

Sequencing matters: Exploring the effects of monomodal and multimodal input order on EFL comprehension

Francisca Maria Ivone and Yusnita Febrianti

English Language Education, Faculty of Letters, Universitas Negeri Malang, Malang, East Java, Indonesia

ABSTRACT

There is limited empirical evidence on how the sequence of a combination of monomodal and multimodal input influences comprehension outcomes amid the increasing interest in the integration of multimedia in listening and viewing comprehension courses. To address this gap, the study investigates how audio-before-video (AbV) and video-before-audio (VbA) sequencing affects listening and viewing comprehension among first-year English department students at an Indonesian university, whether learners' input sequence preferences align with their comprehension performance, and how learning styles relate to both preference and comprehension outcomes. The experimental research employed a single-group repeated measures design, in which 115 participants were exposed to both input sequencing conditions. Data were collected from an English proficiency test, a listening and viewing comprehension test, and a questionnaire. Three non-parametric statistical analyses, i.e., Wilcoxon Signed-Rank, Chi-Square, and Spearman's Correlation, were utilised to examine the collected data. The findings revealed that participants performed significantly better when exposed to the audio-before-video condition (AbV, $M = 77.54$) compared to video-before-audio (VbA, $M = 65.65$), $Z = -6.324$, $p < .001$, $r = .59$, indicating a large effect. However, there was no significant association between input sequence preferences and comprehension performance, $\chi^2(2, N = 112) = 4.779$, $p = .092$. Learning style was also not significantly correlated with input sequence preference ($r_s = -.002$, $p = .980$) or comprehension outcomes in either the AbV ($r_s = 0.037$, $p = 0.693$) or VbA condition ($r_s = 0.081$, $p = 0.389$). The results of this study enrich our understanding of the effects of monomodal and multimodal input sequencing on listening and viewing comprehension. The findings further suggest that listening and viewing instruction may benefit from staged input strategies, such as an initial audio-only exposure followed by video for confirmation, guided note-taking or prediction during the first exposure, and repeated listening and viewing to refine comprehension.

Keywords: Input sequencing; listening comprehension; viewing comprehension; monomodal input; multimodal input

Received:
25 Oktober 2025

Accepted:
13 Agustus 2026

Revised:
16 July 2026

Published:
12 September 2026

How to cite (in APA style):

Ivone, F. M., & Febrianti, Y. (2026). Sequencing matters: Exploring the effects of monomodal and multimodal input order on EFL comprehension. *Indonesian Journal of Applied Linguistics*, 16(2), 373-385. <https://doi.org/10.17509/9b9kg579>

INTRODUCTION

Multimedia resources have become an essential feature of English as a Foreign Language (EFL) instruction that reshapes the nature of language learning input as well as the comprehension skills required of language learners. In recent years, digital platforms such as *BBC Learning English*, *TED Talks*, and *YouTube* have updated language learning by offering learners multimodal input in the form of videos that expose them to a variety of accents, discourse styles, and contextualised language use

(Rodgers & Webb, 2020). Exposure to these and many other digital platforms has led to the extensive use of multimodal input in higher education settings. Consequently, the ability to comprehend a combination of textual, spoken and visual information has become critical for academic success in higher education. In the English Language Teaching (ELT) context, in particular, the significance of multimodal comprehension is twofold as it not only affects language learners' ability to understand academic content but also

*Corresponding author
Email: francisca.maria.fs@um.ac.id

helps develop their core language skills. These consequences have driven the recognition of viewing as a 'new' core receptive language skill alongside listening and reading (Surkamp & Yearwood, 2018) as well as the growing necessity of developing viewing comprehension skills, in addition to listening and reading comprehension skills, as language learners navigate and interpret multimodal input associated with extensive multimedia use (Donaghy, 2020). However, recent ELT research is heavily focused on listening and reading comprehension instead of listening and viewing comprehension. In fact, viewing comprehension is mostly researched in relation to vocabulary development (see Durbahn et al., 2020; Pahlavani & Salehi, 2023) or reading comprehension (see Rogowsky et al., 2016), not listening.

The integration of multimedia formats, such as text, audio, and video, offers considerable advantages for second language (L2) development (Peters & Muñoz, 2020), particularly in enhancing listening comprehension (Yi et al., 2024). Empirical studies have shown that multimedia input supports improvements in listening proficiency (Sejdiu, 2017) and listening fluency (Chang et al., 2018), while also facilitating incidental vocabulary acquisition (Perez, 2020, 2022) and aiding comprehension through contextual cues (Webb & Rodgers, 2009). One of the strengths of multimedia lies in its ability to present information redundantly; for instance, captioned videos can strengthen the connection between spoken and written forms and reinforce understanding (Winke et al., 2010; Vanderplank, 2016). In their review of multimedia modes in second and foreign language learning, Zhang and Zou (2022) emphasise the effectiveness of combining audio with animation to support listening comprehension. Collectively, these findings affirm the pedagogical value of multimodal input in fostering L2 listening development.

From a multimodality perspective, a text is considered multimodal when meaning is realised through more than one semiotic mode, such that meaning is not confined to linguistic elements alone but is realised through a range of semiotic resources, including image, gesture, gaze, sound, and spatial arrangement (cf. Jewitt, 2014; O'Halloran, 2008). Thus, audiovisual input can be operationally defined as multimodal input because spoken language is accompanied by a range of visual information presented through video. Although Jewitt and Kress (2010) argue that texts are rarely purely monomodal, a single-mode input such as audio-only or text-only input can be treated as monomodal when meaning is conveyed primarily through one semiotic mode. This distinction provides a practical basis for differentiating audiovisual from audio-only input in the context of L2 listening and viewing

comprehension, while acknowledging the broader theoretical complexity of multimodality.

