Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 16, No. 1; 2026, March

ISSN 2476-5880
https://www.ijlt.ir

Enhancing Speaking Proficiency Through Graphic-Rich Task-Based Language Testing: A Multimodal Assessment Approach

Nabaa Talib Hashim[1*], Wurood Shaher[2]

ARTICLE INFO

Article History:
Received: September 2025
Accepted: November 2025

KEYWORDS
English as a Second Language (ESL);
graphic-rich prompts;
learner performance;
speaking assessment;
task-based assessment.

ABSTRACT

Speaking tests often depend on text-only prompts, which can narrow what learners say and how they say it. The current research aimed to find out whether adding graphics would support performance within a task-based speaking assessment. Sixty adult ESL learners at CEFR B1–B2 completed the same narrative task under one of two conditions: a graphic-rich prompt or a text-only prompt. The researchers scored the recordings with an analytic rubric adapted from established speaking scales, targeting four constructs: lexical diversity, syntactic complexity, content organization, and fluency. Three trained raters scored all performances after a brief calibration session. The researchers estimated agreement using intraclass correlation coefficients. Group differences were tested with independent-samples t tests. Learners who received the visual prompt performed better on all rubric dimensions. Speech was more coherently sequenced and delivered at a steadier pace, and raters showed consistent judgments. Participants reported that the visuals made planning easier, reduced uncertainty, and kept them on track. Although the work centers on a single narrative task and one proficiency band in one program, the pattern of results points to a practical, low-cost change in assessment design. Well-constructed visual prompts can strengthen performance and increase the perceived authenticity and comprehensibility of task-based speaking tests.

1. Introduction

Speaking is an essential building block of communicative competence and is a specific focus of language assessment (Bachman, 1990; Luoma, 2004). Task-based language assessment aims to assess linguistic performance by means of real-world tasks to elicit authentic language use (Norris, 2009); however, when it comes to the speaking assessment, the prompts or questions for traditional speaking tests are text-based with limited visual or graphic support. In fact, in many contexts where test-takers are asked to speak about a text-based prompt or abstracted topic, such prompts may not sufficiently engage learners or adequately reflect their communicative performance (Fulcher, 2003). The question to be investigated in this study was why visuals (images,

[1] Nabaa Talib, Email: nabaa.talib@uomustansiriyah.edu.iq
[2] Wurood Shaher, Email: wurood-shaher@berkeleycolleg.edu

graphics) are overlooked in speaking test prompts. Visual prompts–whether these be picture prompts, story sequences, or infographics–are intended to support context, cognitive load, and spark richer language use; however, these visuals remain uncommon in standardized speaking assessments. This is an important oversight given the potential of using graphics to (a) enhance the quality of language produced and (b) reduce learner anxiety.

Therefore, this study examined whether graphic-rich prompts could influence performance in ESL speaking assessments. Building on task-based assessment research, which stresses creating tasks that resemble real communication, this study compared two groups of learners. One group completed the speaking task using a sequence of images, while the other worked with the same task presented in text. While multimodal input has been discussed in language testing, there is still little direct evidence showing how visuals shape the quality of learners' spoken responses, the consistency of raters' judgments, or how students themselves perceive such tasks. By taking up this question, this work extends current research on task design in speaking assessment.

The study proposed three main hypotheses. Learners given graphic-rich prompts would likely produce language with greater lexical variety, more complex grammar, better organization, and smoother fluency than learners responding to text-only prompts. the authors also expected rater consistency to hold steady, or possibly improve, when visuals were used. Finally, it was anticipated that learners would find the graphic-rich prompts clearer, more engaging, and less stressful, which would add to the validity of using visuals in speaking assessment. To address these issues, this study pursued three research questions, each tied to a specific objective.

RQ1: Do graphic-rich prompts improve ESL learners' speaking performance compared to text-only prompts?

The first objective was to examine whether images support learners in producing more varied vocabulary, greater syntactic complexity, clearer organization of ideas, and smoother fluency. the authors approached this objective with the assumption that visual context could serve as scaffolding, allowing learners to free up cognitive resources for richer language use.

RQ2: Are speaking scores from tasks with visual prompts rated as consistently as those from text-only tasks?

The objective of this research question was to test whether the reliability of scoring holds steady when images are introduced. Since inter-rater reliability is crucial to fairness, it was also important to determine whether visuals would create discrepancies among raters or, alternatively, whether they might actually make judgments more consistent by grounding performances in a shared storyline.

RQ3: How do learners perceive graphic-rich prompts in speaking assessments?

The final objective of this study was to understand the learner experience. If a task is seen as clear, engaging, and less anxiety-inducing, then it arguably provides a fairer measure of ability. By comparing student perceptions of the two prompt types, the authors aimed to capture the affective side of assessment — whether visuals make tasks not only more effective but also more authentic and motivating for those who take them.

2. Review of Literature

2.1 Task-Based Assessment

Task-based language assessment (TBLA) rests on a straightforward premise: tests should reflect how people actually use language to accomplish goals. Foundational accounts argued that meaningful, goal-oriented tasks reveal ability more clearly than decontextualized items (Brindley, 1989; Long & Norris, 2000), a view reinforced in later syntheses that frame performance as communication in context rather than display of abstract knowledge (Norris, 2016). Recent empirical work has extended this case with classroom and program-level evidence. In a direct comparison with a Present–Practice–Produce model, task-based assessment elicited more functional language use and stronger vocabulary retention (Noroozi & Taheri, 2022). Classroom implementations that require learners to plan and deliver task-focused speaking have likewise reported gains in oral proficiency, including greater lexical reach and confidence relative to conventional activities (Panduwangi, 2021). Parallel developments in testing practice are visible in

high-stakes contexts, where speaking components increasingly incorporate task-based designs to better capture situated communication (East, 2021).

Attention to authenticity has also guided new exam development. A recent digital speaking exam for young learners of Dutch grew out of a needs analysis that mapped real communicative situations to test tasks, demonstrating that realistic scenarios can be standardized even for heterogeneous cohorts (Long, 2005). Despite such momentum, practice does not always match principle. Many speaking tests still rely on minimal, text-only prompts that provide little context and limited opportunities for interaction, a design choice that can depress both engagement and the quality of evidence about ability (Figueras, 2020). The trajectory of findings across these studies points to a workable remedy: align prompts and tasks more closely with the kinds of communicative work learners actually do. Within that effort, graphic-rich materials offer a plausible means to increase authenticity and comparability by supplying shared context for content generation and discourse organization. This rationale underpins the present focus on whether, and how, task-based speaking assessments benefit from visual support.

2.2 Visual and Multimodal Prompts

Recent work treats visuals as part of the speaking task rather than ornament. When a prompt offers a short, coherent sequence of images tied to the communicative goal, test takers plan faster, map a storyline, and keep delivery steady. Studies that asked learners to narrate from four to six ordered frames after one to two minutes of preparation reported more specific details and cleaner event links than text-only instructions, along with gains in fluency and accuracy (Nasri et al., 2019; Mohammad et al., 2020). Learners themselves often describe the benefit in plain language, for example, “easier to think what to say,” which captures the planning support that pictures provide during timed speech (Ahmed, 2020).

