Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 9, 103754 (2026)
https://doi.org/10.29140/lea.2026.103754

From Gaze-Based Evidence of Strategic Processing in Multimodal L2 Speaking Task: Towards Mixed-Method Approach[1]

JUNGHEE BYUN

Seoul National University, Republic of Korea
donne99@snu.ac.kr

Abstract

Despite the growing attention to the use of multimodal prompts in speaking assessments, evidence on the effect of input modes on real-time strategic behavior beyond fixation counts remains underexplored. This eye-tracking study investigated gaze behaviors of 43 adult speakers as they engaged with text, image, and video prompts during the preparation and response phases, focusing on gaze patterns across modes and phases, individual differences, and the correspondence between gaze-based and self-reported strategies. Ten gaze indicators across 10 areas of interest were analyzed using repeated-measures, mixed ANOVAs, principal component analysis, and correlation analysis. Results showed mode- and phase-specific features. Exploratory analysis also indicated individual differences (gender and proficiency) in gaze behavior across modes. Gaze-based strategic components suggest mode dependence of strategic gaze allocation and no significant associations with self-reported strategy use, after controlling for gender, proficiency, and task performance. This finding might suggest a complementary view on the construct of strategic competence. Implications for the cognitive design and validation of a multimodal speaking assessment for L2 learners were discussed.

Keywords: Multimodal L2 speaking assessment, eye-tracking, gaze behavior, strategic competence, cognitive load

[1]The publication of this article was supported by the 2nd Seogang Next-Generation Scholars Support Program in 2024, administered by the Department of English Language and Literature at Seoul National University, South Korea.

Data Availability Statement: *All relevant data are within this paper.*

Introduction

Multimodal language assessment has become prevalent, particularly in performance-based tasks, that incorporate visual stimuli such as text, images, and videos into task prompts. These tasks require test takers to integrate information from multiple sources. This development is especially important in speaking assessment because oral performance depends not only on linguistic knowledge but also on the ability to effectively process information across modes while planning and delivering an oral response. Eye-tracking technology has gained recognition in language assessment research for examining the impact of multimodal inputs on learner engagement and performance (Burton, 2022; Kang et al., 2020; Park, 2018; Yang & Plakans, 2012; Zhang et al., 2014, 2022). This line of research is informed by theoretical perspectives on multimodality (Kress, 2010; Low & Sweller, 2014; Mayer, 2020), which posits that individuals' information processing is influenced by the input format.

As multimodal speaking tasks require test takers to integrate information from text, images, and video while planning and producing an oral response, these tasks impose substantial demands on attention and working memory. These demands can be explained through Cognitive Load Theory (CLT; Plass et al., 2010), which proposes that human working memory has limited capacity. And task performance is shaped by different types of load, including intrinsic load generated by the inherent difficulty of the materials and extraneous load generated by the way instruction and materials are designed and presented. Such cognitive demands of tasks are likely to significantly influence test performance and therefore warrant research on how learners allocate cognitive resources across various input modes (Sweller, 1988, 2005, 2010). From the perspective of the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2005, 2014), test takers must select relevant verbal and pictorial information, organize it into coherent mental representations, and integrate them during task completion.

This CTML framework is relevant to validating multimodal speaking tasks. It is because test takers need to plan and produce an oral response based on their interpretations of visual inputs across different modes. Tracking gaze behaviors throughout this process offers valuable insights into how test takers allocate attention across different task elements, extract relevant information, and manage cognitive load as they prepare for the task. If a particular mode consistently increases processing load or offers an advantage to a particular group, test scores may partially reflect mode-specific features rather than the intended construct. Thus, it is highly important to investigate whether gaze patterns and performance differ across modes, as this may have significant implications for scoring decisions and consequences for test takers, contributing to construct and cognitive validity.

Nevertheless, there is little research on gaze-based indicators of strategic processing across these input modes. Strategic processing refers to the selective allocation, coordination, and monitoring of attention across multimodal input during task performance. As eye-tracking has become well-known as an effective method for investigating learners' attention and cognitive load management in real time, much of the current research remains on receptive skills, such as reading and listening (Bax & Weir, 2012; Bax & Chan, 2019; Kwon & Yu, 2024; Ockey, 2007; Wagner, 2007; Suvorov, 2015). This highlights the need to investigate gaze behaviors in productive tasks, particularly multimodal speaking tasks, where test takers must coordinate multiple input sources while preparing and producing an oral response. Evidence that tasks engage the required cognitive processes is necessary to support construct validity (Kane, 2013). For this validation, eye-tracking can inform how nonverbal input and task type affect the cognitive load required of individual learners. In this context, the current study looks at gaze patterns to measure the cognitive demands of text, image, and video inputs during a speaking task. It further investigates gaze-based indicators of strategic processing as behavioral evidence for interpreting strategic competence.

Despite the growing recognition of eye-tracking in language assessment, several notable gaps persist. As previously highlighted, research has predominantly centered on receptive abilities rather than productive ones, especially in speaking tasks. A meta-analysis revealed that only a few of the 111 eye-tracking studies focused on speaking (Hu & Aryadoust, 2024). Second, in the few existing speaking assessment studies (Burton, 2022; Lee & Winke, 2018; Ma & Winke, 2022), their focus was restricted to fixation-based metrics, thereby neglecting other potentially significant indicators —such as pupil size, blinks, and saccades—that can offer insights into cognitive load and emotional regulation. Third, instruments for evaluating test strategy use require process-oriented evidence in addition to perception-based reports. Self-report tools have predominantly reflected learners' perceptions, providing insight into strategic competence (Zhang et al., 2021). Since strategic behaviors include both conscious and unconscious elements (Bachman & Palmer, 2010; Cohen, 2014), using objective and physiological measures, such as eye-tracking, along with self-report tools can offer a comprehensive picture of how learners deploy strategic behaviors during tasks. Therefore, the current study seeks to identify gaze patterns across modes and to derive gaze-based strategic components from indicators of strategic processing. Lastly, further investigation is needed into the variability of gaze behaviors across modes. More specifically, research should examine the influence of potential factors, such as learner variables like proficiency and gender, on these mode-specific processes (Bax & Weir, 2012; Chan, 2013; Godfroid et al., 2013; McCray & Brunfaut, 2018; Suvorov, 2015). Previous eye-tracking studies (Cazzato et al., 2010; El Haj et al., 2023; Miyahira et al., 2000; Sidhawara et al., 2023) suggest that gender-related differences in gaze behavior may be task-dependent rather than uniform across contexts, which necessitates their inclusion as an exploratory factor in the current study.

To fill these gaps, the present study uses eye-tracking in a multimodal speaking task to look at gaze patterns across modes, find mode-specific gaze patterns, and derive gaze-based strategic components, examining alignment with self-reported strategies. This study aims to enhance our understanding of strategic decision-making in complex, visually mediated speaking tasks by integrating gaze data into the analysis of multimodal test performance. This combined approach to cognitive processing will enhance theoretical frameworks of multimodal assessment and foster the development of cognitively valid, equitable test design. Consequently, this study addresses the following research questions:

RQ1: What are the characteristics of gaze behavior patterns present in each visual input (text, image, and video) within the preparation and response phases of a multimodal speaking task?

RQ2: How do these gaze patterns differ across the three input modes, and do these mode effects vary by learners' individual characteristics (gender and proficiency) and speaking task performance?

RQ3: How do gaze-based strategic components, derived from eye-tracking indicators, differ based on learners' individual characteristics (gender and proficiency) and speaking task performance, and to what extent do they correspond with learners' perceived strategy use?

Literature Review

Visual Input and Cognitive Load During Multimodal Tasks

Recent research based on CLT and CTML demonstrates that multimedia presentations systematically influence cognitive processing and test performance. According to the modality principle of CTML (Castro-Alonso & Sweller, 2021), cognitive efficiency can be maximized when information is distributed across both visual and audio channels. When only visual inputs are provided, as in the current

study, the split-attention effect arises, requiring cognitive processing through a single visual channel. This scenario increases learners' cognitive load within the limited capacity of working memory, as they must integrate information from multiple modalities, potentially imposing extraneous load. Processing different types of information, such as transient information from video and static information like text and images, requires non-native learners to allocate more strategic attention. In contrast, native speakers face less burden in managing cognitive resources and experience fewer linguistic challenges (Abutalebi & Clashen, 2017; Kruger et al., 2014; Mayer et al., 2014; Morishima, 2013). Thus, it can be assumed that different combinations of text, images, and video change the type and amount of cognitive load a task imposes on L2 learners.

For example, Dirkx et al. (2021) found that computer-based math tests incorporating related text and graphic elements led to higher scores, increased attention, and lower cognitive load indicators, as measured by gaze metrics such as longer fixation durations, increased pupil diameters, and higher blink frequency and duration. Suvorov's (2015) video-based English listening test, which included context and content videos, revealed modality-dependent processing demands, eliciting distinct fixation patterns and dwell times. As performance-based tasks increasingly incorporate visual inputs, similar design considerations arise for productive tasks. This underscores the necessity of examining how multimodal inputs affect processing during speaking. Kang et al. (2020) reported that online Korean-speaking assessment using different mode prompts (i.e., picture and text) differed in task difficulty, highlighting the importance of rater awareness regarding the interaction between proficiency and mode. Lee (2020) also found that video-based approaches significantly benefited beginner and intermediate learners but produced a ceiling effect for advanced learners in English-speaking achievement tests. Taken together, these findings suggest that input mode may influence test takers' processing and performance, emphasizing the need for further empirical research on how input type and visual layout affect the development of cognitively valid assessments.

Gaze Behavior as an Indicator of Cognitive and Strategic Processing

Eye-tracking has emerged as an effective method for revealing how learners distribute their attention and manage task demands in real time. Researchers commonly use measures such as fixation duration, saccades, time to first view, revisits, pupil size, and blinks to investigate processes related to attentional selection, depth of processing, and cognitive load (Jarodzka et al., 2015; Sáiz-Manzanares et al., 2021). These measures are useful because they provide behavioral traces of how attention and cognitive resources are allocated during task performance. Meta-analytic work by Hu and Aryadoust (2024), for instance, showed that fixation number and duration may indicate how much visual attention is devoted to a given AOI (Area of Interest) and how long test takers process information there. They can be interpreted as indirect evidence of attentional allocation and depth of processing. Time to first view may reflect how quickly test takers become focused on information that is perceived as important or task-relevant and can thus be linked to selective attention management. Revisit frequency may suggest that test takers return to previously viewed information for checking, verification, and comparison, thereby providing evidence of monitoring. Likewise, the magnitude (or distance) and direction of saccades, defined as rapid eye movements between fixations, can provide evidence of how broadly or narrowly test takers scan visual input and shift attention across AOIs. These metrics may be relevant to visual search and monitoring behaviors. Finally, pupil size and the number or duration of blinks have been associated with changes in cognitive load, fatigue, and emotional state, which may therefore serve as indirect indicators of cognitive load management and affective regulation (Josephson & Myers, 2019; Sáiz-Manzanares et al., 2021; Winberg, 2021). In the present study, these metrics are treated as indirect behavioral evidence for interpreting strategic processing across multimodal input. More detailed operational definitions of the eye-tracking measures for the current study are provided in the Methodology section.

Significant individual variations have often been reported in eye-tracking research on how learners process information depending on factors such as proficiency and gender. For example, advanced learners exhibited more efficient gaze trajectories and shorter, more accurately directed fixations (Godfroid et al., 2013). Likewise, successful readers allocated a greater proportion of their attention to relevant AOIs, invested more time in relevant text segments, and employed effective cognitive strategies, such as eliminating distractors (Bax, 2013; Bax & Chan, 2019). Compared with higher-scoring students, lower-scoring students showed greater cognitive effort during local reading and lower-level cognitive processing, as indicated by eye-tracking analysis of banked gap-fill items (McCray & Brunfaut, 2018).

