Volume 14, Issue 1
Winter and Spring, 2026
pp. 1-29

Journal of Applied Linguistics and Applied Literature: Dynamics and Advances

Assessing the Assessors: A Study of UTAS Level 2 EFL Teachers' Perspectives and Practices in Writing Evaluation and Its Impact on Students' Writing Scores

Zahra Zargaran[1, *] and Mohsen Ghorbanpoor[2]

[1] *Corresponding Author, PhD in ELT, University of Technology and Applied Sciences, Shinas, Oman*
Email: Zahra.zaragaran@utas.edu.om

[2] *MA in ELT, University of Technology and Applied Sciences, AlMusanaah, Oman*
Email: Mohsen.ghorbanpoor@utas.edu.om

Abstract

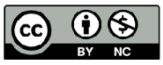
This study examines the perceptions and practices of Teachers who teach Level 2 in using standardised writing assessment rubrics at the University of Technology and Applied Sciences (UTAS) in Oman. A semi-structured interview and think-aloud protocol were employed to examine how teachers interpreted and applied rubric criteria while marking Task 2 writing. Thematic analysis was employed to analyse the qualitative data, revealing both the strengths and limitations of the current rubric, such as vague descriptors and misalignment with A2-level learner expectations. Based on teacher feedback, a modified rubric was developed. To evaluate its impact, a within-subjects design was employed, and a paired-sample t-test was conducted to compare students' writing scores between the two versions of the rubric. Results showed significant improvements in students' overall writing scores, especially in the areas of grammar and vocabulary. The findings underscore the value of teacher-informed rubric design, alignment with instructional goals, and ongoing moderation. This study contributes to enhancing assessment reliability and pedagogical relevance in EFL writing evaluation in higher education institutions such as UTAS in Oman.

Keywords: writing rubric, UTAS, Oman, teaching writing, writing assessment

ARTICLE INFO

Research Article
Received: Saturday, September 13, 2025
Accepted: Thursday, March 12, 2026
Published: Wedensday, April 1, 2026
Available Online: Thursday, March 12, 2026
DOI: https://doi.org/10.22049/jalda.2026.30804.1842

Online ISSN: 2821-0204; Print ISSN: 28208986



Introduction

English as a Foreign Language (EFL) scholars have extensively studied writing assessment in a wide variety of contexts. Writing assessment rubrics primarily serve as standardised evaluation tools which aim to enhance objectivity, consistency, and transparency in grading student writing. But they also serve multiple other purposes including providing valuable feedback to students which guides their learning journey (Nurhayati, 2020), monitoring students' progress over time (Riddle et al., 2016) and enhancing the effectiveness of formative assessment practices (Panadero & Jonsson, 2013).

Despite these benefits, the effectiveness of rubrics is thoroughly dependent on teachers' understanding, acceptance and accurate implementation (Li & Lindsey, 2015). Teachers' impressions of writing assessment criteria significantly affect how these instruments are used in the classroom (Weigle, 2002). In EFL contexts, although teachers typically praise rubrics for their fairness, comprehensiveness, and practicality (Jonsson & Svingby, 2007), research indicates that they often doubt their applicability and dependability, particularly when rubrics are not tailored to particular teaching settings (Reddy & Andrade, 2010). Such differences call for questions regarding the effectiveness of rubrics in fairly evaluating students' writing competency.

Researchers have also evaluated the alignment or misalignment between teachers' perceptions and their assessment practices in this regard. Weigle (2002) highlights that teachers' perceptions of writing assessment tools may significantly influence how these tools are implemented in classroom settings. Lee (2009) indicates that the realities of evaluation practices are not always a perfect reflection of the perceived assessment goals. Therefore, it is reasonable to anticipate that these perceptions have the potential to give rise to discrepancies between the standardised rubrics implemented by an institution and the actual evaluation practices of their teaching staff. By raising alignment with real-world classroom practices, previous studies indicate that including teacher input into rubric design may increase teacher satisfaction and enhance student outcomes (Andrade, 2005). Still, carefully assessing rubric validity and reliability is crucial to avoid inconsistent application (Alderson, 2005).

Diverse, multinational educational contexts can aggravate the possible mismatch between standardised rubrics and teachers' real grading methods. Teachers come from a wide range of contexts (cross-national contexts), hence carrying various cognitive repertoire when using the rubrics which is stated to be a core reason for using the same criteria inconsistently. Therefore, it can be implied that Preparatory Studies Centres (PSC) of the University of Technology and Applied Sciences (UTAS) in Oman could observe a greater gap in the possible discrepancies among students' writing scores. It is probable that the EFL lecturers with diverse educational and ethnic backgrounds affect the deployment of writing rubrics.

The researchers of this study decided to assess and modify the writing rubrics at Level 2 because it is a mediating level from very low to very high writing

abilities. Therefore, the study examined how UTAS EFL teachers who teach Level 2 perceived and used the current writing assessment rubrics. The main goal was to provide a polished Task 2 writing assessment rubric catered to the UTAS GFP Level 2 setting, fostering equitable, dependable, and pragmatic writing assessments. This study also aimed to improve student learning results through more consistent and objective evaluation methods. The study offers insightful analysis pertinent to EFL evaluation procedures in similar worldwide settings by tackling the difficulties of rubric adaptation in a multicultural classroom.

