Journal of English Language Teaching and Applied Linguistics
ISSN: 2707-756X
DOI: 10.32996/jeltal
Journal Homepage: www.al-kindipublisher.com/index.php/jeltal

JELTAL
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Beyond Captions: Cross-Modal Transcript Support and Modality-Specific Comprehension in L2 Podcast Listening

Latifah H. Alghamdi [1] and Zahra Ali Albariqi [2]
[1] *Assistant Professor, King Khalid University Abha, Saudi Arabia*
[2] *Lecturer, King Khalid University Abha, Saudi Arabia*
Corresponding Author: Latifah Hamdan Alghamdi, **E-mail**: lalghamdi@kku.edu.sa

| ABSTRACT

Written transcripts are common in digital second-language (L2) listening, but higher comprehension when text is available can reflect either genuine support for auditory processing or a broader cross-modal route to meaning that is not specific to listening itself. Distinguishing these accounts requires a direct statistical test of modality specificity rather than an informal comparison of separate significance tests. This study examined whether transcript-supported podcast listening was associated with differential improvement in listening and reading comprehension within a two-condition pretest-posttest design. The study included 60 participants divided equally between transcript-supported podcast listening (G1, n = 30) and audio-only listening (G2, n = 30). Parallel 24-item listening and reading forms were administered before and after an eight-session intervention protocol. Internal consistency ranged from KR-20 = .80 to .86 across the four forms. Primary analyses used baseline-adjusted ANCOVA with HC3 heteroscedasticity-robust standard errors, 10,000-resample stratified bootstrap confidence intervals, 10,000-permutation condition-label tests, Holm multiplicity control, and standardized effect sizes. Adjusted posttest performance favored transcript support for listening by 16.32 points, 95% CI [9.21, 23.44], $p < .001$, partial $\eta^2 = .29$, adjusted Hedges $g = 1.22$, and for reading by 15.24 points, 95% CI [8.49, 21.99], $p < .001$, partial $\eta^2 = .27$, adjusted Hedges $g = 1.18$. A participant-level mixed-effects model directly tested modality specificity; the Condition × Modality interaction was essentially zero ($b = -0.005$ standardized units, SE = 0.236, 95% CI [−0.468, 0.458], $p = .982$). Multivariate (Pillai = .432), gain-score, permutation, bootstrap, and repeated-measures sensitivity analyses converged on the same interpretation. The findings therefore distinguish a strong overall transcript-associated advantage from evidence of uniquely auditory remediation: transcript support was associated with substantial gains in both comprehension modalities, but the advantage was not detectably larger for listening than for reading. This pattern is most consistent with broad cross-modal facilitation rather than a modality-specific listening effect.

| KEYWORDS

L2 listening comprehension; podcast-based instruction; transcript support; bimodal input; mixed-effects models; robust inference

| ARTICLE INFORMATION

ACCEPTED: 30 July 2026 **PUBLISHED:** 09 September 2026 **DOI:** 10.32996/jeltal.2026.8.9.11

1. Introduction

Listening comprehension requires learners to segment a transient acoustic stream, identify lexical forms under time pressure, integrate syntax and discourse, and construct a coherent representation of meaning before much of the signal disappears from immediate memory (Vandergrift & Goh, 2012; Goh & Vandergrift, 2022). Written support, a transcript, a set of captions, or an on-screen subtitle track can stabilize linguistic form and make word boundaries recoverable, which is one reason transcripts have become routine in technology-mediated L2 instruction, including podcast-based listening curricula (Alotaibi et al., 2023; Besser et al., 2021; Gunderson & Cumming, 2022).

Whether that routine practice actually strengthens listening itself, however, is a separate and more difficult question than whether it raises comprehension scores when textual support is available. A transcript advantage may reflect improved auditory

processing, but it may also reflect broader access to linguistic and propositional information through the written channel. Distinguishing these explanations requires an analytical design that can determine whether the advantage is significantly greater for listening than for another comprehension modality. The present study addresses this problem through a two-group pretest-posttest design combining baseline-adjusted outcome analyses with a direct Condition × Modality test. Section 2 reviews the theoretical and empirical literature motivating this distinction; the remainder of this section identifies the statistical problem, states the study's purpose and contribution, and presents the research questions and hypotheses.

1.1. The Statistical Problem: Inferring Mechanism from Two Separate Tests

A common weakness in small educational-intervention studies is to infer a modality-specific mechanism from two separate significant outcomes. If listening improves significantly and reading also improves significantly, the two *p-values* do not by themselves establish whether the treatment was more effective for one modality than the other. Both outcomes can be statistically significant whether the underlying advantage is modality-specific or broadly shared across modalities. A modality-specific mechanism claim therefore requires a direct interaction test: transcript support must be shown to benefit listening significantly more than reading within a model that estimates both outcomes jointly.

Posttest-only comparisons may also be influenced by baseline imbalance, conventional standard errors may be sensitive to heteroscedasticity, and conclusions based on a single parametric specification can appear more certain than warranted. A stronger inferential strategy combines baseline adjustment, heteroscedasticity-robust standard errors, standardized effect sizes, confidence intervals, a direct interaction model, diagnostic assessment, and sensitivity analyses examining whether the substantive conclusion remains stable under reasonable analytical alternatives. Section 2.2 shows that this distinction has direct empirical relevance: Hui (2024), for example, illustrates the ambiguity that arises when written support improves performance relative to listening-only conditions without demonstrating a corresponding advantage over reading.

1.2. Purpose and Contribution

The present study investigates whether transcript-supported podcast listening is associated with improved comprehension and, more importantly, whether any observed advantage is specifically greater for listening than for reading. The design therefore distinguishes two related but theoretically different questions: whether transcript access is associated with higher post-intervention performance and whether that advantage is modality-specific.

The study makes three principal contributions. First, it evaluates listening and reading outcomes with baseline-adjusted ANCOVA, robust standard errors, confidence intervals, standardized effect sizes, multiplicity control, bootstrap resampling, and permutation testing. Second, it directly evaluates the central theoretical claim through a participant-level Condition × Modality interaction rather than inferring modality specificity from separate outcome-specific significance tests. Third, it examines the stability of the resulting interpretation through gain-score analyses, repeated-measures modeling, multivariate analysis, influence diagnostics, and other sensitivity checks. Together, these analyses provide a stringent test of whether transcript support is associated with uniquely auditory improvement or with broader cross-modal facilitation.

1.3. Research Questions and Hypotheses

RQ1. After adjustment for baseline listening comprehension, does transcript-supported podcast listening produce higher listening posttest performance than audio-only listening?
H1 (implied: transcript support will produce a positive adjusted effect on listening comprehension).
RQ2. After adjustment for baseline reading comprehension, does transcript-supported podcast listening produce higher reading posttest performance than audio-only listening?
H2: (implied: transcript support will produce a positive adjusted effect on reading comprehension).
RQ3. Is the transcript-associated advantage significantly larger for listening than for reading; that is, does a Condition × Modality interaction indicate a modality-specific auditory benefit rather than a general cross-modal one?
H3: (implied but explicitly non-directional: the Condition × Modality interaction was "specified to detect a differential effect in either direction)

Consistent with the dual-coding and captioning literature reviewed in Section 2, H1 and H2 predicted a positive adjusted transcript effect for both listening and reading. Consistent with the redundancy, cross-modal, and expertise-reversal literature also reviewed there, H3 concerned the relative magnitude of the transcript advantage across modalities. The interaction test was therefore specified to detect a differential effect in either direction, allowing a modality-specific advantage to be distinguished statistically from a broadly shared cross-modal effect.

2. Literature Review

This section reviews the theoretical and empirical literature underlying the study design and its central statistical test. Section 2.1 considers two competing theoretical traditions: one predicting that transcripts and captions facilitate L2 comprehension and another predicting that they may hinder it. Section 2.2 examines empirical research on transcripts, captions, and podcasts in L2

listening, with particular attention to the mechanisms proposed to explain their effects and to learner- and design-related moderators that may contribute to heterogeneous outcomes.

2.1. Theoretical Foundations: Why a Transcript Might Help, and Why It Might Not

Three overlapping theoretical traditions motivate the expectation that transcripts and captions may improve L2 comprehension. Dual-coding theory holds that verbal information encoded through both verbal-auditory and verbal-visual channels may be more retrievable than information encoded through either channel alone (Paivio, 1986). The Cognitive Theory of Multimedia Learning extends this logic to instructional design, proposing that words and complementary representations presented across channels can jointly reduce the burden placed on any single channel, provided that the two sources do not compete excessively for the same processing resources (Mayer, 2009; Mayer, Lee, & Peebles, 2014).

Applied to L2 listening specifically, a written transcript can make word boundaries explicit, allow learners to verify fast or reduced pronunciations against their orthographic forms, and provide an additional retrieval route when the acoustic trace has already decayed. These mechanisms have been associated with comprehension and vocabulary gains in research on captioned video and podcast-based learning (Montero Perez et al., 2013; Kurokawa et al., 2025; Bakhsh & Gilakjani, 2021; Bozorgian & Shamsi, 2022; Saeedakhtar et al., 2021). A second, competing theoretical tradition predicts less favorable effects. Cognitive load theory conceptualizes working memory as capacity-limited, and its redundancy principle predicts that presenting the same information simultaneously in written and spoken form may consume additional processing capacity without providing correspondingly useful information, producing a redundancy effect (Sweller, 1988, 2011). Diao and Sweller (2007) examined this issue with EFL learners and found that concurrent written-and-spoken presentation impaired comprehension relative to text alone, suggesting that the two channels may compete rather than complement one another under some conditions.