Some studies have also investigated gender-related variation in gaze behavior. In multimedia learning contexts, males tended to rely more on pictorial information, while females engaged more thoroughly with both text and images (Sidhawara et al., 2023). In visuospatial planning tasks, males showed greater path optimization and more flexible shifts during task performance, whereas females tended to maintain a more consistent strategy (Cazzato et al., 2010). Similarly, Miyahira et al. (2000) reported different exploratory styles in visual information processing, with females showing longer and fewer fixations, whereas males showed broader overall visual scanning. In autobiographical retrieval, El Haj et al.(2023) found that females' shorter fixations and larger saccades were associated with richer memory retrieval. Taken together, these findings suggest that gender-related differences in gaze behavior are difficult to generalize and appear to be task-dependent. This therefore justifies examining gender as an exploratory factor in differences in gaze allocation and processing in multimodal speaking tasks.

Strategic Competence and Visual Stimuli in Language Assessment

Strategic competence is defined in Bachman and Palmer's (2010) model of communicative language ability as the ability to manage cognitive resources and modify processing strategies in response to task demands and available input. For example, attention and engagement of language learners in listening tasks were influenced by visual stimuli such as images and videos (Ockey, 2007; Suvorov, 2015; Wagner, 2007), indicating that these non-verbal components can affect how we process verbal information. Likewise, in a multimodal speaking assessment, this competence may be reflected in how test takers allocate their attention across task-relevant elements and use these resources to construct their responses. It is through connecting gaze behaviors and speaking performance that we can infer how strategic competence manifests in multimodal tasks and how it may influence score interpretations.

Traditionally, this construct has been investigated by self-reports (Field, 2012; Ockey, 2007; Oxford, 1986; Park & Kim, 2022; Purpura, 1999; Wagner, 2010; Zhang et al., 2021) due to their ability to efficiently sample large participant groups across various tasks. Although informative and efficient, these self-reports capture participants' perceptions rather than their real-time behaviors. In response, recent research has increasingly used eye-tracking to directly explore strategic behavior. For instance, Palmour (2023) found that gaze aversion increased with greater task complexity, while Bax and Chan (2019) reported that the L2 high proficiency group consistently maintained strategic patterns of attention. Nonetheless, more research is needed on the relationship between gaze data and strategic competence, as well as on the influence of multimodal input (e.g., images, text, video) on gaze-based indicators of strategic processing, because different modes can affect different types and levels of cognitive load on learners (Dirkx et al., 2021; Meyer, 2005).

This study discusses Bachman and Palmer's (2010) concept of *strategic competence* as an overarching construct, or higher-order executive process, that governs how language knowledge is deployed in communication or assessment. This construct involves setting goals, assessment, and planning performance to meet varying task demands and available inputs. Strategic competence may be enacted through cognitive processing during task performance, that is, its observable manifestations, such as

focusing attention and combining multiple cues during specific multimodal tasks. From this perspective, *self-reported strategies* and *gaze-based indicators of strategic processing* provide complementary but different kinds of evidence. The term *self-reported strategies* refers to test-takers' conscious, retrospective, and metacognitive awareness of how they manage attention and cognitive resources, as reported by the test-takers after the test. The post-hoc reports may not fully or accurately capture what was happening during the task. On the other hand, *gaze-based indicators of strategic processing* tap into behavioral evidence of cognitive processes through test-takers' real-time eye-movement traces during the test, including fixations, revisits, pupil size, and blinks, which potentially reflect less-conscious aspects of attentional allocation, re-inspection behavior, and cognitive load, respectively. These two approaches capture different aspects of the strategic process. In the analysis, *gaze-based strategic components* refer to exploratory factors derived from observed gaze-based indicators, but they are treated as a conceptual link between eye-movement measures and strategic processing, rather than as direct measures of strategic competence.

Methodology

Research Context and Participants

The speaking skill assessed is explaining the order of events on an environmental topic, focusing on communicating procedural information. Participants must sequentially describe how a recycled invention—made from used cell phones and designed for rainforest rangers—is operated. The study included 43 participants attending a Midwest-based university in the US.

The participants consisted of two main groups: 11 native English speakers (5 females and 6 males) and 32 non-native English speakers, as outlined in Table 1. And they reported no vision or cognitive impairments. The native speaker group was expected to serve as a relative baseline for cognitive processing with greater language familiarity. Their data may provide insight into the strategic behaviors involved in the cognitive processing of visual input during a multimodal task, in which native speakers are less likely to experience language constraints than non-native speakers. Non-native groups may engage in a range of strategies in response to both linguistic and non-linguistic demands of the task. For this reason, the current study uses this comparison to explore how relative language familiarity may shape gaze allocation within the different demands of multimodal task performance.

The non-native group was further divided into 13 advanced learners (6 females, 7 males) and 19 intermediate learners (10 females, 9 males). Participants' ages were collected in groups: 24 participants in the 18–25 age group, 12 participants in the 25–34 age group, six participants in the 35–44 age group, and lastly, 1 participant in the above-45 age group.

Self-reports were used to gather information on their English proficiency levels, with the guideline that TOEFL scores of 95 or above were classified as advanced, and scores between 94 and 80 as

Table 1. *Participants by gender and proficiency level*

Proficiency level	Female	Male	Total
Native speaker	5	6	11
Advanced L2 learner	6	7	13
Intermediate L2 learner	10	9	19
Total	21	22	43

intermediate. The distribution of first language backgrounds was as follows: Hindi (12 participants), Korean (six), Arabic (six), Chinese (four), Japanese (two), Spanish (one), and Greek (one). Majors include humanities such as psychology, English, and social sciences, as well as economics, veterinary studies, human-computer interaction, and engineering. All participants signed an informed consent form before participation. This study was determined to be exempt from review by the Institutional Review Board of the host institution (IRB #24-271).

Research Design

The speaking task consisted of two phases: a preparation phase and a response phase. In the preparation phase, participants first read on-screen information about an invented device that was designed for rainforest protection, accompanied by an image, for 30 seconds. They then watched the same nonverbal video clip—without subtitles—three times, each viewing lasting 25 seconds, to understand the steps that illustrate how the device works (see Figures 1 and 2). In the response phase, shown in Figure 3, participants read the speaking instructions for 20 seconds and then produced their responses for 50 seconds. Table 2 summarizes what test takers were required to do in each phase, how much time was allocated, and which AOIs were defined.

Ten AOIs—four images, three text passages, and three videos—were determined based on their relevance during the preparation and response phases, which were rated by the researcher who developed this task (see Table 2). A particular focus on repeated video-based AOIs was devised to examine how familiarity with dynamic content influenced gaze behaviors and visual attention over time.

AOIs were defined according to three levels of task relevance. First, I chose the most relevant information which directly matched what test takers had to explain in their responses, including the video segments illustrating the device's operation and the on-screen prompt specifying instructions and target audience in the response phase. Second, the next relevant information was chosen to provide test takers with the contextual support for the task topic, such as the device's purpose and its elements presented during the preparation phase. Third, the third relevant AOIs were chosen around the timers displayed

