Learning, 18, Article 009. https://doi.org/10.58459/rptel.2023.180092

[10] Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests. Oxford University Press.

[11] Batty, A. O. (2021). An eye tracking study of attention to visual cues in L2 listening tests. Language Testing, 38(4), 511–535. https://doi.org/10.1177/0265532220951504

[12] Bax, S., & Chan, S. (2019). Using eye-tracking research to investigate language test validity and design. Language Testing, 36(2), 171–192.

[13] Brothers, T., Hoversten, L. J., Dave, S., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. Neuropsychologia, 135, 107225. https://doi.org/10.1016/j.neuropsychologia.2019.107225

[14] Brown, H. D., & Abeywickrama, P. (2019). Language assessment: Principles and classroom practices (3rd ed.). Pearson Education.

[15] Buck, G. (2001). Assessing listening. Cambridge University Press.

[16] Cao, X., & Ma, Z. (2025). The review on eye-tracking studies in L2 assessment. Colombian Applied Linguistics Journal, 27(2), 51–63. https://doi.org/10.14483/22487085.22043

[17] Chalmers, H., Brown, J., & Koryakina, A. (2023). Topics, publication patterns, and reporting quality in systematic reviews in language education: Lessons from the International Database of Education Systematic Reviews (IDESR). Applied Linguistics Review, 14(7), 1239–1267. https://doi.org/10.1515/applirev-2022-0190

[18] Chang, A. C.-S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. System, 41(4), 1080–1093. https://doi.org/10.1016/j.system.2013.06.001

[19] Chapelle, C. A., & Douglas, D. (2006). Assessing language through computer technology. Cambridge University Press. https://doi.org/10.1017/CBO9780511733116

[20] Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. Second Language Research, 32(3), 453–467. https://doi.org/10.1177/0267658316637401

[21] Conklin, K., Pellicer-Sánchez, A., & Carroll, G. (2019). Eye-tracking: A guide for applied linguistics research. Cambridge University Press. https://doi.org/10.1017/9781108233279

[22] Cullipher, S., & VandenPlas, J. R. (2018). Using fixations to measure attention. In J. R. VandenPlas, S. Cullipher, & A. R. Jones (Eds.), Eye tracking for the chemistry education researcher (pp. 53–72). American Chemical Society. https://doi.org/10.1021/bk-2018-1292.ch004

[23] Domínguez Lucio, E., & Aryadoust, V. (2022). Neurocognitive evidence for test equity in an academic listening assessment. Behaviormetrika, 50, 155–175. https://doi.org/10.1007/s41237-022-00171-1

[24] Field, J. (2008). Listening in the language classroom. Cambridge University Press.

[25] Field, J. (2024). Insights into assessing academic listening: The case of IELTS. Cambridge

[26] Finn, A. S., Kraft, M. A., West, M. R., & Leonard, J. A. (2014). Cognitive skills, student achievement tests, and schools. Psychological Science, 25(3), 736–744. https://doi.org/10.1177/0956797613516008

[27] Floyd, J. J. (2010). Listening: A dialogic perspective. In A. D. Wolvin (Ed.), Listening and human communication in the 21st century (pp. 127–140). Wiley-Blackwell. https://doi.org/10.1002/9781444314908.ch5

[28] Flowerdew, J., & Miller, L. (2010). Listening in a second language. In A. D. Wolvin (Ed.), Listening and human communication in the 21st century (pp. 158–177). Wiley-Blackwell. https://doi.org/10.1002/9781444314908.ch7

[29] Fulcher, G., & Davidson, F. (2007). Language testing and assessment: An advanced resource book. Routledge.

[30] Godfroid, A. (2020). Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide. Routledge. https://doi.org/10.4324/9781315775616

[31] Godfroid, A., Finch, B., & Koh, J. (2025). Reporting eye-tracking research in second language acquisition and bilingualism: A synthesis and field-specific guidelines. Language Learning, 75(1), 250–294. https://doi.org/10.1111/lang.12664

[32] Godfroid, A., & Hui, B. (2025). Eye-tracking research in instructed second language acquisition. Language Teaching, 1–31. https://doi.org/10.1017/S0261444825000102

[33] Godfroid, A., & Hui, B. (2020). Five common pitfalls in eye-tracking research. Second Language Research, 36(3), 277–305. https://doi.org/10.1177/0267658320921218

[34] Godfroid, A., Winke, P., & Conklin, K. (2020). Exploring the depths of second language processing with eye tracking: An introduction. Second Language Research, 36(3), 239–249. https://doi.org/10.1177/0267658320922578

[35] Goh, C. C. M. (2018). Metacognition in second language listening. In Teaching listening: Theorizing listening. Wiley-Blackwell. https://doi.org/10.1002/9781118784235.eelt0572

[36] Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. System, 28(1), 55–75. https://doi.org/10.1016/S0346-251X(99)00060-3

[37] Gruba, P. (1998). The role of video media in listening assessment. System, 25(3), 335–345. https://doi.org/10.1016/S0346-251X(97)00026-2

[38] Gruba, P., & Suvorov, R. (2019). Technology and second language listening. In M. A. Peters (Ed.), Encyclopedia of educational innovation. Springer. https://doi.org/10.1007/978-981-13-2262-4_142-1

[39] 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

[40] Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016). The area-of-interest problem in eye-tracking research: A noise-robust solution for face and sparse stimuli. Frontiers in Psychology, 7, Article 1228. https://doi.org/10.3389/fpsyg.2016.01228

[41] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford University Press.