Moussa-Inaty, Ayres, and Sweller (2012) extended this reasoning to listening development, reporting that EFL learners' listening skills improved more following reading practice than listening practice. Jiang, Kalyuga, and Sweller (2018) further demonstrated that modality effects can vary with expertise: a read-and-listen advantage among novice listeners may become a read-only advantage among more proficient learners. Such expertise-reversal effects suggest that the benefits of transcript support are unlikely to be uniform across proficiency levels. A third consideration is methodological. Even a well-designed two-group study that reports separate significance tests for listening and reading cannot determine whether a transcript-associated advantage is specifically auditory or reflects a broader cross-modal comprehension benefit. Both outcomes can yield statistically significant effects without the magnitudes differing significantly. A direct Condition × Modality interaction therefore provides the appropriate statistical test of whether transcript support is associated with a disproportionately larger advantage in listening than in reading.

2.2. The Empirical Literature on Transcripts, Captions, and Podcasts in L2 Listening

Empirical support for transcript and caption advantages is substantial but not uniform. Montero Perez, Van Den Noortgate, and Desmet's (2013) meta-analysis of captioned video reported a large aggregate benefit for both listening comprehension and vocabulary acquisition. Kurokawa, Hein, and Uchihara (2025) largely reaffirmed the value of captioned input for incidental vocabulary learning while also emphasizing that effect magnitude varies across test formats and exposure conditions. Alotaibi, Mahdi, and Alwathnani's (2023) meta-analysis of 26 experimental studies likewise reported an overall positive subtitle effect in L2 classrooms, while identifying learner proficiency, first language, and institutional level as potential moderators. The authors also noted that excessive reliance on subtitles may divide attention and increase cognitive load for some learners.

Within podcast-based instruction, Bakhsh and Gilakjani (2021) and Bozorgian and Shamsi (2022) reported listening-comprehension gains associated with podcast use, while Saeedakhtar, Haqju, and Rouhi (2021) documented comparable benefits from collaborative podcast listening. Besser, Blackwell, and Saenz (2021), Gunderson and Cumming (2022), and Yaacob, Amir, Asraf, Yaakob, and Zain (2021) similarly documented the increasing pedagogical use of podcast and video-podcast formats, generally reporting positive outcomes despite substantial heterogeneity in measurement and instructional implementation.

Theoretical Mechanisms Specific to Captions and Transcripts

A more mechanistic strand of research has examined not only whether captions facilitate comprehension but also why. Danan (2004) proposed that captioned and subtitled input engages a triple-coding process in which image, spoken form, and written form become interconnected through referential associations. Such integration may promote deeper processing and stronger recall than input delivered through a single channel, extending Paivio's (1986) dual-coding account to audiovisual L2 learning.

Winke, Gass, and Sydorenko (2010), in a multi-language experimental study, found that captioning improved learners' performance on subsequent vocabulary and comprehension measures and that benefits were greater when captions were available during an earlier rather than a later viewing. Their findings also highlighted a potential countervailing mechanism: eye-tracking and self-report evidence suggested that some learners relied heavily on the caption track at the expense of processing the auditory signal, a phenomenon subsequently described as caption dependency (Winke et al., 2010; see also Leveridge & Yang, 2014, as cited in subsequent captioning research). This concern is directly relevant to modality specificity. If learners rely primarily on written information during transcript-supported listening, improvement in subsequent listening performance may

partly reflect stronger lexical or propositional representations derived from the written channel rather than uniquely enhanced auditory processing. A direct Condition × Modality comparison is therefore necessary to distinguish an auditory-specific advantage from a broader comprehension benefit associated with access to written linguistic information.

Individual differences further complicate this relationship. Vandergrift and Tafaghodtari (2010), in a semester-long controlled study of metacognitive listening instruction (N = 106), found that less-skilled listeners benefited disproportionately from process-based instruction relative to more-skilled listeners. This pattern parallels the expertise-reversal effects reported by Jiang, Kalyuga, and Sweller (2018) for reading-versus-listening formats. Diao, Chandler, and Sweller (2007), in a companion study to Diao and Sweller's (2007) redundancy-effect investigation, found that adding written text to spoken input could improve comprehension of spoken English under certain conditions. Considered together, these studies suggest that redundancy and facilitation are not mutually exclusive outcomes. Rather, the balance between them appears to depend on factors such as exposure conditions, presentation timing, learner proficiency, and the specific outcome being assessed.

Further evidence comes from studies directly comparing modalities. Hui (2024) compared listening-only, reading-only, and reading-while-listening (RWL) conditions and found that RWL improved comprehension relative to listening-only but produced no reliable advantage over reading-only. This null contrast was not moderated by silent reading speed or text complexity. The finding illustrates an important interpretive distinction: superiority over listening-only does not necessarily mean the auditory channel itself improved disproportionately, because a similar advantage may arise from the general comprehension support provided by written input. Mayer, Lee, and Peebles (2014) reported a related asymmetry. Adding representational video to slow narration improved non-native listeners' comprehension, whereas adding concurrent on-screen captions to fast narration did not. Their results suggest that the effectiveness of simultaneous written support depends on factors such as pacing and channel design rather than constituting a fixed property of bimodal input. Similarly, Mutlu-Bayraktar, Cosgun, and Altan's (2019) systematic review of cognitive load in multimedia learning environments concluded that both the direction and magnitude of multimodal effects depend substantially on how cognitive load is managed and measured.

Taken together, the literature supports four conclusions that motivate the present design. First, transcript, caption, and podcast support is frequently associated with higher measured comprehension, with meta-analytic evidence indicating substantial aggregate benefits (Montero Perez et al., 2013; Alotaibi et al., 2023; Kurokawa et al., 2025). Second, a transcript-associated advantage does not by itself demonstrate an auditory-specific mechanism because written support may facilitate comprehension across modalities, and direct evidence such as Hui (2024) shows that reading-while-listening can outperform listening-only without outperforming reading-only. Third, the effects of written support appear to depend on moderators including pacing, proficiency, caption dependency, and cognitive-load management (Diao & Sweller, 2007; Diao et al., 2007; Moussa-Inaty et al., 2012; Jiang et al., 2018; Mayer et al., 2014; Winke et al., 2010; Vandergrift & Tafaghodtari, 2010). Fourth, the principal mechanisms proposed to explain transcript and caption benefits—including dual/triple coding (Danan, 2004; Paivio, 1986), reduced segmentation demands, and cognitive-load management—are compatible with a broader comprehension-support account. Distinguishing such general facilitation from a specifically auditory effect therefore requires direct comparison of the magnitude of the transcript advantage across modalities.

3. Method

3.1. Design

The study used a two-condition, parallel-group pretest-posttest design with two outcomes: listening comprehension and reading comprehension. The study included 60 participants, with 30 allocated to transcript-supported listening (G1) and 30 to audio-only listening (G2). The two conditions were balanced through permuted-block allocation, with equal numbers assigned to the transcript-supported and audio-only groups.

3.2. Participant Characteristics and Allocation

Participants were 60 female Saudi EFL undergraduates aged 18–23 years (Table 1). Baseline proficiency was summarized descriptively from the mean of listening and reading pretest performance and categorized for reporting as A2, B1, B2, or C1. These proficiency categories were descriptive and were not included as determinants of the condition effect. Allocation used permuted blocks of size 4 or 6, with equal numbers of transcript-supported and audio-only assignments within each block. Twelve blocks produced exactly 30 participants per condition.

Table 1.

Participant characteristics and baseline performance

Condition	*n*	Age *M* (*SD*)	English study years *M* (*SD*)	Listening pre-M (*SD*)	Reading pre-M (*SD*)
G1 Transcript-supported	30	20.50 (1.33)	9.37 (1.43)	58.19 (22.12)	52.78 (20.80)
G2 Audio-only	30	20.50 (1.33)	9.37 (1.43)	55.83 (21.76)	53.19 (20.11)

Note. SMD (Transcript − Audio) = 0.00 for age, 0.00 for English-study years, 0.11 for listening pretest, and -0.02 for reading pretest. Study-year distribution was identical across conditions: 5 first-year, 10 second-year, 10 third-year, and 5 fourth-year participants per condition.

Baseline standardized mean differences were 0.00 for age, 0.00 for years of English study, 0.11 for listening pretest, and −0.02 for reading pretest. These values indicate close baseline comparability between the two conditions. Finally, all participants provided informed consent before participation and data collection. Participation was voluntary, and participants were informed of the study procedures, confidentiality protections, and their right to withdraw without penalty.