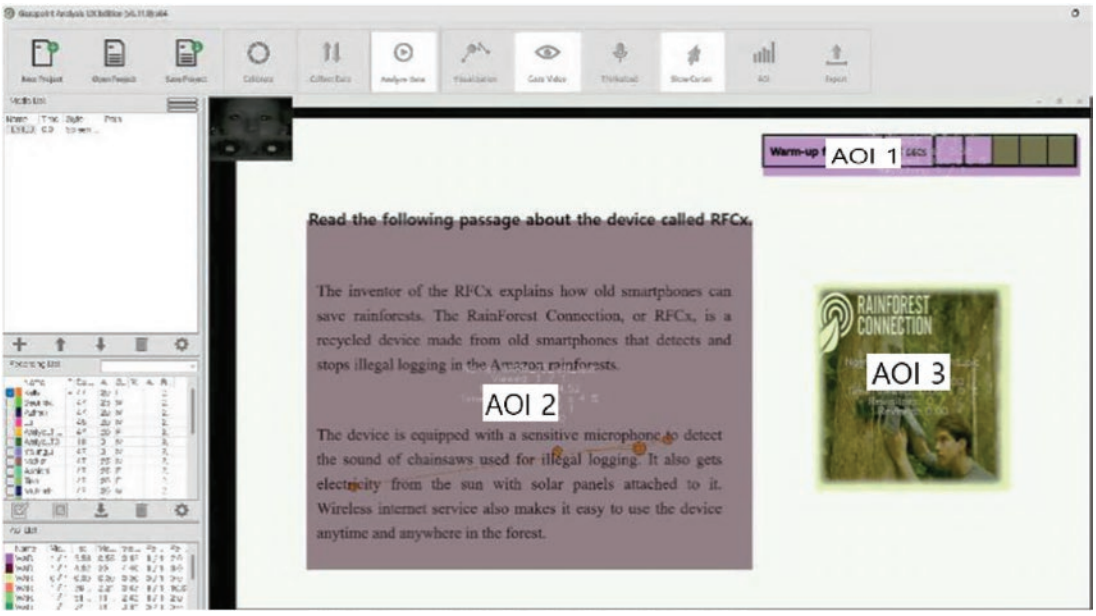

Figure 1. *Screenshot of three AOIs in the preparation phase 1.*

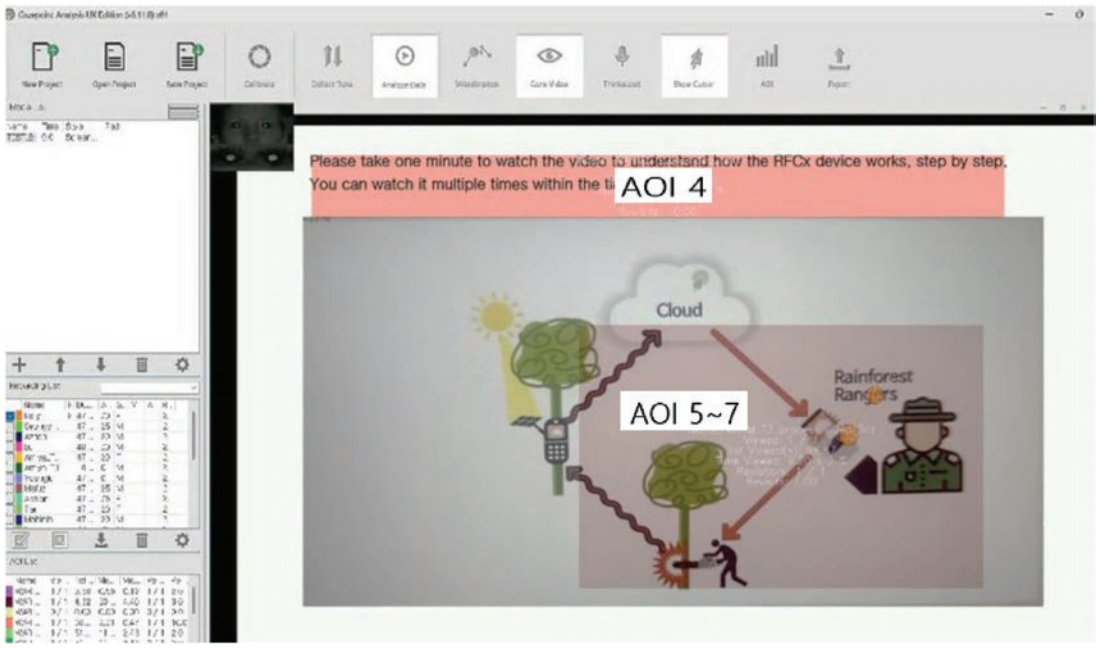

Figure 2. *Screenshot of four AOIs in the preparation phase 2.*

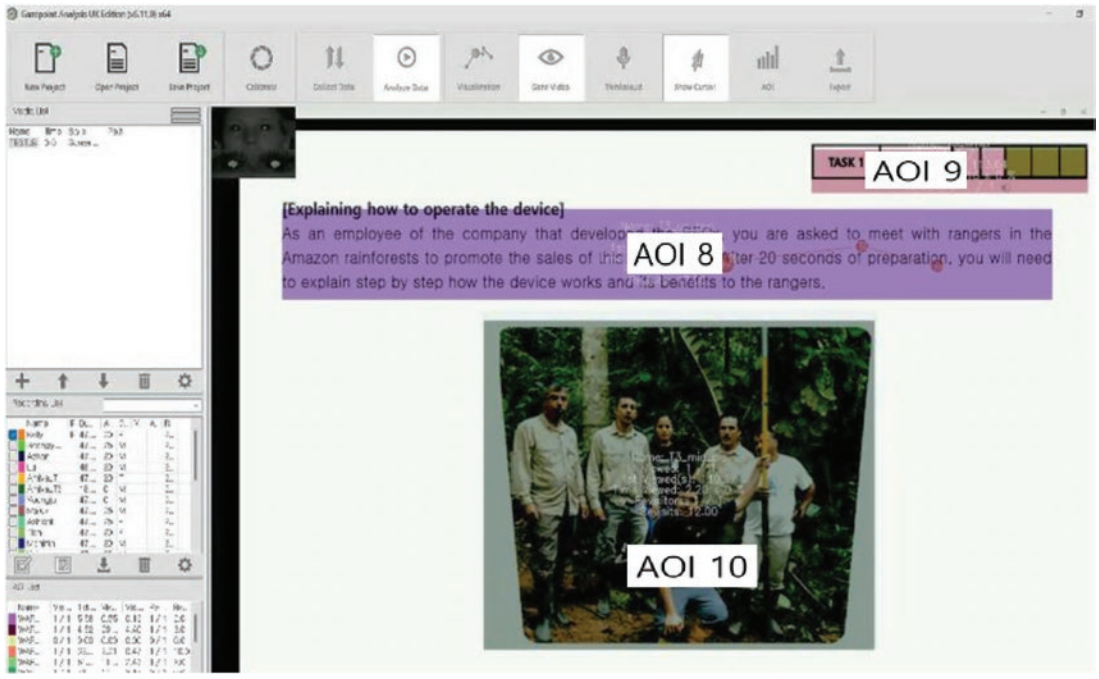

Figure 3. *Screenshot of three AOIs in the response phase.*

in both phases. The timers appeared in the peripheral area, but remained relevant for managing time during task performance.

Test takers were not allowed to take notes during the preparation phase to minimize interruptions to eye-tracking recording and ensure that gaze behavior could be captured as accurately and continuously as possible. This control was intended to reduce the external influence of notetaking on test takers' processing of the multimodal input.

Table 2. *Overview of the multimodal speaking task*

Phase	Segment	What test takers do	Duration (secs)	AOI no. & label	Figure & color
Preparation phase 1	Image Text	Read on-screen information about the invented device and view the accompanying image.	30	▪ AOI 1: Timer ▪ AOI 2: Device picture ▪ AOI 3: Device info	Figure 1, purple Figure 1, yellow Figure 1, brown
Preparation phase 2	Text	▪ Read the instructions specifying what the video clip is about and how many times it will be watched.	75	▪ AOI 4: Video instructions	Figure 2, orange
	Video	▪ Watch a nonverbal video clip without subtitles that shows the device's operation sequence three times		▪ AOI 5~7: 1st, 2nd, and 3rd videos	Figure 2, brown
Response phase	Text	▪ Read the instructions specifying the target audience (forest rangers) and the task (explaining how the device works).	50	▪ AOI 8: Speaking instructions	Figure 3, violet
	Image	▪ Produce a spoken explanation of how the device operates, using information from the preparation phase.		▪ AOI 9: Timer ▪ AOI 10: Target audience picture	Figure 3, purple Figure 3, green

Instrument and Data Collection

The lightweight and portable GP3 Eye Tracker (GazePoint, n.d.) was used to collect eye-tracking data. It recorded video of visual attention in real time with 0.5–1° accuracy at 60 Hz and allowed for 5 or 9-point calibration. The GazePoint Analysis UX edition program (version 6.11.0; Gazepoint, n.d.) was used to stimulate tasks and extract data. Although the instrument has a lower sampling frequency than high-speed eye trackers, a rate of 60 Hz provides acceptable resolution for stably measuring gaze indicators such as fixation, revisits, and pupil size (Holmqvist et al., 2011). The major objective of the current study is to investigate attention allocation and cognitive load in a multimodal input context rather than analyzing microsaccades. Thus, this 60 Hz sampling frequency is considered to ensure reliable research results.

Participant Training and Calibration Procedure

Before collecting data, there was a training session where the participants underwent calibration by maintaining the correct posture while following a moving dot across nine fixed places on the screen, as shown in Figure 4. Calibration was done carefully, with specific changes made for participants wearing glasses, such as raising their seat position.

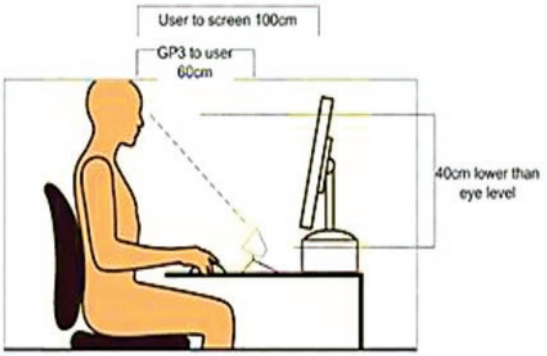

Figure 4. *Seat arrangement for calibration purposes (Gazepoint Analysis User Manual, 2023).*

Task Response

The task was adopted from one that the researcher made for her doctoral dissertation (Byun, 2023). The speaking task required a formal style of language appropriate for IT staff. Test takers had to explain how a device that was invented to protect rainforests from illegal logging in the Amazon works.

Post-Task Strategy Inventory and Interview

A strategy inventory measured cognitive strategies in technology-based English-speaking assessment, based on prior research (Huang, 2016; Zhang & Goh, 2009; Zhang et al., 2021; see Appendix 1). Immediately after completing the task and the self-reported survey containing 39 items on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree), test takers took part in a brief post-task interview about the task experience. The interview was conducted to collect additional information about their perceived task difficulty on a 5-point Likert scale (1 = not difficult at all, 5 = extremely difficult), sources of their perceived difficulty, their use of multimodal input, and their reflections on distinctive gaze behaviors observed during task performance. The interview questions are listed in Appendix 2. And z-standardized total survey scores were computed for comparison with gaze-based strategic component scores obtained from eye-tracking metrics. Internal consistency for the inventory was high, with a Cronbach's α of .876. Although strategy use is multidimensional, the current small sample size made it difficult to model subcategories stably, leading to the use of a standardized overall score as an indicator of the reported strategy use. Collapsing multiple components into a single component will be discussed later as a limitation of the current study.

Task Scoring

Task performance was scored analytically on a scale of 1 to 4 based on fluency, which focused on the flow of participants' speech, pronunciation, and intonation. The scoring also considered the accuracy of lexical and grammatical use, as well as content, which assesses the degree to which responses are appropriate to task requirements and well-organized (see Appendix 3). Two raters with more than two years of experience as teaching assistants at a US-based university scored the responses after training with 12 sample responses (three per score level from 1 to 4), and then individually rated the remaining samples. Interrater reliability of speaking scores by two raters, estimated using a two-way random-effects absolute-agreement intraclass correlation, reflected excellent reliability. Compared to the result of the single-measure ICC(2,1), which was .56 (95% CI = .44−.68, $p < .00$) indicating moderate agreement, the average ICC(2,2) measure came to .91 (95% CI = .86−.95, $p < .00$). Accordingly, the mean of the two raters' scores was used for subsequent analyses.

Eye-Tracking Measures

Gaze indicators included ten metrics for each AOI. Time to first view (average duration until first fixation in milliseconds) captured the time to the first fixation on an AOI, and fixation number and duration (total count and average gaze pause length in milliseconds) refer to how often and how long participants fixated within an AOI. Revisit number indicated the frequency of returning to an AOI after looking elsewhere. Left and right pupil size (diameter in millimeters) are indicators of pupil dilation. Saccade magnitude and direction captured the average distance of eye movements between successive fixations in pixels and the average degrees of visual angles. Lastly, blink number and duration indicate the average frequency and blink duration in milliseconds. More detailed information about major gaze indicators is summarized in Table 3.

Table 3. *Definitions, methods, and interpretations of gaze indicators*

Indicator	Definition and data collection method	Data interpretation
Fixations	• Points where the eyes remain relatively stationary, indicating attention and information processing • Calculate the number, duration, and position of fixations	Indicate the ability to recognize key content or filter out irrelevant information. Mean fixation durations and fixations indicate cognitive load (Jarodzka et al., 2015; Yin & Chen, 2007).
Saccade	• Rapid movements between fixations • Measure amplitude (distance) and speed	Provide insights into reading strategies or search behaviors (Frenck-Mestre, 2005; Gnetov & Kuperman, 2024).
Pupil dilation	• Changes in pupil size • Measure changes in pupil size over time	Imply strategies dealing with increased cognitive load, emotional arousal/ enhanced emotional states, and coping with anxiety factors (Borghini & Hazan, 2018; Schmidtke, 2014).
Blink rate	• Frequency of blinking • Calculate the number and rate of blinks during the assessment period	Changes in blink rate imply strategies dealing with fatigue, cognitive load, emotional state, and coping with anxiety factors (Burton, 2022; Lindberg, McDonough, & Trofimovich, 2021).

Ten gaze indicators used in this study were selected to capture the cognitive processes associated with the ability to plan, monitor, and revise performance in response to task demands and available input, as proposed by Bachman and Palmer (2010). Fixation-based indicators, including time to first view, fixation count/duration, and revisit, reflect the process of planning where and how attention should be allocated and of monitoring whether this attention is on the task-relevant cues. Saccade magnitude and direction capture how test takers scan and monitor their gaze as they explore or confirm necessary information in different spots. Lastly, pupil size and blinking measures are associated with cognitive load, fatigue, and affective arousal, which require test takers to manage cognitive resources and regulate affective responses to multimodal task demands during task performance. This study uses ten indicators as a basis, aiming to derive a more general, smaller set of gaze-based strategic components, which are then compared across input modes. All ten metrics were recorded during both the preparation and the response phases to compare gaze allocation between input processing and speech production. Together with Table 2, these measures provide an overview of how eye-movement data are linked to specific phases and AOIs.

Data Analysis

For RQ1, descriptive statistics of metrics data on AOIs were computed across the three modes. Before conducting ANOVAs, homogeneity of variance was examined using Levene's test for 10 gaze metrics in each input mode, producing 30 tests in total. All metric tests, except the blink duration in image mode, were nonsignificant (*ps* ≥ .05), suggesting no serious violation of the homogeneity assumption. Then repeated-measures ANOVAs were conducted with mode as a within-subjects factor to examine whether gaze behaviors differed across modes.

For RQ2, mixed ANOVAs were run with mode specified as a repeated-measures factor and gender (male and female) and performance group (native, high, and intermediate English learners) as between-subjects factors. The composite score (range 3.7–12) divided participants into three performance groups. Native speakers (n = 11) obtained a mean score of 10.33 (SD = 1.69), whereas the L2 high group (n = 13) and the intermediate group (n = 19) obtained mean scores of 7.69 (SD = 2.38) and 7.13 (SD = 1.88), respectively. Because of relatively small and unequal subgroup sizes, the results should be interpreted as exploratory.

