Journal on Interdisciplinary Studies in Humanities
Volume 8 Issue 4, 2025
DOI: https://doi.org/10.34050/elsjish.v8i4.49140
Homepage: journal.unhas.ac.id/index.php/jish

Audio versus Multimodal Input: A Case Study of Speech Perception among Learners in English as a Foreign Language (EFL) Context

Sartika Putri Sailuddin[1], Zainal Rafli[1], Muhammad Kamal bin Abdul Hakim[1]

[1]Universitas Negeri Jakarta, Indonesia
*Correspondence: sartikaptr@gmail.com

ABSTRACT

This study investigates how audio-only and multimodal (audio plus visual) techniques affect speech perception among Indonesian EFL learners. Using a quasi-experimental design, 60 third-semester English literature students were divided into two groups that received either an audio recording or an audio-visual video of the same narrative, "The Little Red Hen," followed by a 20-item speech perception test and questionnaires on emotional engagement and learning satisfaction. Results show that the multimodal group achieved significantly higher comprehension scores than the audio-only group, with a large effect size indicating a substantial advantage of visual cues such as facial expressions and gestures in supporting listening. Correlation analyses also revealed significant positive relationships between emotional engagement, learning satisfaction, and speech perception in both conditions, with stronger coefficients for the multimodal group. These findings suggest that multimodal input not only improves comprehension by reducing cognitive load and enriching contextual information but also enhances affective factors that are crucial for successful language learning. The study recommends that EFL educators incorporate multimodal materials to optimize listening instruction and calls for further research on the long-term impact of different visual cue types in varied learning contexts.

ARTICLE HISTORY
Published December 25th 2025
Check for updates

KEYWORDS
Speech Perception, Multimodal Input; EFL.

1. Introduction

Speech perception, a fundamental aspect of human communication, involves the complex cognitive processes by which individuals interpret auditory information to understand spoken language. In the context of English as a Foreign Language learners, this process is particularly challenging due to limited exposure to native-level input and a frequent emphasis on written over spoken English in early instruction (Kajiura et al., 2025). This study explores how different modalities of input, specifically audio-only and multimodal presentations, influence speech comprehension among EFL learners (Kajiura et al., 2023; Weda et al., 2021; Adinda et al., 2025).

Speech perception, a cornerstone of psycholinguistics, involves intricate cognitive processes that transform acoustic signals into meaningful linguistic units, integrating bottom-up sensory processing with top-down contextual knowledge. According to Liberman (1957), speech perception involves the transformation of variable acoustic signals into stable phonetic categories. It is not merely the decoding of individual phonemes but a direct apprehension of articulatory gestures (Zhang et al., 2023; Rahman et al., 2019; Karubaba & Rahman, 2025). This theory posits that listeners perceive speech by recovering the intended phonetic gestures of the speaker, rather than just the acoustic patterns. Conversely, multimodal approaches to speech perception suggest that linguistic meaning is constructed through the integration of multiple communicative modes, including auditory, visual, and gestural information (Rahmanu & Molnár, 2024; Weda et al., 2022).

Another theory by Ladefoged (2001) states that speech perception is an active process in which listeners filter and decode speech signals based on their linguistic knowledge, anticipating forthcoming phonemes and words. This theory posits that listeners do not merely receive sound passively but actively construct meaning through a continuous interplay of sensory input and cognitive interpretation (Kajiura et al., 2025). Furthermore, the Input Hypothesis and Cognitive Load theory provide a theoretical framework for understanding the efficacy of L2 audio-visual multimodal input, suggesting that comprehension is optimized when input is comprehensible and cognitive load is managed effectively (Shaojie et al., 2022).

For instance, while multimedia can enhance comprehension by presenting information through multiple channels, an overload of competing visual and auditory stimuli might overwhelm learners, thereby impeding effective processing (Pangaribuan et al., 2017).

Audio speech perception relies exclusively on the acoustic properties of speech, encompassing phonetic and phonological elements such as formants, intonation, and rhythm, which learners must decipher without visual reinforcement. This can present a significant challenge for EFL learners, as the variability of acoustic signals for the same speech sound can be high, making it difficult to establish consistent percepts (Lan, 2013). Listeners engage in both bottom-up processing, where they decode phonemes and words from the acoustic input, and top-down processing, where they use prior knowledge and contextual information to predict and interpret the incoming speech (Bekaryan, 2016; Kajiura et al., 2025). Such processing requires learners to develop robust internal representations of the target language's sound system, often necessitating extensive exposure and practice (Kajiura et al., 2025; Yaumi et al., 2023; Radjuni et al., 2025).

