

JOURNAL OF COMMUNICATION, LANGUAGE AND CULTURE

Evaluating Creativity and Accuracy in Humour-based Language Use

Venosha Ravana^{1*}, Sarala Thulasi Palpanadan², Shalini Saxena³, Mahendra Kumar Dhakad⁴

¹Faculty of Social Science and Humanities, Tunku Abdul Rahman University of Management and Technology,
Kuala Lumpur Campus, Kuala Lumpur, Malaysia

²Centre for Language Studies, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia

³Department of Science and Humanities, Poornima University, Rajasthan, India

⁴Department of Yogic Sciences, Poornima University, Rajasthan, India

*Corresponding author: venosha@taru.edu.my; ORCID iD: 0000-0002-4941-5309

ABSTRACT

Digital tools are increasingly used to deliver language assessments; however, the alignment between peer evaluations and instructor judgements remains underexplored, particularly in creative, humour-based tasks. This mixed-methods case study examined (1) the alignment between peer and instructor scores and (2) how peer and instructor comments aligned with or diverged from the assessment rubric. Thirty undergraduate English-major students, working in seven groups, produced 3- to 5-minute humorous video commercials for a Technology Enhanced Language Learning course. The videos were assessed by the instructor and peers using the same three-criterion rubric: Language, Audio and Visuals, and Task Fulfilment. Peer and instructor scores were compared using Pearson's correlation and a paired-samples *t*-test, while 147 comments were analysed through thematic content analysis ($\kappa = 0.84$). The results revealed a strong positive correlation between peer and instructor scores, $r(5) = 0.873$, $p = 0.010$, and no significant difference in their mean scores, $t(6) = 0.526$, $p = 0.618$. However, their comments revealed divergent priorities. The instructor emphasised linguistic accuracy (71.7%) and technical conventions, whereas peers prioritised creativity (73.6%), humour (68.9%), and audience relatability (55.2%). Interpreted through Amabile's Componential Theory of Creativity, the findings tentatively suggest that combining instructor attention to technical accuracy with peer sensitivity to audience engagement may support a more balanced assessment of creative multimodal assignments. Given the small, single-institution sample, the proposed framework remains exploratory and requires validation in larger and more diverse contexts.

Keywords: peer evaluation, digital assessment, creativity, humour, multimodal assessment

Received: 16 December 2025, **Accepted:** 6 April 2026, **Published:** 30 July 2026

1.0 Introduction

Digital assessments have become increasingly prevalent in language education, utilising technology to measure communicative skills, content understanding, and language proficiency. Unlike traditional exam-based assessments, digital platforms offer interactive, adaptive, and real-time feedback mechanisms that more accurately reflect language use in authentic contexts. Grab (2025) contends that digital assessments may surpass conventional assessments in their ability to capture dynamic language use. For instance, Dziemianko (2022) demonstrated that warning messages in monolingual English learners' dictionaries (MLDs) on error correction practices not only enable close monitoring of students' online activity but also integrate effective digital feedback mechanisms, establishing a significant and positive role for such warnings in MLDs. Similarly, interactive assignments such as video creation and presentations have been reported to provide more flexible, engaging, and effective evaluation methods that support both learners and educators in contemporary language education (Čandrlić et al., 2014).

One widely adopted feature of digital assessments is peer evaluation. Jamrus et al. (2024) explain that peer evaluation enables learners to assess and provide feedback—with instructor guidance—on each other's work using digital tools and platforms. Beyond encouraging learning and introspection, this method has been found to foster communication, critical thinking, teamwork, and self-improvement (Chen et al., 2020). According to Ting and Shukor (2022), peer learning and peer evaluations typically occur in formative or summative assessment contexts, involving the assessment of group work contributions or the review of assignments using explicit rubrics or guided questions.

However, while peer evaluation enhances learner engagement by making assessment an active, student-centred process, it also presents challenges. Stonewall et al. (2024) identified issues such as bias and inaccuracy in grading, largely because students generally lack the training to evaluate others, which may compromise the reliability of the assessment. A particular concern is the potential discrepancy between how students and instructors evaluate creative and open-ended assignments, especially when humour and digital aesthetics are key components. Heidari-Shahreza (2024) emphasises that creativity and humour in language assessments may be evaluated based on novelty or originality (NO) and task appropriateness (TA); however, these components may not be readily apparent to students if they misunderstand the assignment or misinterpret the assessment criteria.

Moreover, instructors may not always be aware of language change and variation prevalent in social media spaces. What appears acceptable to students may be misinterpreted by instructors as erroneous. For example, Gulnazir et al. (2025), who studied language variation in TikTok videos, concluded that creativity is a pertinent aspect of word dissemination on the platform. The word 'rizz'—shortened from charisma, meaning the ability to woo someone—was ultimately named the Oxford Dictionary's Word of the Year (WOTY) in 2023, demonstrating its acceptance in contemporary English communication (Gulnazir et al., 2025). This example illustrates that digital language assessments must be designed with tasks that evaluate the contextual, playful, and pragmatic use of language within conversational or situational settings—not merely correct or incorrect grammar usage. When evaluating the accuracy of evolving lexicon currently in use, peer feedback may offer a more accurate representation of the application of linguistic principles.

The challenges described above are not merely theoretical. Similar integrity and validity issues arise when assessment responsibility is shifted from teachers to peers in digital environments, reflecting broader concerns—raised in both public discourse and educational policy debates—about whether assessments truly capture learner abilities (Manyin, 2024; Teach for Malaysia, 2025). These real-world concerns underscore the need for empirical research that examines whether peer evaluations can reliably assess creative, multimodal language tasks. While digital assessments and peer evaluation are being implemented in classrooms, there is insufficient evidence on whether peer-generated scores align with instructor judgments—and more importantly, whether peers and instructors are evaluating the same qualities when assessing creative work.

Despite the growing body of research on peer evaluation in language learning contexts, two significant gaps remain. First, while several studies have compared peer or instructor scores, most have focused on traditional writing or speaking tasks. Few have examined whether peer-instructor score alignment holds

for creative, multimodal assignments such as video production, where assessment criteria are more subjective and open to interpretation. Second, even when score alignment has been found, prior research has not adequately explored whether peers and instructors prioritise different qualitative factors in their evaluations. Studies such as Jamrus et al. (2024) and Biton (2025) have documented perceptions of peer evaluation, but they have not directly compared the content of peer and instructor feedback to identify divergent evaluative lenses. This distinction is critical because quantitative agreement in scores may mask meaningful qualitative differences with important pedagogical implications.

To address these gaps, the current study examines both the quantitative alignment and qualitative divergence between peer and instructor evaluations in a creative, humour-based video assignment. This assignment a 3 to 5-minute humorous video commercial was selected because it mirrors the types of multimodal, creative tasks increasingly used in digital language classrooms. The task required students to produce content that was both linguistically accurate and creatively engaging, making it well-suited for examining how different evaluators weigh these competing demands. The assessment was conducted in a Technology Enhanced Language Learning course at a Malaysian private university, where students evaluated each other's videos using the same three-criterion rubric (Language, Audio & Visuals, Task Fulfilment) as the instructor.

Specifically, this study is guided by the following research questions:

- a. To what extent do peer-generated scores for the video assignment align with the instructor's scores?
- b. How do peer and instructor comments align with or diverge from the rubric criteria when assessing creative video assignments?

By investigating both score correspondence and how comments align with rubric criteria, this study aims to provide insights into the strengths and limitations of peer evaluation in digital language assessments. The findings contribute to a more nuanced understanding of whether peer and instructor evaluations capture similar or distinct aspects of student performance in creative, multimodal tasks. For digital assessments in language learning contexts, this means moving beyond static definitions and grammatical rules to evaluate the dynamic, creative, and contextually appropriate use of the continually evolving language.

2.0 Literature Review

This section briefly presents insights from recent studies on digital assessments in the context of language learning, the research gaps identified and the theoretical framework used to guide the current study.