For RQ3, principal component analysis (PCA), a data-reduction technique, was used to summarize patterns across multiple covariate gaze metrics within each input mode and to extract the underlying components of gaze behaviors (Holmqvist et al., 2011). This exploratory analysis allows us to derive latent, gaze-based strategic components from observed gaze behaviors, which can be interpreted as reflecting different aspects of strategic processing during the task. And these components can be compared across modes, revealing how cognitive and affective responses differ through specific gaze patterns: for instance, high loadings on fixation-based measures indicate attentional selectivity and monitoring, while pupil size and blinking-related measures indicate cognitive load and affective regulation, respectively. The interpretations will be guided not only by statistics but also by previous eye-tracking research.

The process included varimax rotation with Kaiser normalization. Moreover, t-tests and one-way ANOVA were conducted to compare by gender and performance, followed by a two-tailed partial correlation analysis, controlling for gender, English proficiency, and task performance to assess the alignment of responses from the self-report strategy inventory with the gaze-based strategic component scores. SPSS Statistics (Version 29.0.2.0; IBM Corp., 2023) was used for all analyses. Before any statistical tests to compare modes, Z-score normalization was applied to eye-tracking measures on the eye-tracking data because they were on different scales.

Results

RQ1: Gaze Behavior Patterns for Three Modes

Preparation Phase

Gaze metrics on seven AOIs in the preparation phase are summarized in Table 4. In general, test takers allocated most attention to text AOIs, briefly scanned over image AOIs, and oriented most quickly to video AOIs. When the prep-device info text was presented alongside the prep-device picture image, attention was more strongly focused on the text. Across modes, the text input elicited the highest number of fixations, whereas the image input showed the fewest, with the video mode falling in between ($F(2) = 78.08$, $p < .00$, $\eta^2 = .66$), indicating that textual information requires more sustained visual processing than images or video in this phase.

Table 4. *Descriptive gaze metrics (N = 43)*

Phase	Mode	AOI	Time to first view (ms)	Fixation duration (ms)	Number of fixations	Revisit number	Left pupil (mm)	Right pupil (mm)	Saccade magnitude (pixels)	Saccade direction (pixels)	Blink number	Blink duration (ms)
Preparation	Image	prep-timer	11.34	0.34	1.9	4.00	3.95	3.57	281.11	142.81	19.53	0.04
	Image	prep-device picture	6.76	1.42	4.6	4.30	4.17	4.16	430.96	214.97	20.40	0.08
	Text	prep-device info	1.09	19.57	61.5	18.38	4.11	3.95	257.34	159.93	20.01	0.19
	Text	prep-video instructions	2.34	3.30	14.8	24.99	4.32	4.17	288.53	173.54	17.43	0.18
	Video	1st prep-video	0.17	8.96	27.6	14.98	4.24	4.10	189.38	187.35	14.22	0.17
		2nd prep-video	0.20	10.14	27.1	10.00	4.30	4.16	182.28	187.81	12.71	0.18
		3rd prep-video	0.30	5.88	16.6	8.08	4.28	4.18	168.77	192.84	13.35	0.19
Response	Image	response timer	10.50	1.56	7.9	17.08	4.77	4.50	336.33	135.23	16.68	0.13
	Text	response instructions	2.54	15.44	57.4	58.16	4.82	4.82	314.20	173.62	17.23	0.20
	Image	response-audience picture	2.55	24.39	66.8	30.92	4.75	4.49	231.17	197.70	19.00	0.22

Image input, including the timer and device picture, showed shorter fixation duration and fewer fixations overall, and the timer was typically viewed by test takers at the periphery of the screen. At the same time, higher saccade magnitudes on the device picture suggest that test takers tended to scan this image over a relatively wide area. Taken together, these patterns indicated that image AOIs were treated as supplementary input for checking the device structure and temporal information.

Video input, which was highly relevant to the task, was presented three times and showed distinct patterns of gaze behaviors. Video AOIs had the shortest time to first view (0.17–0.30) and the fewest revisits, indicating that test takers reacted quickly to the dynamic content and watched it continuously. A follow-up repeated-measures ANOVA analysed eye-gaze patterns across the three viewings. The second viewing phase exhibited the longest viewing and fixation durations compared to the first ($F(2,41) = 4.99$, $p = .01$, $\eta^2 = .20$), but these metrics declined during the third viewing. A main effect of time also pointed to the attention to the AOIs peaked at the second viewing ($F(2) = 54.59$, $p < .00$, $\eta^2 = .57$), while fixation number ($F(2) = 77.87$, $p < .00$, $\eta^2 = .65$) and revisits ($F(2) = 30.37$, $p < .00$, $\eta^2 = .42$) decreased significantly across viewings. This pattern suggests that test takers initially scanned a broader range before finding relevant content, as evidenced by a declining trend in saccade magnitude— the distance between the current and previous fixations—($F(1, 42) = 5.86$, $p = .02$, $\eta^2 = .12$).

In summary, during the preparation phase, test takers invested the most attention and time to text AOIs. In particular, the most time was spent on the device-info text to understand the task context and plan their responses, with image AOIs as supplementary visual support. And they relied on repeated video viewings to understand the overall sequence of device operation.

Response Phase

Gaze metrics on the three AOIs for two modes—image and text—during the response phase are summarized in Table 2. In this phase, a text AOI (response instructions) and two image AOIs (the timer and the target audience picture) were presented; no video AOIs appeared. Overall, test takers devoted most of their attention to the response-instructions text and the audience picture, whereas the timer image was viewed as peripheral information. The instructions and audience picture AOIs drew test-takers' attention quickly from the onset of the phase (approximately 2.5 ms for both), while the timer was first fixated later in 10.5 ms, suggesting that test takers first checked what to do and whom they were addressing before monitoring the remaining time.

A similar pattern was observed for fixation duration and number. The audience picture exhibited the longest fixation duration (approximately 24 ms) and the highest fixation number (approximately 65 times), followed by the response-instruction text AOI (approximately 15 ms and 57 times). The timer AOI had the shortest duration and the fewest number (approximately 1.6 ms and 8 times) among the three AOIs. Revisit number also differed: the instructions AOI was revisited most often (approximately 58 times), followed by the audience picture (approximately 31 times) and timer (approximately 17 times). This pattern may suggest that during the speaking response, test takers had to repeatedly revisit the instructions to make sure they included the required elements to fulfill the task while keep looking back at the simulated audience to remain aware of the interlocutor's presence.

Longer saccade magnitudes involving the timer and instructions AOIs, relative to the audience picture, indicate that test takers had to shift their gaze back and forth between the AOIs at the center and the periphery during speaking. Lastly, pupil size and blink measures were higher for the instructions and audience picture AOIs than the timer. It can be interpreted that processing what to say and how to say it to the interlocutor involved greater cognitive load and affective arousal than simply checking the time.

In summary, during the response phase, test takers planned and monitored their speaking responses by referring to the text AOI and the audience image AOI, with occasional revisits to the timer to regulate the length of their response.

Gaze Patterns Within the Same Mode Across the Two Phases

Given that visual attention in speaking tasks is linked to the cognitive and linguistic processes at each task phase, this section compares gaze patterns within the same mode across the preparation and response phases. In image mode, AOIs showed distinct gaze patterns across task relevance and phases: the AOI considered highly task-relevant, such as the response-audience picture, drew substantially more attention than relatively peripheral AOIs (e.g., the prep-timer, prep-device picture, and response-timer).

According to Table 4, the response-audience picture AOI had the longest average fixation duration (24.39 ms), the highest number of fixations (66.8 times), and the shortest average time to first view (2.55 ms), indicating that test takers were immediately drawn to this visual input and treated it as highly relevant from the beginning of the task. This pattern suggests that test takers focused on their gaze on what seemed task-relevant, often visually skipping or briefly looking at less important content. Even with the same relatively peripheral 'timer' images (prep-timer and response-timer), gaze behavior differed by phase; values were higher during the response phase than during the preparation phase. This means that gaze behavior is not only driven by the inherent visual importance of input but is also context-dependent. These findings highlight the dynamic nature of strategic gaze behavior in the image mode, demonstrating that even visually minimal elements can become relevant when they face the immediate need in context.

A similar phase-dependent pattern was observed in the text mode. Test takers tended to give an intensive gaze to texts that were closely tied to task completion (e.g., prep-device info AOI > response instructions AOI > prep-video instructions AOI). The prep-device info and the response instructions, both considered as highly task-relevant, had longer viewing time, fixation duration, and number of fixations than the prep-video instructions. Between the two instructions text AOIs (prep-video instructions and response instructions), test takers gazed longer, more frequently, and revisited more often the response-instructions AOI. In contrast, prep-video instructions received lower fixation and were revisited less often. Similar to the two timer images, the different gaze patterns of the two instructions likely suggest the dynamic nature of strategic behavior, where test takers adjusted their attention allocation based on the contextual demands of task phases.

RQ2: Gaze Patterns Across Three Modes: Effects of Gender, Proficiency, and Cognitive Load

Drawing on the framework outlined in the literature review, which treats gaze behavior as an indicator of cognitive strategic processing, gaze patterns across the three input modes were compared in Table 5 to illustrate how cognitive burden and strategic processing differ.

A mixed ANOVA was designed with the three modes as within-subject factors, gender, and English proficiency in three groups (high, intermediate, and low) as between-subject factors, and post-task strategy report scores as covariates. The low, medium, and high levels indicated in Table 5 do not represent absolute score ranges, but rather relative positions of the three input modes in the same task. To be specific, based on the descriptive statistics in Table 4, when the means for a gaze indicator across the three modes were highest, it was marked as high, and when they were lowest, it was marked as low. And the middle value was marked as medium. Likewise, the same relative coding across three levels was applied to fixation-based, pupil/blink, and saccade-related measures.

Table 5. *Gaze behavior comparisons across three input modes*

Category	Image	Text	Video
Features	• High time viewed, fixation number, and revisit number • Medium pupil sizes • Short fixation duration (.243 sec) • Medium saccade magnitude and small direction • Low blink number and duration	• High time viewed, fixation duration, and number • Large pupil sizes • Medium fixation duration time (.310 sec) • Long saccade magnitude • Medium saccade direction • Highest blink number and duration	• Shortest time to first view • Small pupil sizes • Lower fixation duration/ number, revisit number than text • High fixation duration time (.386 sec) • Small saccade magnitude and long direction • Medium blink number and duration
	→ Medium cognitive load, with participants focusing on medium distances. → Less important images are often ignored, while visual attention increases significantly for images perceived as important.	→ Relatively higher cognitive load → Participants tend to scan over longer distances. → Text captures and maintains more attention than images.	→ Relatively low cognitive load → Participants focus on shorter distances. → Video captures attention rapidly. → Video input is associated with shorter saccades and fewer revisits.
Commonality	In all modes, inputs considered important led to a shorter time to the first view, increased fixation numbers and durations, and a greater number of revisits.		

The results demonstrated substantial disparities in gaze metrics by gender and language background. In all modes, female participants had larger left pupil sizes than males ($F(1) = 4.65$, $p = .04$, $\eta^2 = .11$). Larger pupil sizes correlate with strategies addressing heightened cognitive load and emotional states. Specifically, the biggest pupil size and the relatively high blinking number and duration were seen in text mode (4.69 mm) compared to the other modes, as seen in Table 4. Considering that pupil dilation and increased blinking are generally interpreted as indirect indicators of cognitive load, fatigue, and affective arousal (Josephson & Myers, 2019; Sáiz-Manzanares, et al., 2021; Winberg, 2021), this pattern suggests that, within this task, reading text is a relatively more cognitively demanding task than processing images or videos with video imposing the lowest demand. This is illustrated in Figures 5 and 6. However, it should be noted that this interpretation is based on the relative pattern, in which multiple gaze indicators show similar tendencies across modes, rather than on a single indicator used to draw conclusions about cognitive load.