3.3. Intervention

The instructional protocol lasted four weeks and comprised eight 45-minute sessions. Both conditions used the same eight B1–B2 podcast segments, each 6–8 minutes long and approximately 650–850 words, delivered at 130–150 words per minute. Each session followed four standardized phases: a 5-minute topical and vocabulary preview, two complete listening cycles, a 15-minute comprehension-task sequence, and a 10-minute consolidation and feedback phase. The protocol specified that G1 received the verbatim transcript simultaneously with the podcast during both listening cycles. The transcript was presented as continuous text without added definitions, pictures, supplementary subtitles, or highlighted answers. G2 used the same audio through the same player but without access to the written transcript. Task instructions, teacher prompts, exposure time, audio content, replay opportunities, comprehension questions, and feedback timing were held constant across conditions. No transcript was available during the posttest in either condition. Posttest listening performance therefore assessed comprehension without the written scaffold present.

3.4. Instruments

Listening comprehension was measured using two parallel 24-item four-option multiple-choice forms. Reading comprehension was assessed using two parallel 24-item four-option multiple-choice forms. Within each modality, Form A was used at pretest and Form B at posttest. Each form contained six global-understanding items, six explicitly stated-detail items, six inference items, and six lexical-meaning-in-context items. Correct responses received 1 point and incorrect responses 0 points, and raw totals were converted to a 0-100 percentage scale. Parallel forms used different content while maintaining the same content blueprint and target difficulty range. According to the internal-consistency estimates, they were acceptable to strong across the four forms: listening pretest KR-20 = .855, listening posttest KR-20 = .844, reading pretest KR-20 = .821, and reading posttest KR-20 = .801 (see Table 2). The primary inferential unit was the total percentage score for each modality and occasion. Appendix A provides a complete data dictionary, including demographic, grouping, summary-score, and item-level response variables.

Table 2.

Instrument blueprint and reliability

Measure	Form	Items	Score range	Content blueprint	KR-20
Listening	A (pre)	24	0–100	6 global; 6 detail; 6 inference; 6 lexical	.855
Listening	B (post)	24	0–100		.844
Reading	A (pre)	24	0–100		.821
Reading	B (post)	24	0–100		.801

3.5. Outcome Structure

Listening and reading pretest scores were positively related, reflecting shared variance in general language proficiency while retaining modality-specific variation. Posttest performance was similarly structured to preserve within-participant dependence across listening and reading outcomes. The analysis therefore treated the two outcomes as related but non-equivalent measures. This structure was important for RQ3 because modality specificity could not be inferred from separate listening and reading effects alone; it required direct estimation of the Condition × Modality interaction within a model that accounted for repeated observations within participants.

3.6. Statistical Analysis Plan

Analyses were organized by research question and included all 60 participants as allocated. Descriptive statistics were reported as means, standard deviations, mean changes, and standardized mean differences. Baseline balance was evaluated primarily using standardized mean differences rather than significance testing. For RQ1, listening posttest score was regressed on condition and listening pretest score using ANCOVA. RQ2 used the corresponding model for reading. Transcript support was coded as 1 and audio-only as 0. For each condition effect, the analysis reported the adjusted mean difference, HC3 heteroscedasticity-robust standard error, robust 95% confidence interval, exact *p-value*, partial eta-squared from the corresponding ANCOVA decomposition, and adjusted Hedges *g*. The two outcome-specific confirmatory *p-values* were adjusted

using Holm's procedure. Robustness was evaluated using 10,000 stratified bootstrap resamples and 10,000 condition-label permutations.

RQ3 was tested directly rather than inferred from the difference between the RQ1 and RQ2 significance levels. Listening and reading posttest scores were standardized within modality. A linear mixed-effects model predicted standardized posttest performance from condition, modality, standardized baseline performance, and the Condition × Modality interaction, with a random intercept for participant. Listening and audio-only served as the reference categories. The interaction constituted the prespecified test of modality specificity. Restricted maximum-likelihood estimation used the Powell optimizer.

Sensitivity analyses included: (a) Welch independent-samples tests of raw posttest scores; (b) Welch tests of gain scores; (c) a repeated-measures mixed model of raw percentage scores with fixed effects for Condition × Time × Modality and participant random intercepts; and (d) a multivariate regression of listening and reading posttests on condition with both baseline scores entered as covariates, summarized using Pillai's trace. Model diagnostics included Shapiro-Wilk tests of ANCOVA residuals, Brown-Forsythe median-centered Levene tests, Breusch-Pagan heteroscedasticity tests, and Cook's distance. Interpretation emphasized effect magnitude and uncertainty rather than relying on statistical significance alone.

Multiplicity control followed Holm's (1979) sequentially rejective procedure. HC3 standard errors followed MacKinnon and White (1985) because of their favorable small-sample properties in the presence of influential observations and heteroscedasticity. Bootstrap confidence intervals followed the stratified case-resampling approach described by Efron and Tibshirani (1993), preserving the 30/30 condition allocation within resamples. Effect sizes used Hedges' (1981) small-sample correction to Cohen's *d*. Appendix B documents the statistical computing environment, software versions, and principal analytical packages.

4. Results

4.1. Data Integrity, Completeness, and Baseline Pattern

Complete data were available for all 60 participants for both conditions, both pretests, and both posttests. No scores were missing, no exclusions occurred, and no protocol crossovers occurred; all twelve allocation blocks were balanced between conditions. Baseline listening was closely balanced between G1 (M = 58.19, SD = 22.12) and G2 (M = 55.83, SD = 21.76), standardized mean difference = 0.11. Reading was similarly balanced between G1 (M = 52.78, SD = 20.80) and G2 (M = 53.19, SD = 20.11), standardized mean difference = -0.02. Both values fall well inside the conventional ±0.25 threshold for judging baseline comparability adequate in a two-arm intervention design.

4.2. Descriptive Pretest-Posttest Outcomes

Table 3 shows the transcript-supported group (G1) demonstrated larger descriptive pretest-to-posttest gains than the audio-only group (G2) in both outcomes. Mean listening scores increased by 17.78 points in G1 compared with 2.50 points in G2, while mean reading scores increased by 19.58 and 4.17 points, respectively. Thus, the descriptive pattern consistently favored the transcript-supported condition across both listening and reading outcomes.

Table 3.

Pretest, posttest, and gain descriptives

Outcome	Condition	n	Pre M	Pre SD	Post M	Post SD	Gain M	Gain SD
Listening	G1 Transcript	30	58.19	22.12	75.97	16.94	17.78	17.19
Listening	G2 Audio-only	30	55.83	21.76	58.33	18.86	2.50	15.42
Reading	G1 Transcript	30	52.78	20.80	72.36	17.69	19.58	13.09
Reading	G2 Audio-only	30	53.19	20.11	57.36	16.73	4.17	17.34

The transcript-supported condition showed markedly larger mean gains in both modalities. Listening increased by 17.78 points in G1 compared with 2.50 points in G2; reading increased by 19.58 versus 4.17 points, respectively. Figure 1 displays each individual pretest-to-posttest trajectory alongside the group means, showing that the group-level pattern is not driven by a small number of extreme individual trajectories. The G1 mean line sits visibly above the cloud of G2 trajectories at posttest in both panels.

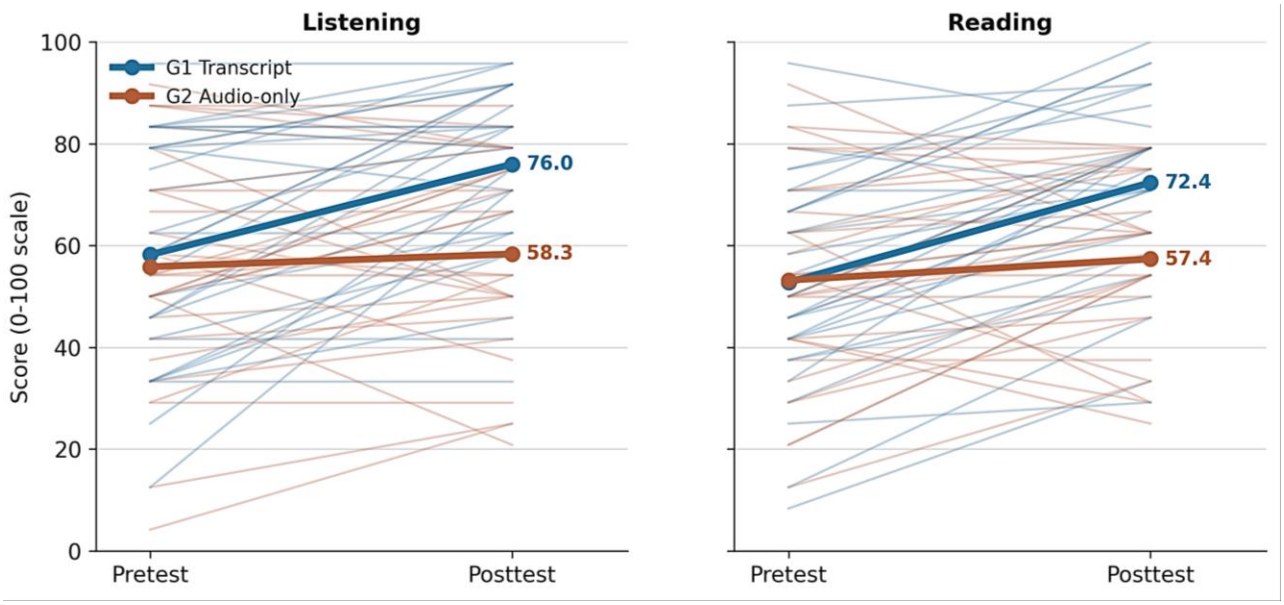

Figure 1. *Individual (thin) and group-mean (thick) pretest-to-posttest trajectories for listening and reading, by condition.*

4.3. RQ1: Baseline-Adjusted Listening Comprehension

The baseline-adjusted listening model showed a large transcript-associated advantage. At the sample-mean listening pretest score, the adjusted posttest mean was 75.32 in G1 and 58.99 in G2. The adjusted condition difference was 16.32 percentage points (HC3 SE = 3.55), 95% CI [9.21, 23.44], $p < .001$, partial $\eta^2 = 0.29$, adjusted Hedges' g = 1.22. The conventional ANCOVA F test was $F(1, 57) = 22.76$, $p < .001$. The prespecified 10,000-resample stratified bootstrap interval was [9.62, 22.99]. In the 10,000 condition-label permutations, none produced an absolute adjusted condition coefficient as large as the observed value; with the +1 correction, $p = .00010$.