2.1 Peer Evaluation in Digital Assessments for Language Learning

Digital assessments for language learning have evolved to utilise innovative, adaptive, and interactive digital tools. For instance, the Duolingo English Test adopts AI and machine learning to dynamically adjust difficulty based on the learner's proficiency. Cheung (2025) also states that higher scores are predicted for students who incorporate AI tools into their language test preparation, using them to strengthen particular areas of weakness and enhance other language skill areas; this is especially true for speaking, where AI pronunciation feedback can be especially helpful, and writing, where grammar and style can be improved with AI-assisted editing. In addition, video-based language assignments in ESL and EFL classrooms have been found to enhance L2 competence through multimodal project-based learning, with students perceiving them as effective for improving proficiency (Ayar, 2023). Analytical rubrics for video speaking assessments boost rater reliability via detailed criteria on content, delivery, visuals, and collaboration.

Regardless, automated scoring systems integrated into most digital language assessments still pose a challenge in terms of credibility and reliability, whereby Edwards et al. (2019) believe that human intervention is still necessary. Multimodal assessment rubrics in ESL/EFL contexts integrate criteria for linguistic, visual, audio, and spatial modes to evaluate compositions such as videos or presentations.

These rubrics often emphasise multimodality, audience awareness, synthesis, and substance; however, raters struggle with ambiguities, leading to inconsistencies influenced by language dominance or rubric design flaws (Chan, 2025). Task-specific rubrics developed collaboratively with raters may improve authenticity and reliability in language application settings.

To provide a ‘human touch’ in digital language assessments, peer evaluations have been frequently integrated as an effective and interactive method (Stonewall et al., 2024). Studies have shown that peer assessment positively impacts writing performance by improving comprehension (Tran & Pham, 2023); however, Hasnani and Mubarak (2020) state that peer assessment can also support the development of students’ speaking performance, helping them develop fluency in communicating ideas and managing constructive criticism.

It is undeniable that a key gap in this area is the challenge of ensuring fairness, reliability, and objectivity in peer evaluations affect the validity of evaluations. Vuogan and Li (2023) found that the qualitative comments made by peer and teacher assessors were similarly effective as the score reliability was equal. The study, conducted in a Chinese EFL classroom, compared instructor and peer comments on selected speech samples of English major students at a college in China. While this was a favourable outcome to indicate peer assessment as an important alternative to or complementary source of teacher assessment in ESL/EFL classrooms, another more recent study contrasts the findings. Biton (2025) reports that students have expressed concern over the validity of peer evaluations, where friendship bias or hesitance to provide critical feedback might compromise the quality and integrity of feedback and grades. This results in anxiety among learners, as such assessments might contribute to formal grading, which influences their overall academic performance. Calibration training, such as stand-alone pre-tasks or exemplar discussions, could align peer marks with those of teachers, reducing harshness or leniency in language assessment contexts. In engineering-related EFL presentations, students were found to mark peers more strictly than staff post-calibration, highlighting the need for iterative training (Mainali et al., 2025).

In addition, the limited empirical research on how peer evaluations capture complex constructs such as creativity and humour in language use—constructs that are inherently subjective and influenced by social context (Heidari-Shahreza, 2024; Maraev et al., 2021)—may contribute to concerns about the reliability of peer evaluations (Biton, 2025; Stonewall et al., 2024). Vuogan and Li (2023), in their meta-analysis of peer feedback in second-language writing, found that peer feedback positively impacts writing performance, though they noted variability in effectiveness across contexts. They also identified the need for further research on how peer feedback can be optimally implemented in language assessment contexts. Eventually, measures must be taken to improve the reliability and validity of peer evaluations by focusing on designing systems that ensure fairness and anonymity and exploring innovative ways to assess nuanced language skills, such as creativity and humour.

2.2 The Role of Creativity and Humour in Language Use

Creativity and humour are important components of language pedagogy, as they enhance learner engagement, motivation, and cognitive development. They foster a positive, calm, and enjoyable learning environment that lowers anxiety and promotes risk-taking, which is crucial for language acquisition in the 21st century (Kessler, 2013; Mai, 2024; Maraev et al., 2021; Pianzola et al., 2020).

Creative tasks in language acquisition develop cognitive flexibility and critical thinking by requiring learners to manipulate language playfully and innovatively (Jones, 2020). According to Heidari-Shahreza (2024), creativity in language assessment may be best assessed in a particular context, which can be historical, socio-cultural, or even demographical. For this, creativity (C) can be calculated using the formula: $C = [O \times TA]_{\text{Context}}$ (Heidari-Shahreza, 2024, p. 2). However, Jones (2020) also noted that most considerations of creativity in language learning and teaching have been framed within the framework of dominant monolingual ideologies, which means that creative language use is viewed as discrete and abstract quotes isolated from one another and from the complex social settings in which they are employed.

On the other hand, Maraev et al. (2021) argued that humour is often an interpretation of interactions which transcends language, culture, and society; hence, a discourse analysis to analyse the mechanisms of inference in dialogues is the best way to understand and assess humour in creative outputs. In a language classroom, it was cleverly suggested that “[t]he ‘ha-ha’ of humour in the classroom may indeed contribute to the ‘aha!’ of learning from the student” (Heidari-Shahreza, 2021, p. 4). This suggests that humour-based language activities may improve retention, stimulate creativity, and reinforce language competencies in an interactive and enjoyable manner. Thus, as proven in the study of Gulnazir et al. (2025) on language use in TikTok videos, to enhance the role of creativity and humour in digital language assessments (DLAs), the design of these assessments must integrate the fluid, inventive nature of social media language which might be more apparent to peers than to instructors.

As mentioned above, peer evaluation fosters interaction and social understanding, which are essential components of humour production and appreciation. Paine et al. (2025) stated that because humour naturally depends on social context and subjective interpretation, which peers may be best able to understand, peer evaluations can be useful for evaluating innovative humour. This echoes Gulnazir et al. (2025), who found that creative DLAs could require learners to complete gaps, construct appropriate dialogue, or critique the usage of creative slang in short video clips or written skits, ensuring that the assessment reflects the communicative intent of humour and artistry of the language.

2.3 Amabile's Componential Theory of Creativity

Amabile's Componential Theory of Creativity (2012) is a comprehensive model that identifies the social and psychological elements required for creativity to explain how an individual creates a creative work (Kessler, 2013). The theory posits that creativity is not a singular trait but rather the result of a confluence of four key components that must all work together to achieve a high level of creative output. The componential theory of creativity specifies that creativity requires a confluence of all its components. As suggested by Kessler (2013), creativity is expected to be highest when an intrinsically motivated person possessing high domain expertise and skill in creative thinking works within an environment that highly supports creativity. The components are considered multiplicative in their effects.

According to the theory, there are four primary components of creativity: three internal and one external. Domain-relevant skills represent the raw materials that individuals draw upon throughout the creative process. Second, creativity-relevant processes encompass cognitive styles and personality characteristics that encourage novel thinking.

Next, intrinsic task motivation refers to the passion or motivation to undertake a problem or task because it is intrinsically interesting, involving, personally challenging, or satisfying. Finally, the social environment, the only external component, includes the work environment and all social forces surrounding the individual. It encompasses both factors that can stimulate creativity and those that can act as obstacles to it. Figure 1 summarises these components based on definitions and analysis from the studies of Fisher et al. (2020) and Gao (2024).