Literature Review

Rubric Effectiveness and Rater Cognition

Meta-analysis studies confirm that well-designed analytic rubrics improve inter-rater reliability and score validity when descriptors are clear and level-appropriate (Jonsson & Svingby, 2007; Brookhart & Chen, 2015). However, reliability collapses when descriptors contain vague quantifiers ("some", "a little", "frequent") or expectations beyond students' developmental stage (Barkaoui, 2007; Li & Lindsey, 2015). Recent validation studies reinforce this statement. For instance, a construct validity analysis of primary trait rubrics in L2 writing showed high internal consistency but highlighted rater overemphasis on irrelevant features like essay length, threatening score validity (Alghizzi & Alshahrani, 2024). Similarly, AI-assisted rubric applications in EFL contexts achieve strong correlations with human raters but falter on nuanced cognition, such as weighing criteria like coherence over mechanics (Doewes et al., 2023). Think-aloud studies consistently show that raters deviate from rubric wording in real time, relying instead on personal constructs (Crusan et al., 2016; Eckes, 2012)

The Perception–Practice Gap

A persistent finding across contexts is the mismatch between what teachers say about rubrics and what they do (Weigle, 2002; Lee, 2009). Yet the majority of studies rely on post-hoc surveys or interviews, which are prone to social-desirability bias and cannot capture moment-by-moment decision-making (Barkaoui, 2007). Concurrent think-aloud protocols remain rare in rubric validation research, despite being the gold standard for exposing cognitive processes (Ericsson & Simon, 1993). While think-aloud protocols have been validated for studying reactivity and veridicality in L2 writing processes (Yang & Zhang, 2023), their application to rater cognition in rubric use is limited to isolated cases, often overlooking real-time discrepancies in scoring. For instance, studies on AI-assisted rating highlight think-aloud protocol's potential for revealing rule-based inflexibility (Jin, 2025), but empirical links to rubric revision remain underexplored.

Teacher-Informed Revision and Score Impact

Although teacher involvement in rubric design is repeatedly recommended (Andrade & Du, 2005; Panadero & Jonsson, 2020), only a handful of studies have measured the effect of revisions on actual student scores, and none at pre-intermediate level using a within-subjects design (see Lim & Sudweeks, 2020 for a B2 example). An experimental study on analytic rubrics in Vietnamese EFL peer/self-

assessment found significant gains in writing proficiency, attributed to clearer descriptors reducing subjectivity, though long-term retention was untested (Phuong et al., 2023). Similarly, integrating AI feedback with teacher input boosted argumentative writing scores by 15–20% in timed tasks, but only when prompts aligned with local rubrics (Le et al., 2024). Another comparative analysis further showed that teacher-revised rubrics in EFL settings improved holistic scores, yet highlighted persistent gaps in vocabulary depth for pre-intermediate learners (Alghizzi & Alshahrani, 2024).

Contextual Misalignment in Gulf Foundation Programs

Standardized rubrics in Saudi Arabia, Qatar, and the UAE frequently import IELTS/CEFR B2-C1 criteria, producing systematic under-scoring of pre-intermediate learners and teacher frustration (Alshakhi, 2019; Goodwin, 2019; Hidri, 2018). An Omani study on EFL foundation programs revealed rubric misalignment with CEFR objectives led to 25% score inflation/deflation due to cultural-linguistic biases, with teachers adapting informally despite institutional mandates (Al-Saadi et al., 2025). In Algerian EFL contexts, instructors identified resource gaps and curriculum-rubric disconnects as key barriers, exacerbating low proficiency transitions (Abderrahmane & Mebitil, 2025). A broader synthesis across GCC nations noted that without localization, rubrics undermine self-assessment efficacy, particularly in settings where L1 interference skews organization criteria (Alhalangy & AbdAlgane, 2023). No study to date has quantified the scoring consequences of this misalignment or tested a locally refined alternative with inferential statistics. The present investigation therefore fills a critical regional and proficiency-level gap.

Theoretical Framework of the Study

Grounded in a conceptual framework that combines rubric-based writing assessment ideas with assessment literacy theory, this study investigates how teachers' views and practices affect student writing outcomes. Teachers' knowledge, abilities, and attitudes required to create, evaluate, and apply assessment instruments successfully define their assessment literacy (Popham, 2009). It covers knowledge of evaluation goals, criteria, scoring systems, and the capacity for insightful comments. Assessment literacy affects how teachers understand rubric criteria, make evaluative judgements, and reconcile standardising with professional judgement in the framework of EFL writing assessment (Davison & Leung, 2009). Clear criteria and performance standards help rubrics, as ordered grading guides, precisely communicate writing quality. As Reynolds-Keefer (2010) contends, effective rubrics promote learning by outlining expectations and guiding student adjustments. Studies show that rubrics used as self-assessment by EFL learners were significantly effective in improving certain aspects of writing performance, i.e., accuracy and coherence (Sabermoghaddam Roudsari et al., 2024). Still, teachers' perspectives, contextual cues, and training help to control the actual rubrics' use (Brown & Harris, 2014). Variations in rubric interpretation can generate variations in grading and feedback, therefore affecting student motivation and development.

The framework holds that teachers' assessment practices, how closely they follow rubric criteria, whether they alter rubrics, and how they provide feedback are shaped by their assessment literacy and contextual constraints. Students' writing

performance and involvement are thus influenced by these criteria in turn. By analysing these linked elements, this study seeks to pinpoint elements that support or impede efficient rubric use in the UTAS Level 2 EFL context, refining assessment methods and student results.