These patterns of performance have been observed across various contexts, from adult English as a Foreign Language (EFL) classroom to beginner Arabic courses, and they remain consistent when picture sequences are accompanied by brief guiding questions that focus learners’ attention on key details like "who did what, when, and why" (Sarmiento-Campos et al., 2022). Expanding on this research, Kwon (2024) examined the role of visual stimuli in listening assessments, moving beyond their use in speaking tasks. Using eye-tracking technology, Kwon found that visual cues not only aid cognitive processing but also increase task engagement, providing further evidence of their effectiveness in enhancing language test performance.

Process and perception evidence help clarify why visuals matter. Pairing imagery with words reduces the burden on working memory and supports retrieval during planning, a mechanism consistent with dual-channel accounts of cognition (Li et al., 2022). Learners tend to experience that support as clarity rather than complexity. In a large undergraduate survey, static picture prompts were rated as easier to interpret and more helpful for outlining ideas than short videos, which some students felt introduced unnecessary background detail that was hard to manage under time limits (Raja Nur Hidayah Yacob et al., 2025). School-based reports show similar benefits for younger test takers: picture-supported speaking lessons, where pupils preview frames, jot a few keywords, and then tell the story, are associated with higher participation and confidence during delivery (Le Huong Hoa et al., 2022; Zahran, 2022). Together, these strands suggest that visuals help at precisely the moment many speakers struggle, the few seconds when ideas must be selected, ordered, and realized as speech.

The advantages are conditional on design. Picture sets that look similar on the surface can elicit different profiles of fluency, lexis, and complexity, which makes piloting and equivalence checks essential before operational use (de Jong & Vercellotti, 2016). Recent development studies outline practical controls that preserve support without cueing content: use culturally transparent images with comparable topical load, add one short cue under each frame, allow brief silent planning, and limit notes to keywords so reading does not replace speaking (Kakitani & Kormos, 2024). Reliability can be maintained when scoring procedures are explicit. In classroom validation, raters calibrated on anchor responses and applied a functional adequacy scale with stable agreement across picture tasks, indicating that visuals can be integrated without compromising score consistency (Koizumi & In’nami, 2022). The overall implication of post-2015 research is therefore

balanced but encouraging. Well-designed visual or multimodal prompts make speaking tasks more comprehensible and engaging while supporting dependable scoring, provided that test specifications, rater calibration, and materials quality are handled with the same care expected of any high-stakes assessment.

2.3 Speaking Performance Constructs

Speaking performance in tests is usually assessed in four areas. In this study, speaking performance is judged in four areas: lexical diversity (word variety), syntactic complexity (how flexible and well-structured the sentences are), content organization (whether the ideas unfold in a clear, sensible order), and fluency (speed and smoothness, with fewer long pauses). These constructs matter in practice. In timed narratives, for example, raters listen for whether a speaker names key participants and events, varies word choice as the story progresses, uses connective language to mark time and cause, and maintains a steady pace without long hesitations. Treating these features as separate lenses produces a more complete picture of what the task elicits and what listeners are actually responding to.

Recent work has refined how these constructs are measured in speech. For lexical diversity, a central concern is length sensitivity, since very short and very long responses can distort simple ratios. A large validation study showed that common indices such as Root TTR and D behave inconsistently across different response lengths, while moving-window measures like MATTR and MTLD were more stable and more closely aligned with proficiency (Kyle et al., 2024). This has practical consequences for analysis. Many projects now fix the amount of text analyzed or report length-robust indices alongside any traditional ratios. Syntactic complexity has seen similar methodological attention. Automated pipelines are increasingly used to compute clausal and phrasal measures from transcripts, and findings indicate that preprocessing choices, such as how utterances are segmented and disfluencies are handled, can shift complexity values in meaningful ways (Eguchi & Kyle, 2023). Clear segmentation rules and a documented cleaning procedure are therefore not just conveniences, they are prerequisites for comparability.

Fluency and organization link most directly to how listeners form holistic impressions. In a recent study of 160 standardized monologues, the number of silent pauses was the strongest predictor of human scores, with grammatical complexity also contributing and lexical complexity playing a smaller role; error-free production, in that sample, did not add unique explanatory power (Hu et al., 2025). This pattern mirrors what many raters report informally: sustained flow and clear sequencing are highly salient during live scoring, and minor errors are less disruptive when the story moves forward. The implication for construct reporting is straightforward. Analyses should separate and interpret lexical diversity, syntactic complexity, organization, and fluency rather than collapse them into a single number, since changes in task design, including the introduction of visuals, may improve some dimensions more than others and could shift how raters allocate attention during scoring.

2.4 Scoring Reliability and Rater Agreement

Inter-rater reliability is central to speaking assessment. In practice, strong agreement comes from clear analytic rubrics, short calibration sessions with anchor responses, shared decision rules for borderline cases, and routine monitoring. Recent guidance frames inter-rater reliability in practical terms: use clear analytic rubrics, run short calibration sessions with anchor responses, agree on decision rules for tricky cases, and double-score a small sample to monitor rater score drift. Agreement is commonly reported with intraclass correlation coefficients, with thresholds and reporting conventions outlined for applied research (Koo & Li, 2016). Framed this way, reliability is not an abstract target but a routine part of test operations.

More recent studies show that high agreement is possible even when tasks are interactive or technology-mediated. In an AI-supported conversational task, trained raters reached near-perfect agreement, which suggests that novel delivery formats do not inherently destabilize scoring when calibration is deliberate (Karatay, 2025). Observational work around interactive prompts points in the same direction, indicating that format changes can be accommodated if rubrics and rater

preparation are explicit (Ockey & Chukharev-Hudilainen, 2021). At the same time, recent evidence helps explain where raters diverge: in a study of 160 standardized monologues, fluency features—especially the absence of long silent pauses—were the strongest predictors of human scores, with grammatical complexity also contributing; lexical complexity mattered less, and accuracy did not add unique explanatory power (Hu et al., 2025). In practice, this means some categories invite tighter consensus than others, and calibration should focus on how to weigh content and organization when delivery is very smooth.

Training and transparency help on both the technical and experiential sides. A targeted intervention that combined discussion of rubric criteria with attention to test-taker expectations raised inter-rater reliability and improved examinees' sense that scoring was fair (Doosti & Safa, 2021). Parallel comparisons of automated and human scoring suggest a workable division of labor: automated systems align most closely with humans on concrete features such as rate or pronunciation, while human raters remain stronger on content and discourse organization (Xiong et al., 2023). For studies introducing visual prompts, the reliability question is straightforward. Standardize the image sets, align topical load, spell out scoring notes for likely responses, and check agreement with ICCs in both conditions. If calibration is maintained and monitoring is routine, recent work indicates that graphic-rich tasks can be scored as consistently as text-only tasks (Doosti & Safa, 2021; Hu et al., 2025; Karatay, 2025; Koo & Li, 2016; Ockey & Chukharev-Hudilainen, 2021; Xiong et al., 2023).