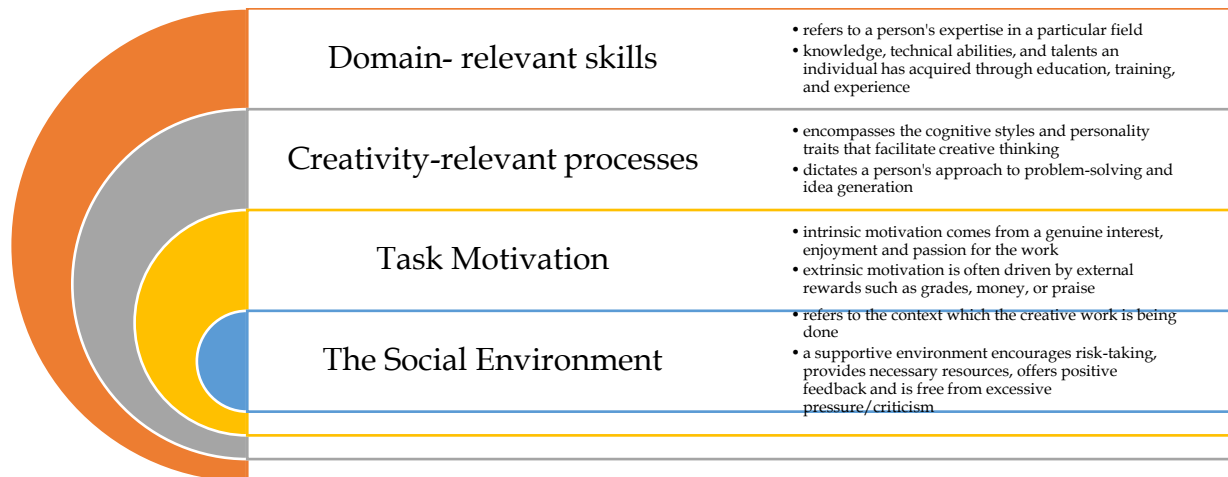
As illustrated in Figure 1, this theory is highly useful because it suggests that creativity is a process that can be understood and nurtured by focusing on these four components, and it can also be assessed using the same components. The theory defines creativity as the production of ideas or outcomes that are both novel and appropriate to some goal (Fisher et al., 2020). When applied to language production and use, a creative linguistic response, whether spoken or written, must be novel and appropriate (valuable, correct, feasible, or fitting) for the specific communicative task or goal. If humour is defined as a linguistic output that is both novel and appropriate (e.g. novel phrasing that is appropriate and fitting for generating laughter in a specific context), then the framework could be applied to assess linguistic cues, comedic timing, and cultural context.

Unfortunately, recent studies utilising the Componential Theory of Creativity in language assessments and digital language assessments are emerging but remain limited in number. Therefore, the current study aims to fill this gap by exploring this theory as a framework for improving teaching and evaluation methods in language education. Recently, Ahmed and Feist (2021) reported that creativity in scientific

and artistic texts can be analysed based on linguistic tools framed within the Componential Theory of Creativity, opening up opportunities to develop language assessment rubrics from the perspective of creativity and humour. Gao (2024) also reported that this theory can be applied to develop and evaluate students' task motivation, expertise, and creativity-relevant skills in vocational language classes.

Figure 1

The Four Components of Amabile's Componential Theory of Creativity



Note. Adapted from Fisher et al. (2020) and Gao (2024).

The relevance of Amabile's theory to the current study lies in its capacity to explain why peer and instructor evaluations might diverge. The theory predicts that creativity emerges from the interaction of domain-relevant skills (knowledge and technical expertise), creativity-relevant processes (novel thinking and risk-taking), and intrinsic task motivation (personal interest and engagement). In the context of a creative language assessment such as the humorous video assignment examined here, instructors may be expected to prioritise domain-relevant skills, as these align with conventional assessment criteria emphasising accuracy and technical correctness. Conversely, peers may be more attuned to creativity-relevant processes and task motivation, as these are more visible and relatable to them as fellow learners and consumers of media content. This theoretical distinction provides a framework for understanding the two gaps identified in the literature: whether peer and instructor scores align quantitatively, and whether their qualitative evaluative priorities differ. The current study therefore applies Amabile's theory to examine both the correspondence and divergence between peer and instructor evaluations in a creativity-based digital language assessment.

3.0 Methodology

This section details the research design, participants, data collection and analysis procedures, and ethical considerations. A case study with a mixed-method design was selected as the methodology to gain an in-depth understanding of the strengths and weaknesses of peer evaluations for digital assessments compared to instructor evaluations. The research questions stated at the end of the Introduction guided data collection, analysis, and discussion.

3.1 Research Design and Justification

A convergent mixed-method design was selected because the research questions required both quantitative comparison (RQ1) and qualitative exploration (RQ2). This design is well-suited for triangulation, allowing the study to determine whether quantitative alignment masks qualitative divergence—a question that neither quantitative nor qualitative methods alone could address (Creswell & Plano Clark, 2018). First, the quantitative component (comparing peer and instructor scores) addressed the *extent* of alignment, while the qualitative component (thematic content analysis of comments) addressed the *nature* of evaluative priorities. Second, the mixed-method approach allowed

for triangulation: if peer and instructor scores aligned quantitatively but diverged qualitatively, this would reveal meaningful differences in evaluation frameworks that a single method could not capture.

It is important to clarify that quantitative analyses were conducted at the group level ($n = 7$ groups), as peer evaluation in this context was a collaborative group activity where all members of a group received the same score. This decision was guided by the assessment design, where marks were awarded collectively. Accordingly, the unit of analysis for the quantitative research question was the group, whereas the unit of analysis for the qualitative research question was the individual comment. This distinction is acknowledged as a limitation in Section 5.3.

3.2 Participants

The course, Technology Enhanced Language Learning, is a third-year core subject in the Bachelor of Arts (English) program. Students in this cohort had achieved a minimum CEFR B2 level (IELTS 6.0 or MUET Band 4.0) based on university entry requirements. As part of their assessments for the course, they were required to submit a video with the following instructions:

Theme for Coursework

The Commercial No One Asked For (But Everyone Needs)

Introduction

This group assignment challenges you to unleash your creativity and digital literacy by producing an engaging video. Your task is to create a humorous and over-the-top commercial for a product or service that seems utterly unnecessary at first glance, but which you'll convince your audience is absolutely essential.

This coursework fulfils Course Learning Outcome 2: Create a video based on a given situation, using online technologies (Bloom's Taxonomy level C6).

Video Requirements

Length: Approximately 3 – 5 minutes long.

Content: The video must consist of spoken and/or written language (e.g., dialogue, voice-overs, catchy slogans, on-screen text).

Elements: Integrate a combination of the following:

Music: Background music, sound effects, jingles.

Subtitles: For dialogue or key information.

Still Images: Photos, graphics, product mock-ups.

Moving Images: Video footage (live-action, animation, screen recordings).

Presentation: Your video will be presented in class.

Evaluation Criteria: Language (30 marks), Use of Audio & Visuals (50 marks) and Task Fulfilment (20 marks).

All members of the same group will receive the same marks.

In Week 8 of the semester, the students were required to present the video in class, whereby other students evaluated and provided comments in their respective coursework groups. There was a total of 7 groups of students for the coursework. In class, the groups provided comments anonymously on the Padlet platform, which could be accessed and read by all students; however, the marks awarded were submitted only to the instructor. It is to be noted that the peer evaluation was only done for the purpose

of this case study and final marks awarded for the actual assessment were fully based on the instructor's assessment in reference to the established rubrics for the coursework.

All students consented to their videos, scores, and peer assessment outcomes being used for the purpose of this case study. The identities of the students were kept anonymous in this case study.

3.3 Data Collection

Two types of data were collected simultaneously. First, the marks awarded by the instructor and peers to each group were compiled in a Microsoft Excel spreadsheet. Second, verbatim comments provided by students for each group's video were collected from the Padlet platform and compiled, along with the instructor's comments, in a Microsoft Word document.

3.3.1 Peer Score Aggregation

Peer scores were calculated as follows: Each of the seven groups was evaluated by all other groups ($n = 6$ peer ratings per group). The average peer score for each group was calculated by taking the mean of these six ratings. Groups did not evaluate their own work. All peers used the same three-criterion rubric as the instructor (Language: 30 marks; Audio & Visuals: 50 marks; Task Fulfilment: 20 marks). The standard deviation of peer scores across groups ranged from 4.2 to 6.8, indicating moderate variation. No calibration training was provided, as peer scores were collected for research purposes only and did not affect course grades.

3.4 Data Analysis

For the analysis of discrepancies in marks awarded, descriptive and inferential statistics were performed using Python-based libraries on JupyterLab, following Davoody and Gharibvand (2024) and Wang et al. (2022). Pearson's correlation coefficient (r) was calculated to determine the strength of the relationship between peer and instructor scores, and a paired t-test was used to assess whether the mean difference between the two sets of scores was statistically significant. As noted above, all quantitative analyses were conducted at the group level ($n = 7$).