Although analytic scoring rubrics are widely advocated for improving reliability and transparency in L2 writing assessment (Jonsson & Svingby, 2007; Brookhart, 2018), three critical gaps remain underexplored. First, most validation and revision studies rely on either teacher perceptions or student outcomes in isolation; few adopt mixed-methods designs that triangulate rater cognition (think-aloud protocols) with actual scoring behavior and subsequent score variation (Barkaoui, 2007; Crusan et al., 2016). Second, despite repeated calls for teacher-informed, context-specific rubric adaptation (Andrade & Du, 2005; Panadero & Jonsson, 2020), empirical evidence of the measurable impact of such revisions on student scores, particularly at pre-intermediate level is scarce. Third, in multilingual foundation programs typical of the Gulf region and other expanding higher education systems, standardized rubrics are frequently modelled on high-stakes tests (e.g., IELTS) and misaligned with local instructional realities and pre-intermediate level expectations (Alshakhi, 2019; Goodwin, 2019), yet almost no studies have quantified the scoring consequences of this misalignment or tested locally refined alternatives.

The present study addresses these gaps by using think-aloud protocols and semi-structured interviews to capture Level 2 EFL teachers' cognitive processes and perceptions of an institution-wide rubric, co-constructing a revised rubric grounded in teacher input and pre-intermediate level instructional objectives, and employing a within-subjects design with paired-samples statistics to determine the effect of the revised rubric on the same students' scores. In doing so, it provides rare empirical evidence of how teacher-informed rubric refinement enhances scoring sensitivity, especially in grammar and vocabulary, domains known to be developmentally critical at pre-intermediate level. Beyond the Omani context, the study's findings offer a replicable assessment process with local syllabi in multilingual preparatory programs worldwide, thereby contributing to the international literature on assessment literacy, rubric validation, and context-sensitive writing evaluation in lower-proficiency tertiary settings.

To address these objectives, the following research questions were formulated:

1. How do Level 2 (A2) EFL teachers at UTAS perceive the comprehensiveness, assessment accuracy, and effectiveness of the current writing assessment rubrics?
2. Are there any discrepancies between EFL Level 2 teachers' perceptions of the writing assessment rubrics used at UTAS and their actual evaluation of student writings?
3. Is there a significant difference between L2 midterm exam writing marks obtained by the current writing assessment rubrics and L2 midterm exam writing marks obtained by the modified rubrics developed based on Level 2 EFL teachers' perceptions?

Method

Participants

This study aimed to detect any improvements in students' writing scores after the teachers used the modified version of the rubric. For this purpose, the researchers used convenience sampling, and 10 teachers in two UTAS branches, Shinas and Al Mussanah, volunteered to participate in this study. The teachers were selected based on their willingness to participate in the study. All teachers had the experience of teaching Level 2 for a minimum of two semesters and came from various teaching contexts including India, Pakistan and Oman. The participants were males and females, averaging 45 years old. The average teaching experience was 10 years. They all were master's holders in teaching English.

Formal ethical approval was not required by University of Technology and Applied Sciences (UTAS) policy for this non-interventional study, which involved only voluntary staff participants and fully anonymized student exam papers. However, the research fully adhered to institutional ethical guidelines. Written informed consent was obtained from all 10 participating teachers prior to data collection. Both the teachers and the senior management of the Preparatory Studies Centers at Shinas and Al Musannah branches were fully consulted and expressed their explicit support for the study. Participation was entirely voluntary, with the right to withdraw at any stage without consequence clearly communicated.

Instruments

Semi-Structured Interview

First, to explore the teachers' cognitive perspectives regarding the writing rubrics in Task 2 of Level 2, a semi-structured interview was conducted. The answers were recorded and analysed. The interview consisted of 5 open-ended questions and was conducted face-to-face. An expert colleague further refined the wording of the questions and assessed the validity and clarity of the interview questions. These questions were piloted and retained with no changes. To ensure that the reliability of the interview was preserved, inter-rater reliability was conducted; two researchers checked the coding and analysis of the interview data. Moreover, both researchers agreed upon the method and the duration of the interview. In addition, piloting the questions ensured the consistency of the interview results.

Think-Aloud Protocol

This study used the think-aloud method to record teachers' thoughts while they were grading student essays using the Level 2 writing rubric. During the scoring process, teachers were asked to verbalise their thoughts aloud. It gave the researchers a better idea of how teachers reasoned, understood the rubric criteria, and evaluated students. This strategy worked well with the semi-structured interviews, as it demonstrated how teachers applied the criteria rather than simply stating their opinions. It made the results more trustworthy and helped make changes to the evaluation tool based on real data.

Task 2: Essay Writing

Task 2 in Level 2 is used to prepare students to write a descriptive essay to describe an object, a person or a place. This writing task assesses students' ability to use proper descriptive adjectives, grammar structures such as the present simple, adverbs of frequency, etc. Students are required to learn how to write a well-developed paragraph by including a topic sentence, supporting details and cohesive devices.

Although the delivery plan emphasises the key words highlighted in the coursebooks used at UTAS (Hubias & Muftahu, 2022), these lexical items are not specifically connected to the writing tasks. Therefore, teachers are open to teaching some topic-related vocabulary on their own using a non-centralised list of vocabulary. The required grammar of the task is not included in the writing syllabus but is a part of the whole grammar syllabus to cover during the semester. Therefore, teachers usually do not follow a specific grammar list required for Task 2; therefore, assessment is usually made referring to the accuracy of the grammar used. The midterm exam writing question was developed by the head office and distributed to all UTAS branches. The question was "write a short essay about your favourite shopping mall".

