International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 7, pp. 645-667, July 2026
https://doi.org/10.26803/ijlter.25.7.29
Received Apr 6, 2026; Revised May 14, 2026; Accepted Jun 29, 2026

Multimodal Design or Communicative Input? ESL Learner Perceptions of TikTok-Inspired Vocabulary Instruction

Jane Xavierine*
INTI International University, Nilai 71800, Malaysia

Norshima binti Zainal Shah
Universiti Pertahanan Nasional, Kuala Lumpur, 57000, Malaysia

Alice Shanthi
Akademi Pengajian Bahasa, Universiti Teknologi MARA (UiTM)
Cawangan Negeri Sembilan, Kampus Seremban
70300, Malaysia

Khayrullayeva Nodira Nematilloyevna
Bukhara State University, 705018, Uzbekistan

Raziya Seydabullaeva
Department of Preschool and Primary Education
University of Innovation Technologies
Uzbekistan, Nukus

Abstract. Short-form video applications such as TikTok offer instructional environments that combine multimodal design with communicative input modes, yet how ESL learners perceive these features in formal classroom contexts remains underexplored. This quantitative descriptive correlational survey study examined the perceptions of 92 ESL learners regarding seven multimodal feature constructs embedded in TikTok-inspired vocabulary instruction. The constructs were derived from a prior qualitative multimodal discourse analysis of 10 authentic TikTok vocabulary videos (Phase 1 of the broader research programme), which informed the design of a validated questionnaire (Cronbach's alpha = .840 to .920) administered within an Intensive English Programme at a private university in Malaysia. Data were analysed using IBM SPSS Statistics version 31. All seven constructs were positively endorsed above the scale midpoint. However, two descriptive patterns emerged. Constructs related to the stabilisation of vocabulary meaning through on-screen text, visual support, and

Citation:
Jane, X., Shah, N, B, Z., Shanthi, A., Nematilloyevna, K, N., & Mamitovich, A, B. (2026). Multimodal Design or Communicative Input? ESL Learner Perceptions of TikTok-Inspired Vocabulary Instruction. *International Journal of Learning, Teaching and Educational Research*, *25*(7), 645–667. https://doi.org/10.26803/ijlter.25.7.29

*Corresponding author: Jane Xavierine; *jane.xavier@newinti.edu.my*

consistent video format received stronger and more consistent endorsement. In contrast, constructs associated with communicative input delivery (narrative framing, audio delivery, embodied cues, and interactive engagement) were endorsed positively, however, with greater individual variability. These patterns informed a tentative exploratory framework distinguishing Design Features from Multimodal Input, offered here as a hypothesis-generating rather than a confirmed framework. Structural validation has been undertaken in a subsequent phase of the research programme using a sample of n = 304 appropriate for structural equation modelling, and is reported separately. The findings provide evidence-based guidance for educators selecting and sequencing TikTok-inspired vocabulary materials, consistent with United Nations Sustainable Development Goal 4.

Keywords: ESL vocabulary instruction; learner perception; multimodal design; short-form video; TikTok-inspired learning; quantitative survey

1. Introduction

Short-form video platforms have fundamentally altered how language learners encounter and engage with new vocabulary outside formal educational settings. TikTok, in particular, has emerged as a widely used informal learning resource, offering compressed instructional content that integrates spoken language, on-screen text, visual imagery, and embodied demonstration within video formats typically under one minute. Within these environments, multimodal features refer to the coordinated deployment of linguistic, visual, auditory, embodied, and spatial modes of communication across a single instructional text. Short-form video supports microlearning, increases student engagement, and improves access to language input (Yang, 2020; Khlaif and Salha, 2021). Recently, researchers have been investigating the extent to which TikTok serves as an informal learning resource and as a platform whose design conventions may be adapted to meet structured classroom needs (Abas et al., 2025; Adiyono et al., 2025; Jumaat et al., 2025).

TikTok's ability to function as a learning tool is a direct result of its multimodal composition. Traditional educational videos do not require the same level of coordination across multiple streams of semiotic information as short-form videos under time constraints. Multimodal theory suggests that meaning is created in digital environments through the coordinated use of linguistic, visual, auditory, embodied, and spatial forms of representation (Kress, 2010; Jewitt, 2009). These forms are interdependent in providing context for new vocabulary introduced in a dense, highly layered communicative event.

Together, the Cognitive Theory of Multimedia Learning and Cognitive Load Theory offer complementary explanations of how learners process and respond to instructional content, depending on the alignment of multimodal components (Sweller et al., 2019; Mayer, 2021). Furthermore, Godwin-Jones (2000) cautioned that while emerging digital technologies should be evaluated based on their cognitive compatibility rather than solely on their ability to engage users, this

caution remains applicable to the incorporation of short-form video into formal ESL classrooms.

Despite growing interest in TikTok as an educational technology, existing research has largely focused on learners' attitudes, motivation, and engagement rather than on the specific multimodal design features learners identify as supporting vocabulary learning in formal classroom contexts (Punkhoom et al., 2025; Yang, 2020). This distinction is important because the instructional viability of design features cannot be deduced merely from learner engagement data. Without construct-level evidence of how learners evaluate specific multimodal features, instructors lack an empirical basis for selecting and sequencing short-form video content for vocabulary instruction.

This study addresses this gap through a quantitative survey of ESL learners' perceptions of seven multimodal feature constructs in TikTok-inspired vocabulary instruction. The study was guided by two research questions: (1) Which of the seven multimodal feature constructs (Narrative Structure and Storytelling, Visual Elements, Audio Elements, Gestural and Spatial Elements, Interactive Elements, Subtitles and Text, and Video Format and Duration) received the strongest and most consistent learner endorsement? (2) What patterns of functional differentiation emerged when learner responses were examined across these constructs?

2. Literature Review

2.1 Multimodal Design in Digital Language Learning

Multimodal theory provides the theoretical foundation for understanding how meaning is produced in digitally mediated instructional settings. Specifically, Jewitt (2009) and Kress (2010) argued that meaning in contemporary communication is not carried by language alone. Rather, it is produced through the coordinated interaction of multiple semiotic modes, including linguistic, visual, auditory, embodied, and spatial resources.

The Multiliteracies Framework put forth by the New London Group (1996) extended the work of Jewitt (2009) and Kress (2010) by recognising that students in digital learning environments create meaning using diverse modal resources, and placed an obligation on instructional designers to determine how to coordinate modal integration into their instructional materials. Moreover, Jewitt et al. (2016) further synthesised these theoretical foundations, demonstrating how multimodal resources function as integrated systems of meaning across diverse communicative and educational contexts.