The distance between fixations, or saccade magnitude, differs across modes ($F(1.67) = 22.28$, $p < .00$, $\eta^2 = .35$). Results include the longest saccades in text mode (290.84 pixels), the second largest in image mode (233.18 pixels), and the shortest in video mode (183.68 pixels). In image mode, the low-performing female group revealed distinctive saccade patterns. It may reflect more extensive visual exploration, markedly differing from that of the high-performing group ($F(2) = 3.74$, $p = .03$, $\eta^2 = .17$), as seen in Figure 7. Video mode exhibited the biggest saccade direction or angles, whereas images had the smallest ($F(1.53) = 38.81$, $p < .00$, $\eta^2 = .49$), which may suggest differences in how visual information was scanned across modes.

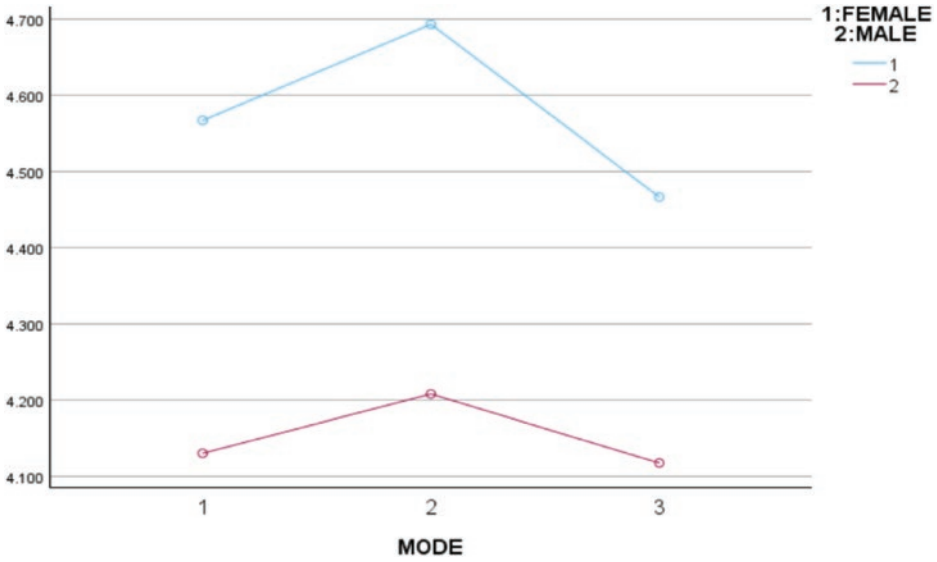

Figure 5. *Means of left pupil size by mode and gender.*

Note. Modes 1, 2, and 3 on the X-axis correspond to image, text, and video, respectively.

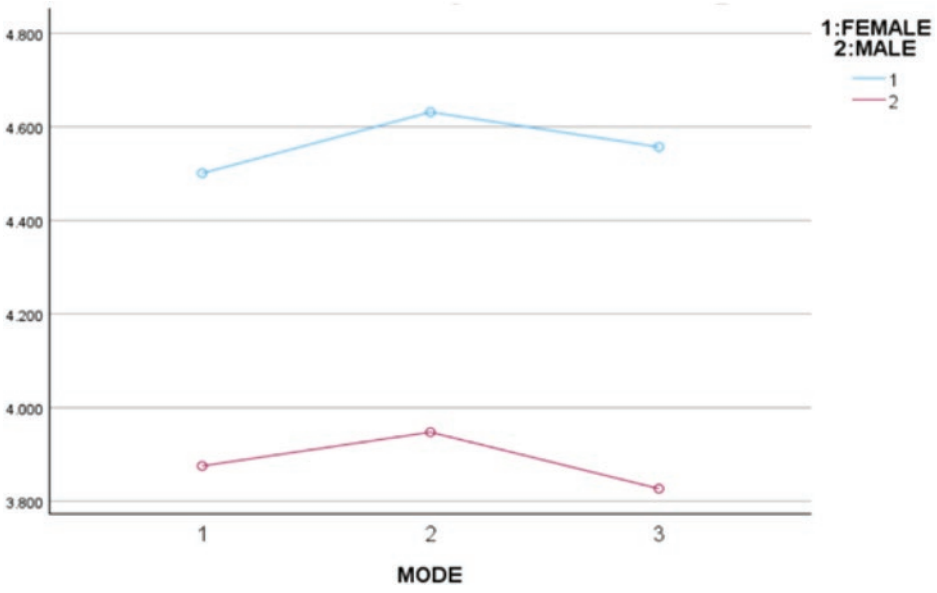

Figure 6. *Means of right pupil size by mode and gender.*

Note. Modes 1, 2, and 3 on the X-axis correspond to image, text, and video, respectively.

The blink number significantly differed across modes ($F(2) = 16.67$, $p < .00$, $\eta^2 = .29$): the highest frequency in text mode, followed by video mode, and the shortest in image mode. Difference in the blink duration was also found across modes ($F(2,40) = 65.92$, $p < .00$, $\eta^2 = .77$): the longest duration in text, followed by image, and the least in video. These data suggest that text mode, which involves more mental effort than the other modes, caused longer blink durations. Nevertheless, this finding should be viewed as indicative rather than conclusive, as blink measures are only indirect markers of cognitive load.

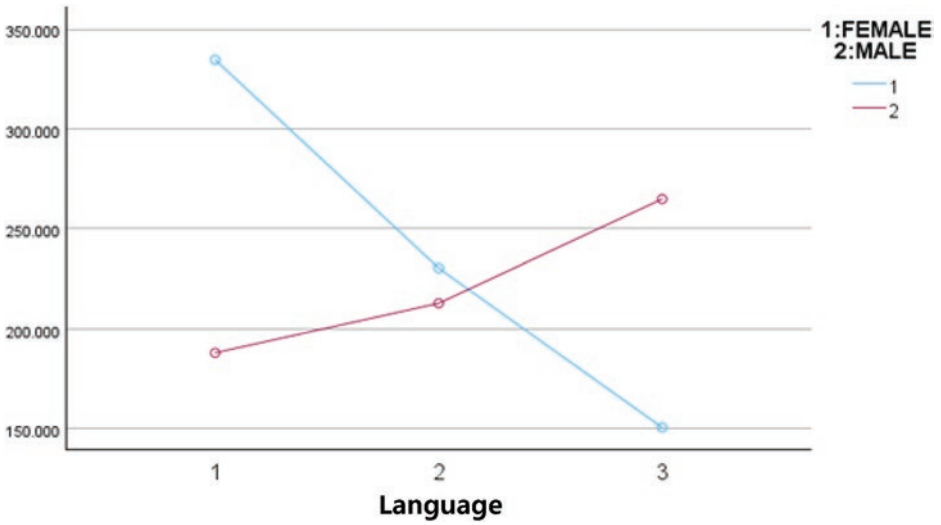

Figure 7. *Means of Saccade magnitude in image mode.*

Note. Languages 1, 2, and 3 on the X-axis of Figure 7 correspond to low L2, high L2 learners, and native speakers, respectively.

RQ3: Gaze-Based Strategic Components and Individual Variables

In image mode, the rotation converged after 5 iterations. Although the Kaiser-Meyer-Olkin (KMO) value (.42) was below the conventional threshold of .50, Bartlett's test of sphericity was significant ($\chi^2(45) = 184.97$, $p < .00$), indicating sufficient correlations among variables to proceed with dimensionality reduction. PCA found five components with eigenvalues exceeding 1, accounting for over 84% of the total variance (Component 1: 26.7%, Component 2: 20.9%, Component 3: 14.6%, Component 4: 11.5%, Component 5: 10.7%).

In text mode, the rotation converged after 8 iterations. The KMO value was .50, which meets the conventional threshold of .50 for acceptable sampling adequacy. Bartlett's test of sphericity was significant ($\chi^2(45) = 239.26$, $p < .00$), indicating sufficient correlations among variables to proceed with dimensionality reduction. PCA found four components with eigenvalues exceeding 1, accounting for 79.6% of the total variance (Component 1: 31.8%, Component 2: 19.3%, Component 3: 15.2%, Component 4: 13.3%).

In video mode, the rotation converged in 6 iterations. The KMO value was .50, which meets the conventional threshold of .50 for acceptable sampling adequacy. Bartlett's test of sphericity was significant ($\chi^2(45) = 250.66$, $p < .00$). PCA identified three components with eigenvalues exceeding 1, accounting for 68.5% of the total variance: Component 1: 32.4%, Component 2: 18.5%, Component 3: 17.6%. The correlation matrices and rotated component matrices for each mode are provided in Appendices 4–6 to illustrate relationships among gaze metric variables in each mode and to support interpretations of the extracted components.

The analysis found both shared and distinct gaze-based strategic components, which are shown in Table 6. Three main gaze-based strategic components appeared in all the input modes: (1) awareness and selectivity, (2) cognitive load management, and (3) compensation and emotional regulation. Long fixations and viewing times reflected attention allocation (awareness and selectivity). Pupil size showed mental effort (cognitive load management), and blinking showed compensatory strategies on how to deal with stress (compensation and emotional regulation).

Table 6. *Gaze-based strategic components*

Mode component	Image	Text	Video
Number of components	5	4	3
Awareness & Selectivity	Component 1: Fixation number, Time viewed	Component 1: Fixation number, Time viewed, Revisit number	Component 1: Saccade magnitude (–), Saccade direction (–), Fixation number, Time viewed
Self-monitoring	Component 2: Saccade direction, Saccade magnitude, Fixation duration	Component 2: Saccade magnitude, Fixation duration	*(Not distinctly separated; partially overlaps with Component 1)*
Cognitive load management	Component 3: Left & Right pupil size	Component 3: Left & Right pupil size	Component 2: Left & Right pupil size, Fixation duration
Compensation & Emotional regulation	Component 4: Blink number, Blink duration	Component 4: Blink number, Blink duration, Saccade direction	Component 3: Blink number, Blink duration
Rechecking	Component 5: "Revisit number	(*Loaded in Component 1*)	(*Not loaded in any component*)

In terms of mode-specific gaze patterns, the image mode added distinctive elements in self-monitoring components related to information processing and strategic rechecking in Components 2 and 5. Component 2 (self-monitoring) accounts for extensive scanning arising from controlled saccadic activity, linked to effective information processing. Component 3 (cognitive load management) and Component 4 (compensation and emotional regulation) were converged by pupil dilation across modes and blink pattern metrics as indicators of cognitive fatigue, respectively. Lastly, Component 5 was extracted from revisiting the AOI for verification. A subsequent *t*-test of component scores suggested a gender-related difference in Component 5 in this data, with a small effect size: male participants showed more frequent revisits to image AOIs ($t(41) = -.50$, $p = .03$, Cohen's d = −.16).

In text mode, Component 1 showed that participants were aware of and selective about the main details they were focusing on. Component 2 showed that participants were self-monitoring by scanning less and fixating for longer durations, as indicated by a reduced saccadic range. Component 3, exhibiting pupil dilation as a marker of cognitive load, reflects cognitive load management, while Component 4 represents the strategy to deal with cognitive fatigue and emotional stimuli by changing blinking and gaze direction.

In video mode, Component 1, converged by fixation and saccade metrics, represented awareness and selectivity. Exploratory analyses also indicated gender and performance differences. In this sample, male participants had higher Component 1 scores than females ($t(41) = -.89$, $p = .02$, Cohen's d = .28). Component 2 revealed pupil dilation as a sign of cognitive load with steady glances. Component 3 connected blinking to managing emotions under cognitive fatigue. The L2 low-performance group showed higher blinking frequency and duration than the high-performance group in this sample ($F(2) = 3.59$, $p = .04$, $\eta^2 = .15$). The results showed that individual differences, such as gender and proficiency, may affect potential variations in gaze-based indicators of strategic processing, as summarized in Table 7.