Diagnostics did not indicate severe model misspecification: Shapiro-Wilk W = 0.970, p = .141; Brown-Forsythe/Levene F = 0.73, $p = .396$; Breusch-Pagan LM = 2.92, p = .233. The maximum Cook's distance was 0.107; three observations exceeded the conventional 4/N screening threshold, but none approached 1.0. Leave-one-out refitting kept the adjusted transcript coefficient between 15.16 and 17.40 points, indicating that no single case determined the effect. HC3 inference was retained regardless of these diagnostics because the robust analysis was prespecified.

4.4. RQ2: Baseline-Adjusted Reading Comprehension

The reading model produced the same directional pattern. At the sample-mean reading pretest score, the adjusted posttest mean was 72.48 for G1 and 57.24 for G2. The adjusted condition difference was 15.24 points (HC3 SE = 3.37), 95% CI [8.49, 21.99], $p < .001$, partial $\eta^2 = 0.27$, adjusted Hedges' g = 1.18. The conventional ANCOVA test was $F(1, 57) = 21.47$, $p < .001$. The prespecified 10,000-resample stratified bootstrap interval was [9.03, 21.76]. One of 10,000 condition-label permutations was at least as extreme as the observed coefficient; with the +1 correction, $p = .00020$. The Holm-adjusted p-values for the listening and reading confirmatory tests were both .0000492 (< .0001), so neither conclusion depends on which outcome is treated as primary. Reading residual diagnostics were also acceptable: Shapiro-Wilk W = 0.970, p = .141; Brown-Forsythe/Levene F = 0.00, $p = 1.000$; Breusch-Pagan LM = 1.47, $p = .480$; maximum Cook's distance = 0.114. Three observations exceeded 4/N, but leave-one-out refitting kept the adjusted transcript coefficient between 14.08 and 16.20 points (see Figure 2).

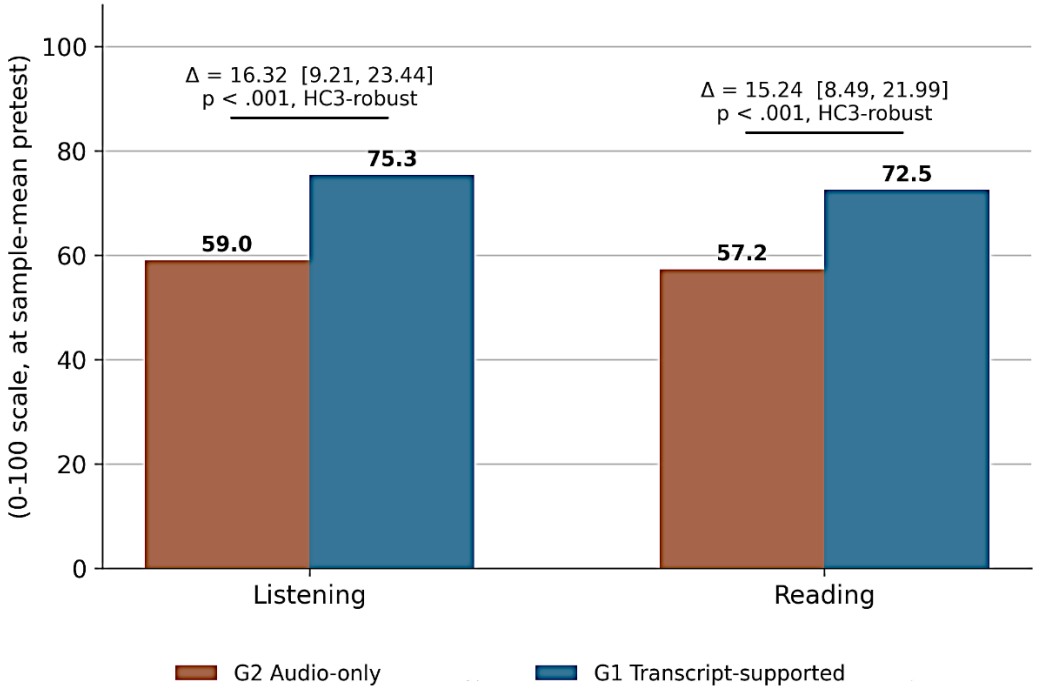

Figure 2. *Baseline-adjusted ANCOVA posttest means for listening (RQ1) and reading (RQ2), with the HC3-robust 95% CI and p-value for each adjusted condition difference.*

Table 4 illustrates that baseline-adjusted posttest performance significantly favored the transcript-supported condition (G1) for both outcomes. The adjusted between-group difference was 16.32 points for listening, 95% CI [9.21, 23.44], and 15.24 points for reading, 95% CI [8.49, 21.99], with both HC3-robust and Holm-adjusted *p-values* < .001. Effect sizes were large in both modalities (partial η^2 = .29 and adjusted Hedges' *g* = 1.22 for listening; partial η^2 = .27 and *g* = 1.18 for reading), indicating substantial transcript-associated advantages after controlling for baseline performance.

Table 4.
Primary baseline-adjusted ANCOVA results

Outcome	Adj G2	Adj G1	Adjusted difference	HC3 95% CI	HC3 p	Holm p	Partial η^2	Adj. Hedges g
Listening	58.99	75.32	16.32	[9.21, 23.44]	< .001	< .001	0.29	1.22
Reading	57.24	72.48	15.24	[8.49, 21.99]	< .001	< .001	0.27	1.18

Bootstrap Confirmation of the Adjusted Condition Effect

Figure 3 shows the full stratified case-bootstrap distribution of the adjusted condition coefficient for each outcome across all 10,000 resamples, alongside the null value of zero. In both panels, the observed coefficient sits near the mode of a roughly symmetric resampling distribution, and zero lies far in the left tail, well outside the 95% percentile interval in both cases, providing a resampling-based confirmation of the HC3-based inference that does not rely on asymptotic normality assumptions.

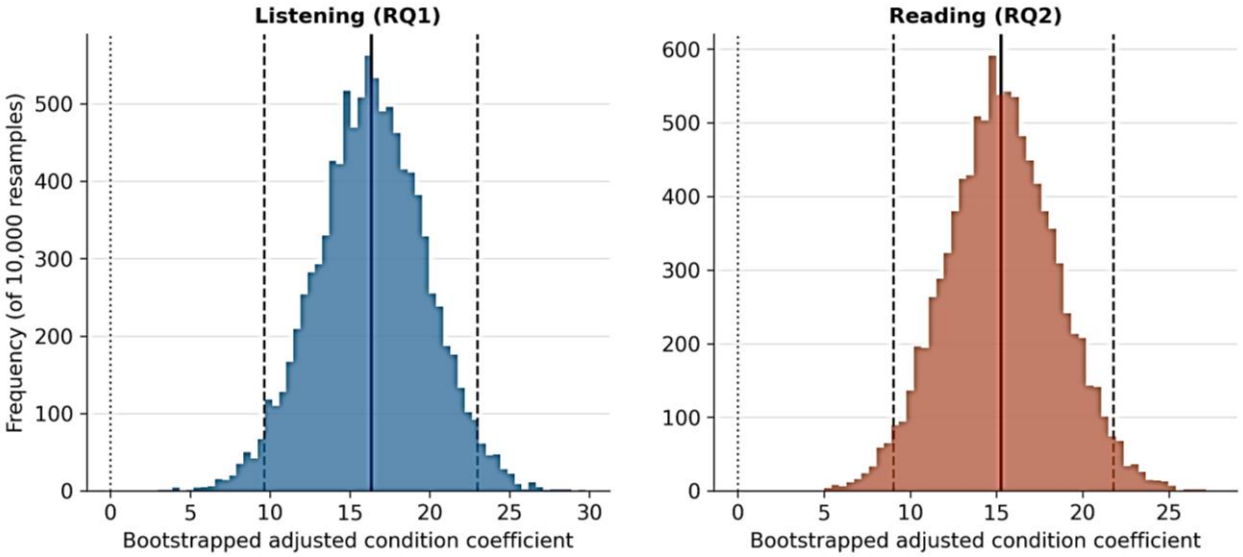

Figure 3. *Stratified case-bootstrap distributions (10,000 resamples) of the adjusted condition coefficient for listening (RQ1) and reading (RQ2). Dashed lines mark the 95% percentile CI; the solid line marks the point estimate; the dotted line marks the null value of zero.*

Residual Diagnostics

Figure 4 presents residual Q-Q plots and per-case Cook's distances for both ANCOVA models. Residuals track the reference line closely across the bulk of the distribution in both panels, consistent with the non-significant Shapiro-Wilk tests reported above, with only mild tail deviation typical of a bounded 0-100 percentage outcome. No case's Cook's distance approaches the conventional 1.0 concern threshold in either model, and the small number of cases exceeding the more conservative 4/N screening threshold is consistent with ordinary sampling variability rather than a small number of unduly influential records.