Writing Rubric Task 2

There are four proficiency levels in the GFP of all UTAS branches. Each level uses a slightly different writing rubric to score students' writings. In all level rubrics, there are four scoring criteria: Task Response, Organisation, Vocabulary, and Grammar. In Level Two, band scores range from 1 to 5. Two Tasks are used to assess pupils' writing skills at all language levels. Task 2 in Level Two requires composing at least three paragraphs totalling 150 words.

Task 2 writing rubric consists of 4 criteria, including Task response, Organisation, Grammar and Vocabulary, and ranges from band 1 to 5 which gives the total score of 20 for the overall writing score (see Appendix I). The rubric is developed and distributed by the UTAS head office, and teachers must follow it. Reliability and validity of the rubric are assessed and confirmed by the UTAS head office. Students' writing is assessed by two assessors and checked for consistency by the Table Head.

Modified UTAS Rubric

Following the analysis of think-aloud protocol data and teacher interviews, the participants' recommendations and suggestions were gathered and modified in accordance with the level-specific materials and the writing assessment criteria created by the head office. The new rubric descriptors were more closely aligned with the grammar structures taught and had a more organised breakdown of criteria across Task Achievement/Response, Organisation, Grammar, and Vocabulary (see Appendix II). A list of specific discourse markers taken from the syllabus is added, and the expectations are aligned with the skills of A2-level learners. The new descriptors place a strong emphasis on both clarity and developmental appropriateness. It means that at higher bands, there can be minor mistakes that do not hinder progress, but the standards are still very high.

Reliability of the revised rubric was confirmed via test–retest correlation. Eight weeks post-moderation, the same 10 teachers independently re-scored the 50 scripts using only the revised rubric. Pearson correlations between the first and second applications were excellent: total score $r(48) = .94$, $p < .001$; Task Response $r = .92$; Organisation $r = .91$; Vocabulary $r = .93$; Grammar $r = .95$ (all $p < .001$), exceeding the .90 threshold for strong reliability in L2 writing assessment (Knoch & Chapelle, 2017).

Procedure

In the General Foundation Program (GFP), students enrolled in Level 2 (A2 CEFR) are taught two writing tasks for the midterm exam and two different tasks for their level exit exam. Task 1 and Task 2 rubrics are developed in the university head office, and branch-level PSC teachers are expected to rigorously follow them.

The interview was carried out before the mid-term exam in a time span of two weeks. Each interview took about 7 minutes on average. All interviews were recorded and analysed using thematic analysis.

All interviews and think-aloud protocols were transcribed verbatim and anonymized. Following Braun and Clarke's (2006) six-phase framework, the two researchers independently read the transcripts several times for familiarization, then generated initial codes line-by-line using both data-driven and theory-driven approaches. Codes were collated into potential themes and reviewed against the coded extracts and the entire data set. Themes were then defined, named, and mapped onto the four rubric criteria. Finally, illustrative quotes were selected. Inter-coder agreement was established through discussion until full consensus was reached; any initial discrepancies were resolved by returning to the raw data. The final thematic map is presented in Tables 1–5.

In the next phase of the data collection, during the mid-term exam, all the same ten teachers were asked to express their mental process of decision-making and evaluation of using the rubrics while scoring Task 2 of the writing papers. Concurrent think-aloud protocol data were audio-recorded and used to explore teachers' perspectives while using the rubric. Think-aloud protocol was used to tap into teachers' practical experience, cognitive processes and reasoning while using the rubric to complement the theoretical and verbalised beliefs mentioned in the interview. This triangulation method increased the validity of the collected interview data.

Researchers randomly selected five students' midterm exam papers from each teacher's marking bundle and recorded teachers' thinking processes while marking the papers. Researchers also made a hard copy of the selected students' marked writing papers for later use. All audio files of teachers' thinking processes were analysed using thematic analysis. The analysis of the data obtained from both the interview and the think-aloud protocol was used to modify the writing rubric. The rubric criteria and band descriptors were modified and used to score the same students' writing papers after 8 weeks. The students' writing papers, which were

randomly selected, did not have scores or any underlining to avoid biased second marking with the modified rubric.

The researchers held a moderation session to explain the changes in the rubric and trained the teachers on how to use it after 8 weeks. A moderation session was held to unify the process of using the modified version of the writing rubric. All teachers were invited to participate in the moderation session, where the modifications applied to the writing rubrics were explained and clarified to the teachers. Some examples were also shown to operationalise the revised assessment tool. Teachers were requested to score the same 50 students' writing papers they had marked during midterm exams to see if the modified version of the scoring rubric would affect students' writing performance. This was done directly after the moderation in the same session so that any possible questions regarding the use of the new rubric could be easily clarified by the researchers. This was followed by a semi-structured focus group discussion where five EFL teachers were randomly selected to share their experiences of using the new rubrics and compare them with the existing assessment criteria. These three phases of the study were conducted consecutively in one setting.

Data Analysis

This study employed a mixed-methods design. The qualitative data (assessor's perceptions of the current Level 2 rubrics' clarity, comprehensiveness, and effectiveness) were collected through structured interviews and a think-aloud protocol, while the quantitative part used a within-subjects comparative component.