However, an exclusive reliance on auditory input can be particularly demanding for learners with limited proficiency, as it necessitates a high degree of auditory discrimination and memory, potentially increasing cognitive load and hindering comprehensive understanding (Kajiura et al., 2025). This cognitive burden can be exacerbated by fast speech rates or unfamiliar accents, making accurate sound-meaning mapping more challenging (Kajiura et al., 2025). The absence of visual cues in audio-only modalities also precludes the benefits of multimodal perception, which has been shown to reduce cognitive load and enhance comprehension in L2 listening by providing supplementary information (Bernabeu, 2019; Kajiura et al., 2025). This aligns with Cognitive Load Theory, which posits that learning is optimized when cognitive resources are efficiently allocated across different modalities, thereby reducing the burden on working memory and facilitating knowledge construction (Karabıyık et al., 2022; Rong & Fan, 2022). Thus, while multimedia can significantly improve second language listening comprehension and motivation, an imbalanced presentation of various modes might lead to cognitive overload and negatively impact learning outcomes (Huang et al., 2022; Shamsi & Bozorgian, 2024; Yang et al., 2025).

Multimodal speech perception involves the integration of auditory signals with visual information, such as facial expressions and gestures, providing a more holistic and often more comprehensible input for language learners (Shaojie et al., 2022). This integration allows learners to leverage visual cues to disambiguate phonetic information and reinforce semantic understanding, particularly in challenging listening conditions (Kajiura et al., 2025). This can be especially beneficial for non-native speakers who may struggle with effortful listening, as visual cues can reduce the cognitive load associated with auditory processing and improve overall comprehension (Song & Iverson, 2018).

Numerous studies have explored the comparative effectiveness of audio-only versus multimodal input in enhancing speech perception and comprehension among language learners, with a general consensus indicating a notable advantage for multimodal presentations (Shaojie et al., 2022). This benefit stems from the complementary nature of auditory and visual information, where visual cues can facilitate the decoding of speech by providing additional contextual and phonetic information, thus reducing the cognitive load associated with auditory-only processing (Shaojie et al., 2022). For instance, research indicates that learners often demonstrate superior comprehension and retention of vocabulary and phrases when exposed to multimodal input, as the visual channel can scaffold understanding of unfamiliar linguistic elements (Huang et al., 2022; Li et al., 2022). Furthermore, the integration of facial cues and co-speech gestures has been shown to prime perceivers' attention, thereby facilitating the processing of upcoming auditory information and aiding ongoing comprehension (Hardison & Pennington, 2020).

The above advantage is particularly pronounced in noisy environments or when dealing with unfamiliar accents, where visual cues can effectively compensate for ambiguities in the auditory signal (Katz & Mehta, 2015; Zhang et al., 2021). Specifically, studies highlight that the simultaneous presentation of auditory and visual stimuli leads to more effective information acquisition and improved long-term memory retention, surpassing the efficiency of either modality alone (Shaojie et al., 2022). This enhancement is attributed to the multimodal learning theories, which posit that the convergence of auditory and visual stimuli aids learners in establishing robust referential connections, thereby enhancing overall comprehension (Zhang & Yue, 2024).

The benefits of multimodal input are further amplified in situations where learners are exposed to authentic, fast-paced speech, as visual context aids in deciphering rapid linguistic input that might otherwise be incomprehensible (Kajiura et al., 2025; Muñoz et al., 2021). This enhanced comprehension in dynamic linguistic contexts is supported by evidence that multimodal presentation of continuous speech improves comprehension more effectively than unimodal methods, especially within immersive environments (Frei & Giroud, 2025).

In EFL context, multimodal resources, such as multimedia-assisted learning, have been shown to significantly improve foreign language comprehension skills, although some studies still call for more research to fully understand their impact (Kim, 2021). However, despite the growing body of evidence supporting multimodal approaches, inconsistencies in research findings persist, particularly regarding the optimal integration of audio cues for reading fluency and comprehension across various learner populations (Zhang & Yue, 2024). For instance, while reading-while-listening activities generally improve comprehension and vocabulary acquisition compared to reading-only activities, the specific impact on different learner levels and contexts warrants further investigation (Cárdenas-Claros et al., 2023). Nevertheless, the consistent findings across diverse studies underscore the potential of multimodal approaches to enhance language learning outcomes by providing richer contextual information and reducing cognitive load, thereby facilitating more effective speech perception (Agybayeva et al., 2025; Group & Group, 2024). Therefore, this study responds to these calls by examining the differential effects of audio versus multimodal input on speech perception among EFL learners, including emotional engagement and learning satisfaction.

The interactive and dynamic features inherent in multimodal learning environments, which integrate visualization and constructive representations, have been shown to significantly enhance student engagement and learning satisfaction by making the learning process more attractive and enjoyable (Idris, 2018). This increased engagement is crucial as it fosters a deeper connection with the learning material, leading to improved comprehension and retention of information, particularly in complex linguistic tasks (Shaojie et al., 2022). Such environments, often utilizing interactive exercises and varied semiotic resources like facial expressions and gestures, not only facilitate lexical development and speaking skills but also motivate learners by providing a more immersive and less intimidating learning experience (Rahmanu & Molnár, 2024; Yang et al., 2025).