The significance of multimodal design is particularly pronounced in second language learning contexts. Jiang and Hafner (2024) demonstrated that how instructional designers frame multimodal composing tasks significantly determines whether multiple modes support or hinder learner comprehension, confirming that the presence of multiple modes alone does not guarantee learning. Similarly, Teng (2023) demonstrated that multimodal vocabulary instruction produced higher engagement and recall only when the visual and

auditory structure supported specific learning objectives. Li and Hafner (2022) reported comparable findings in mobile-assisted learning environments, concluding that positive outcomes depended on clear structural scaffolding rather than on the quantity of modal resources. Duan et al. (2026) similarly discovered that multimodal input and interactive exercises are consistently associated with positive vocabulary retention outcomes in mobile-assisted environments, reinforcing the design principles identified in the present study. Collectively, these studies establish that the functional organisation of multimodal elements, rather than their sheer number, defines the instructional value of multimodal vocabulary learning environments.

2.2 Cognitive processing frameworks for video-based instruction

The Cognitive Theory of Multimedia Learning proposes that learners process multimedia materials through separate visual and verbal channels, and that meaningful learning occurs only when neither channel is overloaded. Mayer and Moreno (2003) and Mayer (2021) identified key design principles, including signalling, coherence, temporal contiguity, and redundancy, to guide the alignment of multimodal components while minimising extraneous cognitive load. Furthermore, Paivio (1986) complemented this framework through dual coding theory, which posits that verbal and imagistic representational systems operate as separate cognitive mechanisms. Note that instructional content that simultaneously activates both systems supports deeper encoding than input confined to a single channel.

Nevertheless, these frameworks present particular challenges for TikTok-style video design. Brevity is the defining feature of short-form content, requiring designers to compress information across multiple modalities within seconds and offering learners a limited opportunity for review or elaborative processing. Traditional design principles were developed for longer instructional formats. They must be independently evaluated for their applicability to temporally compressed social media content (Godwin-Jones, 2000).

2.3 TikTok as a vocabulary learning technology in ESL contexts

Interest in TikTok as an educational technology has grown substantially, with studies reporting positive learner reactions to its microlearning affordances for vocabulary learning (Alshreef and Khadawardi, 2023; Khlaif and Salha, 2021; Yang, 2020). In particular, Tan et al. (2022) specifically examined TikTok's platform features from a pedagogical design perspective, identifying video-based and duet-challenge features as having the greatest potential to promote meaningful and engaging language learning in ESL contexts.

In addition, emerging empirical evidence supports its instructional potential in formal ESL settings. Thanga Rajan and Ismail (2022) determined that Malaysian secondary students whose literature lessons incorporated TikTok outperformed a control group on all measures ($p = .001$), confirming its feasibility as a formal instructional tool. In another Malaysian study, Shanthi et al. (2024) proposed an analytical framework specifically for evaluating TikTok vocabulary videos at the feature level, identifying multimodal categories relevant to instructional design. Shamshul et al. (2024) further confirmed that digital technologies are increasingly

embedded in Malaysian ESL classrooms. Their review emphasised the need for structured pedagogical frameworks to support meaning-making across different content types. Related studies across Thailand and Indonesia have similarly reported positive learner perceptions of TikTok-mediated language content (Punkhoom et al., 2025; Rasyid et al., 2023).

A common limitation across this body of research is its focus on platform-level engagement and general attitudes rather than on specific multimodal design features. Nation (2013), Schmitt (2010), and Webb and Chang (2015) established that vocabulary acquisition depends on repeated, contextually varied encounters with target words. However, no research has examined whether learners identify specific multimodal features in short-form video as cognitively manageable and pedagogically meaningful for acquisition at the construct level.

2.4 Learner perception as design-relevant evidence

Learner perception has been established as a valid basis for evaluating instructional design in technology-enhanced environments, particularly when the aim is to determine whether learners find materials clear, coherent, and cognitively manageable rather than merely engaging (Kucuk and Richardson, 2019). In the Malaysian ESL context specifically, Raman et al. (2023) demonstrated that learner engagement and motivation in technology-enhanced environments are closely tied to the perceived clarity and structural coherence of the instructional design, rather than to the novelty of the technology itself.

On the other hand, Rasyid et al. (2023) surveyed 207 learners in Indonesia regarding their use of English-language TikTok content, reporting positive perceptions of competence across all items (60.1% to 94.7% agreement) and identifying attractiveness, effectiveness, relevance, and motivation as recurring themes. While informative about platform utility, these findings do not address learners' perceptions of specific multimodal design features within formal instructional content.

A theoretical gap therefore exists regarding how multimodal design elements contribute to learners' meaning construction in time-compressed instructional contexts. A practical gap exists regarding valid criteria for educators evaluating short-form video materials for formal ESL vocabulary instruction. In response, this study addresses both gaps by examining how 92 ESL learners in Malaysia perceived seven multimodal feature constructs embedded in TikTok-inspired vocabulary instruction using a validated quantitative instrument grounded in prior systematic analysis of actual instructional video content.

3. Methodology

3.1 Research design

This study adopted a quantitative descriptive correlational survey design to examine ESL learners' perspectives on multimodal elements in TikTok-inspired vocabulary instruction. Note that the decision to employ a survey-based methodology enabled the systematic collection of perception-level data within a specific group of participants. It supported the evaluation of associations among constructs via correlation analyses (Creswell and Creswell, 2018). This design is

well-suited to exploring patterns in learner experience (rather than establishing cause-and-effect or measuring student learning) by describing and interpreting them.

3.2 Setting and participants

The study was conducted at a private university in Malaysia. This private university was selected for its ability to provide a homogeneous linguistic population of ESL learners who received similar, structured English-language instruction. Furthermore, purposive sampling was used to select participants. To be eligible to participate, students needed to be currently enrolled in the intensive English program, have attended sessions featuring TikTok-inspired vocabulary videos, and have sufficient English proficiency to complete the survey instrument. Consequently, a total of 92 learners were included in the final analysis after completing the full sequence of TikTok-inspired vocabulary lessons. Here, voluntary participation occurred without compensation. This is standard procedure for non-interventionist surveys conducted in higher education institutions.

The sample size of 92 was determined by the size of the available cohort completing the full six-week instructional sequence during the data collection window. This sample is adequate for descriptive analysis and Pearson correlation. Pallant (2020) demonstrated that Pearson estimates remain reliable under moderate non-normality for samples exceeding n = 20, and Cohen (1988) recommends samples above 80 to detect moderate effects at conventional alpha levels. The sample, however, is insufficient for confirmatory factor analysis or structural equation modelling, which require a larger, more heterogeneous sample. Hence, structural validation was addressed in a subsequent phase of the research programme.