[42] Holsanova, J. (2014). Reception of multimodality: Applying eye tracking methodology in multimodal research. In C. Jewitt (Ed.), Routledge handbook of multimodal analysis (2nd ed., pp. 285–296). Routledge. https://doi.org/10.13140/2.1.3790.3041

[43] Holzknecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E., & Spöttl, C. (2017). Looking into listening: Using eye tracking to establish the cognitive validity of the Aptis Listening Test (AR G/2017/3). British Council Assessment Research Awards and Grants (ARAGs) Research Reports Online Series.

[44] 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

[45] Hu, X., & Aryadoust, V. (2024). A systematic review of eye-tracking technology in second language research. Languages, 9(4), 141. https://doi.org/10.3390/languages9040141

[46] Hui, B., Wong, S. S. Y., & Au, R. K. C. (2022). Reading aloud listening test items to young learners: Attention, item understanding, and test performance. System, 108, Article 102831. https://doi.org/10.1016/j.system.2022.102831

[47] Imhof, M. (2010). What is going on in the mind of a listener? The cognitive psychology of listening. In A. D. Wolvin (Ed.), Listening and human communication in the 21st century (pp. 97–126). Wiley-Blackwell.

[48] In'nami, Y., & Koizumi, R. (2022). The relationship between L2 listening and metacognitive awareness across listening tests and learner samples. International Journal of Listening, 36(2), 100–117. https://doi.org/10.1080/10904018.2021.1955683

[49] Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87(4), 329–354. https://doi.org/10.1037/0033-

295X.87.4.329

[50] 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

[51] Kho, S. Q. E., Aryadoust, V., & Foo, S. (2022). An eye-tracking investigation of the keyword-matching strategy in listening assessment. Education and Information Technologies. Advance online publication. https://link.springer.com/article/10.1007/s10639-022-11322-y

[52] Kim, J. (2023). Test takers' interaction with context videos in a video-based listening test: A conceptual replication and extension of Suvorov (2015). Language Testing, 40(2), 329–353. https://doi.org/10.1177/02655322221112345

[53] Kim, J. (2024). Second language listeners' emotion and eye gaze: A web based eye tracking study. International Journal of Listening, 39(2), 133–149. https://doi.org/10.1080/10904018.2024.2420098

[54] Kwon, S. K. (2024). A comparative study on audio-only and video-based listening tests: The impact of visual input. In G. Yu & J. Xu (Eds.), Language test validation in a digital age (pp. 67–92). Cambridge University Press & Assessment.

[55] Kwon, S. K., & Yu, G. (2024). The effect of viewing visual cues in a listening comprehension test on second language learners' test taking process and performance: An eye tracking study. Language Testing, 41(3), 649–680. https://doi.org/10.1177/02655322241239356

[56] Lado, R. (1961). Language testing: The construction and use of foreign language tests: A teacher's book. Longmans, Green and Company.

[57] Low, A. R. L., & Aryadoust, V. (2023). Investigating test taking strategies in listening assessment: A comparative study of eye tracking and self-report questionnaires. International Journal of Listening, 37(2), 93–112. https://doi.org/10.1080/10904018.2021.1883433

[58] Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. Language and Cognitive Processes, 27(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

[59] Mulder, K., Brand, S., Boves, L., & Ernestus, M. (2024). Processing reduced speech in the L1 and L2: A combined eye-tracking and ERP study. Language, Cognition and Neuroscience, 39(5), 527–551. https://doi.org/10.1080/23273798.2024.234416

[60] Negi, S., & Mitra, R. (2020). Fixation duration and the learning process: An eye tracking study with subtitled videos. Journal of Eye Movement Research, 13(6), 1–15. https://doi.org/10.16910/jemr.13.6.1

[61] Nguyen, T. H., & Abbott, M. (2017). Promoting process-oriented listening instruction in the ESL classroom. TESL Canada Journal, 34(1), 72–86. https://doi.org/10.18806/tesl.v34i1.1256

[62] Nishikawa, M., Horiguchi, Y., Yu, G., & Luhovyk, O. (2024). The impact of test takers' proficiency on their listen-to-summarize task performance [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-5279657/v1

[63] Oller, J. W., Jr. (1979). Language tests at school: A pragmatic approach. Longman.

[64] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

[65] Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. National Academy Press.