For the qualitative analysis of comments, a thematic content analysis approach was adopted. Thematic content analysis combines the systematic coding and frequency counting of content analysis with the interpretive theme development of thematic analysis (Braun & Clarke, 2006). This approach was selected because it allowed for both the identification of recurring patterns in the comments and the quantification of their prevalence across evaluator groups—addressing the research question of how peer and instructor comments aligned with or diverged from rubric criteria.

All 147 comments (87 peer comments, 60 instructor comments) were transcribed verbatim. The analysis followed a systematic six-phase procedure adapted from Braun and Clarke's (2006) framework for thematic analysis, with additional procedures for quantifying thematic prevalence consistent with content analysis approaches (Erlingsson & Brysiewicz, 2017):

- I. **Phase 1 – Familiarisation:** Two researchers independently read all comments multiple times to gain an overall sense of evaluative patterns. Initial impressions, recurring topics, and potential coding categories were recorded independently.
- II. **Phase 2 – Initial coding:** Inductive codes were generated from the verbatim comments (e.g., 'grammar error,' 'funny narration,' 'audio clarity,' 'unexpected concept'). A codebook was developed iteratively, with each code defined, accompanied by inclusion and exclusion criteria, and illustrated with representative examples. The two researchers coded a random sample of 30 comments (20%) independently to establish coding consistency.
- III. **Phase 3 – Theme generation:** Codes were organised into candidate themes through iterative discussion between the two researchers. This involved grouping related codes (e.g., 'grammar error,' 'pronunciation issue,' 'word choice' were grouped under 'linguistic/technical accuracy') and reviewing them against the research questions.

- IV. **Phase 4 – Theme review:** Candidate themes were reviewed against the full dataset to ensure they accurately represented the data. Themes were refined, merged, or split as needed. Each theme was evaluated for internal coherence (codes within the theme formed a coherent pattern) and external distinctiveness (themes were clearly distinguishable from one another).
- V. **Phase 5 – Theme definition:** Two superordinate themes were defined: 'technical-linguistic accuracy' (encompassing language accuracy, audio-visual quality, and adherence to conventions) and 'creativity-engagement' (encompassing creativity, humour, and audience relatability). A third theme, 'storytelling and narrative flow,' was identified but appeared with moderate frequency in both evaluator groups and was therefore retained as a secondary theme rather than a superordinate one.
- VI. **Phase 6 – Quantification and reporting:** To address the research question of alignment and divergence, the frequency of comments addressing each theme was calculated separately for peer and instructor comments. Percentages were calculated based on the total number of comments within each evaluator group (87 peer comments; 60 instructor comments). This quantification allowed for systematic comparison of evaluative emphasis between the two groups. Findings were documented with representative quotes for each theme, organised by evaluator group.

Finally, the inter-coder reliability was calculated on a randomly selected 20% of comments ($n = 30$). Because comments could address multiple themes, Cohen's kappa was calculated at the code level using a presence/absence matrix. The overall $\kappa = 0.84$, with individual code agreement ranging from 0.79 to 0.89, indicating substantial agreement (Landis & Koch, 1977). Disagreements were resolved through consensus discussion. The codebook remained stable after coding the reliability sample, and no new codes emerged from the remaining comments, suggesting that the coding scheme was sufficiently comprehensive for this dataset. All remaining comments were coded by the primary researcher using the finalised codebook.

3.5 Ethical Considerations and Ensuring Trustworthiness in the Study

All participants provided written informed consent prior to data collection. Students were informed that: (a) participation was voluntary; (b) they could withdraw at any time without penalty; (c) all data would be anonymised prior to analysis, with group identifiers replaced by codes; and (d) peer evaluations were collected for research purposes only and did not affect course grades. The study received institutional ethics exemption as it involved low-risk educational research with informed consent.

To ensure consistency in marking, a second instructor independently scored 100% of the video assignments using the same rubric without reference to the first instructor's marks. The inter-rater agreement between the two instructors was calculated (intraclass correlation coefficient = 0.91), indicating excellent reliability. The final marks used in the analysis were those awarded by the first instructor. The second instructor's independent scoring provided evidence of strong inter-rater consistency. No formal calibration training was provided to the peer evaluators because the peer scores were collected solely for research purposes and did not affect course grades. The absence of calibration training is acknowledged as a limitation in Section 5.3.

4.0 Results and Findings

This section presents the quantitative and qualitative results in the order of the research questions stated in the Introduction.

4.1 RQ1: Quantitative Alignment Between Peer and Instructor Scores

Table 1 presents the total scores awarded by the instructor and peers to each group across three assessment criteria: Language, Audio and Visuals, and Task Fulfilment. It also shows the difference between the instructor's score and the average peer score, thereby addressing RQ1.

Of the seven groups, five had absolute score differences of no more than five marks, whereas two had differences exceeding five marks. According to the grading rubric, a five-mark difference may place a group in a different grade category.

Table 1

Total Marks for Coursework: Video Creation Based on Instructor and Peer Evaluations

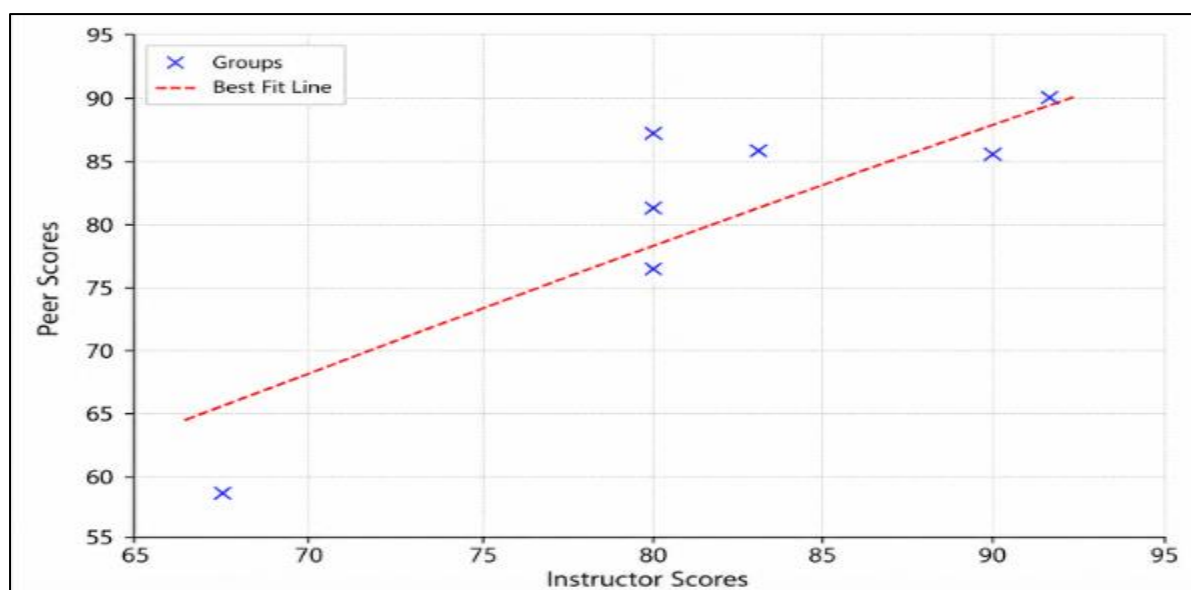
Group	Instructor Total Score	Average Peer Total Score	Difference (Instructor - Peer)
1	80	76.5	3.5
2	83	85.8	-2.8
3	92	90.2	1.8
4	80	87.3	-7.3
5	80	81.3	-1.3
6	68	58.8	9.2
7	90	85.6	4.4

Note. The difference was calculated by subtracting the average peer score from the instructor score. Positive values indicate that the instructor awarded a higher score, whereas negative values indicate that the peers awarded a higher average score.