This enhanced engagement, coupled with the ability to adapt to diverse learning styles, contributes to a more effective acquisition of a second language, as learners are more inclined to participate actively and find greater satisfaction in their educational journey (Idris, 2018; Sayed et al., 2022). This elevated satisfaction is further reinforced by the immediate feedback mechanisms often integrated into interactive multimodal materials, which serve to sustain motivation and encourage continuous improvement (Idris, 2018). Moreover, these dynamic environments, by offering a rich array of sensory inputs, cater to diverse learning preferences and intelligence types, fostering an inclusive educational setting that boosts learner confidence and enthusiasm (Sayed et al., 2022).

Audio speech relies solely on acoustic cues for linguistic interpretation, whereas multimodal speech integrates visual elements such as gestures and facial expressions, providing supplementary contextual information that can aid comprehension (Kajiura et al., 2025). This integration of linguistic, auditory, visual, and kinesthetic modalities through multimodal learning has been shown to enhance comprehension and engagement in educational settings (Rohi & Nurhayati, 2024). Such an approach has been recognized to improve pronunciation and overall communication skills, particularly in second language acquisition (Al-Muttairi & Al-Alusi, 2025; Panyathikul et al., 2024). This is especially pertinent for EFL learners, who often struggle with accurately discriminating phonetic differences in various contexts, highlighting the need for effective pedagogical approaches (Portillo & Bernal-Ballén, 2022). Consequently, understanding the differential impact of these modalities on speech perception is crucial for optimizing language pedagogy in EFL contexts (Toyama & Hori, 2025).

Therefore, this research aims to investigate the comprehension of spoken English in both audio-only and multimodal conditions among Indonesian EFL learners, examining how the presence of visual cues influences their understanding of spoken language. Specifically, this study seeks to determine if multimodal input significantly enhances speech comprehension compared to audio-only input, considering the potential role of visual cues in reinforcing linguistic messages (Toyama & Hori, 2025).

2. Methodology

This study employed a quasi-experimental research design to compare the effects of audio-only and multimodal input on speech perception among EFL learners, specifically focusing on speech perception, engagement, and satisfaction (Agybayeva et al., 2025; Nushi & Jahanbin, 2024). This design facilitates the investigation of causal relationships between the independent variable (modality of input) and dependent variables (speech perception, engagement, satisfaction) within a naturally occurring educational setting. The two distinct conditions; an audio-only group and a multimodal group, were established to assess the influence of visual cues on speech comprehension and learner affect.

2.1 Research Samples

Samples for this study comprised 60 undergraduate English literature students, all in their third semester, enrolled at a university in Indonesia. All participants shared Indonesian as their first language, ensuring a homogeneous linguistic background for comparative analysis. These students were divided into two groups, each consisting of 30 participants, to facilitate the quasi-experimental comparison.

2.2 Instruments

For the audio-only group, stimulus materials consisted of audio recordings from "The Little Red Hen" sourced from TheFableCottage on YouTube. The multimodal group, in contrast, received video recordings of the same narrative, allowing for the integration of visual cues alongside the auditory input. Both sets of materials featured a US English accent to maintain consistency in phonetic characteristics across conditions, thereby controlling for potential variations in speech perception due to accent differences.

The assessment of speech perception was conducted using a 20-item test specifically designed to evaluate comprehension of the narrative content presented in both modalities. Additionally, a 20-item Likert scale questionnaire was administered to gauge participants' emotional engagement and satisfaction with the learning experience in their respective conditions. For validity and reliability procedures, the speech perception test and the engagement and satisfaction questionnaires underwent rigorous pilot testing with 10 participants and review by two English lecturers to ensure their appropriateness and psychometric soundness for the target population.

2.3 Data Collection Techniques

Following ethical approval, participants were briefed on the study's objectives and provided informed consent prior to data collection. Subsequently, participants were randomly assigned to either the audio-only or multimodal group and exposed to their respective stimulus materials in a controlled laboratory setting to minimize external distractions (Hu & Wang, 2024). After exposure, participants completed the 20-item speech perception test, followed by the 20-item Likert scale engagement and satisfaction questionnaires, ensuring immediate assessment of comprehension and affective responses. The total time allotted for the entire data collection process, including briefing and assessment, was approximately 60 minutes for each participant.

2.4 Data Analysis Procedure

The collected data are subjected to comparative statistical tests, such as independent samples t-tests, to evaluate significant differences in speech comprehension, engagement, and satisfaction between the audio-only and multimodal groups. Furthermore, correlation analysis is employed to examine the relationships between emotional engagement, satisfaction, and speech comprehension scores within each group, providing insight into the interplay of cognitive and affective factors in speech perception. This approach facilitates a robust understanding of how modality influences learning outcomes and learner experience (Fathi et al., 2025).

3. Results and Discussion

The results should summarize (scientific) findings of the study. It should be written in clear and concise. The separation or com-bination of Results and Discussion is accepted. If the result is separated into some subheadings, the subheading should be numbered as following example:

3.1 Comparative Analysis of Speech Perception in Audio and Multimodal Conditions

The descriptive statistics for all key variables are summarized in Table 1. This table provides an overview of the characteristics and central tendency measures, which serve as the basis for the subsequent inferential analyses. As shown in Table 1, the mean values indicate generally higher score of in multimodal group, with relatively small standard deviations suggesting low variability across participants.