Table 7. *Individual difference effect on gaze-based strategic components*

Component	Mode	Difference	Evidence
Awareness & Selectivity	Video	Female < Male	$t(41) = -.89$, $p = .02$
Self-monitoring	Image	Female Low > Female High	$F(2) = 3.74$, $p = .03$
Cognitive load management	Image, text, video	Female > Male	$F(1) = 4.65$, $p = .04$
Compensation & Emotional regulation	Video	Low > High	$F(2) = 3.59$, $p = .04$
Rechecking	Image	Female < Male	$t(41) = -.50$, $p = .03$

A subsequent correlation analysis was performed to examine the relationship between the responses from the self-reported strategy inventory (refer to Appendix 7) and the factor scores obtained from gaze-based strategic components, while controlling for individual differences, including gender, English proficiency (native speakers, advanced, and intermediate English learners), and task performance scores. The zero-order (Pearson) correlations between the overall self-report score and the gaze-based factor scores were minimal. The overall self-report score exhibited a weak positive association with Component 4 (Compensation/ emotional regulation) of the image mode ($r = .30$, $p = .05$) and a negative correlation with participants' English proficiency ($r = -.38$, $p = .01$). However, further analysis employing partial correlation indicated no statistically significant evidence of a relationship between perceived and gaze-inferred strategy use across the different modes. In image mode, the correlations varied from rs = .02 to .22 ($p = .09$). In text mode, correlations went from rs = –.13 to .04 ($p = .41$). In the video mode, the correlations varied from rs = –.07 to .05 ($p = .66$). For more details, please see Appendix 7.

Regarding the effects of gender, it is noteworthy to report the post-task interviews that out of 43 participants, only five participants, three L2 male and two L2 female participants, clearly mentioned topic unfamiliarity as a source of task difficulty: for example, "The topic was not general but too specific, so I had difficulty answering it. (Participant #39)", "The task topic related to a specific technology, which I was not familiar with, so I felt need of more information." (Participant #16), and "I couldn't think of IT-related vocabulary, so I paraphrased it."(Participant #8, 38), "I worked hard to remember the operating sequence of the device, which was new to me." (Participant #28).

Discussion

Mode-Specific and Context-Sensitive Gaze Patterns

The findings indicate that the mode of input influences both cognitive and affective engagement. In particular, reading text prompted extended but effortful processing, potentially leading to greater fatigue. Earlier studies support this interpretation by showing that reading text involves a cognitively highly demanding process (Bax & Weir, 2012; Bax & Chan, 2019; Inan et al., 2015; Kaakinen & Hyönä, 2010; Mayer, 2005). Selective scanning in image input, reflected in moderate saccade distances and quick blinking, also suggests an effective strategy to filter unimportant visual information. The video appeared to prompt quick but focused engagement as test takers processed dynamic content through repeated viewings. These mode-dependent gaze patterns provide empirical evidence for cognitive load theory (Sweller, 2005), which distinguishes between intrinsic load, arising from inherent task complexity, and extraneous load, arising from task format and presentation. These findings suggest that

gaze allocation was not arbitrary but rather reflected strategic processing, shaped by task relevance and contextual factors.

In addition to these mode effects, the study also revealed phase-sensitive differences in how each mode was used. In the preparation phase, the text mode was used as a critical resource of information for understanding the task context, including its purpose; the image mode functioned as peripheral support for grasping the device structure and the temporal constraints; and the video mode served as the input for understanding the overall sequence of operating the device. In the response phase, however, the instructions text and audience picture became central resources for planning and monitoring spoken responses, whereas the timer image was treated as peripheral information that helped test takers to regulate the length of their responses. In other words, within each mode, gaze patterns reflected phase-sensitive strategic behavior, as the functions and strategic uses of AOIs differed across the phases. This accounts for attention shifting to peripheral cues (e.g., timers) depending on task phases and for cognitive processing peaking during the second viewing of video stimuli.

The notable context-sensitive nature of gaze allocation may be interpreted in light of Suvorov's (2015) distinction between context and content videos in L2 listening assessment: the former provides situational background, whereas the latter conveys information more directly relevant to task completion. In the present study, test takers may have allocated less attention to the image than to the text during the preparation phase because they perceived it as serving a context-supporting role, whereas during the response phase, they paid more attention to it, perceiving it as serving a target-audience role.

These findings of the present study are also supported by previous research on task-driven visual search (Flecken et al., 2015; Kaakinen & Hyönä, 2010; Lee & Winke, 2018), which draws on the distinction between intrinsic and extraneous load across modes in cognitive load theory (Sweller, 2005). Lee and Winke (2018) found that young non-native children frequently fixated on timers when they hesitated, interpreting this behavior as an indicator of distraction and anxiety among L2 learners. In the present study, gaze behaviors of adult speakers, however, exhibited increased fixation metrics, revisit counts, and pupil size on timers during the response phase, compared to the preparation phase. This pattern indicates that participants intentionally utilized peripheral yet contextually significant cues, such as timers, to manage time and pacing, reflecting strategies for awareness, cognitive load management, and rechecking. However, this interpretation remains tentative, and future research is needed to clarify how the functional role of each input type influences gaze allocation in multimodal speaking tasks.

Gaze-Based Strategic Components and Individual Differences

Three gaze-based strategic components appeared across all modes: awareness and selectivity, cognitive load management, and compensation and emotional regulation. These align with self-regulation models of learning (Zimmerman, 2000), emphasizing learners' use of attention allocation, monitoring, and affective control to meet task demands. Notably, the number and structure of these components varied by modality—five for image, four for text, and three for video. Care should be taken when interpreting these components as fully specified components of strategic competence, as they are derived from a small sample of eye-tracking data. Rather, they should be interpreted as tentative indicators of gaze-based processing.

In addition to the shared gaze-based strategic components, self-monitoring and rechecking components were added in image mode, which is relevant to the effective information processing used to scan broadly and verify information. In contrast, text mode was characterized by a focus on specific details and limited scanning, signaling high cognitive load and resulting in fewer but sustained strategic components—evidence of linear information processing. Video mode, meanwhile, elicited more

integrated gaze patterns aligned with the dynamic nature of the input. These findings are consistent with previous studies showing that as input complexity increases, strategies become more integrated and reactive (Das et al., 2024; Ekin et al., 2025; Martinez-Cedillo et al., 2025).

While Zhang et al. (2021) focused on planning, monitoring, and evaluation through a self-report inventory, this study suggests rethinking the underlying construct of strategic competence to include not only conscious, perception-based descriptions but also unconscious and physiological processes. Looking at both shared and mode-specific gaze patterns provides a fuller picture than relying only on perception-based methods. This may argue that a better understanding of the construct of strategic competence requires both. The approach to combining both resonates with Low and Aryadoust's (2021) argument that both gaze metrics (i.e., visit duration and fixation count) and self-report scores contribute to predicting listening test performance.

Additionally, both mode effects and individual differences were integrated into the analysis of gaze-based strategic components, as illustrated in Table 7. Nonetheless, given the limited sample size, these observations should be interpreted cautiously and require further empirical investigation. In image mode, male test takers gained higher scores on the component associated with revisiting AOIs, whereas some low-performing female test takers tended to show longer scanning durations. One possible interpretation, which should be treated with caution, is that it may have taken them longer to find important cues, leading them to process more information (Sweller, 2005). In video mode, low-performing participants showed higher scores on the component associated with compensation and emotional regulation strategies, whereas males tended to score higher on the component associated with awareness and selectivity. This finding extends Suvorov's (2015) study, which indicated individual differences in L2 learners' attention to video materials in a listening test, although it did not explicitly link this variation to performance. Similarly, findings of the current study agreed with those of Bax and Chan (2019), who found that effective readers fixed their eyes on critical textual locations, while ineffective readers relied more on superficial elements. In summary, these results tentatively suggest that gaze-based strategic processing patterns may vary not only by mode but also across learner characteristics.

Implications for Mixed-Method Approach

This study highlights the need for a mixed-method approach from a methodological perspective. The partial-correlations analysis indicated a small positive zero-order correlation between self-reported strategy use and gaze-based strategic components, together with a small negative correlation with English proficiency. This result probably represents shared variance with participants' individual factors rather than a direct correspondence between perceived and actual strategic behaviors. Because of the response-time difference, self-reported scores may overstate the cognitive effort perceived by lower-proficiency learners, some of whom clearly reported a challenge grasping nonverbal information presented in video mode in the subsequent interviews.

The discrepancy between test takers' perceived strategies and their real-time behaviors may thus provide objective insights into unconscious processes that the report-based approach rarely captures (Chauliaca et al., 2020). For instance, Low and Aryadoust (2021) found that gaze measures could predict final listening test performance, while self-report scores had moderate predictive power. As each data set likely taps into different aspects of the underlying construct of strategic competence, integrating both types of data can lead to a more comprehensive understanding of its complex mechanisms, influenced by individual factors. Thus, the validation of multimodal L2 assessments can be supported by investigating potential disadvantages for certain test-taker groups.

The findings also make it clear that multiple test modes, such as image, text, and video, require test-takers to draw on various information-processing demands and components of strategic competence. The eye-tracking analyses show mode-specific differences in gaze patterns, which can be interpreted as evidence of different strategic approaches to constructing responses. Such variation in strategic use may be further moderated by individual background. For example, task types based on different stimuli (i.e., picture, text, lecture) may yield different difficulties for learners with different proficiency levels. Since gender may subtly affect how test-takers approach modes and allocate attention, the interaction between mode and individual variables requires examination for fairness (Hu & Aryadoust, 2024). If different strategic burdens associated with mode and individual variables influence test scores, the scores will reflect not only the speaking construct to be assessed but also mode-specific gaze patterns and cognitive demands. In this study, regarding the potential role of task topic (a recycled invention) on the gender differences, the post-task interview indicated that only five L2 participants (three males and two females out of 43) reported unfamiliarity with the topic (the invented device for environmental protection) as a source of difficulty. Because this number was small and balanced across gender groups, topic unfamiliarity alone does not appear to sufficiently account for the gender-related patterns in gaze behavior. Instead, the findings may suggest differences in information processing, although this interpretation should be treated cautiously because the current study examined only a single topic and task type. Further research with a wider range of topics and task formats is warranted.

As for practical implications, rater training is necessary to raise raters' awareness of mode-related variables (Kang et al., 2020; Kwon & Yu, 2024), and information on modality in use can be incorporated in the development of rating rubrics. Task prompts may be sequenced to gradually increase processing demands and help manage cognitive load—for example, starting with a video to provide contextual support, then moving to an image to emphasize key ideas and facilitate verification, and concluding with text to allow for in-depth explanations. Spending more time on text-based prompts and adjusting their length and the complexity of their syntactic structures can help test developers and language teachers design multimodal tasks. This approach gives language learners equal opportunities to learn and be assessed by considering proficiency and gender.

To sum up, this study adds to the ongoing discussion of cognitive validity, strategic competence, and multimodal language assessment by exploring the interactions between modality and language learner characteristics to shape gaze behavior and strategic processing. For the development of multimodal L2 assessment theory and practice, this study argues for a mixed-method approach incorporating self-report, performance, and physiological indicators to strengthen empirical evidence by leveraging technological advancements.

Conclusion

This study offers several contributions. First, it examined underrepresented gaze metrics during speaking performance in a multimodal task. In addition to fixation counts and durations, the results underscore the potential of pupil size and blinking behaviors—metrics that have been somewhat overlooked—as significant indicators of learners' capability of cognitive processing and emotional regulation. Adding various metrics expands the range of gaze-based metrics for examining strategy use. Second, the study demonstrated the interaction between modality and learner characteristics, indicating that gaze behavior should be understood in conjunction with learner profiles for informing learners of remedies to improve their strategic behaviors. Finally, combining in a mixed-method approach both conscious and unconscious indicators could support learner strategy portfolios (Pourdana & Tavassoli, 2022; Seong, 2014). Such portfolios may provide language learners with diagnostic feedback on their strategy use and practical guidelines for efficient gaze paths fit for individuals, supported by advancing technologies, thereby helping them become more effective speakers in complicated, multimodal tasks.