The instructor and peer scores followed a similar pattern, with both sets of scores ranking Group 3 highest and Group 6 lowest. Pearson's correlation showed a strong positive association between the scores, $r(5) = .873, p = .010$. A paired-samples t test did not identify a statistically significant difference between the mean instructor and peer scores, $t(6) = 0.526, p = .618$. These findings indicate similar overall scoring patterns, although neither correlation nor a nonsignificant t test establishes exact agreement between the two forms of evaluation.

Figure 2

Relationship Between Instructor and Peer Scores



As illustrated in Figure 2, higher instructor scores were generally associated with higher peer scores. The dashed line represents the best-fitting linear regression equation:

$$\text{Peer score} = 1.18(\text{Instructor score}) - 15.81$$

Points above the line represent groups whose peer scores were higher than predicted by their instructor scores, whereas points below the line represent groups whose peer scores were lower than predicted. The largest discrepancies occurred for Groups 4 and 6, although the scores demonstrated a strong positive overall relationship.

4.2 RQ2: Alignment and Divergence of Comments with Rubric Criteria

To address RQ2, “How do peer and instructor comments align with or diverge from the rubric criteria when assessing creative video assignments?”, all 147 comments posted on Padlet were analysed using thematic content analysis procedures (Erlingsson & Brysiewicz, 2017). These comprised 87 peer comments and 60 instructor comments. The comments were distributed across the seven groups as follows: Group 1 ($n = 20$), Group 2 ($n = 22$), Group 3 ($n = 24$), Group 4 ($n = 19$), Group 5 ($n = 21$), Group 6 ($n = 17$), and Group 7 ($n = 24$). This relatively balanced distribution indicates that no single video disproportionately influenced the identified patterns.

The three rubric criteria were Language (30 marks), Audio and Visuals (50 marks), and Task Fulfilment (20 marks). For the analysis, Language encompassed grammatical accuracy, pronunciation, vocabulary range, and overall clarity. Audio and Visuals encompassed sound quality, music balance, subtitle synchronisation, and production quality. Task Fulfilment encompassed adherence to the commercial format, creativity, originality, and audience engagement.

The codes were grouped into categories and subsequently organised into broader themes. The analysis showed that both instructor and peer comments addressed the rubric criteria but differed in the relative emphasis placed on particular aspects. Therefore, divergence was interpreted primarily as a difference in evaluative emphasis rather than as evaluation outside the rubric. Table 2 presents the dominant themes, representative excerpts with anonymised identifiers, and their frequencies as counts and percentages. It illustrates areas of alignment as well as differences in emphasis between the instructor and peer comments.

Table 2

Thematic Priorities in Peer versus Instructor Evaluations

Theme	Example Quote (with Identifier)	Instructor Comments (n = 60)	Peer Comments (n = 87)
Linguistic/ Technical Accuracy	"Some grammatical errors found on text-on-screen" (Instructor, Group 3); "Pronunciation of 'schedule' needs attention" (Instructor, Group 5)	43 (71.7%)	7 (8.0%)
Audio-Visual Clarity & Production Quality	"Background music overpowers dialogue at 1:30" (Instructor, Group 6); "Subtitles out of sync with narration" (Instructor, Group 2)	35 (58.3%)	11 (12.6%)
Adherence to Commercial Conventions	"Credits need to be included" (Instructor, Group 4); "Missing call-to-action at the end" (Instructor, Group 1)	27 (45.0%)	3 (3.4%)

Theme	Example Quote (with Identifier)	Instructor Comments (n = 60)	Peer Comments (n = 87)
Creativity & Originality	"The concept was so unexpected – never seen anything like this before" (Peer Comment, Group 3)	7 (11.7%)	64 (73.6%)
Humour & Entertainment Value	"Humour carried the entire video. We were laughing throughout" (Peer Comment, Group 2); "The pause before the punchline was perfect" (Peer Comment, Group 5)	2 (3.3%)	60 (68.9%)
Relatability & Audience Engagement	"Really appealed to Gen Z humour – this is exactly how we talk on IG" (Peer Comment, Group 4); "The lecturer cameo was brilliant – so unexpected" (Peer Comment, Group 7)	1 (1.7%)	48 (55.2%)
Storytelling & Narrative Flow	Instructor: "Concept is good, but content takes some time to be established" (Instructor, Group 6); Peer: "The build-up to the product reveal was perfect" (Peer Comment, Group 3)	14 (23.3%)	27 (31.0%)

Note. Percentages exceed 100% because individual comments often addressed multiple themes and were counted in each applicable theme.

As Table 2 shows, instructor comments aligned closely with all three rubric criteria, with 71.7% addressing Language, 58.3% addressing Audio & Visuals, and 45.0% addressing Task Fulfilment. Peer comments, however, diverged in focus: while they did address Task Fulfilment (particularly creativity and audience engagement), they gave minimal attention to Language (8.0%) and Audio & Visuals (12.6%). This pattern indicates that peers disproportionately emphasised one aspect of the rubric (creativity and engagement under Task Fulfilment) while neglecting the other two criteria.

Next, the seven themes presented in Table 2 were derived from the two superordinate themes defined during Phase 5 of the analysis (see Section 3.4). The superordinate theme 'Technical-Linguistic Accuracy' encompassed three themes: Linguistic/Technical Accuracy, Audio-Visual Clarity & Production Quality, and Adherence to Commercial Conventions. The superordinate theme 'Creativity-Engagement' encompassed three themes: Creativity & Originality, Humour & Entertainment Value, and Relatability & Audience Engagement. The theme 'Storytelling and Narrative Flow' appeared with moderate frequency in both evaluator groups and did not fit neatly into either superordinate theme; rather than force it into an existing category, it was retained as a secondary theme to accurately represent the data.

For counting purposes, comments containing more than one theme were counted in each applicable theme. For example, a peer comment stating, "The humour was great and the concept was so creative" would be counted under both 'Humour & Entertainment Value' and 'Creativity & Originality'. This decision was made because the research question focused on the presence of thematic content in comments, not on mutually exclusive categorisation. All coding decisions were documented in the codebook, and any ambiguous comments were discussed between the two researchers until consensus was reached.

4.2.1 Thematic Findings

As shown in Table 2, the instructor's comments aligned closely with the rubric criteria. Comments on linguistic/technical accuracy (43/60 instructor comments, 71.7%), audio-visual clarity (35/60, 58.3%), and adherence to commercial conventions (27/60, 45.0%) dominated instructor feedback. Representative instructor comments included corrective feedback on grammar ("missing comma before the final clause," Instructor, Group 2), pronunciation ("word stress inconsistent," Instructor, Group 5),

and technical execution ("background music too loud at 1:32," Instructor, Group 6). Notably, the instructor rarely commented on creativity (7/60, 11.7%) or humour (2/60, 3.3%), and when such comments appeared, they were often qualified (for example, "Creative concept, but execution needs work," Instructor, Group 4).

In contrast, peer comments diverged substantially from the rubric criteria. Peers prioritised creativity and originality (64/87 peer comments, 73.6%), humour and entertainment value (60/87, 68.9%), and relatability or audience engagement (48/87, 55.2%). Peer comments frequently praised unexpected elements ("the inclusion of lecturers was completely unexpected – that carried the video," Peer Comment, Group 5), comedic timing ("the pause before the punchline was perfect," Peer Comment, Group 5), and generational resonance ("this is exactly how we talk on IG," Peer Comment, Group 4). Only 7/87 peer comments (8.0%) mentioned linguistic accuracy, and no peer comments provided corrective feedback on grammar or pronunciation. Interestingly, peers occasionally acknowledged technical limitations but explicitly dismissed them when creativity was high: e.g., "The audio was a bit weird, but honestly the humour made up for it!" (Peer Comment, Group 3).

Storytelling and narrative flow were the only themes that appeared with moderate frequency in both sets of comments (14/60 instructor comments, 23.3%; 27/87 peer comments, 31.0%), though instructors framed it as a clarity issue ("takes too long to establish the product," Instructor, Group 6) while peers framed it as engagement ("the build-up kept us curious and the climax was so worth it," Peer Comment, Group 3).