In previous studies, multimodal input has been reported to promote L2 learning in general and L2 listening in particular because it aligns with theories of multimodal input, namely Dual Coding Theory and Multimedia Learning Theory. The two theories provide insights into how learners process and retain language through audiovisual materials. Paivio's (1986) Dual Coding Theory suggests that verbal and non-verbal information are processed through distinct yet interconnected cognitive channels, while Mayer's (2009, 2017) Multimedia Learning Theory emphasises how learners can efficiently process multimodal input to maximise learning. Together, the two theories provide a sound theoretical foundation for the use of multimedia in language learning. However, studies on input modalities in EFL learning have shown varied findings. Although the use of multimodal input in language learning contributes to improvements in vocabulary knowledge, and listening and speaking abilities (Li et al., 2022; Li, 2022), there are no significant differences in comprehension retention when monomodal and multimodal input are used (Liu et al., 2018). Similarly, Karabıyık and Aslan (2022) claim that multimodal input (video + audio) improves L2 listening comprehension and reduces cognitive load compared to monomodal input, but according to Li et al. (2022), adding written text may impact long-term retention due to high cognitive load during comprehension.

The use of multimedia resources opens various possibilities for language learning that prompt language educators to ask questions such as how to best present multimedia content to support listening and viewing comprehension. One important consideration is input sequencing or the order in which different forms of input are presented. Cognitive Load Theory (Sweller, 1994) suggests that the sequencing of input affects how efficiently learners process information. As comprehension can be either supported or impeded depending on how input presentation is structured, poorly timed or excessive input can cause cognitive overload, whereas appropriately sequenced input can ease processing and enhance understanding.

Language educators generally agree that thoughtful planning and implementation are key to turning multimedia into a powerful language learning tool (Jamal et al., 2022). However, the issue is not only what learners watch and listen to, but also the order in which they receive different forms of input. Whether language learners should watch before they listen or listen before they watch, however, is not always that straightforward. For instance, without referring to specific types of input, Romain et al. (2025) suggest introducing the most informative cues that can help language learners make more accurate generalisations. However, Plass

et al. (1998) prefer starting with visual input as it gives language learners helpful contextual scaffolding that supports understanding. On the other hand, Winke et al. (2010, 2013) argue that it is better to start with audio because it encourages deeper linguistic processing that sets a stronger foundation for later visual reinforcement. These different views can be explained through Sweller's Cognitive Load Theory and Mayer's Cognitive Theory of Multimedia Learning. Starting with audiovisual input may help learners build context before focusing more closely on the language, whereas starting with audio-only input may encourage them to focus first on the spoken message before visual support is added. The effectiveness of each sequence may therefore depend on how the different forms of input support processing and affect cognitive load.

Despite the growing interest in the use of multimedia in language teaching and learning, previous studies on multimodal input sequencing remain scarce and inconclusive. Sydorenko (2010) suggests that the way information is presented through different input modalities can influence learners' attention to multimodal input and their learning outcomes. In addition to how input is presented, learner-related factors may also influence how learners respond to different forms of multimodal input. One such factor is modality preference. Liu and Todd (2014) found that the effectiveness of dual-modality input in repeated reading depended on several factors, including learners' preferred modality. This suggests that modality preference may play a role in how learners respond to multimodal input. However, these findings should not be taken to mean that matching instruction to a particular learning style necessarily leads to better comprehension. Pashler et al. (2008) found insufficient evidence to support the claim that learners perform better when instruction is matched to their identified learning styles, while Dekker et al. (2012) caution against assuming that learners process information through only one preferred sensory modality. Thus, individual learner preferences may be relevant in the context of multimodal language learning, but their relationship with input sequence and comprehension remains unclear and requires further investigation.

Although previous studies provide valuable insights into the individual components of input sequencing, the combined effects of monomodal and multimodal input presentation order on learners' comprehension performance are largely unknown, as little research has been conducted on input sequencing in language learning and how it may influence listening and viewing comprehension. Due to the widespread use of audiovisual resources in language instruction, understanding the effects of input sequencing is crucial for optimising teaching strategies. Given these gaps in the literature, this

study aimed to investigate the impact of input sequencing on listening and viewing comprehension in a higher education EFL context. To examine exclusively the effects of audio-before-video (AbV) and video-before-audio (VbA) input sequences on the comprehension performance of first-year English department students, to explore the relationship between learners' input sequence preferences and their comprehension outcomes, and to investigate whether or not learning styles (Visual, Kinaesthetic, Auditory, and Multimodal) correlate with input sequence preference and comprehension performance, this study addresses the following research questions:

1. How does the order in which first-year English department students receive input (i.e., audio-before-video vs. video-before-audio) affect how well they comprehend the multimodal content?
2. Are students inclined to perform better when the input sequence matches their personal preference (whether they prefer to listen first, watch first, or have no strong preference)?
3. Is there any correlation between learning styles (i.e., Visual, Kinaesthetic, Auditory, Multimodal), input sequence preference and comprehension outcomes?

This study extends research on multimodal language learning by examining whether the order of audio and audiovisual input affects learners' comprehension. Moreover, beyond its immediate relevance to ELT, it also addresses the broader educational-technology question concerning how input sequencing influences learners' processing of digital multimedia. Despite its pedagogical relevance, input sequencing has received limited attention in ELT research. The study's novelty lies in examining not only the comparative effects of these sequencing conditions but also their relationships with learners' input sequence preferences and learning styles. By doing so, the findings provide empirical bases for instructional design, supporting informed decisions about how multimodal materials may be sequenced to optimise listening and viewing comprehension. Although the study is situated within a local higher-education EFL context which may limit the generalisability of the findings to other educational or cultural settings, the design principles and pedagogical insights derived from the findings may inform both local and global educational settings where multimedia resources are increasingly integrated into language instruction.