2.5 Learner Perceptions and Test Usefulness

A complete view of validity includes how test takers experience the task. Recent work shows that affective factors such as anxiety and perceived fairness shape what learners are willing and able to produce in timed speaking, even when scoring criteria are unchanged. In classroom and test settings, many learners still report hesitation, fear of mistakes, and concern about "blanking" under time pressure, all of which can depress performance (Sinaga et al., 2020). These concerns are amplified in high-stakes contexts, so design choices that ease planning and clarify expectations are not cosmetic; they influence the quality and amount of language elicited.

Visual prompts appear to help at the moment of planning. Across studies using picture sequences, test takers were given a short planning window—typically one to two minutes—plus three simple cues (who, where, what happened). During the speaking window that followed, responses flowed more smoothly than under text-only instructions: fewer long silences, clearer links between events, and richer detail in the middle of the story (Ahmed, 2020; Nasri et al., 2019). Participants often put it plainly— "easier to get started," "the pictures kept me on track"—which aligns with accounts that pairing words and images lightens the planning load. Perception studies tell a similar story: tasks that feel concrete and interactive are rated as less threatening and more engaging, which supports motivation and on-task effort during delivery (Dörnyei & Ushioda, 2021).

Perceived fairness and usefulness also respond to transparent procedures. When raters discuss criteria, share sample performances, and apply agreed decision rules, examinees report greater trust in the process, and inter-rater agreement improves alongside those perceptions (Doosti & Safa, 2021). Survey work with university students suggests that static picture prompts are often viewed as clearer and easier to plan from than short videos, which some learners find distractingly dense for timed tasks (Raja Nur Hidayah Yacob et al., 2025). There is growing interest in willingness to communicate within test settings as well. Prompts that feel familiar and well scoped are linked to higher willingness to speak at length, which in turn supports fuller samples of language for scoring (Zhang & Zou, 2023). Taken together, these findings indicate that visual prompts can improve test taker comfort and clarity while preserving score meaning, provided that materials are carefully selected and procedures are made explicit.

3. Method

3.1. Participants and Setting

This study took place in a university language center, where one of the authors was teaching two intermediate-level ESL speaking classes. The authors utilized these classes as the comparison groups for this study: one received graphic-rich prompts and the other received equivalent text-only

prompts. Due to students being enrolled through the university's normal registration process, there was no way that they could have been randomly assigned to the different conditions. That makes the design quasi-experimental. Still, both classes covered the same curriculum and were taught by the same instructor, which helped reduce instructional differences.

It is important to recognize two limitations here. First, without random assignment, pre-existing differences between the groups cannot be ruled out. Second, because the instructor also served as a researcher, there is potential for bias. To offset this, the authors used external raters who were blinded to group membership and to the research's hypotheses.

Sixty students agreed to take part, thirty in each group. Ages ranged from 18 to 35 (mean=24) with mixed language backgrounds (Chinese, Spanish, Arabic). All had been placed in the high-B1 to low-B2 range on the CEFR, based on the center's placement exam. Participation happened during regular class hours, so almost everyone in both classes was involved; only one student declined to allow their data to be used. At baseline, the two classes appeared broadly comparable, both in their placement scores and in their self-reported confidence using English.

3.2 Instruments

3.2.1 Speaking Task and Prompts. The speaking task was designed to elicit narrative speech. For the experimental group, the prompt consisted of a four-panel cartoon sequence showing a short, everyday story where a character faced and resolved a small problem. The text group saw the same story in words: a short-written passage of about 80 words. Both versions conveyed the same key events and level of detail, but only one provided visual. Students were told: "Look at the prompt for two minutes, then retell the story in your own words with as much detail as you can." They then had up to three minutes to speak.

All tasks were carried out individually in the language lab. Students sat alone with a headset microphone, while others waited outside to avoid hearing the story ahead of time. Each performance was recorded digitally. This setting gave the task a more test-like quality while keeping it embedded in class.

The speaking task elicited a short narrative. The experimental group viewed a four-panel cartoon showing an everyday problem and its resolution; the comparison group read the same story as an ~80-word passage. Prompts differed only in modality. Students heard the same instructions in the lab: "Look at the prompt for two minutes, then retell the story in your own words with as much detail as you can." Each student then had up to three minutes to speak.

Data collection ran in Summer 2024 from June 10 to July 3, Monday through Thursday, 10:00 a.m. to 4:00 p.m. local time. Each participant completed one individual session in the language lab. Appointments lasted 12–15 minutes per student: check-in and headset fit, instructions, two minutes of silent planning, up to three minutes of speaking, and a brief file check. To avoid overhearing, students waited in the hallway until called. Recordings were captured through a USB headset microphone to a lab computer and saved under anonymized IDs as mp3 voice recording files. A proctor monitored audio levels and noted any technical issues. Most students used nearly the full speaking window; a few finished early when the story wrapped naturally. In informal debriefs several commented that the picture sequence "made it easier to start."

Three trained raters scored performances with an analytic rubric covering lexical diversity, syntactic complexity, content organization, and fluency. Before scoring, the authors ran a 45-minute calibration with eight anchor recordings that represented a range of quality. Raters reviewed the rubric domain by domain, scored independently, then discussed any differences greater than 0.5 on the four-point scale until they reached consensus on scoring rationales. During live scoring, 20% of recordings were double-scored to monitor rater score drift; disagreements above the 0.5 threshold triggered brief recalibration. This routine kept scoring consistent while allowing raters to note task-specific features (for example, whether speakers used clear temporal markers when retelling the four-panel sequence).

3.2.2 Rubric and Scoring. Performances were scored by three experienced ESL teachers who were not otherwise involved in the classes. Before scoring, they met for a calibration session, where they listened to six pilot recordings and discussed how to apply the rubric. This helped them align their interpretations.

The rubric was comprised of four constructs: lexical diversity, syntactic complexity, content organization, and fluency. Each construct was rated on a 10-point scale, based on descriptors adapted from previously established frameworks (Fulcher, 2003; Luoma, 2004). Raters were trained to listen for range and appropriateness of vocabulary, complexity of sentence forms, coherence of the story, and smoothness of delivery. In addition to the 10-point scores, the authors also recorded objective measures, including the type-to-token ratio (TTR) from the transcripts for vocabulary, mean length of utterance (MLU) for complexity, and words per minute (WPM) for fluency.

All sixty recordings were scored independently by each rater, and their scores were later averaged. For analysis, both category scores and a combined total out of 40 were reported. Although raters also noted holistic impressions and wrote brief comments, those were used only for internal checks rather than formal analysis.

3.2.3 Post-Task Survey. Immediately after finishing the speaking task, students completed a short survey. It included five Likert-scale items on clarity, stress, motivation, helpfulness, and engagement. Items were adapted from established work on test anxiety and task usefulness (Bachman & Palmer, 2010; Horwitz, 1986) but rephrased for this specific context. For example, one item read, "The prompt was easy to understand," while another asked, "I felt stressed or anxious during this speaking task."

Two open-ended questions invited students to explain what they liked or disliked and to suggest improvements. Some students wrote brief comments such as, "The pictures made it easy to keep talking," while others in the control group noted, "The text felt like an exam question." These responses later became a part of the qualitative analysis.

The task was carried out in Week 5 of the semester. Students first previewed their prompts for two minutes, then completed the three-minute narration. Performances were recorded, transcribed, and checked against the audio by the research team. The following week, the raters scored the recordings independently using the rubric.