These patterns indicate that while peers and instructors agreed on which groups performed best overall (as shown in Section 4.1, $r = 0.873$), their comments reflected the rubric criteria differently. Instructor comments aligned closely with the formal rubric criteria—language accuracy, technical execution, and adherence to conventions. Peer comments diverged from these criteria, instead emphasising creativity-relevant processes and task motivation, components central to Amabile's theory, even when technical execution was imperfect. This divergence suggests that while quantitative scores may align, the qualitative reasoning behind those scores differs substantially between the two evaluator groups.

5.0 Discussion

Amabile's Componential Theory of Creativity posits that creativity emerges from the confluence of three internal components—domain-relevant skills, creativity-relevant processes, and intrinsic task motivation—interacting with the external social environment (Amabile, 2012; Fisher et al., 2020). The findings of this study empirically instantiate this tripartite structure in a language assessment context, revealing that instructor and peer evaluators operationalised these components differently. Instructors anchored their assessments in domain-relevant skills (linguistic accuracy, technical execution, adherence to conventions), while peers foregrounded creativity-relevant processes (originality, humour, relatability) and task motivation (effort, engagement, audience appeal).

5.1 Interpretation of Discrepancies

The quantitative analysis revealed a strong overall alignment between peer and instructor evaluations ($r = 0.873$), with no statistically significant difference in mean scores. This correspondence suggests that peers and instructors broadly agreed on which groups produced the strongest and weakest videos overall. However, this quantitative alignment masks meaningful qualitative divergence. Discrepancies emerged in specific cases: Group 4 received substantially higher peer ratings than instructor ratings (peer average: 87.3 vs. instructor: 80), whereas Group 6 showed the opposite trend (peer average: 58.8 vs. instructor: 68). These divergent cases warrant closer examination through the lens of Amabile's theory.

Next, the thematic content analysis showed that instructor feedback predominantly reflected domain-relevant skills—the technical competencies and knowledge base required for effective video production. Of the instructor comments, 71.7% addressed linguistic/technical accuracy, 58.3% focused on audio-visual clarity, and 45.0% emphasised adherence to commercial conventions (see Table 2). Representative comments included corrective feedback on grammar ("Some grammatical errors found on text-on-screen"), pronunciation ("word stress inconsistent"), and technical execution ("background

music too loud at 1:32"). These patterns indicate that the instructor's evaluative lens prioritised correctness, completeness, and professional standards. Notably, when technical elements were found lacking, the instructor marked the work down even when the video was entertaining—for the instructor, accuracy and technical execution took precedence over entertainment value.

In marked contrast, peer comments diverged substantially from the rubric criteria; peers were far more responsive to creativity-relevant processes and intrinsic task motivation. Of the peer comments, 73.6% praised creativity and originality, 68.9% highlighted humour and entertainment value, and 55.2% commented on relatability and audience engagement. Only 8.0% of peer comments mentioned linguistic accuracy, and no peer provided corrective feedback on grammar or pronunciation. A representative peer comment stated: "The inclusion of lecturers was completely unexpected—that carried the video." (Peer Comment, Group 5). To contextualise this comment, one group had recruited senior lecturers to act as clients in their commercial, which surprised the audience and generated laughter. Another peer noted: "Humour carried the entire video—we were laughing throughout." (Peer Comment, Group 2). Interestingly, peers occasionally acknowledged technical limitations but explicitly dismissed them when creativity was high, as reflected in this comment: "The audio was a bit weird, but honestly the humour made up for it!" (Peer Comment, Group 3). These patterns demonstrate that peers valued the creative processes underlying the work and the motivational drive to produce content that resonated with an audience. This willingness to overlook minor technical flaws when videos were entertaining or relatable highlights peers' appreciation of creativity-relevant processes and task motivation—two core components of Amabile's Componential Theory of Creativity (Gao, 2024; Kessler, 2013).

The divergence in how comments reflected the rubric criteria likely reflects three interconnected factors. Firstly, differential familiarity with the assessment criteria was evident. Students did not receive formal calibration training on the rubric criteria, whereas the instructor was intimately familiar with them. This training gap likely affected the consistency and quality of peer feedback. As noted by Stonewall et al. (2024), students generally lack the training to evaluate others, which may affect the reliability and validity of peer assessments. In the absence of explicit rubric training, untrained peers may naturally gravitate toward what is most salient, accessible, and enjoyable in a video—creativity, humour, and engagement—rather than technical execution. This phenomenon aligns with research by Mainali et al. (2025), who found that calibration training (e.g., stand-alone pre-tasks or exemplar discussions) is essential for aligning peer marks with teacher expectations.

Secondly, the qualification of peer evaluators. The question of whether students are "qualified" evaluators in this context is nuanced. On the one hand, peers are not subject-matter experts in language assessment; they lack the pedagogical training and evaluative experience that instructors possess, which limits their ability to assess linguistic accuracy and technical conventions reliably. On the other hand, peers bring a unique qualification to the assessment of creativity and humour: they are the target audience. Humour naturally depends on social context and subjective interpretation (Maraev et al., 2021; Paine et al., 2025), and peers may be best positioned to evaluate whether creative content resonates with its intended audience. Gulnazir et al. (2025) similarly concluded that creativity in digital language use—as observed on platforms like TikTok—is best assessed within the socio-cultural context in which it is produced and consumed, a context that peers are more attuned to than instructors.

The third factor relates to assessment stakes and motivation. The assessment context (low-stakes, research-only) may have reduced peers' motivation to provide critical corrective feedback. Since peer scores did not affect course grades, peers may have felt less obligated to scrutinise technical shortcomings or to risk potential conflict by offering negative evaluations. This pattern has been observed in prior research (Biton, 2025; Stonewall et al., 2024), where low-stakes settings are associated with leniency bias and reluctance to provide critical feedback. Additionally, the absence of formal grading accountability may have inadvertently encouraged peers to focus on positive, enjoyable aspects of the videos rather than technical deficiencies.

Discrepancies, therefore, occurred when peers appreciated and rewarded creative effort and humour, even when the instructor noted technical shortcomings. Group 4, for instance, was rated higher by peers because of its unexpected concept and engaging humour, despite the instructor's concerns about

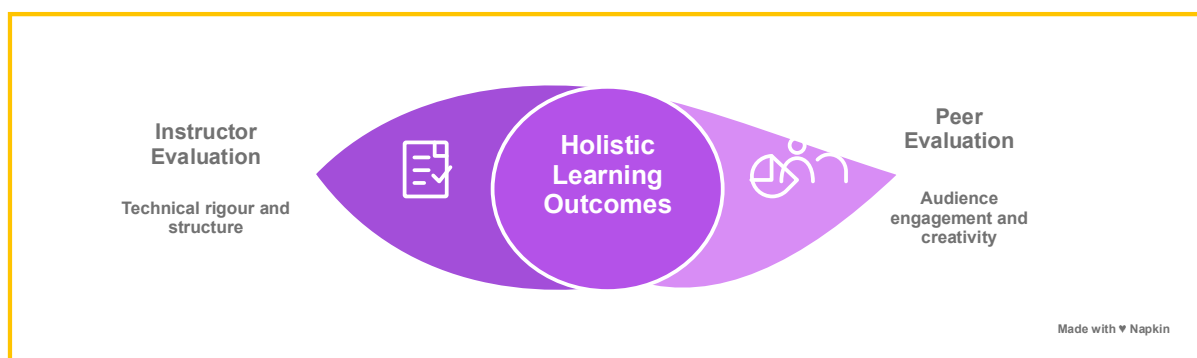
grammar and pronunciation. Conversely, Group 6 received lower peer ratings because its content was less engaging or relatable, despite meeting technical expectations. These patterns are not random; they represent a consistent divergence in evaluation priorities grounded in the different roles, training, and perspectives of peers and instructors. Peers prioritised creativity-relevant processes and intrinsic task motivation, while instructors anchored their evaluations in domain-relevant skills—a distinction that aligns precisely with the components of Amabile's Componential Theory of Creativity.