METHOD

This study employed the experimental design for the purpose of examining the effect of input sequencing

on listening and viewing comprehension among first-year English department students. It pursued to explore how two input sequencing conditions, i.e., *video-before-audio* (VbA) and *audio-before-video* (AbV), influence learners' comprehension performance, while also considering two moderating variables, i.e., perceived learning style and input sequence preference.

Research Design

The study adopted a single-group repeated measures design, in which all participants experienced both conditions of input sequencing, i.e., VbA, which gave them an opportunity to watch a video before listening to the corresponding audio and AbV, which let them listen to the audio before watching the corresponding video. Watching in this study means the participants were provided access to aural, visual, and textual input presented in video format. Whereas, listening grants them access to aural input in audio format. This research design allowed for direct comparison of the comprehension scores under both input sequencing conditions

within the same group of participants. Because the same students completed both conditions, differences between participants were less likely to influence the comparison.

The *independent variable* in this study was the input sequence, with two conditions, i.e., VbA and AbV. The *dependent variables* were the comprehension scores, which were measured using a comprehension test administered after each input sequence condition. The test was in four parts. To establish content validity, the test was reviewed by experts who evaluated the relevance and appropriateness of the test items in relation to the input materials and the targeted comprehension skills. The first and third parts were used in the VbA input sequence condition, while the second and fourth parts employed the AbV input sequence condition. Table 1 details the input sequence conditions and comprehension test parts. The study also incorporated *moderating variables*, i.e., input sequence preference and learning style, which were expected to influence how the participants responded to the two input sequencing conditions.

Table 1
Comprehension Test and Input Sequence Conditions Outline

Part	Topic	Video/Audio Length	Input Sequence Conditions	Test Item Number
1	Environment	2:08	VbA	1, 2, 3, 4, 5
2	Learning and Creativity	7:23	AbV	6, 7, 8, 9, 10, 11, 12
3	Happiness	3:08	VbA	13, 14, 15
4	Gratitude	4:28	AbV	16, 17, 18, 19, 20

Note. AbV = audio-before-video, VbA = video-before-audio

Setting and Participants

The participants were 115 first-year English department students aged 18 to 19 from a public university in Indonesia. They were selected through convenience sampling because the researchers had access to the intact classes and needed to work within the available time. This approach is common in educational research (Creswell, 2015). However, while convenience sampling allows for efficient

data collection in naturally occurring classroom contexts of this study, it may limit the generalisability of the findings beyond the specific population studied, as the sample may not fully represent the broader population of EFL learners. Table 2 presents the participants' demographic characteristics, preferred learning styles, input sequence preference, and comprehension performance.

Table 2
Participants' Demographics, Learning Styles, and Preferences (N = 115)

Gender	CEFR Level	Learning Style	Input Sequence Preference	Comprehension Performance
Female	A1	Visual	AbV	AbV > VbA
74.8%(f= 86)	6.1% (f= 7)	23.5% (f= 27)	37.4% (f= 43)	68.7% (f= 79)
Male	A2	Auditory	VbA	VbA > AbV
25.2% (f= 29)	18.3% (f= 21)	3.5% (f= 4)	60.0% (f= 69)	20.0% (f= 23)
	B1	Kinaesthetic	No preference	VbA = AbV
	43.5% (f= 50)	5.2% (f= 6)	2.6% (f= 3)	11.3% (f= 13)
	B2	Multimodal		
	25.2% (f= 29)	67.8% (f= 78)		
	C1			
	7.0% (f= 8)			

Note. f= frequency, AbV = audio-before-video, VbA = video-before-audio, CEFR A1 = Beginner, CEFR A2 = Pre-Intermediate, CEFR B1 = Intermediate, CEFR B2 = Upper-Intermediate, CEFR C1 = Advanced

The participants' gender, perceived learning style, and input sequencing preference were collected through a questionnaire and are summarised in Table 2. An online CEFR-aligned listening test provided a baseline measure of their English proficiency. The questionnaire grouped participants according to the type of input they said they preferred. Preferences for images and video were labelled visual, preferences for spoken input auditory, and preferences for movement or hands-on activity kinaesthetic. Participants who preferred to use both visual and auditory information were included in the multimodal group. These groupings were later compared with the participants' input sequence preferences and comprehension scores.

Data Collection

Initially, the participants completed a listening proficiency test to measure their baseline English proficiency. This baseline measure was obtained using the *British Council EnglishScore* mobile application, a standardised CEFR-aligned assessment that provides an objective estimate of learners' English proficiency levels. Following the proficiency test, participants were exposed to both input sequencing conditions (i.e., AbV and VbA) in a counterbalanced order to control for potential order effects. The study did not involve an extended instructional treatment period. Instead, the sequencing conditions were embedded directly within the experimental tasks. After each sequencing condition, the students answered questions about the audio or video they had just received. Once all tasks were finished, they completed a questionnaire covering demographic information, input sequence preference, and preferred learning style.

Data Analysis

The collected data were analysed using a combination of statistical tests to examine the effects of input sequencing on listening and viewing comprehension and to explore relationships between learning styles, input sequence preferences, and comprehension outcomes. First, an assumptions test was conducted to decide whether parametric or non-parametric tests should be used in the data analysis. A test of normality using Kolmogorov-Smirnov was performed to assess the normality of the distribution of comprehension scores of both conditions. Both scores' p-value < 0.001 presented in Table 3 indicate that the data significantly deviates from normality.