All interviews and think-aloud protocols were recorded, transcribed, and anonymised. Then, using Braun and Clarke's (2006) six-phase thematic analysis framework, recurring themes were identified and categorised to align with the core components of the rubric, namely Task Achievement/Response, Organisation, Grammar, and Vocabulary. Organisation, tabulation and frequency calculation of the coded data were done in Microsoft Excel. To check the data normality, the *Shapiro-Wilk* was calculated, which led to the use of the paired samples *t*-Test to allocate the difference in using the original and the modified version of the writing rubric. IBM Statistical Product and Service Solutions (SPSS) Statistics version 27 was used to determine if there was a significant difference in the marking outcome of the assessors using the existing and modified rubrics.

Results

Qualitative Findings from Interviews

Tables 1-4 below display participants' perceptions of the Level 2 writing assessment rubric. Overall, the teachers were content with the rubric's foundational structure. However, they also highlighted some shortcomings, challenges, and potential areas for improvement, believing that both assessment fairness and instructional relevance of the current rubric could be enhanced.

Limitations in Comprehensiveness, Accuracy, and Effectiveness

The most noticeable concern among 90% of the respondents was the use of unclear terms such as *"a little lack of clarity", "most",* and *"only some",* which give way to subjectivity in task achievement (TA) and task response (TR). They argued that this ambiguity is the root cause of inconsistent interpretations and student assessment outcomes among raters. Similarly, non-operational terms, including *"frequent"* and *"rare"* in grammar and vocabulary range indicators, were pointed out by 40% of teachers. Another noteworthy observation was that 70% of the EFL teachers considered the expectation of error-free writing for Band 5 to be unrealistic for A2-level learners. They strongly felt this expectation was not supported by early learning stages and could be discouraging for the learners.

Lack of clarity in the use of discourse markers was another frequently cited issue, mentioned by 50% of participants. Teachers argued that failure to define a "range" results in excessive or unnatural use of the transition markers. They posit that such an approach creates a tendency for both the students and teachers to connect all sentences in the text just to meet the perceived criteria. Moreover, 40% of the assessors deemed task achievement in guided narrative tasks too mechanical. These EFL teachers at UTAS asserted that the rubric fails to distinguish between superficial copying and students' use of the prompts to elicit genuine ideas. Other relevant oversights include a lack of critical thinking (20%), independent creative writing (20%), and no added value for paraphrasing. Exclusion of some organization features, such as paragraph unity (30%), further illustrates its incomprehensiveness.

Table 1
Limitations and Drawbacks in Comprehensiveness, Accuracy and Effectiveness

Points Mentioned	Total	Percentage
1. Subjectivity in TA/TR because of terms like "a little lack of clarity," "most," and "only some."	9	90%
2. Expects zero errors for band 5, which is unrealistic for Level 2 students (vocabulary and grammar expectations are not level adaptive, negative impact on marks)	7	70%
3. Lacks detail in discourse markers (unclear range)	5	50%
4. Task achievement is too mechanical in guided tasks (No differentiation between students who copy prompts and those who use them creatively.)	4	40%
5. Lacks detail in grammar range indicators (vague terms like "rare," "some," and "frequent.")	4	40%
6. Lacks detail in vocabulary range indicators (vague terms like "some," and "frequent.")	4	40%
7. Lacks other organization features (such as unity and cohesive devices)	3	30%
8. Rubric is too lenient (Minimum marks are too high, even for poor writing.)	3	30%

9.	Lacks critical thinking.	2	20%
10.	Limits independent thinking, creativity or advanced grammar attempts.	2	20%
11.	Promotes excessive use of discourse markers, leading to unnatural writing	2	20%
12.	Does not encourage paraphrasing.	2	20%
13.	L1 familiarity causes discrepancy between native English and Arab assessors (impeding / not impeding mistakes)	2	20%
14.	Allows a wide inter-rater discrepancy (6 marks)	2	20%
15.	Lacks clarity on what is considered on/partially on/off topic.	1	10%
16.	Encourages organised writing but limits reordering the prompts.	1	10%
17.	Overlooks inclusion of paragraph features (topic sentence).	1	10%
18.	Lacks detail in vocabulary register, tone and style (not task specific)	1	10%

Perceived Benefits of the Rubric

Despite these limitations, the rubric comes with many strengths shown in Table 2. All the teachers (100%) acknowledged that the core writing components, including task achievement/response, organisation, grammar, and vocabulary, are covered, which sets a common standard across the institution. Most of them (70%) found its structured framework a coherent guide for student writing evaluation. Fairness (30%), reduction in bias (10%), and the potential to serve as a guide for students (10%) were among other benefits.

However, while one assessor considered leniency in the lower band levels, where even minimally developed responses receive passing marks, three others found it problematic, arguing it could lead to inflated scores for underperforming students.