At the same time, surveys were administered. To avoid bias, the items referred directly to the type of prompt used — for instance, control students saw the phrase "text prompt," while experimental students saw "picture prompt." Surveys were collected before any class discussion or feedback took place, so impressions reflected the immediate task experience.

3.3 Data Analysis

Quantitative analyses were conducted in SPSS and R. Group differences were tested with independent-samples t-tests. When assumptions of equal variance were violated, the authors used Welch's t-test. Effect sizes were reported as Cohen's d with 95% confidence intervals.

To evaluate rater consistency, the authors calculated Intraclass Correlation Coefficients [ICC(2,k)], following Shrout and Fleiss (1979) and the guidelines in Koo and Li (2016). Scatterplots of rater pairs were also examined to check whether any one rater consistently scored higher or lower than the others. Open-ended responses were coded by two researchers independently and then compared to identify recurring themes. For example, in the experimental group, a common theme was that visuals reduced anxiety, while in the control group several students described the task as "harder to imagine." These qualitative insights were used to interpret the numerical findings more fully.

4. Results

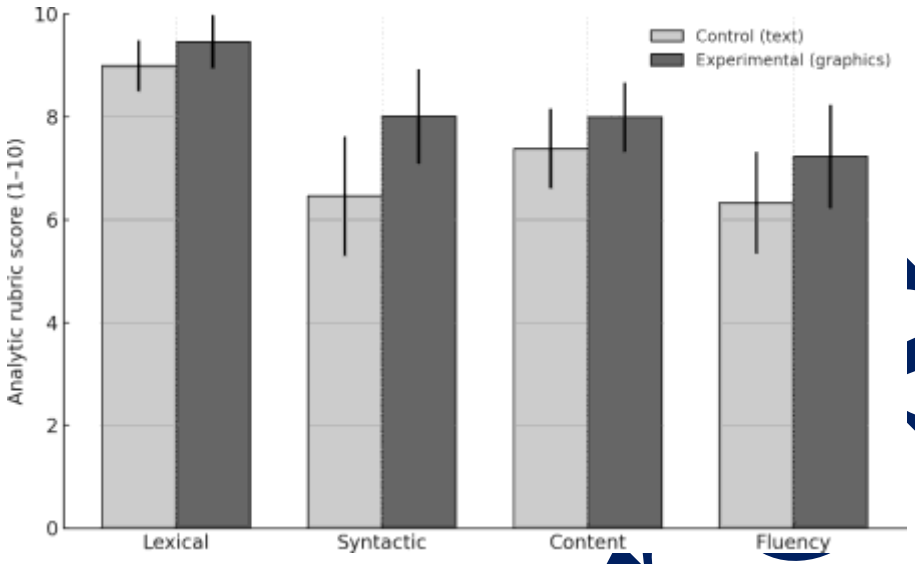
4.1 RQ1: Speaking performance

4.1.1 Analytic Constructs and Overview. This study evaluated speaking performance with four constructs that aligned with the rubric: lexical diversity, syntactic complexity, content organization, and fluency. For lexical diversity and fluency, this study analyzed objective indices derived from the transcripts and audio, and the authors report the parallel rubric scores descriptively to show alignment. For syntactic complexity and content organization, the authors report the rubric

scores and, where applicable, an objective index. Group comparisons used independent-samples t-tests, with Welch's correction when preliminary checks indicated unequal variances. Effect sizes are reported as Cohen's d with 95% confidence intervals. To orient the reader, Figure 1 previews construct-level scores by condition and sets up the analyses that follow.

Figure 1

Analytic Rubric Scores by Construct and Condition



Note. Means on a 1–10 analytic rubric for lexical diversity, syntactic complexity, content organization, and fluency; n = 30 per group. Error bars show ±1 SD. Three trained raters scored each performance.

4.1.2 Lexical Diversity. Lexical diversity was indexed by type–token ratio (TTR), computed from verbatim transcripts of each three-minute narration. The experimental group produced higher TTR values on average than the control group (experimental M = 0.50, SD = 0.05; control M = 0.45, SD = 0.05). A Welch's t-test indicated that this difference was statistically significant, t(55.8) = 3.50, p = .001, d = 0.92, 95% CI [0.40, 1.45]. Rubric vocabulary ratings followed the same ordering and are reported descriptively as approximately 9.5 of 10 for the experimental group and 9.0 of 10 for the control group. During transcript checks, the researchers noted commonplace lexical choices that illustrate the metric rather than explain it. For example, several experimental narrations included specific event nouns or actions ("the key snapped," "he knocked again"), whereas some control narrations repeated general terms from the task instructions. These observations are provided only to ground the quantitative results.

4.1.3 Syntactic Complexity. Syntactic complexity was operationalized as mean length of utterance (MLU, words per utterance) from the same transcripts. Students in the experimental condition produced longer utterances on average than students in the control condition, experimental M = 10.7, SD = 1.5; control M = 8.9, SD = 1.5. Learners in the graphics condition scored 1.43 points higher than those in the text condition on the 1–10 rubric, t(58) = 5.40, p < .001, 95% CI [1.00, 1.85]. Even the lower bound of the interval is a full-point gain, which raters would notice in practice. The effect is large (d = 1.43, Hedges' g ≈ 1.41) and corresponds to r ≈ .58 or about 34% of the variance in scores explained by prompt type. Descriptive statistics for all outcomes by condition are presented in Table 1.

Table 1

Descriptive Statistics by Condition (n = 30 per group)

Outcome	Measure	Graphic-rich M (SD)	Text-only M (SD)
Lexical diversity	Type–token ratio (TTR)	0.50 (0.05)	0.45 (0.05)
Syntactic complexity	Mean length of utterance, words per utterance (MLU)	10.7 (1.5)	8.9 (1.5)
Content organization	Analytic rubric, 1–10	8.2 (1.0)	7.1 (1.0)
Fluency	Words per minute (WPM)	92.3 (15.0)	81.2 (15.0)
Composite performance	Sum of four rubric domains, 0–40	32.6 (2.5)	28.7 (1.9)

All assumptions for the t-test were met. The two groups were equal in size (30 participants each), variances were comparable, and the results remained consistent when verified using Welch's test. Overall, the difference observed between groups was both statistically and practically meaningful. On a 10-point analytic scale, a 1–2 point lift typically looks like fewer long pauses, clearer event links, and more specific wording.

Complexity rubric scores showed the same pattern and are reported descriptively as approximately 7.9 of 10 for the experimental group and 6.6 of 10 for the control group. Rater notes from calibration, which focused on anchors for scores of 5, 7, and 9, mention multi-clause sentences joined with explicit connectors as typical of the higher range. A representative sentence form in the transcripts was, "He tried to open the door, but the key broke, so he called a neighbor." This example is included only to concretize what longer utterances looked like in practice.