Table 3
Kolmogorov-Smirnov Test Result

Conditions	Statistic	df	Sig.
AbV	.197	115	<.001
VbA	.182	115	<.001

Due to the nature and distribution of the data, non-parametric tests were used in this study. First, Wilcoxon Signed-Rank test was performed to compare the comprehension scores of participants under the two input sequencing conditions. This test was appropriate because the same participants were tested in both conditions, allowing for a direct comparison of their performance across conditions. After that, effect size was calculated as $r = |Z|/\sqrt{N}$, where Z is the standardised test statistic and N is the number of paired observations. The effect size was then interpreted based on Cohen's (1988) conventional benchmarks. Next, Chi-Square test was used to examine the associations between categorical variables, including input sequence preference and comprehension performance. Finally, a correlation analysis using Spearman's correlation was conducted to examine potential relationships between learning styles, input sequence preference, and comprehension outcomes. A significance level of $\alpha = 0.05$ was employed to all statistical tests.

FINDINGS AND DISCUSSION

This section presents the findings and discussion of the study, focusing on how two input sequences, i.e., listening before watching (AbV) versus watching before listening (VbA), influence comprehension and how these effects relate to preferred learning styles and learner preferences and are organised based on the study's three research questions. The findings are supported by statistical analyses and discussed concerning relevant theories and previous studies to make sense of what the data might mean. The sub-section starts with findings on the impact of input sequence on overall comprehension performance and reveals notable differences between the two input sequence conditions. The next sub-section explores the alignment between learners' preferred input sequence and their actual performance to highlight interesting discrepancies between learners' perceptions and their comprehension outcomes. The last sub-section investigates the relationships among input sequence preference, comprehension outcomes, and preferred learning styles. Some trends that offer insights into the interaction of cognitive preferences and multimodal learning emerge and are discussed.

How Input Sequence Affects Comprehension Outcomes

The descriptive analysis results presented in Table 4 reveal that the participants generally performed better when they listened to the audio before watching the video (AbV) rather than the other way around (VbA) as the mean of the AbV condition is considerably higher ($M = 77.54$) compared to that in the VbA condition ($M = 65.65$). The observed mean difference between the two input sequencing

conditions provides preliminary evidence that the audio-before-video sequence may support more effective multimodal processing and comprehension of audiovisual content. Scores were more spread out in the VbA condition ($SD = 15.25$) than in the AbV condition ($SD = 9.57$). This means that performance under VbA was less consistent across participants. Although some participants achieved a score of 100 in both conditions, the wider spread of VbA scores shows that others found this sequence more difficult. Despite these findings, as this study used a single-group repeated-measures design, the differences between the two input sequencing conditions were interpreted with caution. Although the order of the conditions was counterbalanced to reduce possible order, practice, and fatigue effects, the absence of a control group means that the effects of repeated exposure cannot be fully ruled out. Therefore, the findings are not interpreted as strong causal evidence but rather as indicating an association between input sequencing and comprehension performance. Future studies using

control or comparison groups are needed to further examine the causal effects of multimodal input sequencing.

Table 4

Descriptive Analysis of Comprehension Scores

	AbV	VbA
N	115	115
Mean	77.5357	65.6522
Std. Deviation	9.56941	15.25042
Range	50.00	75.00
Minimum	50.00	25.00
Maximum	100.00	100.00
Sum	8916.60	7550.00

Table 5 shows a significant difference between the AbV and VbA scores ($p < .001$). The null hypothesis was therefore rejected, with participants performing better in the AbV condition. The findings confirm that the order of input presentation had a significantly different impact on learners' comprehension.

Table 5

Hypothesis test summary based on Wilcoxon Signed-Rank Test Results

Null Hypothesis	Test	Sig. ^{a,b}	Decision
The median of differences between AbV and VbA equals 0.	Related-Samples Wilcoxon Signed Rank Test	<.001	Reject the null hypothesis.

^aThe significance level is .05.

^bAsymptotic significance is displayed.

The hypothesis test summary of the Related-Samples Wilcoxon Signed Rank test result presented in Table 6 provides additional confirmation that the findings strongly suggest that the overall participants performed significantly better when exposed to AbV, reinforcing the idea that listening to the audio first may facilitate comprehension more effectively than watching the video first, $Z = -6.324$, $p < .001$, with an effect size of $r = .59$, indicating that the difference between the two input sequencing condition was large. Looking at the significant p-value ($p < .001$) also confirms that the difference between the AbV and VbA conditions was not attributable to chance.

Table 6

Related-Samples Wilcoxon Signed Rank Test Summary

Summary	Detail
Total N	115
Test Statistic	736.000
Standard Error	298.935
Standardized Test Statistic	-6.324
Asymptotic Sig. (2-sided test)	<.001
Effect size (r)	.59

The Cognitive Load Theory (Sweller, 1994) provides an explanation for why the AbV group outperformed the VbA group. When language learners were exposed to both audio and visual input

simultaneously from the start, they were likely to experience cognitive overload that may trigger them to struggle to divide attention between processing the language and interpreting visual information. Mayer's Multimedia Learning Theory also offers an explanation to this phenomenon as it suggests that effective multimedia learning occurs when verbal and visual parts are presented in a way that facilitates comprehension without reducing cognitive resources. In this study, the AbV condition reduced extraneous load by first isolating auditory input. Because EFL learners have limited working memory capacity, this input isolation encourages them to focus on one cognitive channel before adding another.

The findings further suggest that in the EFL context, the potential benefits of multimodal language learning based on Paivio's (1986) Dual-Coding Theory, which has long been grounded on the idea that providing input in multiple modalities results in enhanced comprehension, may not always be true as starting with monomodal input, like audio, before introducing additional visual elements is shown to have helped language learners in this study to develop better aural comprehension. The findings also suggest that the AbV sequencing that encourages language learners to selectively focus on the auditory input first leads to better comprehension when the visual information is later added, when they are ready for more complex input

presentation. These are consistent with the Multimodal Ensembles Theory, which claims that the progression from simple to complex modal combinations (Baldry & Thibault, 2006) assists learners in processing meaning more efficiently.