5.2 Implications for Language Assessments and Pedagogy

The findings of this study have several implications for digital assessment for language learning. First, they suggest that a dual-perspective assessment framework—combining instructor rigor with peer sensitivity to engagement—may offer a more balanced approach to evaluating creative, multimodal assignments than either perspective alone. This framework as illustrated in Figure 3, is grounded in Amabile's theory and ensures that students receive feedback on both technical accuracy (domain-relevant skills) and audience appeal (creativity-relevant processes and task motivation). Such a complementary system may help students recognise that successful media products require both professionalism and imaginative appeal.

Figure 3

A Dual Perspective Framework for Creativity-based Digital Assessments



Note. Conceptual framework for dual-perspective digital assessments for language learning (developed by the authors based on study findings and Amabile's Componential Theory of Creativity)

Second, the divergence observed highlights the importance of calibration training for peer evaluators. As Mainali et al. (2025) demonstrated, calibration can reduce harshness or leniency in peer marking and align peer scores more closely with instructor expectations. Without such training, peer evaluations may remain unreliable for high-stakes decisions.

Third, these findings suggest that peer evaluation is best suited for formative, low-stakes purposes or as a supplementary data source, rather than as a replacement for instructor judgment. The risks of bias, leniency, and inconsistency in untrained peer scoring (Biton, 2025) caution against using peer evaluations for summative grading unless robust calibration and moderation procedures are in place.

5.3 Limitations of the Current Study

Several limitations should be acknowledged. First, the quantitative analyses were based on only seven group-level score pairs. This small analytical sample limits statistical power and may produce unstable correlation and mean-difference estimates. Consequently, the nonsignificant result of the paired-samples t test should not be interpreted as evidence that instructor and peer scores were equivalent. Second, the use of aggregated peer scores meant that variation among individual peer evaluators was not examined. Third, the single-institution, single-instructor, and single-assignment design limits the generalisability of the findings to other educational contexts, assessors, and assessment formats. Fourth, the absence of peer-evaluator calibration may have affected the consistency and quality of peer scoring and feedback. The low-stakes nature of peer evaluation may also have influenced students' motivation

to provide critical comments. Finally, the Padlet comments were nested within groups and may have included multiple contributions from the same students. Therefore, the comment frequencies should be interpreted descriptively rather than as independent observations suitable for inferential comparison.

6.0 Conclusion and Recommendations for Future Research

This study examined assessment priorities in instructor and peer evaluations within a creativity-based digital language assessment and aligned these findings with Amabile's Componential Theory of Creativity. Within the scope of this small-scale case study (30 students, 7 groups, single institution), the findings indicate that peer and instructor evaluations were quantitatively aligned ($r = 0.873$) but qualitatively distinct. Instructors prioritised domain-relevant skills, particularly language accuracy and adherence to conventions, whereas peers emphasised creativity-relevant processes and intrinsic task motivation.

These findings tentatively suggest that a dual-perspective framework—combining instructor rigor with peer sensitivity to engagement—may offer a more balanced approach to assessing multimodal language assignments than either perspective alone. However, given the exploratory nature and limited scope of this study, these conclusions should be considered preliminary. The framework is proposed here as a theoretically grounded hypothesis for future research, rather than a proven model for wide adoption.

Pedagogically, these findings suggest that digital language assessment rubrics might benefit from explicit criteria that capture both technical accuracy and creative engagement. However, peer evaluation should be reserved for low-stakes formative purposes or as a supplementary data source, not as a replacement for instructor judgment, given risks of bias and unreliability in untrained peer scoring (Biton, 2025).

A key methodological implication arising from this study is that quantitative score alignment should not be interpreted as evidence of evaluative equivalence. As the qualitative analysis revealed, peers and instructors may agree on overall rankings while applying fundamentally different criteria. Researchers validating peer assessment instruments should therefore complement statistical comparisons with qualitative analyses of feedback content.

Future research should therefore: (a) move beyond simple score comparisons to incorporate qualitative analyses of feedback content when validating peer assessment instruments; (b) experimentally compare trained and untrained peer scoring to assess the value of calibration in aligning both scores and evaluative priorities; (c) examine whether peer-instructor divergence predicts subsequent learning gains or influences student motivation; (d) develop and validate rubrics that explicitly integrate creativity and accuracy criteria, thereby acknowledging the dual demands of multimodal assignments; and (e) replicate this study with larger, more diverse samples across multiple institutions and assessment contexts to test the generalisability of the dual-perspective framework proposed here.

Acknowledgement

The authors would like to extend their heartfelt gratitude to the Department of English Studies at the Tunku Abdul Rahman University of Management and Technology for their support.

Conflict of Interest

The authors have declared that no competing interests exist.

Author Contribution Statement

VR: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Visualization, Writing: original draft, Writing: review and editing. STP: Methodology, Validation,

Writing: review and editing. SS: Investigation, Resources, Validation, Writing: review and editing. MKD: Investigation, Resources, Validation, Writing: review and editing.

Funding

This research received no external funding.

Ethics Statement

This study received institutional ethics exemption because all student participants provided informed consent prior to data collection for the specific purpose of this research. Students consented to the use of their videos, scores, and peer assessment outcomes. No retrospective analysis of pre-existing data was conducted; all data were collected prospectively with explicit consent. The study did not involve any experimental manipulation, and all data were anonymized prior to analysis. Under the university's ethical review guidelines, such low-risk educational research qualifies for exemption from full IRB review.

Data Access Statement:

Research data supporting this publication are available upon request to the corresponding author.

Author Biography

Dr Venosha Ravana is an Assistant Professor at Tunku Abdul Rahman University of Management and Technology, Malaysia. Her areas of expertise include instructional design and technology, technology-enhanced language learning, and teacher development. She is actively involved in research supervision and publication in peer-reviewed journals. She currently manages the PhD in Education programme at her institution.

Associate Professor Dr Sarala is an English language lecturer whose areas of expertise include teaching English as a second language, writing instruction, mixed methods research, and teacher education. She is also involved in developing modules and programmes for school and university students. She has published books, book chapters, newspaper articles, and research articles in indexed journals.

Professor Dr Shalini Saxena is a Professor of English and Soft Skills at Poornima University, Jaipur, with more than 19 years of teaching and research experience. She holds a PhD from the LNM Institute of Information Technology and has authored a book and published more than 25 research articles in peer-reviewed journals. Her research interests include gender studies, cultural narratives and mythological retellings, postcolonial ecocriticism, and innovative pedagogy.

Dr Mahendra Kumar Dhakad is an experienced academic with more than 18 years of teaching experience in higher education. He is currently an Associate Professor at Poornima University and holds a PhD in Education. His research interests include educational and sports psychology, guidance and counselling, curriculum studies, and inclusive education. He has published more than 20 research articles in peer-reviewed journals.

References

Ahmed, S. T., & Feist, G. J. (2021). The language of creativity: Validating linguistic analysis to assess creative scientists and artists. *Frontiers in Psychology*, 12, Article 724083. <https://doi.org/10.3389/fpsyg.2021.724083>