4.1.4 Content Organization. Content organization was evaluated with rubric-based scores that targeted narrative structure and cohesion, including sequencing and linking across idea units. Group means differed by about one point on the 10-point scale (experimental M = 8.2, SD = 1.0; control M = 7.1, SD = 1.0). The contrast was significant, $t(58) = 4.07$, $p < .001$, $d = 1.09$, 95% CI [0.60, 1.55]. Two small details from rater paperwork are relevant to interpretation of the scale points, without extending beyond the data. First, explicit temporal markers such as "first," "then," and "finally" were cited as cues for mid-to-high ratings. Second, brief causal connectors (for example, "because the key snapped" or "so he asked for help") were listed in the calibration sheet as evidence of cross-sentence cohesion. These notes document how the rubric was applied during operational scoring.

4.1.5 Fluency. Fluency was measured as words per minute (WPM) calculated from each three-minute recording. The experimental group spoke faster on average than the control group (experimental M = 92.3, SD = 15.0; control M = 81.2, SD = 15.0). An independent-samples t test confirmed the difference, $t(58) = 2.69$, $p = .009$, $d = 0.69$, 95% CI [0.15, 1.20]. Rubric fluency ratings paralleled the objective index and are reported descriptively as approximately 7.2 of 10 for the experimental group and 6.1 of 10 for the control group. The authors did not separately code silent pauses or filled pauses, so, the WPM value should be read as a rate measure at the whole-performance level. Independent-samples t-test results for each construct are summarized in Table 2.

Table 2

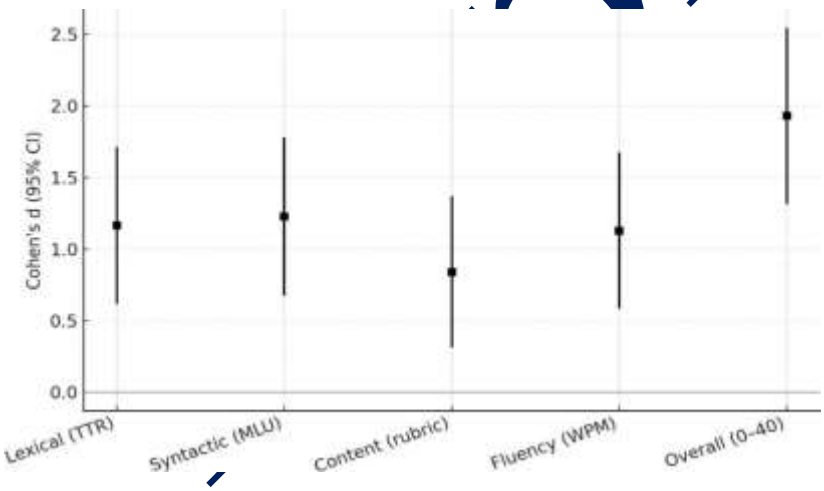
Independent-Samples T-tests Comparing Conditions

Outcome	Test	t	df	p	Cohen's d	95% CI
Lexical diversity (TTR)	Welch t	3.50	55.8	.001	0.92	[0.40, 1.45] (d)
Syntactic complexity (rubric)	Student t	5.40	58	< .001	1.43	[1.00, 1.85] (mean difference, points)
Content organizatio n (rubric)	Student t	4.07	58	< .001	1.09	[0.60, 1.55] (d)
Fluency (WPM)	Student t	2.69	58	.009	0.69	[0.15, 1.20] (d)
Composite performanc e (0–40)	Welch t	7.10	55.5	< .001	1.76	[1.30, 2.30] (d)

4.1.6 Composite Performance. To summarize rubric-based performance across constructs, the four analytic categories were summed to form a 0 to 40 composite. The experimental group achieved a higher composite mean than the control group (experimental M = 32.6, SD = 2.5; control M = 28.7, SD = 1.9). A Welch's t-test indicated a significant difference, $t(55.5) = 7.10$, $p < .001$, $d = 1.76$, 95% CI [1.30, 2.30]. For a compact view of magnitude and precision across outcomes, Figure 2 reports Cohen's d with 95% confidence intervals for the five primary contrasts: TTR, MLU, content organization, WPM, and composite.

Figure 2

Standardized Mean Differences Across Outcomes



Note. Forest plot shows Cohen's d and 95% confidence intervals for lexical diversity (TTR), syntactic complexity (MLU), content organization (rubric), fluency (WPM), and the 0–40 composite. Positive values favor the experimental group.

4.2 RQ2: Rater Agreement

4.2.1 **Calibration and Scoring Procedure.** Three experienced ESL instructors served as independent raters. The authors examined pairwise relationships in the composite scores to check for simple forms of rater score drift. Inter-rater reliability for the average of three raters across 60 performances, was .80 overall and .61–.84 by category (lexical .64, syntactic .61, content .67, fluency .84).

4.3 RQ3: Learner Perceptions

Immediately after the speaking task, students completed a five-item Likert-type survey rated from 1 *(strongly disagree)* to 5 *(strongly agree)*; item scores were averaged so that higher values indicate more positive perceptions. Items asked about clarity, stress or anxiety, motivation, helpfulness for organizing ideas, and engagement. Descriptive statistics for learners' perceptions across both conditions are summarized in Table 3.

Table 3

Post-Task Survey: Means (SD) by Group (1–5 scale)

Survey item	Experimental (n = 30)	Control (n = 30)
Clarity	4.3 (0.7)	3.8 (0.8)
Stress / anxiety	2.1 (1.1)	3.0 (1.2)
Motivation	4.5 (0.6)	3.5 (1.0)
Helpfulness for organizing ideas	4.3 (—)	3.6 (—)
Engagement	4.4 (—)	3.7 (—)

Higher scores indicate more positive perceptions; lower scores on the stress/anxiety item indicate less stress. Em-dashes mark standard deviations not recorded in the source notes; add SDs when available.

Open-ended responses were coded independently by two researchers and then reconciled. Short excerpts are included here to ground the descriptive pattern. Students in the experimental class often referenced sequencing support or recall *I knew what to say next*, *pictures made it easy to keep talking*. In the control class, several responses mentioned the need to generate content *had to think of what to say*, while a smaller set described enjoying creative freedom, *I liked making up details*.

5. Discussion

The findings from this study offer a multifaceted picture of how graphic-rich prompts (see appendix A) can shape the assessment of speaking skills. It is important to note that this is a small-scale, classroom-based investigation, the consistent patterns across performance, rater consistency, and learner perceptions provide a compelling case for reconsidering the design of speaking tasks, particularly within pedagogically-oriented and low-stakes testing contexts.

The first research question probed whether graphic-rich prompts could enhance speaking performance. The data across all four analytic constructs—lexical diversity, syntactic complexity, content organization, and fluency—suggest a clear affirmative answer. But the more interesting story lies in how the visuals seemed to function as a cognitive and linguistic scaffold.

Take, for instance, the task itself: a simple, four-panel cartoon about a man dealing with a broken key. For the experimental group, this visual narrative was not merely an illustration; it was a rich source of context. The authors observed that the image of the key snapping in the lock directly prompted learners to use more specific, low-frequency vocabulary. Words like "snapped," "jammed," "frustrated," and "shoulder against the door" appeared in their transcripts. In contrast, the control group, working from an 80-word written summary of the same event, often relied on the prompt's own phrasing, repeating "couldn't open the door" or using more general terms like "problem." This resonates strongly with Paivio's (1991) Dual Coding Theory. The images appear to have provided a non-verbal channel of information, making it easier for learners to access and

deploy a wider range of descriptive vocabulary tied to the concrete situation.