Given the importance of modality processing and cognitive load in multimodal comprehension, the findings also suggest that because the AbV condition enabled language learners to construct a mental representation of the content from the aural input before incorporating more complex visual elements, it resulted in enhanced comprehension. Vandergrift (2007) reasons that early exposure to auditory input enhances comprehension because it allows language learners to process linguistic information before integrating additional multimodal elements. This is because starting with audio establishes a foundation for the subsequent visual reinforcement that comes with videos (Winke et al., 2010, 2013). This notion of initial focus on linguistic information is also promoted by Mayer and Moreno (2003) in their Off-Loading concept that advocates the restriction of visual channel demand by emphasising auditory input.

Although the use of subtitles or visual clues may enhance comprehension due to the rich input, redundant input, such as onscreen text may impede understanding or cause cognitive overload. The findings in this study highlight the need to be selective when using multimedia resources in language teaching and learning. Audiovisual resources such as videos often allow access to written texts in the form of captions, subtitles, or animated scribbles. However, they do not always enhance comprehension as Zheng et al. (2022) and Burin et al. (2024) warn us that visualisations can inhibit comprehension when they are decorative and redundant rather than functional. Yet, Fenesi and Kim (2014) stress that non-redundant designs can improve comprehension and retention. So, spending time listening to the audio without the written text helps language learners focus on the aural input and

avoid overreliance on written text (Chang & Millett, 2014). Alternatively, employing complementary instructional strategies such as pre-task activities or allowing limited first-language use (Warrick, 2021) can be used in language instruction to further reduce cognitive strain and support language learners' gradual engagement with multimodal texts.

The findings highlight the importance of sequencing multimodal content in EFL learning contexts. They imply that AbV is the input sequence model that facilitates more efficient multimedia processing and comprehension compared to VbA. Consequently, language teachers and learners may want to start with listening to the audio before watching a video. This is because initial listening to auditory input is shown to have provided language learners with the cognitive focus that can enhance better comprehension when more layers of multimedia input are introduced later. However, this study involved students from only one Indonesian university. Their results may be linked to the way they are taught and to how familiar they are with audiovisual materials. Learners in other institutions or countries may respond differently. For that reason, the audio-first sequence should be tested with other age groups, subjects, and learning settings before it is applied more widely.

Alignment of Input Sequence Preferences with Comprehension Performance

The second research question investigates the alignment between input sequence preference and comprehension performance. The descriptive analysis results presented in Table 7 show that the degree of alignment varied across preference groups. Among the 115 sample participants, the majority (60%) favoured the VbA input sequence, and slightly more than one-third (37%) preferred AbV as their input sequence preference. A very small number (3%) did not have a specific input preference sequence when asked.

Table 7

Input Sequence Preference and Comprehension Performance Crosstabulation (N= 115)

			Comprehension Performance			Total
			AbV > VbA	VbA > AbV	VbA = AbV	
Input Sequence Preference	AbV	Count	33	4	6	43
		Expected Count	29.5	8.6	4.9	43
	VbA	Count	44	18	7	69
		Expected Count	47.4	13.8	7.8	69
	No Preference	Count	2	1	0	3
		Expected Count	2.1	.6	.3	3
Total		Count	79	23	13	115
		Expected Count	79.0	23.0	13.0	115

Among the participants who preferred AbV input sequencing, the majority (77%) performed better when listening before watching, while only 9% performed better with the other input sequence condition (VbA). A small number (14%)

performed equally well in both conditions. When looking closely into those who preferred VbA, a higher proportion (64%) performed better with AbV, suggesting a misalignment between preference and comprehension performance.

Confirming the misalignment, only 18 students, who preferred watching before listening, performed better in their chosen VbA condition. There is a low percentage (11%) of participants who performed well in both conditions, regardless of their input sequence preference. Finally, among those with no specific preference, two favoured AbV and one preferred VbA. Because the no preference group is

too small, the three participants in this group were not included in the Chi-Square test so as not to spoil the findings.

Table 8 shows that the chi-square test did not provide evidence of a significant association between input sequence preference and comprehension performance, $\chi^2(2, N = 112) = 4.779, p = .092$.

Table 8

Chi-Square Tests Result (N=112)

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	4.779 ^a	2	.092
Likelihood Ratio	5.199	2	.074
N of Valid Cases	112		

^a1 cells (16.7%) have expected count less than 5. The minimum expected count is 4.99.

The likelihood-ratio result was also non-significant ($p = .074$). Thus, the findings further support the lack of a strong relationship between the two variables. Regarding assumption checks, only 16.7% of the cells had an expected count less than 5, with the minimum expected count being 4.99. The expected cell counts were considered acceptable for the analysis. With $p = .092$, the analysis did not provide enough evidence to reject the null hypothesis. The Chi-Square test shows whether

input sequence preference is associated with comprehension performance, but it does not indicate the strength of that association. To see how strong the association was, we examined Cramér's V (see Table 9). The value was .152, which falls within Cohen's (1988) small-effect range. Phi was also low at .216. Overall, the relationship between preference and performance was weak and not statistically significant ($p = .254$).

Table 9

Symmetric Measures of Cramér's V and Phi Coefficients Analysis Result

		Value	Approximate Significance
Nominal by Nominal	Phi	.216	.254
	Cramer's V	.152	.254
N of Valid Cases		115	

One striking outcome of the data analysis is the misalignment between learners' preferences and their comprehension performance. While it might seem intuitive that language learners comprehend better when input sequence matches their preference, the dataset in this study reveal a more complex fact. Among those who preferred AbV, 77% indeed performed better in that condition. However, of those who preferred VbA, a surprising 64% still performed better when they listened before watching, directly contradicting their expectations. Statistically, the relationship between input preference and comprehension performance was not significant ($p > .05$), and the small effect size (Cramér's $V = 0.152$) suggests only a weak association. In short, learner preference was not a reliable predictor of their comprehension performance.