- Amabile, T. M. (2012). *Componential theory of creativity* (Working Paper No. 12-096). Harvard Business School.
- Ayar, Z. (2023). Exploring the efficacy of video project assignments in English language learning through the lens of tertiary-level EFL students. *Dil ve Edebiyat Araştırmaları*, 28, 470–481. <https://doi.org/10.30767/diledeara.1343292>
- Biton, Y. (2025). Student reflections on peer assessments: Benefits and challenges in a mathematics class. *Educational Process: International Journal*, 14, Article e2025003. <https://doi.org/10.22521/edupij.2025.14.3>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Čandrlić, S., Katić, M. A., & Dlab, M. H. (2014, May). Online vs. paper-based testing: A comparison of test results. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 775–780). IEEE. <https://doi.org/10.1109/MIPRO.2014.6859649>
- Chan, S. W. C. (2025). Assessing multimodal compositions in EAP context: Raters' perceptions and operationalization of assessment criteria. *Language Assessment Quarterly*, 22(4–5), 555–581. <https://doi.org/10.1080/15434303.2025.2605656>
- Chen, H., Park, H. W., & Breazeal, C. (2020). Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education*, 150, Article 103836. <https://doi.org/10.1016/j.compedu.2020.103836>
- Cheung, E. (2025, March). *Enhancing Duolingo English Test performance: An analysis of online resources, learner strategies, and the emerging role of AI* [Unpublished manuscript]. ResearchGate.
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- Davoody, L. R., & Gharibvand, L. (2024). Qualitative and quantitative data analysis and visualization using Python. In *Proceedings of the Western Users of SAS Software 2024 Conference* (Paper 194-2024). WUSS. https://www.wuss.org/proceedings/2024/194/194_FINAL_paper_pdf.pdf
- Dziemianko, A. (2022). Warnings in learners' dictionaries: Do they help to correct errors and learn usage? *International Journal of Lexicography*, 35(2), 239–265. <https://doi.org/10.1093/ijl/ecac001>
- Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social identity theory in human-robot interactions. *Computers in Human Behavior*, 90, 357–362. <https://doi.org/10.1016/j.chb.2018.08.027>
- Erlingsson, C., & Brysiewicz, P. (2017). A hands-on guide to doing content analysis. *African Journal of Emergency Medicine*, 7(3), 93–99. <https://doi.org/10.1016/j.afjem.2017.08.001>
- Fisher, C. M., Ananth, P., & Caliskan, O. D. (2020). A winding road: Teresa Amabile and creative process research. In R. Reiter-Palmon, C. M. Fisher, & J. Mueller (Eds.), *Creativity at work* (pp. 35–46). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-61311-2_4
- Gao, Q. (2024). To enhance students' creativity based on componential theory of creativity: A case study of vocational Chinese class, Beijing Polytechnic, China. In *Proceedings of the 2024 2nd International Conference on Language, Innovative Education and Cultural Communication* (pp. 480–487). Atlantis Press. https://doi.org/10.2991/978-2-38476-263-7_58

- Grab, M. Ö. (2025). Assessing language proficiency in the digital age: Exploring innovative methods for assessment and feedback in ESL classrooms. *Journal of Digital Educational Technology*, 5(2), Article ep2510. <https://doi.org/10.30935/jdet/16286>
- Gulnazir, N., Zainuddin, S. Z., & Salehuddin, K. (2025). Language innovation on social media: A case study of *rizz* on TikTok. *GEMA Online Journal of Language Studies*, 25(1), 119–136. <https://doi.org/10.17576/gema-2025-2501-07>
- Hasnani, H., & Mubarak, M. H. (2020). The use of peer assessment in enhancing ESP students' speaking ability. *Ethical Lingua: Journal of Language Teaching and Literature*, 7(1), 74–81. <https://doi.org/10.30605/25409190.135>
- Heidari-Shahreza, M. A. (2021). Humor in the language classroom: Pedagogical benefits and practical considerations. *TESOL Journal*, 12(2), Article e572. <https://doi.org/10.1002/tesj.572>
- Heidari-Shahreza, M. A. (2024). Playful creativity in TESOL: Toward humor-integrated language learning (HILL). *TESL-EJ*, 27(4), 1–23. <https://doi.org/10.55593/ej.27108a2>
- Jamrus, M. H. M., Zamri, A. A., Norazizan, M. H., & Kamarulzaman, M. H. (2024). Using peer evaluation as a tool in English language learning in Malaysian classroom. *International Journal of Research and Innovation in Social Science*, 8(9), 121–132. <https://doi.org/10.47772/IJRISS.2024.809010>
- Jones, R. H. (2020). Creativity in language learning and teaching: Translingual practices and transcultural identities. *Applied Linguistics Review*, 11(4), 535–550. <https://doi.org/10.1515/applirev-2018-0114>
- Kessler, E. H. (2013). Componential theory of creativity. In E. H. Kessler (Ed.), *Encyclopedia of management theory*. SAGE. <https://doi.org/10.4135/9781452276090.n50>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Mai, T. C. T. (2024). Impacts of humor as a communication strategy for teachers in the EFL classrooms. In *Proceedings of the International Conference on Research in Education and Science* (Vol. 10, pp. 871–881).
- Mainali, B., Tse, N. M. K., Mardonova, M., & Benavente Molina, I. (2025). Calibrating peer assessment: Lessons from student and staff marking comparisons. In *Proceedings of the Australasian Association for Engineering Education 2025 Conference*. Note: Add page range, conference location, publisher, and DOI/URL if available.
- Manyin, M. (2024, August 25). Quality of Malaysia's education sliding, says former Sarawak minister. *The Vibes*. <https://www.thevibes.com/articles/news/103205/quality-of-malaysias-education-sliding-says-former-sarawak-minister>
- Maraev, V., Breitholtz, E., Howes, C., Larsson, S., & Cooper, R. (2021). Something old, something new, something borrowed, something taboo: Interaction and creativity in humour. *Frontiers in Psychology*, 12, Article 654615. <https://doi.org/10.3389/fpsyg.2021.654615>
- Paine, A. L., Hashmi, S., Fink, E., Mitchell, P., & Howe, N. (2025). Humorous peer play and social understanding in childhood. *Communications Psychology*, 3(1), Article 76. <https://doi.org/10.1038/s44271-025-00252-3>
- Pianzola, F., Rebora, S., & Lauer, G. (2020). Wattpad as a resource for literary studies: Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PLOS ONE*, 15(1), Article e0226708. <https://doi.org/10.1371/journal.pone.0226708>
- Stonewall, J. H., Dorneich, M. C., & Rongerude, J. (2024). Evaluation of bias in peer assessment in higher education. *International Journal of Engineering Education*, 40(3), 543–556.

- Teach For Malaysia. (2025, April 28). *Beyond exams: Rethinking assessments towards purposeful learning*. Ed Pulse. <https://teachformalaysia.org/beyond-exams-rethinking-assessments-towards-purposeful-learning/>
- Ting, C., & Shukor, N. A. (2022). Effects of peer learning on pupils' learning performance and creativity in English writing skills using digital storytelling. *Sains Humanika*, 14(3–2), 105–115. <https://sainshumanika.utm.my/index.php/sainshumanika/article/view/2025>
- Tran, O. T. T., & Pham, V. P. H. (2023). The effects of online peer feedback on students' writing skills during coronavirus pandemic. *International Journal of Instruction*, 16(1), 881–896. <https://doi.org/10.29333/iji.2023.16149a>
- Vuogan, A., & Li, S. (2023). Examining the effectiveness of peer feedback in second language writing: A meta-analysis. *TESOL Quarterly*, 57(4), 1115–1138. <https://doi.org/10.1002/tesq.3178>
- Wang, H.-I., Manolas, C., & Xanthidis, D. (2022). Statistical analysis with Python. In D. Xanthidis, C. Manolas, O. K. Xanthidou, & H.-I. Wang (Eds.), *Handbook of computer programming with Python* (pp. 373–408). CRC Press. <https://doi.org/10.1201/9781003139010-9>
- Ting, C., & Shukor, N. A. (2022). Effects of Peer Learning on Pupils' Learning Performance and Creativity in English Writing Skills using Digital Storytelling. *Sains Humanika*, 14(3–2), 105–115. <https://sainshumanika.utm.my/index.php/sainshumanika/article/view/2025>
- Tran, O. T. T., & Pham, V. P. H. (2023). The Effects of Online Peer Feedback on Students' Writing Skills during Corona Virus Pandemic. *International Journal of Instruction*, 16(1), 881–896. <https://doi.org/10.29333/iji.2023.16149a>
- Vuogan, A., & Li, S. (2023). Examining the effectiveness of peer feedback in second language writing: A meta-analysis. *TESOL Quarterly*, 57(4), 1115–1138. <https://doi.org/10.1002/tesq.3178>
- Wang, H.-I., Manolas, C., & Xanthidis, D. (2022). Statistical analysis with Python. In D. Xanthidis, C. Manolas, O. K. Xanthidou, & H.-I. Wang (Eds.), *Handbook of computer programming with Python* (pp. 373–408). CRC Press. <https://doi.org/10.1201/9781003139010-9>