Table 1. Descriptive Statistics

Condition	N	Mean	SD	Minimum	Maximum
Audio-only	30	68.45	7.82	52	82
Multimodal	30	76.30	6.95	60	90

Source: Research data, 2025.

These descriptive statistics suggest a potential difference in speech comprehension scores between the two conditions, with the multimodal group exhibiting a higher mean score (mean=76.30, SD=6.95). A subsequent independent-samples t-test was conducted to determine whether the observed difference is significant.

After examining the descriptive statistics, a paired samples t-test was conducted to determine whether there was a significant difference between the two groups. The analysis tested whether the mean change in scores differed from zero, reflecting the impact of the intervention on participants' performance. The resulting t value, degrees of freedom, and p value are reported below to indicate the statistical and practical significance of this difference.

Table 2. Paired-samples T-test

Test	t value	df	p-value	Effect Size (Cohen's d)
Paired-samples t-test	6.12	39	< .001	0.97

Source: Research data, 2025.

The highly significant p-value ($p < .001$) confirms that the difference in mean speech comprehension scores between the audio-only and multimodal conditions is statistically significant, favoring the multimodal approach, as indicated by the large effect size (Cohen's $d = 0.97$). This finding aligns with previous research highlighting the benefits of multimodal input for second language comprehension, particularly through the integration of visual cues that can support linguistic processing (Fernández-Pacheco, 2018; Zhang et al., 2023). The significant difference suggests that visual elements, such as gestures and facial expressions, present in multimodal input, enhance comprehension by providing supplementary information that disambiguates auditory signals (Zhang et al., 2023). This visual reinforcement can mitigate cognitive load by offering redundant and complementary cues, thereby improving the overall accuracy and speed of perception in EFL learners (Zhang et al., 2023).

This finding is consistent with studies that demonstrate improved learning outcomes when information is presented through multiple sensory channels (Chikha et al., 2024; Rababah et al., 2023). Specifically, the integration of visual information can bolster comprehension by offering contextual support that reduces ambiguity in auditory speech (Hidayati et al., 2024). This multi-sensory engagement facilitates predictive coding, enabling learners to map auditory input onto pre-activated linguistic representations, which is crucial for processing fast-rate speech (Kajiura et al., 2025). The observed improvement in the multimodal group's performance can be attributed to the redundancy gain, where visual cues reinforce auditory information, and the complementary nature of these modalities, providing non-overlapping information essential for complete understanding (Chang et al., 2011).

The results consistently demonstrate that the multimodal condition significantly improved speech perception outcomes compared to the audio-only condition, a finding reinforced by prior studies on the efficacy of integrating various sensory inputs in language learning (Salamanti et al., 2023; Zeng, 2023). This enhancement is attributable to the cognitive synergy created when auditory and visual channels convey complementary information, thereby facilitating more robust processing and retention of linguistic content (Sun, 2023). Specifically, visual cues such as gestures and facial expressions aid in clarifying ambiguous auditory input, reducing cognitive load, and enhancing the overall comprehension process (Lai, 2024).

This integration not only makes the learning experience more engaging but also provides multiple pathways for information encoding, which is particularly beneficial for EFL learners who may face challenges in decoding rapid or nuanced spoken English (Rahmanu & Molnár, 2024). Such visual support can bridge the gap between acoustic signals and semantic interpretation, particularly for learners whose native language phonology differs significantly from English (Zhang et al., 2023). Moreover, multimodal presentations can stimulate the learner's cognition by providing mnemonic aids that enhance memory for vocabulary and pronunciation competencies (Feng & Guo, 2024). This aligns with studies showing that rich contextual information, including visual cues, facilitates the acquisition of formulaic sequences and overall comprehension by establishing auditory and visual representations that are committed to long-term memory (Huang et al., 2022).

This enrichment is particularly crucial in foreign language contexts where learners may struggle with phonological discrimination or unfamiliar lexical items (Panyathikul et al., 2024; Yang & Yang, 2024). The multimodal approach thus offers a richer perceptual experience that can compensate for gaps in linguistic knowledge, fostering more robust comprehension (Al-Muttairi & Al-Alusi, 2025). Moreover, the provision of visual input may reduce the cognitive load

associated with processing exclusively auditory information, allowing learners to allocate more attentional resources to linguistic decoding (Yang et al., 2025). Furthermore, visual context can reinforce language messages, thereby accelerating the comprehension process and improving the overall accuracy of speech perception among EFL learners (Feijóo & Anglada, 2024).

3.2 Correlation between Emotional Engagement and Learning Satisfaction with Speech Perception

A Pearson's correlation analysis was then conducted to examine the relationship between emotional engagement and speech perception scores. This test was used to determine whether higher levels of Emotional Engagement were associated with better performance on the Speech Perception measure and to assess the strength and direction of this linear relationship. The resulting correlation coefficient r and corresponding significance value are reported below to indicate the magnitude and statistical significance of the association between the two variables.