This study also has several limitations. The use of a single monologic task format restricts the generalizability of results to other speaking contexts and to the strategic behaviors of more diverse groups. Various task types and participant groups are needed for future research. Although notetaking was not allowed to ensure continuous capture of gaze behavior, this decision may have increased the task's working memory load. Test takers had to remember information from both the reading and video-viewing phases without external support. Consequently, their cognitive processes may have reflected memory-related burden alongside mode-specific processing demands. This possibility should therefore be taken into account when interpreting the findings. Given the small sample size (N = 43), the exploratory PCA results on 10 gaze indicators (RQ3) should be viewed with caution and replicated with larger datasets. There is also a need for further work on how self-reported data and real-time gaze data can complement each other in capturing the interaction between conscious and unconscious mechanisms underlying strategy use across modes. Analysis of RQ2 and RQ3 was conducted by aggregating gaze metrics across the task phases, which limits phase-specific interpretations. A further technical limitation is the eye-tracking instrument's sampling frequency (60 Hz), which may not capture finer details of gaze behavior. Future studies may need to use higher-resolution eye-tracking instruments to investigate more subtle gaze traces and cognitive processes.

Acknowledgement

I would like to express sincere gratitude to the anonymous reviewers for their thoughtful and constructive comments. Their careful reading and valuable suggestions greatly motivated me to work harder to improve this study in the publication process.

References

Abutalebi, J., & Clashen, H. (2017). Memory retrieval and sentence processing: Differences between native and non-native speakers. *Bilingualism: Language and Cognition*, *20*(4), 657–658. doi:10.1017/S136672891700027X

Byun, J.-H. (2023). *Developing and validating a mobile augmented reality(MAR)-mediated English speaking assessment for Korean EFL high school learners* [Doctoral dissertation, Seoul National University]. http://www.dcollection.net/handler/snu/000000178360

Bachman, L. F., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441–465. https://doi.org/10.1177/0265532212473244

Bax, S., & Chan, S. (2019). Using eye-tracking research to investigate language test validity and design. *System 83*, 64–78. https://doi.org/10.1016/j.system.2019.01.007

Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE Reading test. *Cambridge ESOL Research Notes 47*, 3–14.

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Front Neurosci 12*, 152. doi: 10.3389 /fnins.2018.00152

Burton, J. D. (2022). Gazing into cognition: Eye behavior in online L2 speaking tests. *Language Assessment Quarterly, 20*(2), 190–214. DOI: 10.1080/15434303.2022.21436 80

Castro-Alonso, J. C., & Sweller, J. (2021). The modality principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 211–220). Cambridge University Press.

Cazzato, V., Basso, D., Cutini, S., & Bisiacchi, P. (2010). Gender differences in visuospatial planning: An eye movements study. *Behavioural Brain Research, 206*(2), 177–183. https://doi.org/10.1016/j.bbr.2009.09.010.

Chauliaca, M., Catryssea, L., Gijbelsa, D., & Donchea, V. (2020). It is all in the "Surv-Eye": can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research 8*(3), 26–39.

Cohen, A. D. (2014). *Strategies in learning and using a second language.* Routledge.

Das, A., Wu, Z., Skrjanec, I., Feit, A.M., (2024). Shifting focus with HCEye: Exploring the dynamics of visual highlighting and cognitive load on user attention and saliency prediction. *arXiv* 8, 236. https://doi.org/10.48550/arXiv.2404.14232

Dirkx, K. J. H., Skuballa, I., Manastirean-Zijlstra, C. S., & Jarodzka, H. (2021). Designing computer-based tests: design guidelines from multimedia learning studied with eye tracking. *Instructional Science 49*, 589–605. https://doi.org/10.1007/s11251-021-09542-9

Ekin, M., Krejtz, K., Duarte, C., Tuchowski, A. T., Krejtz, I. (2025). Prediction of intrinsic and extraneous cognitive load with oculometric and biometric indicators. *Science Reports 15*, 5213. https://doi.org/10.1038/s41598-025-89336-y

El Haj, M., Boutoleau-Bretonnière, C., Guerrero Sastoque, L., Lenoble, Q., Moustafa, A. A., Chapelet, G., Sarda, E., & Ndobo, A. (2023). How do women and men look at the past? Large scan path in women during autobiographical retrieval—a preliminary study. *Brain Sciences*, *13*(3), 439. https://doi.org/10.3390/brainsci13030439

Field, J. (2012). *The cognitive validity of the lecture-based questions in the IELTS Listening paper.* In L. B. Taylor, & C. J. Weir (Eds.), *IELTS Collected Papers 2: Research in reading and listening assessment*. Cambridge University Press.

Flecken, M., Gerwien, J., Carroll, M., & Stutterheim, C. V. (2015). Analyzing gaze allocation during language planning: a cross-linguistic study on dynamic events. *Language and Cognition, 7*(1), 138–166. doi: 10.1017/langcog.2014.20

Frenck-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language: a review of methodologies and experimental findings. *Second Language Research*, *21*(2), 175–198.

Gazepoint. (n.d.). *Gazepoint Analysis UX Edition (Version 6.11.0) [Computer software]*. Gazepoint. https://www.gazept.com/

Gazepoint. (n.d.). *Gazepoint GP3 Eye Tracker* [Hardware]. Gazepoint. https://www.Gaze pt.com/

Gazepoint. (2023). Gazepoint analysis user manual. Gazepoint. https://www.gazept.com/

Gnetov, D., & Kuperman, V. (2024). Reading proficiency predicts spatial eye-movement control in the first and second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 50*(8), 1315–1328. https://doi.org/10.1037/xlm0001325

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition, 35*(3), 483–517.

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

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

Huang, H.-T. D. (2016). Exploring strategy use in L2 speaking assessment. *System, 63,* 13–27.

IBM Corp. (2023). *IBM SPSS Statistics for Windows (Version 29.0.2.0)* [Computer software]. IBM Corp.

Inan, F. A., Crooks, S. M., Cheon, J., Ari, F., Flores, R., Kurucay, M., & Paniukov, D. (2015). The reverse modality effect: Examining student learning from interactive computer-based instruction. *Computers & Education, 82*, 234–246.

Jarodzka, H., Janssen, N., Kirschner, P.A. and Erkens, G. (2015). Avoiding split attention in computer-based testing. *British Journal of Educational Technology 46*(4), 803–817. https://doi.org/ 10.1111/bjet.12174

Josephson, S., & Myers, M. (2019). Augmented reality through the lens of eye tracking. *Visual Communication Quarterly, 26*(4), 208–222.

Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during reading. *Journal of experimental psychology. Learning, Memory, and Cognition*, *36*(6), 1561–1566. https://doi.org/10.1037/a0020693

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kang, S., Park, H., Lee, S., Min, B., Hong, E., & Ahn, H. (2020). A study of variables in the online Korean speaking tests for academic purpose. *Journal of the International Network for Korean Language and Culture, 17*(1), 35–62. DOI: 10.15652/ink.2020.17.1.035

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

Kruger, J.-L., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2. *Journal of Eye Movement Research*, *7*(5), 1–15. https://doi.org/10.16910/jemr.7.5.4

Kwon, S. K., & Yu, G. (2024). The effect of viewing visual cues in a listening comprehension test on second language learners' test-taking process and performance: An eye-tracking study. *Language Testing, 41*(3), 649–680.

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing, 35*(2), 239–269.

Lee, Y. (2020). A study of using flipped learning on differentiated-level college learners' perceptions, class satisfaction, and reading and English-speaking Achievement. *Korean Association For Learner-Centered Curriculum And Instruction, 20*, 181–206.

Lindberg, R., McDonough, K., & Trofimovich, P. (2021). Investigating verbal and nonverbal indicators of physiological response during second language interaction. *Applied Psycholinguistics*, *42*(6), 1403–1425. doi:10.1017/S014271642100028X

Low, A. R. L., & Aryadoust, V. (2021). Investigating test-taking strategies in listening assessment: a comparative study of eye-tracking and self-report questionnaires. *International Journal of Listening, 37*(2), 93–112. https://doi.org/10. 1080/ 1090 4018. 2021.1883433

Low, R., & Sweller, J. (2014). The modality principle in multimedia learning. in R. E. Mayer (ed.), *The Cambridge handbook of multimedia learning* (pp. 227–246). Cambridge University Press.

Ma, W., & Winke, P. (2022). An investigation of the impact of jagged profile on L2 speaking test ratings: Evidence from rating and eye-tracking data. *Language Assessment Quarterly, 19*(4), 394–421. DOI: 10.1080/15434303.2022.2078720

Martinez-Cedillo, A. P., Gavrila, N., Mishra, A., Geangu, E., & Foulsham, T. (2025). Cognitive load affects gaze dynamics during real-world tasks. *Experimental Brain Research 243*(4), 82–91. https://doi.org/10.1007/s00221-025-07037-4

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. Cambridge University Press.

Mayer, R. E. (2020). Modality Principle. In *Multimedia Learning* (3rd ed.), pp. 281–304. Cambridge University Press.

Mayer R. E., Lee H., & Peebles A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology, 28*, 653–660. doi: 10.1002/acp.3050

McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*, *35*(1), 51–73.

Miyahira, A., Morita, K., Yamaguchi, H., Morita, Y. and Maeda, H. (2000), Gender differences and reproducibility in exploratory eye movements of normal subjects. psychiatry and clinical neurosciences, *54*, 31–36. https://doi.org/10.1046/j.1440-1819.2000.00632.x

Moreno, R., Mayer, R. (2007). Interactive multimodal learning environments. *Educ Psychol Rev 19*, 309–326. https://doi.org/10.1007/s10648-007-9047-2

Morishima, Y. (2013). Allocation of limited cognitive resources during text comprehension in a second language. *Discourse Processes*, *50*(8), 577–597. https://doi.org/10.1080/0163853X.2013.846964

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24*(4), 517–537.

Oxford, R. L. (1986). *Strategy inventory for language learning*. Various versions. Oxford Associates.

Park, E. & Kim, J. (2022). The effects of the learning variables and self-regulated learning strategy affecting academic achievement of a general English online course. *English Language Assessment, 17*(2), 205–228.

Park, M. (2018). Innovative assessment of aviation English in a virtual world: Windows into cognitive and metacognitive strategies. *ReCALL, 30*, 196–213.

Plass, J. L., Moreno, R., & Brünken, R. (Eds.). (2010). *Cognitive load theory*. Cambridge University Press.

Pourdana, N. & Tavassoli, K. (2022). Differential impacts of e-portfolio assessment on language learners' engagement modes and genre-based writing improvement. *Lang Test Asia 12*, 7. https://doi.org/10.1186/s40468-022-00156-7

Sáiz-Manzanares, M. C., Pérez, I. R., Rodríguez, A. A., Arribas, S. R., Almeida, L., & Martin, C. F. (2021). Analysis of the learning process through eye-tracking technology and feature selection techniques. *Applied Sciences, 11*(13), 6157.

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: evidence from pupillometry. *Frontiers in Psychology*, *5*(137). doi:10.3389/fpsyg.2014.00137

Seong, Y. (2014). Strategic competence and L2 speaking assessment. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 14*(1), 13–24.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 463–483.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 19–29). Cambridge University Press.

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.004

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology, 11*(6), 67–86.

Winberg, M. (2021). *Implementation of a mobile device eye-tracking system using augmented reality.* [Degree project in media technology, Second cycle, 30 credits]. Stockholm, Sweden.

Yang, H., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly, 46*, 80–103.

Zhang, D. & Goh, C. C. M. (2009). Strategy knowledge and perceived strategy use: Singaporean students' awareness of listening and speaking strategies. *Language Awareness, 15*(3), 199–219.

Zhang, L., Goh, C., & Kunnan, A. (2014). Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: A multi-sample SEM approach. *Language Assessment Quarterly, 11*, 102–76.