Table 2

Benefits of Using a Unified Writing Rubric

Points Mentioned	Total	Percentage
1. Rubric covers key components (task achievement, organisation, grammar, vocabulary).	10	100%
2. Provides a structured framework for grading.	7	70%
3. Ensures fairness and consistency across teachers.	3	30%
4. Encourages students to focus on organisation, grammar, and vocabulary.	1	10%
5. Reduces biased assessment	1	10%
6. Helps students understand expectations.	1	10%
7. Allows student support due to lenient criteria	1	10%

Recommendations for Rubric Enhancement

Interviewees also offered their insights on improving Level 2 writing assessment rubrics. Considering the natural appearance of mistakes in the early stages of language learning, the most recommended change (70%) was to revise Band 5 criteria and allow minimal grammatical or lexical lapses. The purpose of this modification is to make expectations more achievable. Forty per cent recommended defining specific descriptors for ambiguous indicators such as *"some"* or *"few,"* and reflecting task-specific features (30%) to account for differences between narrative and descriptive writing. Teachers also suggested different task types could have differentiated weightage of criteria (40%) to allow creativity and critical thinking (20%) as well as unity and cohesion (20%) in the organisation and task response components of task 2 questions, while more focus could be placed on vocabulary and grammar accuracy and range in task 1 questions that are guided. Providing annotated sample responses (40%) to promote inter-rater reliability and incorporating special considerations for students with special needs (10%) were among other notable suggestions.

Table 3

EFL Teachers' Input in Semi-structured Interviews- Recommendations

Points Mentioned	Total	Percentage
1. Allow a range of errors for band 5 (e.g., no more than 5 errors).	7	70%
2. Provide more specific descriptors for terms like "most," "some," and "few."	4	40%
3. Allow change in criteria weightage based on tasks (more weight on TR and Org in task 2, and more weightage on VOC and GR in task 1)	4	40%
4. Provide detailed guides and sample papers for teachers.	4	40%
5. Develop task-specific rubrics for different writing tasks.	3	30%
6. Reward creativity and critical thinking.	2	20%
7. Expand organisation criteria to include paragraph unity and cohesion.	2	20%
8. Develop separate rubrics for students with special needs (e.g., dyslexia).	1	10%
9. Develop a checklist instead of a table	1	10%
10. Include criteria for handwriting clarity.	1	10%
11. Address the impact of AI tools like ChatGPT on student writing.	1	10%

Other Contextual Considerations

Interviewees highlighted a few contextual issues that are noteworthy. Several teachers (30%) maintained that the overall expectations of the rubrics are

higher than the achievement level of the students. The unwarranted inclusion of performance indicators such as "no punctuation errors", even though all punctuation rules are not taught and practised in Level 2 learning materials, further widens the gap between high marks and actual student attainment. Other complications could be biased grading depending on whether the assessor shares L1 with students (20%) and the inability to differentiate top-performing students (10%). Whereas post-exam moderation sessions help reduce discrepancies (20%), they are deemed time-consuming and difficult for inexperienced teachers (20%). One assessor criticised designing the rubrics in a Western academic framework, replicating IELTS, arguing its misalignment with the local educational context (10%).

Table 4

EFL *Teachers' Input in Semi-structured Interviews – Other Considerations*

Points Mentioned	Total	Percentage
1. Rubrics expect indicators that Level 2 students cannot produce (do not fully reflect the learning objectives of L2).	3	30%
2. Teachers' familiarity with students can bias grading.	2	20%
3. Moderation sessions reduce discrepancies.	2	20%
4. Time-consuming moderation and challenges for novice teachers.	2	20%
5. Handwriting quality influences grading, but it is not addressed.	1	10%
6. Higher-achieving students perceive unfairness in grading.	1	10%
7. The rubric is designed for a Western academic context (IELTS), which may not align with the needs of local students.	1	10%

Data from the Think-Aloud Protocol

Table 5

Mapping Level 2 Writing Marking Criteria-Task 2

Criteria	Assessment Indicators	Total	Percentage
Task Response	Addressing the task (all/most/some parts)	10	100%
	Presenting a developed response with supported ideas.	8	80%
Organization	Organising ideas into individual paragraphs (logically)	10	100%
	Arranging ideas coherently	6	60%

Grammar	Having an overall/clear progression of ideas	8	80%
	Using a range of discourse markers appropriately / accurately	10	100%
	Using the range of structures required for the task (full range / most/some/very few/extremely limited)	8	80%
	Considering the range of grammar mistakes	10	100%
Vocabulary	Considering punctuation errors	10	100%
	Considering whether the mistakes impede communication	8	80%
	Considering the correct choice of vocabulary	8	80%
	Considering the range of spelling errors	10	100%
	Considering the appropriacy of register or style with the nature of the task	4	40%
Other Rubric Indicators	Considering whether the mistakes impede communication	8	80%
	Considering the relevance of the answer	4	40%
	Considering word count	4	40%

The think-aloud protocols revealed that teachers followed the Level 2 writing rubric systematically when marking students' Task 2 papers, and their verbal reflections aligned with the key categories: Task Response, Organisation, and Grammar. As shown in Table 5, under Task Response, all teachers (100%) commented on whether students addressed the prompt fully, with clear reference to the criterion "Addressing the task (all/most/some parts)". Most teachers (80%) also emphasised the importance of supporting ideas with evidence, corresponding to the "Presenting a developed response with supported ideas" indicator. Statements such as "this point lacks support" or "examples strengthen the argument" were common. In terms of organisation, teachers frequently noted logical paragraphing and the use of discourse markers. The indicators "Organising ideas into individual paragraphs (logically)" and "Using a range of discourse markers appropriately" (both 100%) were often cited. However, "Arranging ideas coherently" (60%) was less frequently mentioned, suggesting this may receive less attention due to its lower weighting. For Grammar, all teachers consistently evaluated both structural variety and correctness. Full attention was given to indicators like "Considering the range of grammar mistakes" and "Considering punctuation errors" (both 100%). Teachers frequently identified grammatical issues such as verb tense errors, article misuse, and punctuation mistakes. Overall, the data suggested that teachers place greater emphasis on the more heavily weighted indicators in the rubric, guiding their assessment focus during paper evaluation.