Table 1: Demographic Profile of Participants (*N* = 92)

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	38	41.3
	Female	54	58.7
	Total	**92**	**100.0**
Age range	18 to 20 years	29	31.5
	21 to 23 years	47	51.1
	24 to 26 years	13	14.1
	27 years and above	3	3.3
	Total	**92**	**100.0**
English proficiency level	Pre intermediate	92	100.0
	Total	**92**	**100.0**

3.3 Video Corpus and Selection Criteria

The survey instrument was developed from Phase 1 of the broader research programme, in which multimodal discourse analysis was conducted on 10 purposively selected TikTok vocabulary videos (five institution-produced, five creator-produced) using ATLAS.ti. Notably, the analysis was guided by a five-mode framework: linguistic, visual, auditory, embodied, and spatial. Recurring multimodal features and orchestration patterns identified across the corpus were translated into seven provisional construct domains for Phase 2 instrument

development. The expert panel that validated the instrument comprised five reviewers with expertise in multimodal discourse analysis and ESL instruction at the tertiary level.

Table 2 summarises the four sets of inclusion criteria that were used to select these videos.

Table 2: Inclusion Criteria for TikTok Video Corpus Selection

Criterion	Requirement
Instructional focus	Explicit and primary focus on English vocabulary teaching; vocabulary instruction must be the main rather than incidental purpose of the video.
Multimodal character	Content must combine at least two semiotic modes, including spoken language, on-screen text, visual elements, embodied demonstration, or spatial organisation; videos relying solely on spoken word or written text were excluded.
Learner suitability	Content must be appropriate for pre-intermediate ESL learners, assessed based on vocabulary level, pace of explanation, clarity of delivery, and degree of contextual support provided; suitability judgements were verified by an experienced ESL practitioner familiar with the pre-intermediate level.
Technical quality	Videos must have been publicly available on TikTok at the time of data collection and must have sufficient audio and visual quality to permit reliable identification of multimodal components.

Note. Corpus size was determined by analytical saturation. No substantively new multimodal patterns or coding categories emerged from the eighth video onward.

Questionnaire items were generated from the recurring multimodal features identified in the earlier video analysis. Consequently, the resulting instrument was organised into seven construct domains: Narrative Structure and Storytelling, Visual Elements, Audio Elements, Gestural and Spatial Elements, Interactive Elements, Subtitles and Text, and Video Format and Duration. Each construct contained six learner-facing items, producing a total of 43 items. Additionally, items were written from the learner's perspective, framed positively, and measured on a five-point Likert scale ranging from Strongly Disagree to Agree Strongly.

Subsequently, the initial item pool underwent two rounds of structured evaluation by five reviewers with expertise in multimodal discourse and ESL. Each reviewer rated the instrument on a five-point scale across eight evaluation criteria, including the degree of correspondence between items and research objectives, the inclusion of relevant multimodal aspects, the clarity for ESL students, and the absence of bias in wording. According to COSMIN (Mokkink et al., 2010) principles, content validity was assessed using the Content Validity Index. Ratings of 4 or higher indicated "acceptance" for the S-CVI/Ave. Round 1 yielded an S-CVI/Ave of .90, with four criteria at I-CVI = 1.00 and four at I-CVI =

.80. Meanwhile, items not achieving full agreement were revised before Round 2, which demonstrated total agreement across all criteria (all I-CVIs = 1.00, S-CVI/Ave = 1.00), confirming alignment with the study's purpose.

Table 3: Item-Level and Scale-Level Content Validity Index for the Survey Instrument Across Two Rounds of Expert Review (N = 5 experts)

Evaluation criterion	I-CVI Round 1	I-CVI Round 2
Alignment with research objectives	1.00	1.00
Coverage of key multimodal and design features	0.80	1.00
Suitability of Likert scale items for analysis	1.00	1.00
Clarity of language for ESL learners	0.80	1.00
Suitability of questionnaire format	1.00	1.00
Appropriateness of length and complexity	0.80	1.00
Absence of leading or biased wording	0.80	1.00
Suitability for data collection and analysis	1.00	1.00
S-CVI/Ave	0.90	1.00

Following expert validation, the questionnaire was pilot-tested with a comparable ESL cohort to assess clarity and internal consistency before the main administration. In the main study sample of 92 participants, internal consistency remained strong across all subscales, with Cronbach's alpha values ranging from .840 for Video Format and Duration to .920 for Subtitles and Text, all exceeding the accepted threshold of .70 (Hair et al., 2019).

Table 4: Internal Consistency Reliability of Questionnaire Subscales

Construct	No. of items	Cronbach's alpha
Subtitles and Text	6	.920
Interactive Elements	6	.916
Gestural and Spatial Elements	6	.914
Visual Elements	6	.890
Narrative Structure and Storytelling	6	.876
Audio Elements	6	.865
Video Format and Duration	7	.840

3.4 Data collection procedure

The relevant institutional authority granted ethical approval for the study before data collection began. Informed of the voluntary nature of their participation and the anonymous nature of their answers, participants could withdraw at any time without consequences. Accordingly, all participants provided written consent before the questionnaires were administered.

Correspondingly, data were collected after participants completed a standardised sequence of classroom activities that incorporated TikTok-style vocabulary videos

within their intensive English programme. Specifically, each participant was shown one of six different videos per week for six weeks. The video for each week was selected based on the unit of study for that week, corresponding to the units outlined in the coursebook (Cutting Edge - pre-intermediate) used by the students. The six unit areas studied during the six weeks were: daily routines and habits, childhood memories, places visited while traveling, memorable experiences and personal characteristics, employment and future aspirations, and the use of narrative language. Approximately 8 to 10 target vocabulary items were introduced in each unit. These vocabulary items were not taught in isolation but were embedded within the broader curriculum.

Table 5: Six-Week Vocabulary Intervention: Unit Topics and Target Vocabulary

Week	Unit Topic	TikTok lesson focus	Target vocabulary
1	Daily routines and habits	Describe your daily routine like a native speaker	wake up, make my bed, putting on make-up, doing your hair, have breakfast, scroll through your phone
2	Childhood memories	Talking about childhood memories using *used to*	remember, forget, used to, childhood, memory, habit, recall, remind
3	Travel destinations	How to sound natural when describing a city or country	beautiful, crowded, peaceful, exciting, ancient, modern, lively, scenic, remote, local
4	Life experiences and personal qualities	How to sound natural when describing a city or country	experience, achieve, challenge, success, confident, creative, hard-working, generous
5	Work and future goals	How to express your future goals in English	career, goal, dream, future, succeed, opportunity, motivate, focus
6	Storytelling language	Five tips for telling a story in English	say, tell, suddenly, finally, then, before that, after that, meanwhile

Note. All vocabulary items were drawn from the corresponding unit of Cutting Edge (3rd ed., pre-intermediate; Cunningham et al., 2013)

Each lesson followed a consistent sequence: topic introduction, two video viewings, a guided explanation, and a short practice task. Consequently, coursebook activities were used to reinforce the target vocabulary.