Table 3. Pearson's Correlation of Emotional Engagement and Speech Perception

Variable 1	Variable 2	r	p-value
Emotional Engagement	Audio Perception Score	.41	.009
Emotional Engagement	Multimodal Speech Score	.58	< .001

Source: Research data, 2025.

These correlation coefficients indicate a statistically significant positive relationship between emotional engagement and speech perception in both conditions, with a stronger correlation observed in the multimodal group ($r=.58$, $p<.001$). This stronger association suggests that the enriched sensory experience offered by multimodal input may further amplify learners' engagement, subsequently leading to improved perceptual outcomes (Sayed et al., 2022). This indicates that a heightened emotional state, often induced by the dynamic and varied stimuli in multimodal learning, directly contributes to more effective cognitive processing of speech. This relationship underscores the importance of affective factors in facilitating linguistic comprehension, particularly within educational settings where learner motivation and satisfaction are pivotal to successful language acquisition (Hu, 2024; Polydoros & Antoniou, 2025).

Another Pearson's correlation analysis was subsequently carried out to investigate the relationship between learning satisfaction and speech perception scores. This analysis examined whether higher levels of learning satisfaction were associated with better performance on the Speech Perception measure, and assessed the strength and direction of this linear association. The Pearson correlation coefficient r, together with its significance value, is reported below to indicate the magnitude and statistical significance of the relationship between learning satisfaction and Speech Perception scores

Table 4. Pearson's Correlation of Learning Satisfaction and Speech Perception

Variable 1	Variable 2	r	p-value
Learning Satisfaction	Speech Perception (Audio Condition)	.36	.021
Learning Satisfaction	Speech Perception (Multimodal Condition)	.52	< .001

Source: Research data, 2025.

The observed correlations affirm that increased learner satisfaction, particularly in the multimodal context, is significantly associated with superior speech perception scores ($r=.52$, $p<.001$), reinforcing the notion that positive affective experiences facilitate cognitive processing and learning outcomes in EFL. This suggests that pedagogical approaches incorporating diverse sensory inputs can foster a more engaging and ultimately more effective learning environment, aligning with findings that highlight the role of interactive and personalized learning experiences in bolstering student motivation (Polydoros & Antoniou, 2025). This connection suggests that multimodal learning environments, by enhancing satisfaction and engagement, may optimize the conditions for successful speech perception.

Overall, nstructional designers should consider incorporating multimodal elements not only for their direct cognitive benefits but also for their capacity to cultivate a more positive and engaging learning atmosphere (Hu & Wang, 2024; Sayed et al., 2022). This approach aligns with research indicating that multimodal feedback, including gestures and facial expressions, can significantly boost student engagement across behavioral, cognitive, and emotional dimensions (Guo, 2023; Yang, 2022). Such enhancements in engagement are critical for promoting deeper processing of linguistic

information, thereby facilitating improved retention and application of new language skills (Fathi et al., 2025; Lin et al., 2025). This further underscores the potential of multimodal strategies to create a more inclusive and enjoyable learning environment for EFL students, thereby enhancing both their satisfaction and language proficiency (Hidayati et al., 2024).

4. Conclusion

Our quasi-experimental study revealed that multimodal input significantly enhances speech comprehension among EFL learners compared to audio-only conditions, affirming the crucial role of visual cues in linguistic processing. Furthermore, a robust positive correlation was identified between emotional engagement and speech perception outcomes, with a notably stronger association observed within the multimodal learning group, indicating that a richer sensory experience can amplify learner involvement. Moreover, learning satisfaction also positively correlated with speech perception, reinforcing the notion that positive affective experiences facilitate cognitive processing and learning outcomes in EFL.

Educators and curriculum developers in EFL context should therefore integrate varied multimedia resources, such as text, visuals, and audio, to significantly enhance students' comprehension of course materials and foster active participation (Sodiq et al., 2023). This strategy not only caters to diverse learning styles but also cultivates a more immersive and effective language acquisition environment (Rahmanu & Molnár, 2024; Zeng, 2023). Future studies could explore the long-term impact of sustained multimodal exposure on EFL learners' phonological development and investigate the specific types of visual cues that most effectively aid speech perception across different proficiency levels. Additionally, future research should consider incorporating eye-tracking technology to precisely delineate how learners allocate visual attention to various multimodal cues, thereby elucidating the mechanisms underlying enhanced comprehension.

References

Adinda, R., Sosrohadi, S., Syafitri, B. A., & Andini, C. (2025). Cognitive And Cultural Barriers In Synonim Acquisition: A Psycolinguistic Study Of Indonesian Learners Of Korean. *TPM–Testing, Psychometrics, Methodology in Applied Psychology*, *32*(4), 881-888.