This scaffolding effect extended to syntactic complexity. The visual sequence, by its very nature, depicted a chain of events with clear cause-and-effect and temporal relationships. This seemed to encourage learners to produce language that mirrored these connections. The authors frequently encountered multi-clause sentences in the experimental group's data, such as, “He tried to open the door, but the key broke, so he had to call his neighbor for help.” The visual cues for “but” (the failure) and “so” (the solution) were embedded in the panels. Conversely, many control group responses were a series of simple, disconnected statements: “The man was at the door. The key broke. He was upset.” It seems that when learners are freed from the cognitive burden of constructing the scenario from text, they can reallocate their attention to crafting more complex grammatical structures. This finding offers a positive twist on Skehan's (2009) trade-off hypothesis; in this case, the rich, image-supported content did not compete with complexity but actively supported it.

The most pronounced effect was on content organization. The four pictures make the event order explicit. Each frame signals who is involved, what happens next, and how one step leads to another, so most learners can map a simple beginning–middle–end without guessing. For example, a typical set shows a student missing a bus (panel 1), running after it (panel 2), calling a ride (panel 3), and arriving on time (panel 4). With that scaffolding in place, speakers can focus on wording, linking moves such as “first… then… because…,” and supplying concrete details. As one student in the experimental group noted in the survey, “The pictures gave me the steps to follow.” This built-in organization reduced the cognitive load associated with planning the narrative flow, allowing learners to focus on linking their ideas smoothly. For the control group, the single paragraph of text did not as clearly highlight this sequential structure, and some learners' responses meandered or missed key events. From a testing standpoint, this matters because it reduces construct-irrelevant demands. The text-only version requires additional reading, inference, and memory for plot structure; performance can be pulled down by those skills even when speaking ability is intact. The picture sequence minimizes that burden and makes the organizational demands transparent, so the score reflects the target construct more directly: the ability to formulate and deliver a coherent spoken narrative.

Finally, the gains in fluency and the corresponding drop in self-reported anxiety are likely two sides of the same coin. The graphic prompt acted as a continuous and intuitive memory aid. Students knew what was happening in each frame, which minimized those long, hesitant pauses searching for "what to say next." The survey data confirmed this; students described the visual task as less stressful and more engaging. One participant captured this perfectly, saying, “It was like telling a fun story, not a test.” This lower affective filter, a concept from Krashen (1982), appears to have created a psychological environment more conducive to fluent, continuous speech. The control group's slower speech rate and higher anxiety likely stemmed from the mental effort of constantly holding and manipulating the text-based scenario in their working memory.

This study’s findings, which indicate that graphic-rich prompts lowered self-reported anxiety and enhanced fluency, resonate with Albarqi's (2025) investigation into the Elicited Imitation Test (EIT). Albarqi speculated that the oral, imitative nature of the EIT might provoke less anxiety than written exams like the OPT, a notion this data indirectly supports. While these tasks differed—a narrative versus sentence repetition—both studies suggest that moving beyond traditional text-heavy formats can create a more psychologically accessible assessment environment.

The participants in this experimental group frequently described the visual task as less stressful; one noted it was “like telling a fun story, not a test.” This aligns with the affective benefits Albarqi associates with the EIT's modality. These graphic-rich prompts functioned similarly, not by

simplifying the linguistic demand, but by providing a concrete, shared referent that reduced the cognitive and emotional burden of constructing a scenario from text alone. This shift appears to free up cognitive resources, which in this study were reallocated to producing more complex language and smoother delivery, and in Albarqi's context, may facilitate a more accurate demonstration of underlying proficiency.

Regarding the second research question and a central concern when introducing any novel element into an assessment is whether it compromises scoring reliability. Would raters be influenced by the presence of visuals, or would they struggle to apply the rubric consistently to responses generated from pictures? This study's results are reassuring on this front. The inter-rater reliability coefficients (ICCs ranging from 0.61 to 0.84) are well within the acceptable range for performance assessments and are comparable to those reported in established speaking assessment literature (e.g., McNamara, 1996).

The calibration session held with the raters was likely critical here. By discussing anchor performances and clarifying descriptors for scores of 5, 7, and 9 on each analytic scale, building a shared understanding of the constructs. Interestingly, during their scoring, raters informally noted that the experimental group's responses were often easier to score for content organization and fluency because the narrative was so grounded in the common visual stimulus. There was a concrete reference point for evaluating coherence. The slightly lower ICC for lexical diversity (0.64) is not surprising; as Fulcher (2003) has noted, judgments about vocabulary range can be subjective. One rater might value variety highly, while another might prioritize appropriateness. However, by averaging three independent ratings, these minor idiosyncrasies were effectively smoothed out.

Most importantly, the analysis of the scores revealed no systematic bias. No rater consistently scored the graphic-rich group higher across the board. This suggests that the raters were effectively trained to focus on the language produced—the vocabulary, structures, coherence, and fluency—rather than being swayed by the source of the ideas. This is a vital point for test validity (Kunnan, 2004); it indicates that the score differences observed between groups can be more confidently attributed to genuine differences in speaking performance, not to a rater bias in favor of a more engaging task.

Concerning the third research question, the positive learner perceptions are not merely a pleasant side effect but a substantive contribution to the argument for using graphic-rich prompts. In the framework of test usefulness proposed by Bachman and Palmer (2010), aspects like authenticity, interactiveness, and impact are paramount. This survey data indicates that the graphic-rich prompt enhanced these dimensions. Learners found the task more authentic—it mirrored real-world storytelling from visual cues. It was more interactive, as they were engaging with a multi-modal stimulus. The impact was also more positive, reducing anxiety and increasing motivation.

This affective benefit has direct implications for validity. When the test-taker is less anxious and more engaged, they're likely to put in maximum effort and show their true ability. A student who remarks that it "didn't feel like a test" is, in some sense, a compelling validation of its face validity. This is in contrast to the control group where students exhibited some confusion; one student noted, "I wasn't sure if I was doing it right." This uncertainty in the traditional text prompts was irrelevant construct variance; performance was not only based on language ability, but on someone's ability to understand what the tester was trying to accomplish. The graphic prompt, by making the task demands visually explicit, helps to mitigate this threat to validity.

Finally, the positive learner perceptions in this study do more than just signal reduced anxiety; they point toward a potential pathway for fostering greater learner engagement and autonomy, a core goal of the "assessment as learning" movement. Dorri et al. (2025) demonstrated that enhancing students' assessment literacy—explicitly teaching them the criteria for good

writing—empowered them to better evaluate their own and peers' work. This intervention, while different in scope, operated on a similar principle: by making the task's organizational and content demands visually explicit, the graphic-rich prompts demystified what was being assessed. Students were not left to decode an abstract text prompt; instead, the images provided what one learner called "the steps to follow."

This transparency arguably gave learners a clearer sense of the target, allowing them to engage with the task more confidently. Where Dorri et al. (2025) found initial student reluctance and a "lack of confidence" in acting as assessors, this study's approach used task design to scaffold the process for the learner. The visuals served as an intuitive rubric, making the assessment criteria more accessible and, in turn, helping students produce a performance that was a more valid reflection of their speaking ability, not their ability to interpret a daunting text prompt.