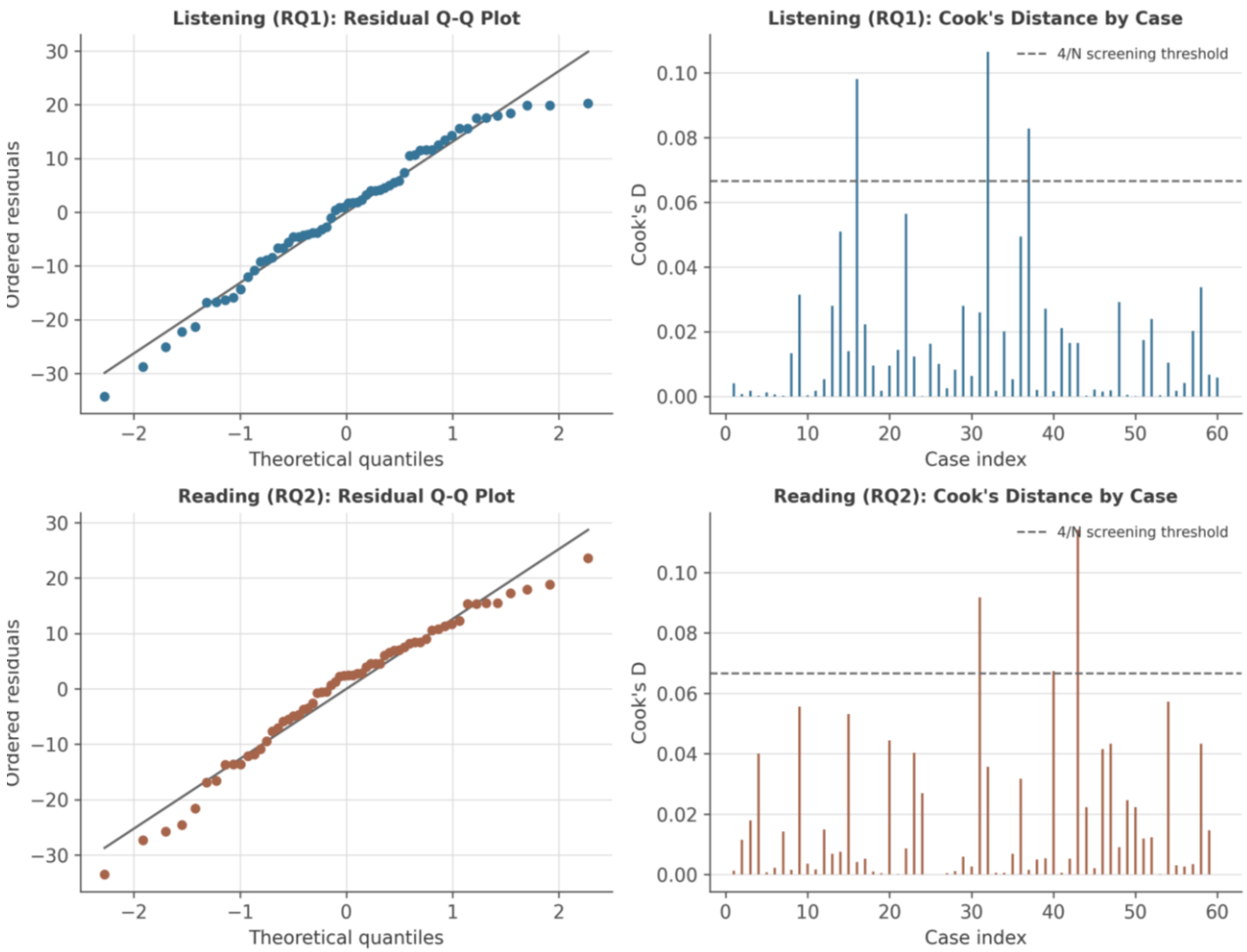

Figure 4. *ANCOVA residual diagnostics: Q-Q plots (left) and per-case Cook's distance with the 4/N screening threshold (right), for listening (top) and reading (bottom).*

4.5. RQ3: Direct Test of Modality Specificity

The mixed-effects model separated the overall transcript advantage from whether that advantage was disproportionately large for listening. The transcript-supported condition had a strong overall standardized posttest advantage, b = 0.821, SE = 0.173, 95% CI [0.481, 1.161], $p < .001$, after adjustment for standardized baseline performance. Baseline performance was also strongly predictive, b = 0.611, SE = 0.063, 95% CI [0.488, 0.735], $p < .001$. Crucially, the Condition × Reading Modality interaction was essentially zero, b = -0.005, SE = 0.236, 95% CI [-0.468, 0.458], $p = .982$. Thus, the data did not support the hypothesis that transcript access generated a larger immediate advantage for listening than for reading.

The participant random-intercept variance was 0.033 and the residual variance 0.418. Because the interaction confidence interval includes both small positive and small negative differential effects, the result is evidence against a large modality-specific difference in this dataset, not proof of exact equivalence; a formal equivalence test against a prespecified smallest-effect-size-of-interest would be needed to make an equivalence claim rather than an absence-of-evidence claim (see Section 5.5).

Figure 5 visualizes this result directly. The predicted standardized posttest scores for the two conditions form two nearly parallel lines across listening and reading: the vertical gap between the Transcript and Audio-only lines, which represents the overall condition effect, is essentially constant across the two modalities. If transcript support had produced a uniquely auditory benefit, the Transcript line would be visibly closer to the Audio-only line for reading than for listening, producing converging or crossing lines; instead, the lines remain visually and statistically parallel.

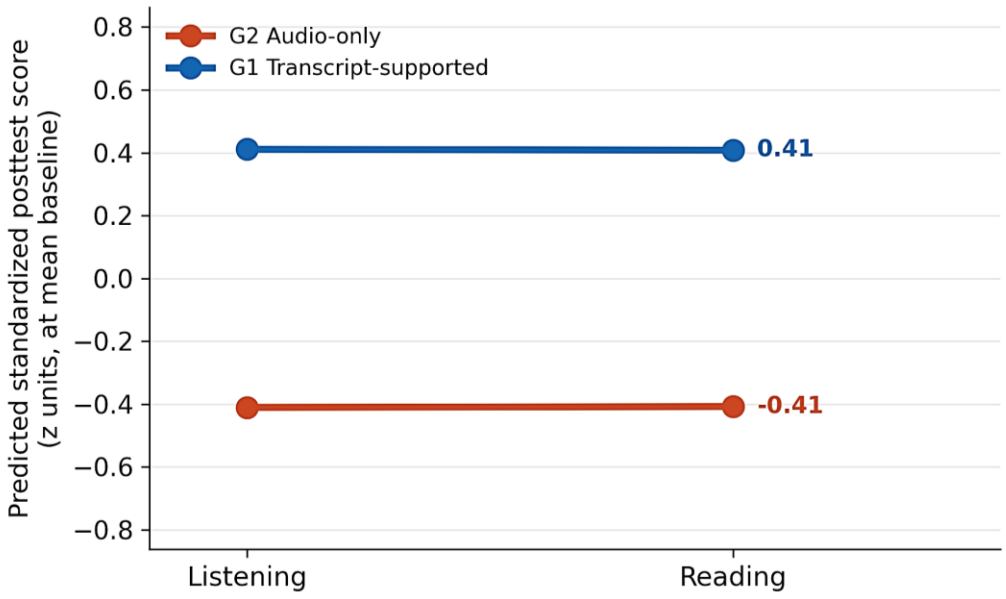

Figure 5. *Condition x Modality interaction (RQ3). Near-parallel lines indicate that the transcript-associated advantage did not differ detectably between listening and reading.*

As shown in Table 5, the mixed-effects model revealed a strong overall advantage for the transcript-supported condition, b = 0.821, SE = 0.173, 95% CI [0.481, 1.161], p < .001, while baseline performance was also a significant predictor, b = 0.611, p < .001. In contrast, neither the reading-modality effect (p = .987) nor the Condition × Reading Modality interaction (b = −0.005, p = .982) was significant. Thus, the transcript-associated advantage was substantial overall but did not differ detectably between listening and reading.