These findings challenge a long-held assumption that instruction should match students' preferences. They are consistent with Kirschner and van Merriënboer's (2013) argument that learners do not always judge accurately which conditions support their learning. Lodge et al. (2015) also note that the effect of a particular modality depends on the task. Familiarity may partly explain the mismatch found in this study. Since students

regularly watch videos, they may feel more comfortable and confident with video-first input, even when it does not lead to better comprehension. Preference, therefore, may reflect familiarity rather than actual learning effectiveness. Lastly, preferences may reflect comfort, confidence, or even entertainment, which are regarded as more beneficial than cognitive benefits. These findings highlight the need for critical reflection and experimentation in EFL teaching and learning. Rather than relying solely on what learners think works best, language educators should guide students to explore a range of input strategies and sequences and reflect on their learning outcomes so that they know what and when the strategies work best.

For students in this study, the sequence they preferred was not always the one that produced better scores. Teachers should therefore consider learners' actual performance when choosing an input sequence and help students reflect on which conditions support their comprehension better.

Relationships between Learning Styles, Input Preference, and Comprehension Outcomes

The demographics of the participants in this study, as presented in Table 2, shows that a majority

identify themselves as having a multimodal learning style, only a very small percentage identify themselves as auditory learners, yet most of them perform well in AbV condition. The Spearman's rho correlation analysis results presented in Table 10

revealed no significant relationships between input sequence preference, learning style, and comprehension outcomes across the two conditions, i.e., AbV and VbA.

Table 10
Summary of Spearman's Rho Correlation Test Result

		Input Sequence Preference	Learning Style	AbV	VbA
Input Sequence Preference	Correlation Coefficient	1.000	-.002	-.123	.046
	Sig. (2-tailed)	.	.980	.191	.626
	N	115	115	115	115
Learning Style	Correlation Coefficient	-.002	1.000	.037	.081
	Sig. (2-tailed)	.980	.	.693	.389
	N	115	115	115	115
AbV	Correlation Coefficient	-.123	.037	1.000	.148
	Sig. (2-tailed)	.191	.693	.	.115
	N	115	115	115	115
VbA	Correlation Coefficient	.046	.081	.148	1.000
	Sig. (2-tailed)	.626	.389	.115	.
	N	115	115	115	115

Note. AbV = Audio before Video, VbA = Video before Audio

Input sequence preference did not strongly predict comprehension outcomes as the correlation between input sequence preference and AbV outcomes was weak and negative ($r_s = -0.123$, $p = 0.191$), while the correlation with VbA outcomes was also weak and non-significant ($r_s = 0.046$, $p = 0.626$). The findings once again confirm that learners' stated preferences for input sequencing were not reliable indicators of their actual comprehension outcomes.

Similarly, learning style showed no significant correlation with comprehension outcomes. The correlation between learning style and AbV outcomes was $r_s = 0.037$, $p = 0.693$, and for VbA outcomes, it was $r_s = 0.081$, $p = 0.389$, both indicating weak, non-significant associations. This suggests that the participants' self-identified learning styles did not meaningfully predict how well they comprehended the multimodal input in either input sequence condition. However, learners' learning styles in this study were measured through self-reported questionnaire data which reflect perceived rather than objectively assessed preferences and may be influenced by individual interpretation or response bias.

Finally, there was only a weak association between comprehension outcomes in the two conditions. The correlation between AbV and VbA performance was $r_s = 0.148$, $p = 0.115$, suggesting that a student's success in one input sequence condition did not strongly predict their performance in the other. The findings confirm that neither input sequence preference nor learning style are significantly correlated with comprehension outcomes. Performance in one condition therefore gave little indication of how a participant would perform in the other. The variation in scores may instead relate to factors not measured in this study,

such as working memory, attention, prior knowledge, or familiarity with the topic.

What can be highlighted from the findings is the weak, non-significant correlation between perceived learning styles and comprehension performance in both the AbV and VbA conditions and how they are consistent with the growing scepticism of traditional learning style theories, especially the idea that instruction should be customised to learners' learning styles (Pashler et al., 2008). Despite the popularity of this belief in the educational field, more and more studies suggest that learners do not necessarily perform better when taught in their so-called preferred style, e.g., auditory or visual (Rogowsky et al., 2014). This is partly because effective learners tend to flexibly adapt their processing strategies, rather than relying on fixed learning habits (Dörnyei, 2005) and partly because compensatory techniques may bridge the gap. For instance, a learner who identifies themselves as a visual learner may still succeed in the AbV condition by mentally visualising the content while listening. Moreover, comprehension success depends not just on the formats of the input, i.e., audio and visual formats, but also on how learners manage attentional load, task demands, and cognitive effort (Liu & Todd, 2014). Thus, reasons why learning styles do not predict comprehension outcomes in the EFL context are that cognitive strategies may matter more than learning styles and preferences.

While language learners may have input sequence preferences and perceived learning styles, these do not signify better multimodal comprehension outcomes. Instead, it appears that the participants' comprehension performance is more strongly influenced by how they engage with the content and their ability to manage and integrate

information across modalities. The study's findings further reinforce the idea that a flexible, strategy-based approach to language learning is more favoured than rigidly matching instruction to learning style labels. Thus, not categorising language learners into rigid learning styles and teaching them cognitive and metacognitive strategies that aid them in processing information across multiple modalities can empower them to manage input from multiple channels and make informed choices about how they learn best in different contexts.

CONCLUSION

By successfully answering the three research questions, the study provides valuable empirical evidence on how input sequencing influences audiovisual comprehension, the misalignment between preference and performance, and the partial role of learning styles in predicting comprehension. Firstly, the findings show that learners performed considerably better when they listened to the audio before watching the video (AbV) rather than the reverse (VbA). They reinforce the idea that AbV facilitates comprehension more effectively than VbA as the difference between the two input sequence conditions was statistically significant.