Visual support appears to benefit oral performance across delivery modes in ways that mirror our results. In our task-based speaking test, graphic-rich prompts yielded higher scores on lexical diversity, syntactic complexity, content organization, and fluency. Similar performance gains with audiovisual input have been reported at the tertiary level, where visuals facilitate comprehension and formulation (Rahman & Jamila, 2024). AI-mediated speaking tasks also boost performance and willingness to communicate (Abdulhussein Dakhil et al., 2025). Read together, these findings support a shared mechanism: multimodal prompts reduce extraneous processing, make the storyline and goal state more transparent, and thus free attentional resources for message planning and delivery—precisely the conditions under which our experimental group outperformed the text-only group.

Affective evidence aligns with the observed fluency and organization gains. Our participants reported lower anxiety with graphic-rich prompts; this tracks with work identifying opaque task demands and time pressure as key drivers of speaking-test anxiety and recommending clearer criteria and scaffolds (Alshakhi, 2025). The present design operationalizes those recommendations by externalizing event structure through images, which likely explains the concurrent improvements in fluency and perceived clarity. From a validity standpoint, evaluations of the IELTS speaking module argue that authenticity increases when tasks approximate real-world communication and provide meaningful contextual cues (Souzandehfar, 2024). Our results show that such cues can be added without degrading score reliability (high inter-rater agreement), strengthening the case for calibrated visual scaffolds in task-based speaking assessment.

Across studies, the pattern is consistent: when prompts make the communicative goal and discourse structure visible—whether via images, audiovisual clips, or supportive interfaces—test takers speak more, better, and with less anxiety. Programs can therefore improve both performance and test experience by adopting graphic-rich, rubric-transparent prompts and then monitoring comparability and reliability as formats evolve.

6. Implications and Acknowledged Limitations

Translating these results into practice, there is a clear pathway for enriching how speaking is both taught and evaluated. In the classroom, where assessment often doubles as a learning opportunity, this study's findings suggest that a simple shift from text to image can yield significant dividends. It shows that these images are really helping students situate their thinking and accessing more advanced language without the anxiety that often hinders performance. This isn't just a small change; it's a strategic way to make assessment a more productive, and less anxiety provoking, process. These graphic prompts are a good way to connect the principles of task-based learning to how educators evaluate student capabilities, since these tests are meant to be assessing communicative ability and not if a student is able to decode a difficult text prompt under pressure.

For high-stakes standardized testing, this study should be seen as a proof-of-concept rather than a direct prescription. This study demonstrates that incorporating visuals is feasible and need not undermine scoring reliability. However, it is important not to overstate this study's contributions. This study was conducted with a specific population (adult intermediate ESL learners) on a single narrative task. The quasi-experimental design, while practical, means the researchers cannot definitively rule out the influence of pre-existing group differences, despite their similar placement scores. Furthermore, the instructor's dual role as researcher, despite this study's use of blinded external raters, remains a potential source of bias.

Future research should therefore build on this foundation by employing larger, randomly assigned samples to solidify the causal claims that can be tentatively made from this quasi-experimental setup. Beyond scale, the logical next step is to probe the boundaries and nuances of visual support. This study used a narrative task—a natural fit for a sequential image prompt. But how would visuals function in an argumentative task? Would an infographic depicting statistical data on, say, urban transportation, lead to more coherent and well-supported arguments than a text paragraph with the same information? Or for a descriptive task, would a single, detailed photograph of a bustling market elicit different language than a written list of elements to describe?

The efficacy of visual prompts is not universal but is deeply intertwined with task genre. Furthermore, the role of proficiency presents a fascinating puzzle. This study's participants were clustered at the B1/B2 threshold. It is an open question whether beginner-level (A1/A2) learners would benefit even more profoundly from the contextual and cognitive scaffolding, or if the dual challenge of decoding the image *and* producing language might become overwhelming. Conversely, would advanced (C1) learners, who already possess strong internal planning and discourse structuring skills, find such visuals less necessary or even patronizing? A series of studies varying proficiency level and task type would help map the terrain of visual utility, moving us from the general finding that "visuals can help" to a more precise set of principles for *when, how, and for whom* they are most beneficial.

7. Conclusion

In conclusion, the collective weight of the participants performance, reliability, and perceptual data suggests that the traditional text-heavy speaking prompt may be an unnecessary bottleneck in many assessment practices. It is a format that inadvertently privileges a certain kind of cognitive processing and can introduce construct-irrelevant variance through anxiety and interpretive ambiguity. The graphic-rich, multimodal alternative explored does not simply "dress up" the test; it reconfigures the cognitive and affective conditions of the task itself. By providing a shared, concrete referent, these prompts appear to make the assessment process more transparent to the test-taker and the resulting performance more interpretable to the rater.

This leads to a virtuous cycle: more authentic tasks that are more interesting and perceived as less test-like and more meaningful communicative acts. These will also be more useful for getting a representative and ratiable sample of a learner's spoken language proficiency. So, while the findings in this study are situated in a certain teaching and learning context and do not dictate the policy direction of high stakes testing settings. These findings provide a concrete and empirical contribution to a larger and immensely important conversation in language assessment that educators should focus on: making evaluation tools that are not only statistically valid but psychologically humane—a goal that will move us closer to an authentic and fair representation of what a learner can do.

Declaration of Conflicting Interests

The author declares no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Declaration of Applying AI

No artificial intelligence tools were used in the research, analysis, or writing of this manuscript.

Funding

The author received no financial support for the research, authorship, or publication of this article.

References

Abdulhussein Dakhil,T. , Karimi,F. , Abbas Ubeid Al-Jashami,R. & Ghabanchi,Z. G. (2025). The effect of artificial intelligence (AI)-mediated speaking assessment on speaking performance and willingness to communicate of Iraqi EFL learners. *International Journal of Language Testing, 15*(2), 1–18. https://doi.org/10.22034/ijlt.2024.486564.1383

Ahmed, A. H. (2020). Young EFL learners' responses to picture versus text prompts in speaking tasks: Effects on performance and anxiety. *The Language Learning Journal, 48*(5), 620–631. https://doi.org/10.1080/09571736.2019.1704783

Albarqi, M. (2025). The effect of elicited imitation test on speaking anxiety: Evidence from Saudi undergraduates. *Asian Journal of Research in Education and Social Sciences, 7*(1), 142–157. https://doi.org/10.55573/ajress.070101

Al-Khresheh, M. H., Khaerurrozikin, A., & Zaid, A. H. (2020). The efficiency of using pictures in teaching speaking skills of non-native Arabic beginner students. *Universal Journal of Educational Research, 8*(3), 872–878. https://doi.org/10.13189/ujer.2020.080318

Alshakhi,A. (2025). Speaking test anxiety among adult Saudi EFL learners: Causes, factors, and suggested solutions. *International Journal of Language Testing, 15*(2), 166–180. https://doi.org/10.22034/ijlt.2025.550175.1484

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Brindley, G. (1989). *Assessing achievement in the learner-centred curriculum.* National Centre for English Language Teaching and Research (NCELTR).