3.5 Data analysis

Data were analysed using IBM SPSS Statistics version 31 (Field, 2018). Skewness and kurtosis were computed to assess normality before correlation analyses. Skewness ranged from -0.730 to -1.325 and kurtosis from 0.803 to 2.550, indicating moderately negatively skewed distributions consistent with ceiling effects in

perception surveys. Two constructs exceeded the ±1.00 skewness threshold: Narrative Structure and Storytelling (skewness = -1.325) and Visual Elements (skewness = -1.292). Pearson correlation was nonetheless retained given the sample size of 92. Pallant (2020) demonstrated that Pearson estimates remain reliable under moderate skewness for samples exceeding n = 20. Note that correlations were computed for all 21 pairwise construct combinations and tested at p ≤ .05.

4. Results

4.1 Descriptive statistics: Learner endorsement across constructs

The first research question asked which multimodal feature constructs received the strongest and most consistent learner endorsement. Given the social media pedagogy literature's emphasis on engagement and interactivity as TikTok's primary instructional value (Khlaif and Salha, 2021; Yang, 2020), an engagement-first pattern – in which interactive and participatory features would dominate endorsements – served as the expected baseline against which the findings were interpreted. Figure 1 illustrates the mean endorsement scores and standard deviations across all seven constructs.

Table 6: Descriptive Statistics for Multimodal Feature Constructs (N = 92)

Construct	*M*	*SD*	Skewness	Kurtosis
Subtitles and Text	3.92	0.86	−1.039	1.551
Visual Elements	3.90	0.84	−1.292	2.509
Narrative Structure and Storytelling	3.84	0.91	−1.325	2.550
Video Format and Duration	3.78	0.79	−0.958	1.997
Gestural and Spatial Elements	3.70	0.88	−0.730	0.803
Audio Elements	3.70	0.89	−0.900	0.820
Interactive Elements	3.65	0.90	−0.735	1.215

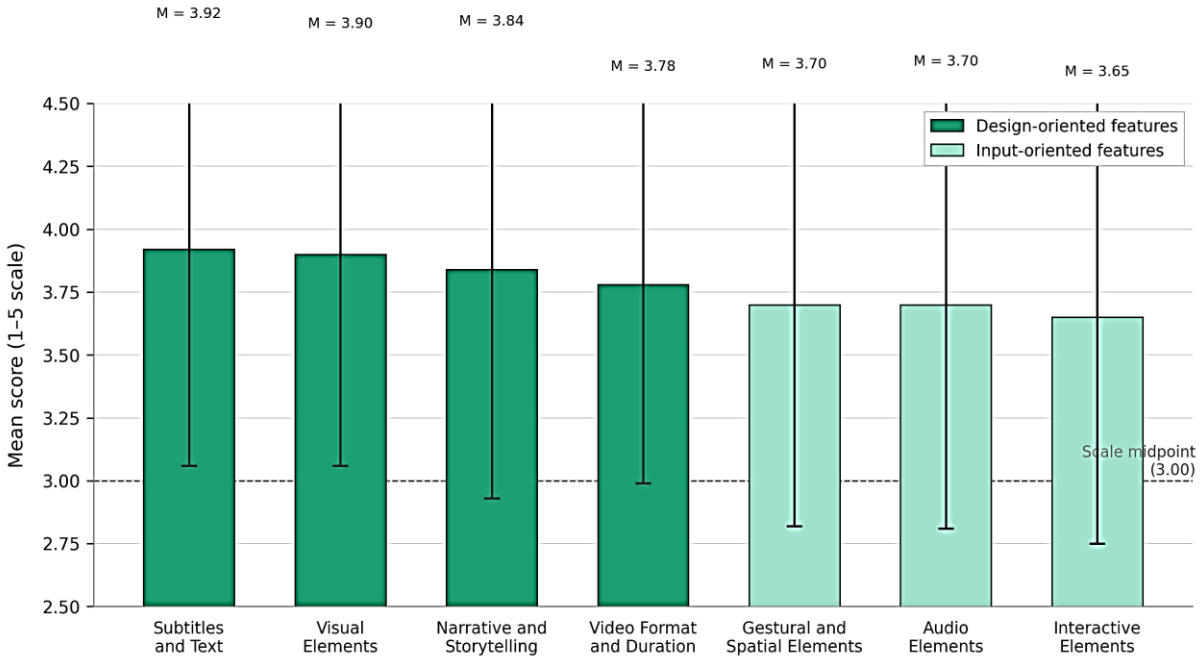

Figure 1: Mean Endorsement Scores and Standard Deviations for Seven Multimodal Feature Constructs (N = 92)

Note. Error bars represent ±1 standard deviation. The dashed line indicates the midpoint of the scale (3.00). Darker bars represent design-oriented features; lighter bars represent input-oriented features. Constructs are ordered from highest to lowest mean score.

4.2 Functional differentiation across constructs

The second research question examined what pattern of functional differentiation emerged across constructs when learner response patterns were considered collectively. The descriptive statistics presented in Table 6 reveal two observational patterns in the endorsement data that warrant attention at this stage of analysis. Nevertheless, these should be read as descriptive rather than confirmatory, given the analytical limitations of Phase 2 discussed in Section 4.3.

The first pattern concerns four constructs: Subtitles and Text, Visual Elements, Video Format and Duration, and Narrative Structure and Storytelling. These constructs sat at the upper end of the mean range (M = 3.78-3.92). They showed comparatively lower standard deviations, indicating that endorsement was relatively consistent across the sample. Importantly, these features share a common instructional function: they make vocabulary meaning visible, stable, and cognitively accessible within the constraints of short-form video through explicit on-screen text, sustained visual reference, and consistent structural formatting.

Collectively, these features constitute the architectural conditions of TikTok-inspired vocabulary instruction. They establish the visual and structural environment in which lexical meaning becomes observable before learners engage with the video's communicative content. Within this first pattern, item-level inspection of the Subtitles and Text construct showed that endorsement was

strongest for items relating to the inclusion of captions or subtitles, on-screen definitions presented during explanation, and the perceived role of subtitles in connecting spoken pronunciation with orthographic form. This suggests that learners valued subtitles as a tool for form-to-meaning mapping rather than as a generic visual feature.