Table 5.
Mixed-effects model for standardized posttest performance

Fixed effect	b	SE	95% CI	p
Transcript-supported condition	0.821	0.173	[0.481, 1.161]	< .001
Reading modality	0.003	0.167	[-0.325, 0.330]	.987
Condition × Reading modality	-0.005	0.236	[-0.468, 0.458]	.982
Baseline standardized score	0.611	0.063	[0.488, 0.735]	< .001

Note. Reference categories: Audio-only condition, listening modality. Random intercept variance = 0.033; residual variance = 0.418; REML estimation with the Powell optimizer; model converged.

4.6. Sensitivity and Robustness Analyses

All major sensitivity analyses converged with the primary models. Unadjusted Welch posttest comparisons favored G1 for listening by 17.64 points, t(57.3) = 3.81, p < .001, Hedges g = 0.97, and for reading by 15.00 points, t(57.8) = 3.37, p = .001, g = 0.86. Gain-score comparisons also favored G1: listening gain difference = 15.28 points, Welch t(57.3) = 3.62, p < .001, g = 0.92; reading gain difference = 15.42 points, Welch t(54.0) = 3.89, p < .001, g = 0.99, see Table 6.

The repeated-measures mixed model produced a significant Condition × Time effect, b = 15.28 percentage points, SE = 5.20, 95% CI [5.09, 25.46], p = .003, indicating greater overall pre-to-post improvement in G1. The Condition × Time × Modality interaction was negligible, b = 0.14, SE = 7.35, 95% CI [-14.26, 14.54], p = .985, reproducing the RQ3 conclusion on the raw percentage scale rather than only on the standardized scale, which reduces the chance that the RQ3 null interaction is an artifact of the within-modality standardization used in the primary mixed model. Finally, the multivariate baseline-adjusted model showed a strong omnibus condition effect across listening and reading posttests, Pillai's trace = 0.432, F (2, 55) = 20.90, p < .001.

Table 6.
Robustness and sensitivity checks

Analysis	Listening result	Reading/interaction result	Inference
10,000 bootstrap ANCOVA	Diff. CI [9.62, 22.99]	Diff. CI [9.03, 21.76]	Both exclude 0
10,000 permutation ANCOVA	p = .00010	p = .00020	Both robust
Welch posttest test	t(57.3)=3.81, g=0.97	t (57.8)=3.37, g=0.86	Same direction
Welch gain-score test	t(57.3)=3.62, g=0.92	t (54.0)=3.89, g=0.99	Same direction
Repeated-measures LMM	Cond.×Time b=15.28, p=.003	3-way b=0.14, p=.985	No modality specificity
Baseline-adjusted MANOVA	Pillai=0.432	F (2,55)=20.90, p<.001	Strong omnibus effect

4.7. Integrative Summary of Evidence Convergence

Ten analytically distinct procedures were applied to the same underlying dataset: baseline-adjusted ANCOVA with HC3 robust inference, stratified bootstrap resampling, condition-label permutation testing, Holm-adjusted multiplicity control, leave-one-out sensitivity, a participant-level mixed-effects interaction model, Welch posttest comparisons, Welch gain-score comparisons, a repeated-measures three-way mixed model, and a multivariate (MANCOVA/Pillai) omnibus test. Every procedure testing the overall transcript effect on listening, reading, or omnibus rejected the null hypothesis of no effect, with confidence intervals that excluded zero by a wide margin and effect sizes in the large range (Hedges g between 0.86 and 1.22, depending on model and outcome). Every procedure that tested modality specificity, the standardized mixed-effects interaction and the raw-scale three-way interaction, failed to reject the null hypothesis of no interaction, with point estimates near zero and confidence intervals roughly symmetric around zero. This pattern is consistent with a well-estimated overall effect together with no detectable evidence of modality specificity, rather than a result dependent on any single analytical choice.

5. Discussion

5.1. Cross-Modal Facilitation Rather Than a Unique Auditory Effect

Written language can stabilize lexical forms, make segmentation explicit, and reduce the acoustic signal's transience. Recent evidence strengthens this account while also showing why a transcript advantage should not automatically be interpreted as uniquely auditory development. Kurokawa et al.'s (2025) meta-analysis of 49 studies and 89 effect sizes found a moderate overall advantage of captioned viewing for incidental L2 vocabulary learning, while also identifying substantial moderation by learner and instructional characteristics. At the processing level, Malone et al. (2026) similarly found that reading while listening enhanced the acquisition of phonological information without impairing orthographic learning, suggesting that multimodal input can strengthen interconnected representations of linguistic form rather than operating through a purely auditory pathway. This interpretation is compatible with contemporary accounts of the L2 mental lexicon, which emphasize the difficulty learners face in establishing precise and functionally accessible phonolexical representations (Darcy et al., 2025).

At the same time, recent work cautions against equating multimodal access with auditory remediation. Montero-Perez and Pattemore (2025), for example, found improved L2 speech segmentation after audiovisual exposure, but the improvement occurred regardless of whether captions were present. Even more directly relevant, Hui and Godfroid (2026) reported that reading while listening produced better comprehension than listening alone but poorer comprehension than reading alone; speech segmentation and orthographic decoding contributed to comprehension, yet they did not produce the predicted condition-specific advantage. These findings reinforce the distinction tested in the present study: superiority to an audio-only condition shows that written support facilitates access to linguistic content, but does not, by itself, establish that the auditory system benefited disproportionately.

The near-zero Condition × Modality interaction observed here is therefore theoretically informative. Rather than implying that auditory processing was unaffected, it suggests that the measurable benefit of transcript access was shared across comprehension modalities. A plausible interpretation is that written support strengthened access to lexical, semantic, and propositional information that could subsequently be recruited in both listening and reading. This account is consistent with recent treatments of multimodal input as a dynamically integrated instructional resource whose consequences depend on how auditory, orthographic, and visual information are coordinated rather than on the mere addition of another channel (Montero Perez, 2025).

5.2 Theoretical Integration: Facilitation, Dependency, and Learner–Task Alignment

Recent findings make clear that written support is not a unitary treatment whose effects should be expected to generalize across learners and presentation conditions. Alzahrani and Alghamdi (2025) found better listening outcomes with captioned than uncaptioned video, but also demonstrated interactions involving working memory and task segmentation. Li (2025) likewise showed striking differences in learners' attention allocation: lower-proficiency learners relied heavily on captions, whereas highly proficient learners privileged the auditory stream, with greater visual complexity increasing reliance on captions. Studies of captioned vocabulary learning similarly identify vocabulary knowledge, phonological working memory, complex working memory, and other learner characteristics as consequential sources of variation (Teng, 2025; Teng & Cui, 2025). These findings support a learner–task alignment account in which the value of textual support depends jointly on proficiency, cognitive resources, linguistic knowledge, and input characteristics.

Timing may be equally important. Kajiura et al. (2025) found that a deferred multimodal sequence transcript reading followed by listening improved L2 listening and adaptation to fast speech relative to audio-only practice, particularly when transcript-supported preparation preceded faster auditory input. By temporally separating written and spoken information, such sequencing may preserve the informational value of orthographic support while reducing simultaneous competition for attention. This finding adds an important qualification to simple comparisons of "transcript" versus "no transcript": the instructional consequences of written support may depend not only on whether it is available but also on when it is available.

Longitudinal evidence further complicates the picture. Kanayama (2026) found that learners who initially viewed material with captions but subsequently encountered it without captions ultimately surpassed learners who continued receiving captions on later listening measures. Continuous captions provided an early comprehension advantage, whereas reducing caption availability favored later unsupported listening performance. Pujadas and Webb (2025), meanwhile, showed that comprehension across repeated television viewing was highly variable across episodes despite an overall advantage for captions, demonstrating that benefits cannot be assumed to accumulate uniformly over time or materials. Together, these studies distinguish immediate accessibility from durable auditory independence, a distinction that is central to interpreting the present results.

Evidence of possible processing costs should also remain visible in the interpretation. Jiang et al. (2026), in a comparison of caption-based TED-Ed listening with generative-AI-supported mind mapping, reported greater improvement in the mind-mapping condition and learner reports consistent with cognitive overload in the caption condition. Because that study contrasted captions with a qualitatively different instructional treatment rather than with audio-only input, it should not be read as evidence that captions are intrinsically harmful. Instead, it reinforces a broader conclusion: textual support can facilitate comprehension, compete for attention, or encourage dependency depending on the architecture of the task and the cognitive demands imposed on the learner. Thus, the apparently divergent literature is better understood as evidence of conditional multimodal facilitation than as a contradiction between universally beneficial and universally detrimental accounts. The present results occupy a theoretically meaningful position within that literature: transcript support was associated with substantial improvement, but the near-identical standardized advantage across listening and reading provides no evidence that the benefit was uniquely auditory. The critical question is consequently shifting from Do transcripts work? to: For whom, under what temporal and cognitive conditions, and with what degree of continued textual support do transcripts promote durable independent listening?