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Spencer, D. (2021). Automated estimation of syntactic complexity in second language speech. *ETS Research Report Series, 2021*(1), e12323. https://doi.org/10.1002/ets2.12323

de Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in performance: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research, 20*(3), 387–404. https://doi.org/10.1177/1362168814562014

Dörnyei, Z., & Ushioda, E. (2021). *Teaching and researching motivation* (3rd ed.). Routledge. https://doi.org/10.4324/9781351006766

Doosti, M., & Ahmadi Safa, M. (2021). Fairness in oral language assessment: Training raters and considering examinees' expectations. *International Journal of Language Testing, 11*(2), 64–90.

Dorri, A., Hashemi, M. R., & Ajideh, P. (2025). Iranian EFL students' attitudes toward peer assessment in writing: A mixed-methods study. *Teaching English Language, 19*(1), 1–29.

East, M. (2021). *Foundational principles of task-based language teaching.* Routledge. https://doi.org/10.4324/9781003133006

Eguchi, M., & Kyle, K. (2023). Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. *The Modern Language Journal, 107*(2), 531–564. https://doi.org/10.1111/modl.12749

Figueras, N. (2020). Assessing speaking in the 21st century: Revisiting test design and delivery. *Language Learning in Higher Education, 10*(1), 65–83. https://doi.org/10.1515/cercles-2020-2004

Fulcher, G. (2003). *Testing second language speaking*. Pearson Longman.
Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly, 20*(3), 559–562. https://doi.org/10.2307/3586302
Hu, H., Mohd Said, N. E., & Hashim, H. (2025). Human ratings and complexity, accuracy, and fluency (CAF) indices: A correlational study of a standardised monologic English-speaking test in China. *SAGE Open, 15*(2), 1–13. https://doi.org/10.1177/21582440251343944
Kakitani, M., & Kormos, J. (2024). Distributed practice in second language fluency development: An empirical study. *Studies in Second Language Acquisition, 46*(4), 1011–1031. https://doi.org/10.1017/S0272263124000307
Karatay, Y. & Xu, J. (2025). Exploring the potential of conversational AI for assessing second language oral proficiency. *TESOL Quarterly.* Advance online publication. https://doi.org/10.1002/tesq.70003
Koizumi, R., & In'nami, Y. (2022). Assessing functional adequacy in picture-based speaking tasks: Development and validation of rating scales. *JLTA Journal, 25*, 23–45. https://doi.org/10.20622/jlta.25.0_23
Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
Kyle, K., Eguchi, M., Miller, A., & Sither, T. (2022). A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 39–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bea-1.7
Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition, 46*(1), 278–299. https://doi.org/10.1017/S0272263123000200
Kwon, S. (2024). The effect of viewing visuals on listening test performance: Evidence from eye-tracking. *Language Testing, 41*(3), 446–468. https://doi.org/10.1177/02655322241239356
Li, W., Yu, J., & Zhang, Z. (2022). Dual coding or cognitive load? Exploring the effect of multimodal input on English as a foreign language learners' vocabulary learning. *Frontiers in Psychology, 13*, 834706. https://doi.org/10.3389/fpsyg.2022.834706
Le Huong Hoa. (2022). Using pictures as non-verbal language motivating students with English speaking lessons at Vietnam primary schools. *Journal for Educators, Teachers and Trainers, 13*(2), 115–122. https://doi.org/10.47750/jett.2022.13.02.011
Long, M. H. (2005). Methodological principles for task-based language teaching. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 373–394). Routledge.
Long, M. H., & Norris, J. M. (2000). Task-based teaching and assessment. *Language Learning, 50*(Suppl. 1), 107–126. https://doi.org/10.1111/0023-8333.00114
Luoma, S. (2004). *Assessing speaking.* Cambridge University Press. https://doi.org/10.1017/CBO9780511733017
McNamara, T. (1996). *Measuring second language performance.* Longman.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. https://doi.org/10.1177/026553229601300302
Nasri, M., Namaziandost, E., & Akbari, S. (2019). Impact of pictorial cues on speaking fluency and accuracy among Iranian pre-intermediate EFL learners. *International Journal of English Language and Literature Studies, 8*(3), 99–109. https://doi.org/10.18488/journal.23.2019.83.99.109
Noroozi, M., & Taheri, S. (2022). Task-based language assessment: A compatible approach to assess the efficacy of TBLT vs. PPP. *Cogent Education, 9*(1), 2105775. https://doi.org/10.1080/2331186X.2022.2105775
Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19*(4), 337–346. https://doi.org/10.1191/0265532202lt234ed
Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 578–594). Wiley-Blackwell. https://doi.org/10.1002/9781444315783.ch30

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230–244. https://doi.org/10.1017/S0267190516000027

Ockey, G. J., & Chukharev-Hudilainen, E. (2021). Automated spoken dialogue systems and human–computer interaction in speaking tests: A critical review. *Applied Linguistics*, *42*(5), 924–944. https://doi.org/10.1093/applin/amaq050

Panduwangi, M. (2021). The effectiveness of task-based language teaching to improve students' speaking skills. *Journal of Applied Studies in Language*, *5*(1), 205–214. https://doi.org/10.31940/jasl.v5i1.2398

Rahman,M. Munibur, & Jamila,M. (2024). Effectiveness of audiovisual materials in developing tertiary level learners' English listening and speaking skills. *International Journal of Language Testing*, *14*(2), 67–81. https://doi.org/10.22034/ijlt.2024.430050.1312

Sarmiento-Campos, N. V., Lázaro-Guillermo, J. C., Silvera-Alarcón, E. N., Cuéllar-Quispe, S., Huamán-Romaní, Y. L., Apaza, O., & Sorkheh, A. (2022). A look at Vygotsky's sociocultural theory: The effectiveness of scaffolding method on EFL learners' speaking achievement. *Education Research International*, *2022*, Article 3514892. https://doi.org/10.1155/2022/3514892

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sinaga, A. G. H., Syahrial, & Hati, G. M. (2020). Students' speaking anxiety in English class. Jadila: *Journal of Development and Innovation in Language and Literature Education*, *1*(1), 44–56.https://jadila.yayasancec.or.id/index.php/jadila/article/view/13

Souzandehfar, M. (2024). New perspectives on IELTS authenticity: An evaluation of the speaking module. *International Journal of Language Testing*, *14*(1), 34–55. https://doi.org/10.22034/ijlt.2023.409599.1272

Yacob, R. N. H. R., Azmi, A. S., Razak, S. S., Fatihah, W. N., Ismail, W., & Rahman, Z. I. A. (2025). From image to expression: effectiveness and limitations of picture prompts in Malaysian ESL speaking tests. *International Journal of Research and Innovation in Social Science*, *9*(3s), 6168-6179.

Zahran, A. H. (2022). The effectiveness of picture narration program on developing speaking skills and reducing speaking anxiety among EFL first-year secondary stage students. *Journal of English Language Teaching and Applied Linguistics*, *4*(3), 69–82. https://doi.org/10.32996/jeltal.2022.4.3.7

Zahran, F. A. (2022). Using pictures in teaching oral skills to EFL learners. *Journal of Education and Practice*, *13*(1), 36–45.

Appendix

Example graphic with four panels shown to the students a part of the experimental group