Quantitative Data Analysis

To examine whether the use of the modified writing assessment rubric led to significant improvements in student writing performance, paired-samples *t*-tests were conducted comparing students' total writing scores under the original rubric (Current) and the modified rubric (Modified).

Table 6

Descriptive Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Current	50	12.10	2.30	0.33
Modified	50	13.94	2.13	0.30

Table 6 presents the average total writing score using the original rubric, which is 12.10 points, with a standard deviation of 2.30, indicating moderate variability in student performance under this assessment condition. However, applying the modified writing rubric, the mean score increased to 13.94 points, with a slightly lower standard deviation of 2.13, which suggests a consistent improvement in student scores across participants. This increase of approximately 1.84 points in the mean score explains that the modified rubric may have facilitated a more favourable or sensitive evaluation of student writing quality.

Table 7

Tests of Normality (Shapiro-Wilk for Difference Scores)

Statistic	df	Sig.
Diff-Total	0.977	50

Examining the distribution of the difference scores (Table 7), calculated by subtracting each student's score under the original criteria from their score under the changed rubric, was used to test the assumptions needed for parametric analysis, that is, normality. The Shapiro-Wilk test yielded a statistic of W = 0.977 with a non-significant p-value of 0.372, indicating no significant deviation from a normal distribution. Consequently, the data satisfied the premise of normality, thereby confirming the suitability of conducting a paired-samples t-test for the next inferential analysis.

Table 8

Paired Samples Correlations

Pair	N	Correlation	Sig.
Current & Modified	50	0.81	0.000

To assess the relationship between student scores under the original and modified writing assessment rubrics, a Pearson correlation analysis was conducted. The result (Table 8) illustrated a strong positive correlation between the total scores assigned using the original rubric and those assigned using the modified rubric, $r(48) = 0.81$, $p < 0.001$. This significant association implies that both scoring systems regularly showed individual variations in student writing performance. Stated differently, students who performed well under the original criteria often did so under the updated rubric, therefore proving the consistency and dependability of the score across situations. However, despite this consistency, the modified rubric yielded systematically higher scores, indicating its potential to capture improvements or nuances in student writing more sensitively.

Table 9

Paired Samples Test

	Mean	Std. Dev.	Std. Error Mean	95% CI Lower	95% CI Upper	t	df	Sig. (2-tailed)
Current - Modified	-1.84	1.38	0.20	-2.25	-1.43	-9.20	49	0.000

A paired-samples t-test was conducted to determine whether there was a statistically significant difference in students' total writing scores when assessed using the original rubric compared to the modified rubric. The analysis, in Table 9, revealed a significant increase in scores under the modified rubric (M = 13.94, SD = 2.13) relative to the original rubric (M = 12.10, SD = 2.30). The mean difference of 1.84 points was statistically significant, $t(49) = -9.20$, $p < 0.001$, with a 95% confidence interval for the mean difference ranging from -2.25 to -1.43. This finding suggests that the revised rubric may offer a more sensitive and comprehensive evaluation of student writing performance, as it produced notably better writing scores. Confirming that scores rose with the modified rubric, the negative t-value shows the direction of the difference—that is, modified minus current.

Table 10

Effect Size (Cohen's d)

Mean Difference	SD Difference	Cohen's d
1.84	1.38	1.33

In addition to statistical significance and to quantify the magnitude of the observed difference in writing scores between the original and modified rubrics, Cohen's d was calculated. Effect-size measures are essential in educational and language-assessment research because even small p-values can arise from trivial differences when sample sires are moderate or large (Plonsky & Oswald, 2014; Larson-Hall & Plonsky, 2015).

Table 10 shows the effect size, which was found to be d = 1.33, which is considered a very large effect according to conventional benchmarks (Cohen, 1988). This large effect size indicates that the improvement in student scores when using the modified rubric is not only statistically significant but also practically meaningful. In other words, the modified rubric substantially enhanced the assessment sensitivity, resulting in a pronounced increase in students' writing performance scores. Such a large effect underscores the potential educational impact of adopting the modified rubric in writing assessment practices. Moreover, it indicates that the revised rubric did not merely produce statistically higher scores but substantially altered the evaluation of student writing performance in a manner that is educationally meaningful.

Table 11

Differences in the Scores from Modified Rubric Criteria

Factor	Mean Difference	SD (approx)	Interpretation
Task Response	0.44	1.00	Smallest improvement
Organization	0.52	1.00	Moderate improvement
Grammar	0.68	1.10	Largest improvement
Vocabulary	0.60	1.05	Second largest improvement

Analysis of the differences between the original and modified rubrics across 50 students revealed systematic patterns in scoring changes. According to Table 11,

the Grammar component was identified as the largest mean improvement (mean difference = 0.68), indicating that the modifications to grammatical assessment criteria had the most substantial impact on student scores. It was followed by Vocabulary (mean difference = 0.60), Organisation (mean difference = 0.52), and Task Response (mean difference = 0.44). These findings suggest that the revised Grammar criteria in the modified rubric were particularly effective at capturing nuances in students' grammatical accuracy that the original rubric may have overlooked.