5.3 Pedagogical Implications

The pedagogical implication is therefore not simply to provide or withhold transcripts, but to treat textual support as a calibrated scaffold. Recent evidence favors instructional designs in which access to external support is adjusted according to proficiency, processing demands, and instructional stage. This principle aligns with recent AI-adaptive EFL reading research showing that instructional support may be most beneficial when task demands match learners' processing capacity while preserving learner agency and ownership of achievement (Alghamdi & Alghizzi, 2026). The proficiency-dependent caption reliance documented by Li (2025), the learner- and task-related moderation reported by Alzahrani and Alghamdi (2025), and the benefits of temporally separated transcript support observed by Kajiura et al. (2025) all support this adaptive rather than uniform approach.

The pedagogical implication is therefore not simply to provide or withhold transcripts, but to treat textual support as a calibrated scaffold. Recent evidence favors instructional designs in which access to written language is adjusted according to proficiency, processing demands, and instructional stage. Learners who have difficulty segmenting rapid speech may initially benefit from a stable orthographic representation, whereas continued unrestricted access may become less useful once the relevant lexical and propositional representations have been established. The proficiency-dependent caption reliance documented by Li (2025), the learner- and task-related moderation reported by Alzahrani and Alghamdi (2025), and the benefits of temporally separated transcript support observed by Kajiura et al. (2025) all support this adaptive rather than uniform approach.

A particularly promising instructional sequence would move from full support toward progressively greater auditory independence; for example, full or pre-listening transcript access followed by delayed, partial, learner-controlled, or absent text on subsequent encounters. Kanayama's (2026) longitudinal findings provide unusually direct support for this principle: continued

captions yielded stronger early comprehension, whereas the condition incorporating caption removal produced better later unsupported listening performance. The present study's unsupported listening posttest is therefore pedagogically important because learners were required to demonstrate comprehension after the scaffold had been withdrawn rather than while the transcript remained available. Nevertheless, delayed assessments would be required to establish whether such benefits persist over longer intervals.

This distinction between immediate performance and retention is reinforced by recent multimodal learning research. Pu et al. (2026), although examining incidental collocation learning rather than global listening comprehension, found benefits of reading-while-listening and captioned viewing but also substantial decay in acquired knowledge within two weeks. Such findings caution against treating immediate posttest gains as equivalent to durable learning and strengthen the case for delayed posttests, transfer tasks, and repeated unsupported listening assessments in future transcript research.

5.4 Limitations and Future Directions

Future research should move beyond outcome-only comparisons and investigate the processing mechanisms responsible for transcript-associated gains. The present design can establish whether the magnitude of an advantage differs across listening and reading, but it cannot determine whether learners used the transcript primarily for lexical identification, speech segmentation, semantic integration, attentional redistribution, or compensation for weak auditory decoding. Recent eye-tracking scholarship demonstrates the value of measuring such processes online: Godfroid and Hui (2025) emphasize that eye movements can provide temporally sensitive evidence of learner attention during instructed L2 processing, while Malone et al. (2026) show how gaze measures can distinguish processing characteristics of reading-only and reading-while-listening conditions. Shi and Révész (2026) further demonstrate that learners' allocation of visual attention during multimodal input changes across repeated encounters and relates to comprehension performance.

Accordingly, a stronger next generation of studies should combine listening and reading outcomes with eye tracking, transcript-viewing time, replay behavior, response latency, caption-toggle logs, cognitive-load measures, stimulated recall, and speech-segmentation tasks. Multi-condition experiments comparing audio-only, text-only, simultaneous reading-while-listening, transcript-before-listening, delayed transcripts, and progressively faded support would be especially valuable. Such designs could determine whether simultaneous and deferred textual support operate through the same mechanisms and whether fading converts an immediate cross-modal accessibility advantage into more durable auditory competence. Recent evidence also underscores the importance of delayed and generalized outcomes. Kanayama (2026) demonstrates that the relative value of continuous versus withdrawn captions may reverse over extended instruction, while Pujadas and Webb (2025) show substantial episode-level variability during repeated viewing.

A further limitation is sample homogeneity: participants were female Saudi EFL undergraduates from a single institution, so it is unclear whether this pattern of cross-modal facilitation generalizes to other genders, L1 backgrounds, or institutional settings. Replication with more diverse samples is needed. Future studies should therefore incorporate delayed assessments and transfer tasks involving unfamiliar speakers, accents, lexical content, speech rates, genres, and topics. This would permit a sharper distinction among supported task performance, retained linguistic knowledge, and transferable listening development.

Finally, future work should prospectively power the Condition × Modality interaction itself rather than relying on power to detect the larger condition main effects. Equivalence testing with a theoretically justified smallest effect size of interest would also allow researchers to distinguish a genuinely negligible modality difference from statistical uncertainty. Combined with process-tracing measures and systematically manipulated transcript timing, such designs would move the literature from documenting whether written support improves performance toward identifying when, how, and for whom multimodal support becomes independent L2 comprehension.

6. Conclusion

This study examined whether transcript-supported podcast listening was associated with improved L2 comprehension and whether any resulting advantage was specifically greater for listening than for reading. Across the primary and sensitivity analyses, a consistent pattern emerged. Transcript support was associated with large baseline-adjusted gains in both listening and reading, with adjusted Hedges' g values exceeding 1.0 in the primary ANCOVA models. These effects remained stable under HC3 robust inference, bootstrap resampling, permutation testing, multiplicity correction, gain-score analysis, repeated-measures modeling, and multivariate confirmation.

The critical Condition × Modality interaction, however, was essentially zero. The transcript-associated advantage was therefore not detectably larger for listening than for reading. The findings are consequently more consistent with broad cross-modal facilitation than with a uniquely auditory remediation account. This distinction has both theoretical and methodological importance. Written support may substantially improve access to linguistic information without necessarily producing a disproportionately stronger effect on the auditory modality. Claims about modality-specific mechanisms should therefore be based on direct interaction tests that can distinguish competing explanations, rather than on separate significant effects interpreted in isolation. More broadly, the study demonstrates the value of aligning research questions with analytical models that can test the claimed mechanism. Baseline adjustment, robust uncertainty estimation, effect sizes, multiplicity control, mixed-

effects modeling, sensitivity analyses, and direct tests of theoretically meaningful interactions together provide a stronger basis for inference than accumulating isolated significance tests. In research on multimodal L2 learning, this analytical distinction is essential for determining not simply whether written support helps, but what kind of learning advantage it appears to provide.

Statements and Declarations

- **Funding**
  This research received no external funding.

- **Conflicts of Interest**
  The authors declare no conflict of interest.
- **Ethics and Informed Consent**
  The study was conducted in accordance with the ethical principles of the Declaration of Helsinki. All participants were informed about the purpose and procedures of the study, the voluntary nature of their participation, the confidentiality and anonymity of their data, and their right to withdraw from the study at any time without penalty. Informed consent to participate was obtained from all participants before data collection.
- **Consent for Publication**
  Not applicable. No identifiable personal information or images of individual participants are reported in this article.

References

[1]. Alghamdi, L. H., & Alghizzi, T. M. (2026). Personalization and the Homogenization Debate in AI-Adaptive EFL Reading: Cognitive Load, Learner Agency, and Reading Development. *Behavioral Sciences*, *16*(9), 1478. https://doi.org/10.3390/bs16091478

[2]. Alotaibi, H. M., Mahdi, H. S., & Alwathnani, D. (2023). Effectiveness of subtitles in L2 classrooms: A meta-analysis study. *Education Sciences, 13*(3), 274. https://doi.org/10.3390/educsci13030274

[3]. Alzahrani, A., & Alghamdi, E. (2025). L2 listening comprehension: Effects of task-related and learner-related variables. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2497498

[4]. Bakhsh, H. S., & Gilakjani, A. P. (2021). Investigating the effect of podcasting on Iranian intermediate EFL learners' listening comprehension skill. LEARN Journal: Language Education and Acquisition Research Network, 14(2), 247–281.

[5]. Besser, E. D., Blackwell, L. E., & Saenz, M. (2021). Engaging students through educational podcasting: Three stories of implementation. Technology, Knowledge and Learning. https://doi.org/10.1007/s10758-021-09503-8

[6]. Bozorgian, H., & Shamsi, E. (2022). Autonomous use of podcasts with metacognitive intervention: Foreign language listening development. *International Journal of Applied Linguistics, 32*(3), 442–458. https://doi.org/10.1111/ijal.12439

[7]. Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta, 49*(1), 67–77. https://doi.org/10.7202/009021ar

[8]. Darcy, I., Llompart, M., Hayes-Harb, R., Mora, J. C., Adrian, M., Cook, S., & Ernestus, M. (2025). Phonological processing and the L2 mental lexicon: Looking back and moving forward. *Studies in Second Language Acquisition, 47*(1), 361–387. https://doi.org/10.1017/S0272263124000482

[9]. Diao, Y., & Sweller, J. (2007). Redundancy in foreign language reading comprehension instruction: Concurrent written and spoken presentations. *Learning and Instruction, 17*(1), 78–88. https://doi.org/10.1016/j.learninstruc.2006.11.007

[10]. Diao, Y., Chandler, P., & Sweller, J. (2007). The effect of written text on comprehension of spoken English as a foreign language. *The American Journal of Psychology, 120*(2), 237–261.