Zhang, W., Zhang, L. J., & Wilson, A. J. (2021). Supporting learner success: Revisiting Strategic competence through developing an inventory for computer-assisted speaking assessment. *Frontiers in Psychology, 12*, 689581. https://doi.org/ 10.3389/ fps yg.2021.689581

Zhang, W., Zhang, L. J., & Wilson, A. J. (2022). Strategic competence, task complexity, and foreign language learners' speaking performance: A hierarchical linear modelling approach. *Applied Linguistics Review, 15*, 1121–1149.

Zimmerman, B. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25*(1), 82–91. https://doi.org/10.1006/ceps.1999.1016.

Appendix 1

Test-Taking Strategy Inventory for L2 Speaking Assessment (Adapted from Huang, 2016)

The following behaviors are considered good for improving your ability to speak in English. Please rate each of the following statements by the extent to which they are true of you by ticking (√) one on a scale from 1 (Almost never true of me) to 5 (Almost always true of me).

No.	Statement
1	When I was given time for preparation, I thought about how to organize my answer.
2	I knew the goals of the tasks and what they required me to do.
3	I understood the purpose of the visual cues for the tasks.
4	When I was given time for preparation, I rehearsed my answer.
5	I understood the steps required to complete tasks.
6	When reading texts for tasks, I skipped words I did not know.
7	When I saw pictures, I moved from the center to the periphery.
8	When I saw the pictures in the warm-up activity, I used my background knowledge to connect two items.
9	When responding to questions, I tried to respond to the part of the question that I understood.
10	When reading questions, I would guess their meaning if I didn't understand.
11	When reading questions for responding, I guessed what they might ask if I did not understand them.
12	When reading questions for responding, I tried to understand the gist and not every word.
13	I answered by using familiar sentence patterns.
14	When responding to questions, I tried to elaborate on answers as much as I could.
15	When responding to questions, I would incorporate my relevant experiences.
16	I tried to use complete sentences rather than words or phrases to give my answers.
17	When reading the visual text, I tried to catch the main idea rather than every word.
18	When reading the questions for responding to questions, I tried to figure out the keywords.
19	I tried to simplify my answer.
20	I tried to slow down.
21	I used words and phrases to give my answer rather than complete sentences.
22	I encouraged myself to do my best.
23	I reminded myself not to be nervous.
24	I comforted myself that it was fine if I did not do well.
25	I told myself that this test was a good opportunity to assess my English oral proficiency level.
26	When I was given time for preparation, I thought about the vocabulary word to use.
27	When preparing for tasks, I thought about the English names of the objects in visual cues (text, image, and video).
28	When I was given time for preparation, I thought about the sentences to use.
29	I paid attention to the visual cues (text, audio, video) to answer the questions accurately.
30	When preparing for tasks, I rehearsed unfamiliar words and phrases in my mind.

(continue)

(continued)

No.	Statement
31	When preparing tasks, I browsed through the visual cues (text, audio, video) several times.
32	When preparing for Task 1, I rehearsed the process of working the device in my mind.
33	When preparing to answer Task 1, which involves explaining the sequence, I planned the order in which to explain how the device works.
34	I relied on different visual cues (text, audio, video) depending on the tasks.
35	I used visual cues to determine how to approach each task.
36	I organized what I was going to say before speaking.
37	I paid attention to the time allowed for preparation.
38	I guessed the meaning of the unknown words or expressions by relying on the visual cues (text, audio, video) in the tasks.
39	When explaining how the device works, I would paraphrase the English names of the objects that I did not know.

Appendix 2

Interview Questions

The interview included the following prompts:

1. How difficult was the task for you? Please select one number on a 5-point Likert scale (1 = not difficult at all, 5 = extremely difficult) and explain why you selected that rating.

2. What do you think made the task easy or difficult for you?

3. Can you remember how you approached Phase 1? For example, which one did you look at first: the directions, the text, or the picture? Please explain how you understood the information during this phase.

4. Can you explain how you understood the working steps of the device from Phase 2, where you watched the video without captions, and how this helped you prepare your response?

5. (Optional when distinctive gaze behaviors were observed.) What was happening at that moment, and what were you thinking during that part of the task?

Appendix 3

Speaking Rating Scale (Anonymous, 2023)

	Accuracy	Fluency	Content
	• Range and accurate grammar and lexical use • Complexity and facility of syntactical structures • Repetitions and error corrections	• Natural flow, Intonation, and accent • Comprehensibility of pronunciation • Frequency of pauses and hesitations	• The response includes awareness of the interlocutors (rainforest rangers as potential device users), procedural information detailing four steps, and the device's benefits.
4	Proper choice and use of language forms for the task Diverse and complex syntactical structures Diverse and effective (appropriate) choice of vocabulary Evidence of being aware of keywords Few grammatical errors	Full awareness of rhythm and intonation Highly comprehensible pronunciation Almost no pause	Responses include all requested information.
3	Proper choice and use of language forms to the task Attempts to use complex syntactical structures but the accuracy of the responses mainly comes from repetitive sentence structures Noticeable presence of vocabulary describing shapes, status, and actions Correct use of keywords and topic-relevant terminology in sentences Minor errors that don't affect comprehensibility	Comprehensible pronunciation One or two consistently distinct accents A few pauses or hesitations	Responses do not indicate that the forest rangers are aware of themselves as potential device users. Responses include at least three steps out of the device's four-step procedure.
2	Lack of diversity and facility of syntactical structures Avoidance of the keywords new to test takers and a lack of understanding of relevant terminology related to the topic Insufficient length of sentence(short and simple sentences) Presence of significant grammar errors that negatively affect comprehensibility and communication, and that violate basic grammaticality	Frequent pauses and hesitations result into overall slow pace.	Responses include half of the four steps (at most two steps) of the procedure.
1	Incomplete sentences Struggle to develop discourse at the sentence level Clear evidence of avoiding and misunderstanding some keywords blocks the comprehension of the task	Long pauses Many of the words used are difficult to understand.	Test takers fail to explain most of the procedure due to a lack of understanding.

Appendix 4

Correlation Matrix for Gaze Metrics in Image Mode

Variables	1	2	3	4	5	6	7	8	9	10
1. Left pupil	–									
2. Right pupil	.62	–								
3. Fixation duration	–.41	–.11	–							
4. Saccade magnitude	.22	.27	.35	–						
5. Saccade direction	–.04	.16	.46	.54	–					
6. Blink number	–.08	.09	.14	.33	.06	–				
7. Blink duration	–.02	.11	.30	.36	.23	.41	–			
8. Fixation number	–.21	–.03	–.01	–.38	–.14	–.08	–.11	–		
9. Time viewed	–.28	–.02	.29	–.32	–.09	–.18	–.13	.85	–	
10. Revisit number	.06	–.13	.05	.10	.22	–.04	.24	.06	–.19	–

Rotated Component Matrix for Gaze Metrics in Image Mode

	Component				
Variable	**1**	**2**	**3**	**4**	**5**
Left pupil	–.22	–.14	.88	–.09	.12
Right pupil	.09	.15	.89	.11	–.15
Fixation duration	.20	.78	–.35	.18	–.08
Saccade magnitude	–.35	.67	.29	.32	.01
Saccade direction	–.08	.85	.09	–.02	.19
Blink number	–.10	.01	–.01	.89	–.15
Blink duration	–.02	.26	.03	.75	.30
Fixation number	.94	–.15	–.03	–.02	.11
Time viewed	.95	.08	–.10	–.12	–.17
Revisit number	–.03	.09	–.04	.04	.96

Appendix 5

Correlation Matrix for Gaze Metrics in Text Mode

Variables	1	2	3	4	5	6	7	8	9	10
1. Left pupil	–									
2. Right pupil	.61	–								
3. Fixation duration	.08	.27	–							
4. Saccade magnitude	.09	–.14	–.61	–						
5. Saccade direction	.15	.43	.20	–.25	–					
6. Blink number	–.15	.09	–.21	.03	.34	–				
7. Blink duration	–.18	–.03	–.01	.20	.21	.47	–			
8. Fixation number	.20	.32	.32	–.19	.20	.11	.31	–		
9. Time viewed	.21	.28	.53	–.34	.12	–.07	.22	.92	–	
10. Revisit number	.20	.23	.04	.15	.28	.18	.26	.64	.42	–

Rotated Component Matrix of Gaze Metrics in Text Mode

	Component			
Variable	**1**	**2**	**3**	**4**
Left pupil	.16	–.12	.86	–.22
Right pupil	.16	.19	.85	.16
Fixation duration	.27	.82	.10	–.11
Saccade magnitude	.03	–.92	.00	–.02
Saccade direction	.02	.31	.45	.67
Blink number	.02	–.11	–.05	.86
Blink duration	.42	–.16	–.27	.64
Fixation number	.93	.20	.13	.09
Time viewed	.87	.41	.08	–.08
Revisit number	.72	–.22	.23	.23

Appendix 6

Correlation Matrix for Gaze Metrics in Video Mode

Variables	1	2	3	4	5	6	7	8	9	10
1. Left pupil	1.00	.51	–.13	.02	–.11	–.31	–.26	.00	–.01	.02
2. Right pupil	.51	1.00	–.32	.25	.20	.02	–.13	.14	–.00	.20
3. Fixation duration	–.13	–.32	1.00	–.41	–.21	–.33	.05	–.06	.42	–.37
4. Saccade magnitude	.02	.25	–.41	1.00	.90	.07	–.15	–.47	–.61	.29
5. Saccade direction	–.11	.20	–.22	.90	1.00	.10	–.04	–.43	–.49	.29
6. Blink number	–.31	.02	–.33	.07	.10	1.00	.49	.21	.02	–.10
7. Blink duration	–.26	–.13	.05	–.15	–.04	.49	1.00	.28	.23	–.12
8. Fixation number	.00	.14	–.06	–.47	–.43	.21	.28	1.00	.79	.07
9. Time viewed	–.01	–.00	.42	–.61	–.49	.02	.23	.79	1.00	–.34
10. Revisit number	.02	.20	–.37	.29	.29	–.10	–.12	.07	–.34	1.00

Rotated Component Matrix Gaze Metrics in Video Mode

	Component		
Variable	**1**	**2**	**3**
Left pupil	.19	.55	–.59
Right pupil	.05	.82	–.17
Fixation duration	.30	–.69	–.26
Saccade magnitude	–.85	.34	.05
Saccade direction	–.80	.22	.14
Blink number	–.02	.15	.85
Blink duration	.20	–.10	.75
Fixation number	.79	.31	.34
Time viewed	.87	–.09	.12
Revisit number	–.29	.50	.01

Appendix 7

Zero-Order and Partial Correlations Between the Total Self-Report Strategy Score and Gaze-Based Factor Scores by Mode

Mode	Pair (self-report total ↔)	r	p	df	partial r	df	p
Image	Awareness & Selectivity	.20	.20	41	.16	38	.34
	Self-monitoring	.08	.62	41	.21	38	.21
	Cognitive load management	.02	.92	41	.07	38	.69
	Compensation & Emotional regulation	.30	.05	41	.29	38	.07
	Rechecking	−.06	.72	41	−.05	38	.74
Text	Awareness & Selectivity	−.18	.25	41	−.05	38	.75
	Self-monitoring	−.02	.92	41	.04	38	.79
	Cognitive load management	−.07	.68	41	−.05	38	.76
	Compensation & Emotional regulation	−.05	.76	41	−.05	38	.75
Video	Awareness & Selectivity	−.06	.69	41	.02	38	.91
	Cognitive load management	−.08	.62	41	−.04	38	.81
	Compensation & Emotional regulation	.06	.69	41	.00	38	.99
Gender	1: Female, 2: Male	−.03	.90	41			
Proficiency	3: Native, 2: High, 3: Mid	−.38	.01	41			
Performance	Speaking score	.02	.90	41			