[66] Plakans, L., & Park, G. M. (2024). Listening in multimodal tasks. In L. Plakans & G. M. Park

(Eds.),

[67] The Routledge handbook of second language acquisition and listening (pp. 1–12). Routledge.

[68] Polat, M. (2020). Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. Novitas-ROYAL (Research on Youth and Language), 14(2), 76–.

[69] Pratiwi, K., & Andriyanti, E. (2019). External factors causing students' difficulties in listening. Journal of English Language Teaching and Linguistics, 4(2), 229–240. https://doi.org/10.21462/jeltl.v4i2.282

[70] 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

[71] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

[72] Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search (The 35th Sir Frederick Bartlett Lecture). The Quarterly Journal of Experimental Psychology, 62(8), 1457–1506. https://doi.org/10.1080/17470210902816461

[73] Richards, J. C. (2008). Teaching listening and speaking: From theory to practice. Cambridge University Press.

[74] Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. Studies in Second Language Acquisition, 35(2), 213–235. https://doi.org/10.1017/S0272263112000861

[75] Rost, M. (2011). Teaching and researching listening (2nd ed.). Routledge. https://doi.org/10.4324/9781315833705

[76] Rukthong, A. (2020). MC listening questions vs. integrated listening-to-summarize tasks: What listening abilities do they assess? System, 94, Article 102439. https://doi.org/10.1016/j.system.2020.102439

[77] Scharenborg, O., & van Os, M. (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. Speech Communication, 108, 53–64. https://doi.org/10.1016/j.specom.2019.03.001

[78] Schmidt, E., & Pastorino-Campos, C. (2024). Eye tracking and EEG in language assessment. In G. Yu & J. Xu (Eds.), Language test validation in a digital age. Cambridge University Press.

[79] Suvorov, R. (2015a). Interacting with visuals in L2 listening tests: An eye-tracking study (AR-A/2015/1). University of Hawai'i at Mānoa.

[80] Suvorov, R. (2015b). 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

[81] Suvorov, R. (2018). Investigating test taking strategies during the completion of computer delivered items from the Michigan English Test (MET): Evidence from eye tracking and cued retrospective reporting.

[82] Suvorov, R. (2022). Listening: Exploring the underlying processes. In L. Gurzynski-Weiss & Y. Kim (Eds.), Instructed second language acquisition research methods (pp. 257–279). Routledge.

[83] Suvorov, R. (2024). The use of eye tracking in validating L2 listening assessments (pp. 43–66). In G. Yu & J. Xu (Eds.), Language test validation in a digital age. Cambridge University Press.

[84] Suvorov, R., & Irgin, P. (2026). Technology in listening assessment. In M. Reed & J. M. Levis (Eds.), The handbook of second language listening (pp. 185–198). Wiley. https://doi.org/10.1002/9781394312375.ch14

[85] Taylor, L., & Geranpayeh, A. (2011). Journal of English for Academic Purposes, 10(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

[86] Tóthová, M., & Rusek, M. (2025). Eye tracking in science education research: Comprehensive literature review. Science & Education. Advance online publication. https://doi.org/10.1007/s11191-025-00644-1

[87] Vandergrift, L. (2011). Second language listening: Presage, process, product, and pedagogy. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (Vol. 2, pp. 455–471). Routledge.

[88] Vandergrift, L., & Goh, C. C. M. (2012). Teaching and learning second language listening: Metacognition in action. Routledge.

[89] Van Engen, K. J., & McLaughlin, D. J. (2018). Eyes and ears: Using eye tracking and pupillometry to understand challenges to speech recognition. Hearing Research, 369, 56–66. https://doi.org/10.1016/j.heares.2018.04.013

[90] Wagner, E. (2013). Assessing listening. In A. J. Kunnan (Ed.), The companion to language assessment (Vol. 1). Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla094

[91] Weir, C. J. (2005). Language testing and validation: An evidence-based approach. Palgrave Macmillan.

[92] Winke, P., & Lim, H. (2014). The effects of testwiseness and test taking anxiety on L2 listening test performance: A visual (eye tracking) and attentional investigation (IELTS Research Report No. 3). British Council, Cambridge English Language Assessment & IDP: IELTS Australia.

[93] Winke, P. M., Godfroid, A., & Gass, S. M. (2013). Introduction to the special issue: Eye-movement recordings in second language research. Studies in Second Language Acquisition, 35(2), 205–212. https://doi.org/10.1017/S0272263112000860

[94] Zhai, J., & Aryadoust, V. (2022). The metacognitive and neurocognitive signatures of test methods in academic listening. Frontiers in Psychology, 13, 930075. https://doi.org/10.3389/fpsyg.2022.930075

[95] Zhang, C. (2023). Using eye-tracking and retrospective verbal reports to explore the cognitive processes of banked gap-filling: A case study featuring methodological triangulation. Language Testing in Asia, 13, Article 22. https://doi.org/10.1186/s40468-023-00234-4