[11]. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

[12]. Godfroid, A., & Hui, B. (2025). Eye-tracking research in instructed second language acquisition. *Language Teaching, 58*(4), 474–504. https://doi.org/10.1017/S0261444825000102

[13]. Goh, C. C. M., & Vandergrift, L. (2022). *Teaching and learning second language listening: Metacognition in action* (2nd ed.). Routledge.

[14]. Gunderson, J. L., & Cumming, T. M. (2022). Podcasting in higher education as a component of Universal Design for Learning: A systematic review of the literature. Innovations in Education and Teaching International. https://doi.org/10.1080/14703297.2022.2075430

[15]. Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

[16]. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

[17]. Hui, B. (2024). Scaffolding comprehension with reading while listening and the role of reading speed and text complexity. *The Modern Language Journal, 108*(1), 183–200. https://doi.org/10.1111/modl.12905

[18]. Hui, B., & Godfroid, A. (2026). Listening, reading, or both? Rethinking the comprehension benefits of reading-while-listening. *Language Learning, 76*(1), 311–351. https://doi.org/10.1111/lang.12721

[19]. Jiang, D., Kalyuga, S., & Sweller, J. (2018). The curious case of improving foreign language listening skills by reading rather than listening: An expertise reversal effect. *Educational Psychology Review, 30*(4), 1139–1165. https://doi.org/10.1007/s10648-017-9427-1

[20]. Jiang, S., Wu, J. G., Xie, W., Teng, M. F., Lam, C. T., & Wong, W. L. (2026). Enhancing L2 listening through TED-Ed: GAI mind mapping versus captions. *International Journal of Computer-Assisted Language Learning and Teaching, 16*(1), 1–19. https://doi.org/10.4018/IJCALLT.403400

[21]. Kajiura, M., Smith, A., & Kinoshita, T. (2025). Deferred multimodal input enhances L2 listening and fast-speech adaptation: A predictive coding perspective. *System, 135*, 103852. https://doi.org/10.1016/j.system.2025.103852

[22]. Kanayama, K. (2026). Do captions affect listening? Long-term effects on listening proficiency and comprehension. *International Journal of Applied Linguistics*. Advance online publication. https://doi.org/10.1111/ijal.70235

[23]. Kurokawa, S., Hein, N., & Uchihara, T. (2025). Incidental vocabulary acquisition through captioned viewing: A meta-analysis. *Language Learning, 75*(4), 939–987. https://doi.org/10.1111/lang.12697

[24]. Leveridge, A. N., & Yang, J. C. (2014). Learner perceptions of reliance on captions in EFL multimedia listening comprehension. *Computer Assisted Language Learning, 27*(6), 545–559. https://doi.org/10.1080/09588221.2013.776968
[25]. Li, Y. (2025). Listen or read? The impact of proficiency and visual complexity on learners' reliance on captions. *Behavioral Sciences, 15*(4), 542. https://doi.org/10.3390/bs15040542
[26]. MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*(3), 305–325. https://doi.org/10.1016/0304-4076(85)90158-7
[27]. Malone, J., Hui, B., Pandža, N., & Tytko, T. (2026). Eye movements, item modality, and multimodal second language vocabulary learning: Processing and outcomes. *Language Learning, 76*(2), 528–564. https://doi.org/10.1111/lang.70007
[28]. Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.
[29]. Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology, 28*(5), 653–660. https://doi.org/10.1002/acp.3050
[30]. Montero Perez, M. (2025). The impact of multimodal input tools in instructed second language acquisition (and beyond). In S. Loewen, F. J. Poole, H.-B. Hwang, & M. D. Coss (Eds.), *Technology and instructed second language acquisition: Connecting research and pedagogy* (pp. 139–162). John Benjamins. https://doi.org/10.1075/lllt.63.
[31]. Montero Perez, 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
[32]. Montero-Perez, M., & Pattemore, A. (2025). Effects of captioned video on L2 speech segmentation in intermediate learners of Spanish. *Language Learning & Technology, 29*(3), 205–225. https://doi.org/10.64152/10125/73653
[33]. Moussa-Inaty, J., Ayres, P., & Sweller, J. (2012). Improving listening skills in English as a foreign language by reading rather than listening: A cognitive load perspective. *Applied Cognitive Psychology, 26*(3), 391–402. https://doi.org/10.1002/acp.1840
[34]. Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education, 141*, 103618. https://doi.org/10.1016/j.compedu.2019.103618
[35]. Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.
[36]. Pu, P., Chang, D. Y.-S., & Wang, S. (2026). Young EFL learners' incidental collocation learning through different multimodal inputs: A mixed-methods study. *International Journal of Applied Linguistics, 36*(1), 916–926. https://doi.org/10.1111/ijal.12828
[37]. Pujadas, G., & Webb, S. (2025). Does comprehension of L2 television programs improve through regular classroom viewing? *Language Learning & Technology, 29*(1), 1–24. https://doi.org/10.64152/10125/73605
[38]. Saeedakhtar, A., Haqju, R., & Rouhi, A. (2021). The impact of collaborative listening to podcasts on high school learners' listening comprehension and vocabulary learning. *System, 101*, 102588. https://doi.org/10.1016/j.system.2021.102588
[39]. Shi, D., & Révész, A. (2026). The effects of repeating video-lecture-based tasks on learners' L2 multimodal processing: An exploratory study. *Studies in Second Language Acquisition, 48*(2), 400–423. https://doi.org/10.1017/S0272263125101460
[40]. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
[41]. Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Academic Press.
[42]. Teng, M. F. (2025). Effectiveness of captioned videos for incidental vocabulary learning and retention: The role of working memory. *Computer Assisted Language Learning, 38*(1–2), 206–234. https://doi.org/10.1080/09588221.2023.2173613
[43]. Teng, M. F., & Cui, Y. (2025). Incidental vocabulary learning from captioned viewing: A forest plot of word- and learner-related factors. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688251372987
[44]. Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.
[45]. Vandergrift, L., & Tafaghodtari, M. H. (2010). Teaching L2 learners how to listen does make a difference: An empirical study. *Language Learning, 60*(2), 470–497. https://doi.org/10.1111/j.1467-9922.2009.00559.x
[46]. Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology, 14*(1), 65–86.
[47]. Yaacob, A., Amir, A. S. A., Asraf, R. M., Yaakob, M. F. M., & Zain, F. M. (2021). Impact of YouTube and video podcast on listening comprehension among young learners. *International Journal of Interactive Mobile Technologies, 15*(20), 4–19. https://doi.org/10.3991/ijim.v15i20.23701

Appendix A: Data Dictionary and Codebook

The analytical dataset contains data from 60 participants and 111 variables: 15 demographic, grouping, and summary-score variables and 96 item-level response variables corresponding to 24 items across four assessment forms.

Table A1. Codebook for summary and outcome variables

Variable	Type / Range	Description
id	Integer, 1–60	Participant identifier.
block	Integer, 1–12	Permuted allocation-block number
group	Text	G1 Transcript-supported or G2 Audio-only
condition	Text	Transcript or Audio
tr	0 / 1	Condition indicator used in statistical models (1 = Transcript)
age	Numeric, years	Age, 18–23 years
study_year	Integer, 1–4	Undergraduate study year
english_years	Numeric	Years of formal English study
baseline_cefr	A2 / B1 / B2 / C1	Descriptive proficiency category based on mean pretest performance
listening_pre	0–100	Listening pretest percentage score
listening_post	0–100	Listening posttest percentage score
reading_pre	0–100	Reading pretest percentage score
reading_post	0–100	Reading posttest percentage score
listening_gain	Numeric	Listening posttest minus pretest score
reading_gain	Numeric	Reading posttest minus pretest score
Lpre_01–24	0 / 1	Listening-pretest item responses
Lpost_01–24	0 / 1	Listening-posttest item responses
Rpre_01–24	0 / 1	Reading-pretest item responses
Rpost_01–24	0 / 1	Reading-posttest item responses

The descriptive baseline-CEFR distribution was A2 = 15, B1 = 23, B2 = 15, and C1 = 7. CEFR categories were calculated from mean pretest performance and used descriptively rather than as covariates in the primary models.

Appendix B.

Statistical Computing Environment

The statistical analyses were conducted in Python using established numerical and statistical libraries. Table B1 summarizes the principal software components used for data management, statistical estimation, diagnostic testing, and sensitivity analyses.

Table B1. Statistical computing environment

Component	Version	Principal use
Python	3.12.3	Statistical computing environment
NumPy	2.4.4	Numerical operations and resampling
pandas	3.0.2	Data management and restructuring
SciPy	1.17.1	Statistical tests and numerical routines
stats models	0.14.6	ANCOVA, robust covariance estimation, mixed-effects models, and diagnostic procedures

The principal analyses included baseline-adjusted ANCOVA, HC3 heteroscedasticity-robust covariance estimation, Holm multiplicity adjustment, stratified bootstrap confidence intervals, condition-label permutation tests, linear mixed-effects modeling, Welch comparisons, repeated-measures modeling, and multivariate analysis. Bootstrap and permutation analyses used 10,000 iterations. Repeated analyses were conducted under fixed computational settings to ensure that the numerical results reported in the manuscript were reproducible to the stated precision. Statistical outputs were retained at full numerical precision during computation and rounded only for presentation in the manuscript. The reported tables and figures were derived from the same analytical results used for the corresponding textual descriptions.