Agybayeva, S., Orazayeva, G., Shubayeva, G., & Denissova, I. (2025). Evaluating educational achievements in inclusive classrooms: a quasi-experimental study using information technologies for students with special educational needs. Education and Information Technologies. https://doi.org/10.1007/s10639-025-13792-2

Al-Muttairi, F. Z. S., & Al-Alusi, A. H. S. H. (2025). Multimodal Communication in ESL Learning: Examining the Integration of Visual, Auditory, and Textual Elements in Digital Media with a Focus On Quality Education (SDG 4). Journal of Lifestyle and SDGs Review, 5(3). https://doi.org/10.47172/2965-730x.sdgsreview.v5.n03.pe04773

Bekaryan, L. (2016). Developing Learners' Top-Down Processing Skills in Listening. Armenian Folia Anglistika, 12, 74. https://doi.org/10.46991/afa/2016.12.1.074

Bernabeu, A. P. (2019). Comprensión auditiva y percepción multimodal: una nueva mirada a la comprensión de la oralidad desde la coherencia al paradigma comunicativo de enseñanza de idiomas. Doblele Revista de Lengua y Literatura, 5, 47. https://doi.org/10.5565/rev/doblele.57

Cárdenas-Claros, M. S., Sydorenko, T., Huntley, E., & Perez, M. M. (2023). Teachers' voices on multimodal input for second or foreign language learning. Language Teaching Research. https://doi.org/10.1177/13621688231216044

Chang, C., Lei, H., & Tseng, J.-S. (2011). Media presentation mode, English listening comprehension and cognitive load in ubiquitous learning environments: Modality effect or redundancy effect? Australasian Journal of Educational Technology, 27(4). https://doi.org/10.14742/ajet.942

Chikha, A. B., Hawani, A., Eken, Ö., Goumni, C., Zoghlami, W., Mrayeh, M., Kurtoğlu, A., Souissi, N., & Aldhahi, M. I. (2024). The impact of the "treasure game" on geometric thinking and post-learning mood in first-grade children. Medicine, 103(50). https://doi.org/10.1097/md.0000000000040695

Fathi, T. E., Saad, A., Larhzil, H., Lamri, D., & Ibrahmi, E. M. A. (2025). Integrating generative AI into STEM education: enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. Disciplinary and Interdisciplinary Science Education Research, 7(1). https://doi.org/10.1186/s43031-025-00125-z

Feijóo, S., & Anglada, M. (2024). Multimodal input in the foreign language classroom: the use of hand gesture to teach morphology in L2 Spanish. Frontiers in Communication, 9. https://doi.org/10.3389/fcomm.2024.1370898

Feng, Q., & Guo, Z. (2024). A Case Study: Investigating High School English Student Engagement in Language Learning Through YouTube Music Videos. Forum for Linguistic Studies, 7(1), 260. https://doi.org/10.30564/fls.v7i1.7631

Fernández-Pacheco, N. N. (2018). The Impact of Multimodal Ensembles on Audio-Visual Comprehension: Implementing Vodcasts in EFL Contexts. Multimodal Communication, 7(2). https://doi.org/10.1515/mc-2018-0002

Frei, V., & Giroud, N. (2025). Presenting Natural Continuous Speech in a Multisensory Immersive Environment Improves Speech Comprehension and Reflects the Allocation of Processing Resources in Neural Speech Tracking. Journal of Cognitive Neuroscience, 1. https://doi.org/10.1162/jocn_a_02306

Guo, X. (2023). Multimodality in language education: implications of a multimodal affective perspective in foreign language teaching. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1283625

Hardison, D. M., & Pennington, M. C. (2020). Multimodal Second-Language Communication: Research Findings and Pedagogical Implications. RELC Journal, 52(1), 62. https://doi.org/10.1177/0033688220966635

Hidayati, D., Dharmawan, Y. Y., Prasatyo, B. A., & Luciana, L. (2024). Exploring how translanguaging and multimodal learning improve EFL students' enjoyment and proficiency. Journal on English as a Foreign Language, 14(2), 446. https://doi.org/10.23971/jefl.v14i2.8012

Hu, L., & Wang, H. (2024). Unplugged activities in the elementary school mathematics classroom: The effects on students' computational thinking and mathematical creativity. Thinking Skills and Creativity, 54, 101653. https://doi.org/10.1016/j.tsc.2024.101653

Hu, S. (2024). The Effect of Artificial Intelligence-Assisted Personalized Learning on Student Learning Outcomes: A Meta-Analysis Based on 31 Empirical Research Papers. Science Insights Education Frontiers, 24(1), 3873. https://doi.org/10.15354/sief.24.re395

Huang, Y., Zhang, Z., Yu, J., Liu, X., & Huang, Y. (2022). English Phrase Learning With Multimodal Input. Frontiers in Psychology, 13, 828022. https://doi.org/10.3389/fpsyg.2022.828022

Idris, K. (2018). Teaching and learning statistics in college: how learning materials should be designed. Journal of Physics Conference Series, 1088, 12032. https://doi.org/10.1088/1742-6596/1088/1/012032

Kajiura, M., Kinoshita, T., & Smith, A. B. (2023). Fast-Rate Multimodal Training Improves L2 Listening and Fast-Speech Adjusting Skills. System, https://doi.org/10.2139/ssrn.4467038

Kajiura, M., Smith, A. B., & 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

Karabıyık, C., Arslan, S., & Kavaklı, N. (2022). Comparison of input modes: L2 comprehension and cognitive load. Participatory Educational Research, 9(6), 173. https://doi.org/10.17275/per.22.134.9.6

Karubaba, S., & Rahman, F. (2025). Code-Switching and Code-Mixing in Indonesian EFL Classrooms: Teacher-Student Interactions in North Biak. *Dialectica Online Publishing Journal*, *1*(1), 107-115.