The second pattern concerns Audio Elements, Gestural and Spatial Elements, and Interactive Elements, all of which sat at the lower end of the mean range (M = 3.65 to 3.70) with comparatively higher standard deviations. Although all three constructs received positive endorsement above the scale midpoint, responses showed greater individual variability than those in the first pattern. This suggests that learners' perceived value of these features is more individually modulated, likely reflecting differences in processing style, prior experience, and communicative preference. This variability is consistent with Lim (2021), who demonstrated that embodied resources, such as gesture and movement, serve as systematic meaning-making tools in instructional contexts. However, their reception depends substantially on individual learner dispositions and prior experience.

Interactive Elements recorded the lowest mean across the seven constructs (M = 3.65, SD = 0.90). A closer inspection of the item-level patterns suggests that learners were more receptive to closely scaffolded interactivity, such as quizzes and repetition prompts, than to more generative demands, such as commenting with their own examples or creating original sentences. This distinction is important because it implies that interactivity was valued when it remained structured and low in production demand but was less strongly embraced when it required fuller learner self-generation.

Taken together, these two observational patterns were used to inform a tentative functional distinction in the proposed framework: features that operate as structural design conditions are distinguished from features that serve as modes of communicative input. This tentative distinction is offered as an exploratory organising principle for framework development rather than as an empirically confirmed factor structure. As discussed in Section 4.3, the narrow mean range across all seven constructs (3.65 to 3.92, a spread of only 0.27 points) and the uniformly high inter-construct correlations indicate that Phase 2 alone does not establish fine-grained dimensional separation. Accordingly, confirmatory structural testing of this distinction has been conducted in a subsequent phase of the broader research programme, using a redesigned instrument and a sample of n = 304, appropriate for structural equation modelling, and is reported separately.

4.3 Correlational relationships among constructs

To examine the relational dimension of the second research question, Pearson correlation coefficients were calculated for all 21 pairwise combinations of the seven constructs. All constructs showed negative skewness, with values ranging from -0.730 to -1.325, indicating departures from perfect normality consistent with the ceiling clustering observed in the descriptive data. The elevated skewness observed in Narrative Structure and Storytelling (skewness = -1.325) and Visual Elements (skewness = -1.292) reflects the tendency of the sample to endorse these

constructs particularly strongly, resulting in a left-skewed response distribution. These values are consistent with the broader ceiling effect pattern observed across all seven constructs and do not indicate a methodological anomaly. Nonetheless, Pearson correlation was retained given the sample size of 92 and its established robustness to moderate violations of normality at this sample size (Pallant, 2020). Table 7 presents the full correlation matrix.

Table 7: Pearson Correlations Among Multimodal Feature Constructs (N = 92)

Construct	NS	VE	AE	GS	IE	ST	VF
Narrative and Storytelling (NS)							
Visual Elements (VE)	.868**						
Audio Elements (AE)	.747**	.788**					
Gestural and Spatial (GS)	.781**	.800**	.758**				
Interactive Elements (IE)	.817**	.833**	.713**	.797**			
Subtitles and Text (ST)	.793**	.841**	.672**	.755**	.804**		
Video Format and Duration (VF)	.742**	.779**	.652**	.760**	.805**	.853**	

*Note. ** Correlation is significant at the p < .01 level (2-tailed). NS = Narrative Structure and Storytelling; VE = Visual Elements; AE = Audio Elements; GS = Gestural and Spatial Elements; IE = Interactive Elements; ST = Subtitles and Text; VF = Video Format and Duration.*

All 21 pairwise correlations were positive, strong, and statistically significant ($p < .01$). In particular, across the full matrix, the mean inter-construct correlation was $r = .781$ (range .652 to .868), indicating uniformly strong association across all seven domains. This pattern indicates that learners who endorsed one multimodal feature dimension tended to endorse others at a similarly positive level. Hence, this reflects a broadly integrative orientation toward multimodal design in TikTok-inspired vocabulary instruction rather than a selective preference for individual design elements.

In addition, the magnitude of the correlations provides additional context for the endorsement patterns described in Section 4.2. The largest correlation among all pairwise comparisons was observed between Subtitles and Text and Video Format and Duration ($r = .853$), followed closely by Subtitles and Text with Visual Elements ($r = .841$) and Narrative Structure and Storytelling with Visual Elements ($r = .868$). The strength of these associations suggests that learners perceived these features as functionally integrated rather than as independent design decisions: a consistent video format, visible textual support, and high visual clarity are experienced as a coherent instructional unit rather than a collection of separate choices.

Correlations involving Audio Elements were generally smaller than those involving other constructs. The lowest pairwise coefficients across the entire matrix were observed for Audio Elements with Video Format and Duration ($r = .652$) and Audio Elements with Subtitles and Text ($r = .672$). This pattern reflects a functional distance between how auditory information is processed and how structural visual design is perceived. On-screen text and visual elements can be scanned, spatially located, and revisited within the viewing frame. Audio input

cannot be processed in the same way. As a result, learners appear to receive audio features as part of the natural communicative delivery of instruction rather than as a distinct and deliberate design choice, consistent with Punkhoom et al. (2025), who observed that EFL learners demonstrated greater awareness of visual and textual features in TikTok content than of auditory ones.

Two implications follow from the correlational results taken together with the descriptive statistics. First, the strong positive associations across all construct pairs indicate that learners experienced the seven multimodal features as components of an integrated instructional experience rather than as separate or competing dimensions. This is consistent with the multimodal theoretical premise that meaning in TikTok-inspired vocabulary instruction emerges through coordinated mode use rather than through isolated semiotic channels.

Second, and critically, the uniformly high inter-construct correlations taken alongside the narrow mean range of 0.27 points (3.65 to 3.92) indicate that Phase 2 alone does not provide strong evidence of fine-grained dimensional separation among the seven constructs. The seven domains were designed as broad, learner-facing construct categories appropriate for pre-intermediate ESL learners, not as the final latent structural specification of a measurement model. Therefore, structural distinctness of the proposed Design Features and Multimodal Input dimensions was tested in the subsequent phase of the broader research programme, using a redesigned instrument and a sample of n = 304, which was appropriate for confirmatory structural modelling.

Based on the Phase 2 descriptive and correlational findings as a whole, an initial exploratory model was proposed to represent a tentative functional distinction between features that primarily stabilise and structure vocabulary meaning and those that primarily contribute to the communicative delivery of input. Figure 2 summarises this learner-informed initial model. This model is offered as an exploratory organising structure and should not be interpreted as a confirmed two-factor measurement model.