Secondly, the findings also suggest a noteworthy mismatch between input sequence preference and comprehension performance. While the majority of those who preferred AbV performed better in that condition, a surprisingly high percentage of learners who preferred VbA also performed better with AbV. Yet the statistical analysis confirmed that there is no significant correlation between input sequence preference and comprehension performance. It can be concluded that learners' preferences were not reliable indicators of their actual comprehension performance.

Finally, no significant relationships were found between input sequence preference, learning styles, and comprehension performance. These findings imply that learners' self-identified learning styles did not predict their success in the two input sequence conditions because neither individual preference nor learning style played a key role in comprehension. The findings further highlight the potential influence of other cognitive or contextual factors in multimodal learning.

The results show that adding more modes at the same time does not necessarily improve comprehension. For classroom use, teachers could begin with the audio and introduce the video afterwards, especially when the main aim is to help students understand the content. At the curriculum level, input sequencing can be treated as a deliberate design choice, with sequenced multimodal tasks that allow learners to process auditory and visual

information progressively. In digital and app-based learning environments, the findings suggest that providing flexible sequencing options, such as audio-first modes, may better support learners' comprehension, especially in self-directed contexts.

Despite limitations related to context, design, and self-reported measures, the findings offer pedagogically relevant insights into multimodal input sequencing. By highlighting input sequencing as an important factor in learners' comprehension of multimodal materials, the findings provide practical guidance for instructional design in EFL classrooms, teacher education, and digitally mediated learning environments. In particular, the findings suggest that thoughtful sequencing of multimodal input, especially audio-before-video (AbV), may support more effective comprehension across a range of educational contexts where multimedia resources are increasingly used. Accordingly, language teachers and instructional designers are encouraged to consider employing various multimodal input sequencing, more particularly AbV, when designing, creating, and using language learning materials, ensuring that input presentation aligns with cognitive principles rather than learner preferences alone. Future research may further extend this study by exploring some additional cognitive factors, such as working memory and attentional control, to better understand the mechanisms behind multimodal input processing and sequencing.

AI USE ACKNOWLEDGEMENT

The authors used ChatGPT (OpenAI) to help improve the language and clarity of the manuscript. All suggested revisions were checked by the authors, who remain responsible for the final content of the article.

CONTRIBUTION OF THE AUTHORS

The first author was mainly responsible for designing the overall research, determining the methodology, conducting the main data analyses for the study, and writing the manuscript. The second author helped with data collection, partial data analysis, and manuscript writing. Both authors reviewed and approved the final version of the manuscript.

REFERENCES

- Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and analysis: A multimedia toolkit and coursebook*. Equinox.
- Burin, D. I., González, F. M., Marrujo, J. G., & Martínez, M. A. (2024). Online instructional videos: Effects of presentation format on comprehension and metacomprehension. *The*

- Journal of Experimental Education*, 94(1), 46–66.
<https://doi.org/10.1080/00220973.2024.2376606>
- Chang, A., Millett, S., & Renandya, W. A. (2018). Developing listening fluency through supported extensive listening practice. *RELC Journal*, 50(3), 422–438. <https://doi.org/10.1177/0033688217751468> (Original work published 2019)
- Chang, C.-S., & Millett, S. (2014). The effect of extensive listening on developing L2 listening fluency: Some hard evidence. *ELT Journal*, 68(1), 31–40.
<https://doi.org/10.1093/elt/cct052>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Creswell, J. W. (2015). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (5th ed.). Pearson.
- Dekker S, Lee NC, Howard-Jones P and Jolles J (2012) Neuromyths in Education: Prevalence and Predictors of Misconceptions among Teachers. *Front. Psychology* 3(429).
<https://doi.org/10.3389/fpsyg.2012.00429>
- Donaghy, K. (2020). *Advancing learning: The fifth skill—‘viewing’*. Macmillan Education.
<https://www.onestopenglish.com/professionaldevelopment/advancing-learning-the-fifth-skill-viewing/557577.article>
- Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Lawrence Erlbaum Associates, Inc.
- Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, 88, Article 102166.
<https://doi.org/10.1016/j.system.2019.102166>
- Fenesi, B., & Kim, J. A. (2014). Learners misperceive the benefits of redundant text in multimedia learning. *Secondary Education Psychology*, 5, 1–7. <https://doi.org/10.3389/fpsyg.2014.00710>
- Jamal, M., Syed, S. A., & Metwally, A. A. (2022). Investigating realities and misconceptions about multimedia learning. *International Journal of Science and Research (IJSR)*, 11(4), 673–677.
<https://doi.org/10.21275/SR22409093304>
- Jewitt, C., & Kress, G. (2010). Multimodality, literacy and school English. In D. Wyse, R. Andrews, & J. Hoffman (Eds.), *The Routledge international handbook of English, language and literacy teaching* (pp. 342–352). Routledge.
<https://doi.org/10.4324/9780203863091-33>
- Jewitt, C. (2014). Multimodal approaches. In S. Norris & C. D. Maier (Eds.), *Interactions, images and texts: A reader in multimodality* (pp. 127–136). De Gruyter Mouton.
<https://doi.org/10.1515/9781614511175.127>
- Karabıyık, C., & Arslan, S. (2022). Comparison of input modes: L2 comprehension and cognitive load. *Participatory Educational Research (PER)*, 9(6), 173–191.
<http://dx.doi.org/10.17275/per.22.134.9.6>
- Kirschner, P. A., & van Merriënboer, J. J. G. (2013). Do learners really know best? Urban legends in education. *Educational Psychologist*, 48(3), 169–183. <https://doi.org/10.1080/00461520.2013.804395>
- Li, S. (2022). The efficacy of multimodal input on ESL learners’ listening and speaking abilities. *Sino-US English Teaching*, 19(8), 263–269.
<https://doi.org/10.17265/1539-8072/2022.08.001>
- 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, Article 834706.
<https://doi.org/10.3389/fpsyg.2022.834706>
- Liu, Y. T., & Todd, A. G. (2014). Dual-modality input in repeated reading for foreign language learners with different learning styles. *Foreign Language Annals*, 42(4), 684–706.
<https://doi.org/10.1111/flan.12113>
- Liu, Y., Jang, B. G., & Roy-Campbell, Z. (2018). Optimum input mode in the modality and redundancy principles for university ESL students’ multimedia learning. *Computers & Education*, 127, 190–200.
<https://doi.org/10.1016/j.compedu.2018.08.025>
- Lodge, J. M., Hansen, L., & Cottrell, D. (2015). Modality preference and learning style theories: rethinking the role of sensory modality in learning. *Learning: Research and Practice*, 2(1), 4–17.
<https://doi.org/10.1080/23735082.2015.1083115>
- Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press, 2(2). <https://doi.org/10.1017/CBO9780511811678>
- Mayer, R. E. (2017). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 43–71). Cambridge University Press.
<https://doi.org/10.1017/cbo9781139547369.005>
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52.
http://dx.doi.org/10.1207/S15326985EP3801_6
- O’Halloran, K. L. (2008). Systemic functional-multimodal discourse analysis (SF-MDA):