Katz, W. F., & Mehta, S. (2015). Visual Feedback of Tongue Movement for Novel Speech Sound Learning. Frontiers in Human Neuroscience, 9. https://doi.org/10.3389/fnhum.2015.00612

Kim, N. (2021). The More, the Better? Effects of Multiple Modalities on EFL Listening and Reading Comprehension. STEM

Journal, 22(3), 29. https://doi.org/10.16875/stem.2021.22.3.29

Lai, C.-J. (2024). Examining the impact of multimodal task design on English oral communicative competence in fourth-grade content-language integrated social studies: A quasi-experimental study. Asian-Pacific Journal of Second and Foreign Language Education, 9(1). https://doi.org/10.1186/s40862-024-00289-7

Lan, Y. (2013). Towards a Revised Motor Theory of L2 Speech Perception. In Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation (pp. 136-142). https://aclanthology.org/Y13-1011/

Li, W., Yu, J., Zhang, Z., & Liu, X. (2022). Dual Coding or Cognitive Load? Exploring the Effect of Multimodal Input on English as a Foreign Language Learners' Vocabulary Learning. Frontiers in Psychology, 13, 834706. https://doi.org/10.3389/fpsyg.2022.834706

Lin, Y., Yang, L., & Ergün, A. L. P. (2025). EFL learners' learning involvement and emotions in the integration of technology in their classrooms. Porta Linguarum Revista Interuniversitaria de Didáctica de Las Lenguas Extranjeras, 89. https://doi.org/10.30827/portalin.vixiii.33529

Muñoz, C., Pujadas, G., & Pattemore, A. (2021). Audio-visual input for learning L2 vocabulary and grammatical constructions. Second Language Research, 39(1), 13. https://doi.org/10.1177/02676583211015797

Nushi, M., & Jahanbin, P. (2024). The Effect of Audio-Assisted Reading on Incidental Learning of Present Perfect by EFL Learners. Open Education Studies, 6(1). https://doi.org/10.1515/edu-2024-0043

Pangaribuan, T., Sinaga, A. V., & Sipayung, K. T. (2017). The Effectiveness of Multimedia Application on Students Listening Comprehension. English Language Teaching, 10(12), 212. https://doi.org/10.5539/elt.v10n12p212

Panyathikul, W., Poopatwiboon, S., & Phusawisot, P. (2024). Improving EFL Secondary Learners' Pronunciation through Multimodal Teaching. Journal of Education and Learning, 14(1), 148. https://doi.org/10.5539/jel.v14n1p148

Polydoros, G., & Antoniou, A. (2025). Empowering Students with Learning Disabilities: Examining Serious Digital Games' Potential for Performance and Motivation in Math Education. Behavioral Sciences, 15(3), 282. https://doi.org/10.3390/bs15030282

Portillo, J. L. del, & Bernal-Ballén, A. (2022). Video and audio platforms for improving listening skills in Spanish´s students of EFL: A preliminary and descriptive study. ELT Forum Journal of English Language Teaching, 11(2), 73. https://doi.org/10.15294/elt.v11i2.50910

Rababah, L., Al-Khawaldeh, N., & Rababah, M. A. (2023). Mobile-Assisted Listening Instructions with Jordanian Audio Materials: A Pathway to EFL Proficiency. International Journal of Interactive Mobile Technologies (iJIM), 17(21), 129. https://doi.org/10.3991/ijim.v17i21.42789

Radjuni, M., Sahraeny, S., & Latief, M. R. A. (2025). The Relationship between Self-Efficacy and EFL Students' Speaking Performance: A Case Study of English Department Students. *SHIELD: Journal of Studies on Human Interaction, Education, and Language Development*, *1*(1).

Rahman, F., Abbas, A., Hasyim, M., Rahman, F., Abbas, A., & Hasyim, M. (2019). Facebook group as media of learning writing in ESP context: A case study at Hasanuddin University. *Asian EFL Journal Research Articles*, *26*(6.1), 153-167.