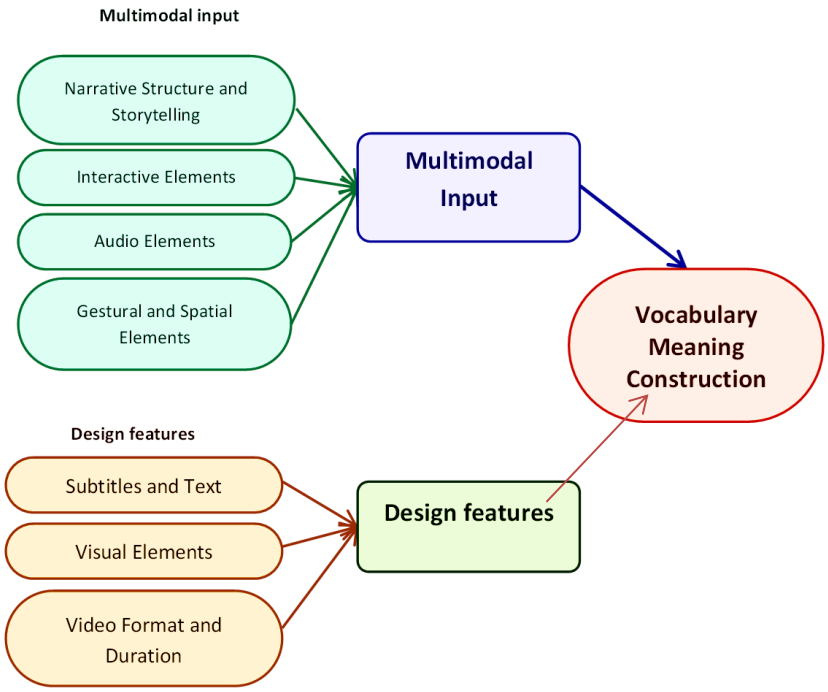

Figure 2. Initial Learner-informed Model Derived from Phase 2 Descriptive and Correlational Analysis

5. Discussion

Three main findings emerged: differential endorsement across the seven constructs, two descriptive patterns in that endorsement, and an exploratory framework organisation derived from those patterns. All seven constructs were positively endorsed above the scale midpoint. However, the patterns of means, standard deviations, and inter-construct correlations revealed a functional distinction that informed the proposed framework. These findings are discussed in relation to existing theory before implications for framework development and classroom practice are considered.

5.1 The design-oriented grouping: Stability, visibility, and cognitive anchoring

The stronger endorsement of Subtitles and Text, Visual Elements, and Video Format and Duration is both theoretically consistent and empirically meaningful. All three features serve the same instructional function: making vocabulary meaning visible, stable, and cognitively accessible within the time constraints of short-form video. Notably, Subtitles and Text in particular (M = 3.92, with among the smallest standard deviations) appear to have functioned as a non-negotiable instructional element rather than an optional supplement.

Theoretically, this finding aligns with Mayer's (2021) principles of redundancy and signalling, which predict that simultaneous verbal and visual representations reduce extraneous cognitive load and improve channel integration. It is also consistent with Paivio's (1986) dual coding theory: when learners encounter rapidly delivered spoken input, on-screen text provides a stable visual anchor for both phonological and orthographic processing.

5.2 The input-oriented grouping: Individual variation and implicit reception

Students endorsed the second grouping (Audio Elements, Gestural and Spatial Elements, Interactive Elements, and the contextualizing dimension of Narrative Structure/Storytelling) positively, but with substantially greater variation. This pattern is consistent with the idea that these features comprise the multimodal substance of instruction as individually experienced rather than as universally applicable conditions. In other words, students vary considerably in how they respond to vocabulary delivered through voice, gesture, narrative framing, or interactive prompting, likely because of differences in processing style, prior experience, and communicative preference.

Audio Elements demonstrated the lowest inter-construct correlations across the matrix ($r = .652$ and $.672$ with design-oriented constructs), reflecting a functional distance between auditory processing and structural visual design. On-screen text and visual elements can be scanned and revisited within the viewing frame; audio cannot. Consequently, learners appear to receive audio features as part of natural communicative delivery rather than as a deliberate design choice, consistent with Punkhoom et al. (2025), who observed that English as a Foreign Language (EFL) learners were more aware of visual and textual TikTok features than auditory ones.

The lower ranking of Interactive Elements compared to Subtitles, Text, and Visual Elements clearly contradicts the assumptions underlying much of social media pedagogy research: namely, that social media platforms inherently promote learner engagement. Given that learners are presumably expressing their values through TikTok-style social media platforms. For example, the participatory/interactive features would be most highly valued by learners in structured ESL classroom contexts. Meanwhile, learners in structured ESL classroom contexts appear to adopt a clarity-first orientation, valuing features that provide stability and access to vocabulary meaning, not those that encourage learner participation. Thus, this finding aligns with the study's structured instructional context. For example, ESL learners assessing multimodal design through a pedagogical lens, not an engagement/platform lens.

5.3 Framework implications and alignment with prior work

The two descriptive patterns observed in Phase 2 provide a learner-centred, exploratory empirical basis for an initial framework that distinguishes between Design Features and Multimodal Input. This distinction is grounded theoretically in Kress's (2010) multimodality framework, which differentiates between the material conditions of a text and the semiotic resources deployed within it (Jewitt, 2009). It is practically justified by the finding that learners perceived and responded to these two dimensions differently while endorsing both positively. Importantly, this framework is offered as exploratory and hypothesis-generating rather than as a confirmed structural model. The narrow mean range and uniformly high inter-construct correlations in Phase 2 preclude strong claims about dimensional separation, and structural validation is reported separately.

5.4 Limitations

The study has several limitations. First, the sample was drawn from a single Malaysian private university and comprised pre-intermediate learners. The two-pattern distinction may not extend to other proficiency levels, public university settings, or EFL contexts outside Malaysia. The findings should be treated as context-bound. Although N = 92 was adequate for descriptive and correlational analysis, purposive sampling precludes statistical generalisation to the broader ESL learner population.

Second, all data were collected via a single self-report questionnaire at a single time point, raising the possibility of common-method bias (Podsakoff et al., 2003). Procedural remedies were applied during instrument design: assurance of anonymity, neutral non-leading item wording verified through expert review, and separate questionnaire sections for each construct. Moreover, a Harman single-factor test conducted on the companion Phase 3 instrument (n = 304) indicated that the first unrotated factor accounted for 10.40% of total variance, well below the commonly cited 50% threshold. This provides indirect support that common method variance is unlikely to be the dominant source of association in the broader research programme.

Third, the study measured perceptions rather than vocabulary-acquisition outcomes. Features identified as supportive may not universally yield learning gains. Future research should incorporate longitudinal designs and experimental comparisons to address these limitations.