- Constructing ideational meaning using language and visual imagery. *Visual Communication*, 7(4), 443–475.
<https://doi.org/10.1177/1470357208096210>
- Pahlavani, P., & Salehi, A. S. (2023). Viewing comprehension and vocabulary size as predictors of vocabulary learning strategies among Iranian EFL learners. *International Journal of English Language, Education and Literature Studies (IJEEL)*, 2(6), 24–38.
<https://doi.org/10.22161/ijeel.2.6.4>
- Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.
<https://archive.org/details/mentalrepresenta0000paiv/page/n5/mode/2up>
- Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, 9(3), 105–119.
<https://doi.org/10.1111/j.1539-6053.2009.01038.x>
- Perez, M. M. (2020). Incidental vocabulary learning through viewing video: the role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, 42(4), 749–773.
<https://doi.org/10.1017/S0272263119000706>
- Perez, M. M. (2022). Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching*, 55(2), 163–192.
<https://doi.org/10.1017/S0261444821000501>
- Peters, E., & Muñoz, C. (2020). Introduction to the special issue: Language learning from multimodal input. *Studies in Second Language Acquisition*, 42(3), 489–497.
<https://doi.org/10.1017/S0272263120000212>
- Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology*, 90(1), 25–36.
<https://doi.org/10.1037/0022-0663.90.1.25>
- Rodgers, M. P. H., & Webb, S. (2020). Incidental vocabulary learning through viewing television. *International Journal of Applied Linguistics*, 171(2), 191–220.
<https://doi.org/10.1075/itl.18034.rod>
- Rogowsky, B. A., Calhoun, B. M., & Tallal, P. (2014). Matching learning style to instructional method: Effects on comprehension. *Journal of Educational Psychology*, 107(1), 64–78.
<https://doi.org/10.1037/a0037478>
- Rogowsky, B. A., Calhoun, B. M., & Tallal, P. (2016). Does modality matter? The effects of reading, listening, and dual modality on comprehension. *SAGE Open*, 6(3), 1–9.
<https://doi.org/10.1177/2158244016669550>
- Romain, L., Milin, P., & Divjak, D. (2025). Order effects in second language learning. *Language Learning*, 75(3), 623–665.
<https://doi.org/10.1111/lang.12675>
- Sejdiu, S. (2017). Are listening skills best enhanced through the use of multimedia technology. *Digital Education Review*, 32, 60–72.
<https://doi.org/10.1344/der.2017.32.60-72>
- Surkamp, C., & Yearwood, T. (2018). Receptive competences—reading, listening, viewing. In Surkamp, C., Viebrock, B. (Eds.), *Teaching English as a foreign language*. J. B. Metzler, Stuttgart.
https://doi.org/10.1007/978-3-476-04480-8_6
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312.
[https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology*, 14(2), 50–73.
<https://doi.org/10.12968/lnic.2010.14.2.50>
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40, 191–210.
<https://doi.org/10.1017/S0261444807004338>
- Vanderplank, R. (2016). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. Palgrave Macmillan.
<https://link.springer.com/book/10.1057/978-1-137-50045-8>
- Warrick, A. (2021). Strategies for reducing cognitive overload in the online language learning classroom. *International Journal of Second and Foreign Language Education*, 1(2), 25–37.
<https://doi.org/10.33422/ijfsle.v1i2.124>
- Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics*, 30(3), 407–427.
<https://doi.org/10.1093/applin/amp010>
- 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.
<http://dx.doi.org/10.125/44203>
- Winke, P., Sydorenko, T., & Gass, S. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, 97(1), 254–275.
<http://www.jstor.org/stable/23361749>
- Yi, Q., Dong, Z., & Qiao, H. (2024). Enhancing EFL listening and speaking skills: Strategies and practice for implementing multimedia and multi-modal approaches. *Journal of Humanities, Arts and Social Science*, 8(9), 2211–2217.
<https://doi.org/10.26855/jhass.2024.09.032>
- Zhang, R., & Zou, D. (2022). A state-of-the-art review of the modes and effectiveness of multimedia input for second and foreign

- language learning. *Computer Assisted Language Learning*, 35(9), 2790–2816.
<https://doi.org/10.1080/09588221.2021.1896555>
5
- Zheng, Y., Ye, X., Hsiao, J. H. (2022). Does adding video and subtitles to an audio lesson facilitate its comprehension? *Learning and Instruction*, 77, 1–13.
<https://doi.org/10.1016/j.learninstruc.2021.101542>