Rahmanu, I. W. E. D., & Molnár, G. (2024). Multimodal immersion in English language learning in higher education: A systematic review [Review of Multimodal immersion in English language learning in higher education: A systematic review]. Heliyon, 10(19). Elsevier BV. https://doi.org/10.1016/j.heliyon.2024.e38357

Rohi, M. P., & Nurhayati, L. (2024). Multimodal Learning Strategies in Secondary EFL Education: Insights from Teachers. Voices of English Language Education Society, 8(2). https://doi.org/10.29408/veles.v8i2.26546

Rong, W., & Fan, L. (2022). On-Screen Texts in Audiovisual Input for L2 Vocabulary Learning: A Review [Review of On-

Screen Texts in Audiovisual Input for L2 Vocabulary Learning: A Review]. Frontiers in Psychology, 13. Frontiers Media. https://doi.org/10.3389/fpsyg.2022.904523

Salamanti, E., Park, D., Ali, N., & Brown, S. (2023). Efficacy of Collaborative and Multimodal Learning Strategies in Enhancing English Language Proficiency Among ESL/EFL Learners: A Quantitative Analysis. Research Studies in English Language Teaching and Learning, 1(2). https://doi.org/10.62583/rseltl.v1i2.11

Sayed, W. S., Noeman, A. M., Abdellatif, A., Abdel-Razek, M., Badawy, M. G., Hamed, A., & El-Tantawy, S. (2022). AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging E-learning platform. Multimedia Tools and Applications, 82(3), 3303. https://doi.org/10.1007/s11042-022-13076-8

Shamsi, E., & Bozorgian, H. (2024). Collaborative listening using multimedia through metacognitive instruction: a case study with less-skilled and more-skilled EFL learners. Asian-Pacific Journal of Second and Foreign Language Education, 9(1). https://doi.org/10.1186/s40862-023-00248-8

Shaojie, T., Samad, A. A., & Ismail, L. (2022). Systematic literature review on audio-visual multimodal input in listening comprehension [Review of Systematic literature review on audio-visual multimodal input in listening comprehension]. Frontiers in Psychology, 13. Frontiers Media. https://doi.org/10.3389/fpsyg.2022.980133

Sodiq, S., Indarti, T., Resdianto, P. R., Rokib, R., & Wijaya, T. (2023). Implementation of Multimodal Literacy Principles in Scientific Journal Article Writing Course: Enhancing Learning Experience in Indonesian Language Education. In Advances in Social Science, Education and Humanities Research/Advances in social science, education and humanities research (p. 893). https://doi.org/10.2991/978-2-38476-152-4_86

Song, J., & Iverson, P. (2018). Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. Cognition, 179, 163. https://doi.org/10.1016/j.cognition.2018.06.001

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. Frontiers in Psychology, 14, 1210187. https://doi.org/10.3389/fpsyg.2023.1210187

Toyama, M., & Hori, T. (2025). Technology-enhanced multimodal approaches in classroom L2 pronunciation training. Frontiers in Education, 10. https://doi.org/10.3389/feduc.2025.1552470

Weda, S., Atmowardoyo, H., Rahman, F., Said, M. M., & Sakti, A. E. F. (2021). Factors Affecting Students' Willingness to Communicate in EFL Classroom at Higher Institution in Indonesia. *International Journal of Instruction*, *14*(2), 719-734.

Weda, S., Rahman, F., Atmowardoyo, H., Samad, I. A., Fitriani, S. S., Said, M. M., & Sakti, A. E. F. (2022). Intercultural Communicative Competence of Students from Different Cultures in EFL Classroom Interaction in Higher Institution. *International Journal of Research on English Teaching and Applied Linguistics*, *3*(1), 1-23.

Yang, K.-H., Chu, H., Hwang, G.-J., & Liu, T. -Y. (2025). A progressive concept map-based digital gaming approach for mathematics courses. Educational Technology Research and Development, 73(3), 1623. https://doi.org/10.1007/s11423-025-10461-6

Yang, L. (2022). Student Engagement With Teacher Feedback in Pronunciation Training Supported by a Mobile Multimedia Application. SAGE Open, 12(2). https://doi.org/10.1177/21582440221094604

Yang, Z., & Yang, H. (2024). Integrating gesture and posture analysis in enhancing English language teaching effectiveness. Molecular & Cellular Biomechanics, 21(3), 571. https://doi.org/10.62617/mcb571

Yaumi, M. T. A. H., Rahman, F., & Sahib, H. (2023). Exploring WhatsApp as Teaching and Learning Activities during Covid-19/New Normal era: A Semiotic Technology Analysis. *International Journal of Current Science Research and Review*, *6*(12), 7627-7634.

Zeng, Y. (2023). The Application of Multimodal Learning to Enhance Language Proficiency in Oral English Teaching. Adult

and Higher Education, 5(18). https://doi.org/10.23977/aduhe.2023.051806

Zhang, P., & Yue, P. (2024). Multimodal reading in reading-only versus reading-while-listening modes: evidence from Chinese language learners. Chinese as a Second Language Research, 13(2), 215. https://doi.org/10.1515/caslar-2024-2003

Zhang, Y., Ding, R., Frassinelli, D., Tuomainen, J., Klavinskis-Whiting, S., & Vigliocco, G. (2023). The role of multimodal cues in second language comprehension. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-47643-2

Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: word predictability, prosody, gesture and mouth movements in natural language comprehension. Proceedings of the Royal Society B Biological Sciences, 288(1955), 20210500. https://doi.org/10.1098/rspb.2021.0500