6. Conclusion

This study examined ESL learners' perceptions of seven multimodal feature constructs embedded in TikTok-inspired vocabulary instruction at a private university in Malaysia, using a validated quantitative survey administered to 92 Intensive English Programme learners. All seven constructs received positive endorsement above the scale midpoint. Nonetheless, the pattern of endorsement revealed a descriptively meaningful distinction rather than a uniform preference profile. Constructs that anchor and structure vocabulary meaning through on-screen text, visual support, and consistent video format were endorsed more strongly and more consistently across the cohort. Meanwhile, constructs constituting the communicative character of instructional input, including narrative framing, audio delivery, embodied cues, and interactive prompts, were endorsed positively but with greater individual variability.

Taken together, these two-pattern distinctions informed an exploratory, preliminary framework organisation distinguishing Design Features from Multimodal Input. The framework is offered here as hypothesis-generating rather than structurally confirmed. The narrow mean range across all seven constructs (3.65 to 3.92) and the uniformly high inter-construct correlations (mean r = .781) indicate that Phase 2 alone cannot establish fine-grained dimensional separation. Accordingly, structural validation of the proposed organisation was conducted in a subsequent phase of the broader research programme using a redesigned instrument and a sample of n = 304, appropriate for structural equation

modelling, and is reported separately. The present study's contribution is the demonstration that the multimodal features identified through prior qualitative analysis in Phase 1 carry meaningful learner endorsement and provide a generative, learner-informed basis for framework development.

Theoretically, the findings provide learner-informed empirical evidence for organising short-form video vocabulary instruction around a tentative distinction between structural design conditions and communicative input modes. It extends the application of multimodal theory and multimedia learning frameworks to the specific constraints of temporally compressed instructional content. In practice, the findings offer educators and instructional designers evidence-based guidance for evaluating and selecting TikTok-inspired vocabulary materials: features that provide cognitive clarity and structural stability should be treated as foundational design conditions, before supplementing with embodied, auditory, or interactive features. Given the exploratory nature of the proposed framework, these recommendations should be treated as principled starting points for instructional decision making rather than as prescriptive design rules.

Future research should examine whether the two-pattern distinction replicates across different ESL and EFL contexts, proficiency levels, and institutional settings. Experimental studies examining whether design-oriented features produce measurably stronger vocabulary retention outcomes than input-oriented features would provide important complementary evidence. Furthermore, longitudinal designs tracking whether perceptions change with extended exposure to TikTok-inspired instruction would add further stability to these findings.

Given that the present sample consisted of pre-intermediate learners from a single Malaysian private university, replication with learners at higher proficiency levels and across diverse national contexts is needed before broader generalisations can be made. Overall, these findings contribute to the broader goal of developing equitable and evidence-based digital language learning environments consistent with Sustainable Development Goal 4.

7. Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper. The research was conducted independently, and no financial, commercial, or institutional relationships influenced the design, data collection, analysis, interpretation, or reporting of the findings.

8. Acknowledgments

The authors wish to acknowledge the use of ChatGPT to assist with grammatical editing during the preparation of this manuscript. The tool was used to improve clarity, structure, and readability of the text. No data were generated or analysed by the AI system, and all research design decisions, data analysis, interpretations, and intellectual contributions remain solely those of the authors. The final manuscript accurately represents the authors' original scholarly work.

9. References

Abas, I. H., Krishnamurthi, N., Rasli, A., & Gusteti, M. U. (2025). A Delphi study on factors influencing school students' adoption of social media as a learning platform in Malaysia. *International Journal of Evaluation and Research in Education, 14*(3), 1743–1751. https://doi.org/10.11591/ijere.v14i3.32939

Adiyono, A., Al Matari, A. S., Patimah, L., & Nurdiyansyah, A. A. (2025). Can AI-optimized YouTube videos enhance Islamic religious education? A quantitative study on student learning outcomes. *Jurnal Pendidikan Agama Islam, 22*(1), 175–194. https://doi.org/10.14421/jpai.v22i1.11100

Alshreef, N. R., & Khadawardi, H. A. (2023). Using TikTok as a tool for English vocabulary learning in the EFL context. *English Language Teaching, 16*(10), 125. https://doi.org/10.5539/elt.v16n10p125

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE.

Cunningham, S., Moor, P., & Redston, C. (2013). Cutting edge intermediate (3rd ed.). Pearson Education.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Duan, H., Yunus, M. M., & Ismail, H. H. (2026). Mobile-assisted English vocabulary learning: A systematic review of platforms, learning outcomes, and implications (2019-2025). *International Journal of Learning, Teaching and Educational Research, 25*(2), 65–90. https://doi.org/10.26803/ijlter.25.2.4

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE.

Godwin-Jones, R (2000). Literacies and technology tools/trends. *Language Learning & Technology, 4*(2), 10–16. https://doi.org/10.64152/10125/25094

Hafner, C. A. (2014). Embedding digital literacies in English language teaching: Students' digital video projects as multimodal ensembles. *TESOL Quarterly, 48*(4), 655–685. https://doi.org/10.1002/tesq.138

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review, 31*(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203

Jewitt, C. (Ed.). (2009). The Routledge handbook of multimodal analysis. Routledge.

Jewitt, C., Bezemer, J., & O'Halloran, K. (2016). Introducing multimodality. Routledge.

Jiang, L., & Hafner, C. A. (2024). Digital multimodal composing in L2 classrooms: A research agenda. *Language Teaching, 57(*2), 153–170. https://doi.org/10.1017/S0261444824000107

Jumaat, N. A. F., Shanthi, A., Suppiah, P. C., Johar, E. M., & Arumugam, N. (2025). TikTok for education: Implications for language learning in a second language context. *Asian Journal of University Education, 21*(3), 734–749. https://doi.org/10.24191/ajue.v21i3.50

Khlaif, Z. N., & Salha, S. (2021). Using TikTok in education: A form of micro-learning or nano-learning? *Interdisciplinary Journal of Virtual Learning in Medical Sciences, 12*(3), 2–7. https://doi.org/10.30476/ijvlms.2021.90211.1087

Kress, G. (2010). Multimodality: A social semiotic approach to contemporary communication. Routledge.

Kucuk, S., & Richardson, J. C. (2019). A structural equation model of predictors of online learners' engagement and satisfaction. *Online Learning, 23*(2), 196–216. https://doi.org/10.24059/olj.v23i2.1455

Li, Y., & Hafner, C. A. (2022). Mobile-assisted vocabulary learning: Investigating receptive and productive vocabulary knowledge of Chinese EFL learners. *ReCALL, 34(*1), 66–80. https://doi.org/10.1017/S0958344021000161

Lim, F. V. (2021). Designing learning with embodied teaching: Perspectives from multimodality. Routledge.

Mayer, R. E. (2021). Multimedia learning (3rd ed.). Cambridge University Press.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Nation, I. S. P. (2013). Learning vocabulary in another language (2nd ed.). Cambridge University Press.

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66*(1), 60–92. https://doi.org/10.17763/haer.66.1.17370n67v22j160u

Paivio, A. (1986). Mental representations: A dual coding approach. Oxford University Press.

Pallant, J. (2020). SPSS survival manual: A step by step guide to data analysis using IBM SPSS (7th ed.). Open University Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Punkhoom, W., Sutthisan, S., Suksawas, W., Raksasat, N., & Jehma, H. (2025). TikTok-enhanced language learning: Evaluating short-form video content impacts on Thai EFL students' vocabulary growth. *rEFLections, 32*(3), 1768–1792.

Raman, K., Hashim, H., & Ismail, H. H. (2023). Enhancing English verbal communication skills through virtual reality: A study on engagement, motivation, and autonomy among English as a second language learners. *International Journal of Learning, Teaching and Educational Research, 22*(12), 237–261. https://doi.org/10.26803/ijlter.22.12.12

Rasyid, F., Hanjariyah, H., & Aini, N. (2023). TikTok as a source of English language content: Perceived impacts on students' competence: Views from Indonesia. *International Journal of Learning, Teaching and Educational Research, 22*(10), 340–358. https://doi.org/10.26803/ijlter.22.10.19

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. Palgrave Macmillan.

Shamshul, I. S. M., Ismail, H. H., & Nordin, N. M. (2024). Using digital technologies in teaching and learning of literature in ESL classrooms: A systematic literature review. *International Journal of Learning, Teaching and Educational Research, 23*(4), 180–194. https://doi.org/10.26803/ijlter.23.4.10

Shanthi, A., Jumaat, N. A. F., Suppiah, P. C., Johar, E. M., & Arumugam, N. (2024). From trending to teaching: A framework to analyse TikTok videos for vocabulary instruction. *International Journal of Social Science Research, 12*(2), 136–150. https://doi.org/10.5296/ijssr.v12i2.21904

Sweller, J., van Merrienboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tan, K. H., Rajendran, A., Muslim, N., Alias, J., & Yusof, N. A. (2022). The potential of TikTok's key features as a pedagogical strategy for ESL classrooms. *Sustainability, 14*(24), Article 16876. https://doi.org/10.3390/su142416876

Thanga Rajan, S., & Ismail, H. H. (2022). TikTok use as strategy to improve knowledge acquisition and build engagement to learn literature in ESL classrooms.

International Journal of Learning, Teaching and Educational Research, 21(11), 33–53. https://doi.org/10.26803/ijlter.21.11.3

Teng, M. F. (2023). Incidental vocabulary learning from captioned videos: Learners' prior vocabulary knowledge and working memory. *Journal of Computer Assisted Learning, 39*(2), 517–531. https://doi.org/10.1111/jcal.12756

Webb, S., & Chang, A. C. S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition, 37*(4), 651–675. https://doi.org/10.1017/S0272263114000606

Yang, H. (2020). *Secondary-school students' perspectives of utilizing TikTok for English learning in and beyond the EFL classroom*. In Proceedings of the 2020 3rd International Conference on Education Technology and Social Science (ETSS 2020) (pp. 162–183). Clausius Scientific Press.

Appendix 1

Learner Perception Questionnaire on TikTok Vocabulary Features

Dear Participants,

This questionnaire explores students' perceptions of multimodal features in TikTok vocabulary learning videos. This is not a test. There are no right or wrong answers. Your responses will remain confidential and will be used solely for research purposes.

Instructions:

This questionnaire consists of seven sections (A–G). Each section contains six items. Please indicate your level of agreement using the following scale:

1 = Strongly Disagree
2 = Disagree
3 = Not Sure
4 = Agree
5 = Strongly Agree

1. **Section A: Narrative Structure and Storytelling**

1. I enjoy watching vocabulary presented within a story or narrative in TikTok videos.
2. I like TikTok videos that introduce vocabulary through real-life scenarios.
3. I prefer TikTok videos that have a clear beginning, middle, and end structure for teaching vocabulary.
4. Vocabulary presented within a story helps me remember the words better.
5. Real-life scenarios help me understand the meaning of vocabulary better.
6. A clear narrative structure improves my vocabulary learning.

2. **Section B: Visual Elements**

1. I enjoy animated or cartoon visuals in TikTok vocabulary videos.
2. I like videos that show real objects or situations related to vocabulary words.
3. I enjoy on-screen text effects such as highlighted words or animations.
4. Animated visuals help me remember new words.
5. Real objects or situations are effective for my vocabulary learning.
6. On-screen text effects help me notice and remember vocabulary better.

3. **Section C: Audio Elements**

1. I find it easier to learn when there is no background music.
2. I like it when creators repeat vocabulary words clearly.
3. I enjoy sound effects that emphasize key vocabulary.
4. The absence of background music helps me focus.
5. Repetition improves my pronunciation learning.
6. Sound effects help me remember key vocabulary.

4. **Section D: Gestural and Spatial Elements**

1. I enjoy it when creators use hand gestures or body language.
2. I like it when creators act out vocabulary meanings.
3. Gestures and movements make meanings clearer.
4. Body language helps me understand vocabulary better.
5. Acting out meanings improves my vocabulary learning.
6. Facial expressions and movements help me remember new vocabulary.

5. Section E: Interactive Elements

1. I like videos that include quizzes or challenges.
2. I enjoy when videos ask viewers to repeat words or respond.
3. I prefer videos that encourage comments using new words.
4. Interactive quizzes improve my word retention.
5. Repetition or response prompts are effective for learning.
6. Creating my own sentences helps me remember vocabulary.

6. Section F: Subtitles and Text

1. I like it when vocabulary videos include subtitles.
2. I prefer when definitions appear on screen during explanation.
3. Example sentences on screen help me understand vocabulary.
4. Subtitles help me connect pronunciation with spelling.
5. On-screen definitions improve my understanding.
6. On-screen example sentences help me understand vocabulary use in context.

7. Section G: Video Format and Duration

8. I prefer shorter videos under one minute that focus on fewer words.
9. I like the vocabulary series more than standalone videos.
10. I prefer videos that follow a consistent format.
11. Shorter videos are effective for my learning.
12. The vocabulary series helps me learn better than standalone videos.
13. Videos with clear explanations and visuals are engaging.

Thank you for completing this questionnaire.