The significant increases in Vocabulary scores suggest that the amended rubric improved the evaluation of lexical resources and use as well. Although students' scores showed a light but satisfactory increase in Organisation and Task Response, the rather slight variations imply that the changes to these elements, albeit helpful, had a less clear impact on scoring results. The consistent pattern of improvement across all four assessment components supports the overall effectiveness of the modified rubric, with a mean total score increase of 2.28 points per student. This comprehensive enhancement across multiple dimensions of writing assessment demonstrates that the modified rubric provides a more sensitive and potentially more accurate evaluation of student writing performance, with particular strengths in assessing grammatical competence and vocabulary usage.

Discussion

This study, in the qualitative phase, examined the perceptions of EFL teachers at UTAS, Shinas and Al Mussanah, regarding the Level 2 writing evaluation rubric. It showed that teachers had a deep awareness of both the rubrics' pros and cons. The vague terms in the rubric, such as "some," "few," and "a little lack of clarity," were a major problem for 90% of the participants. They thought these terms made grading more subjective and inconsistent. Research has demonstrated that unclear criteria can cause raters to be untrustworthy (Barkaoui, 2007; Weigle, 2002). It shows how important it is to use operationalised descriptors to make sure that different raters agree with one another. Also, some said that expecting Band 5 students to write without mistakes was unreasonable for A2-level students. It aligns with the suggestions of Alderson (2005) and Fulcher and Davidson (2007), who propose that early-stage learners should not be punished for making mistakes that are part of their development. Some teachers also noted that the rubric encourages mechanical writing and the unnatural overuse of discourse markers. This finding aligns with other research on how form-focused rubrics can hinder genuine language production (Hamp-Lyons, 2007).

Although the rubric had some issues, instructors appreciated that it had a clear structure and included all the essential components of writing. They believed this made it fairer and more standardised, which Brown and Hudson (2002) stated was important for effective language evaluations. However, calls for improvements to rubrics, such as *adding task-specific criteria, indicators of creative and critical thinking, and annotated sample responses*, indicate that an increasing number of people in the field of assessment agree that rubrics need to be revised to support both teaching and accurate evaluation (Brookhart, 2018). There were also problems with

the context, especially with how the rubric seemed to be based on the IELTS, which may not accurately reflect the language and cultural realities of the local context. Shohamy (2001) supports this criticism and calls for localised assessments to make language testing fairer and more relevant.

The think-aloud data further supported the qualitative interview results, indicating that participants were often confused and inconsistent when using the original rubric. Teachers often stopped, asked questions about imprecise phrases, and admitted they were unsure when trying to match student performance with vague descriptors. These problems with thinking demonstrate that the rubric is unclear and does not align with the material being taught. This triangulated information strengthened the case for the requirement of a teacher-informed rubric that clarifies expectations, makes scoring more reliable, and supports writing evaluation based on strong teaching principles within the foundation context.

Another purpose of this study was to find out how adjustments to the criteria for writing evaluation affected the writing skills of Level 2 students. The main goal was to find out which parts of the rubric had the biggest effect on raising scores. The results showed that the new rubric made a considerable difference in overall writing scores. All four areas examined (Task Response, Organisation, Grammar, and Vocabulary) showed positive mean differences. The Grammar part showed the biggest average gain among the 50 students, followed by Vocabulary, Organisation, and Task Response. The observed improvements in student writing scores following the implementation of the modified rubric are consistent with a growing body of research emphasising the value of clear, detailed, and criterion-referenced rubrics in writing assessment (Brookhart, 2018; Jonsson & Svingby, 2007). Rubrics that articulate explicit expectations and performance descriptors have been shown to enhance both the reliability and validity of scoring, as well as to provide more actionable feedback for students (Andrade & Du, 2005; Panadero & Jonsson, 2013). In this study, the largest gains in the Grammar component suggest that the revised rubric provided more specific or nuanced criteria for evaluating grammatical accuracy, enabling teachers to recognise and reward improvements in this domain. This finding aligns with research by Lim and Sudweeks (2020), who found that targeted rubric modifications can lead to more sensitive and discriminating assessments of student writing features.

The substantial improvement in Vocabulary scores further supports the notion that the modified rubric enhanced the assessment of lexical resources and usage. Previous studies have highlighted the importance of vocabulary in academic writing, and the challenges teachers face in consistently evaluating this aspect (Nation, 2001; Read, 2000). The new rubric has solved these problems by making the vocabulary descriptors more precise. It has led to higher and more consistent marks for this part. This outcome is in line with Mosquera's (2017) emphasis on changes in rubrics for performance improvement.

Improvements in Organisation and Task Response, while present, were less pronounced than those observed for Grammar and Vocabulary. It may be due to the original rubric already providing sufficient clarity in these areas or the modifications

made being less substantial. Nonetheless, the positive mean differences across all components indicate that the rubric revision process was broadly effective. As Sadler (2009) notes, even incremental improvements in rubric clarity can have meaningful effects on both assessment outcomes and instructional alignment.

The focus group discussion also revealed that the assessors considered the new Level 2 writing rubric at UTAS to be clearer, more consistent, and better suited to the classroom goals. These reactions resonate with findings by Jonsson and Svingby (2007), who argue that providing precise descriptors for each assessment criterion promotes both objectivity and reliability and helps raters to interpret the performance indicators more consistently. Another appreciated change was the shift away from a proficiency test, i.e. IELTS, to assessment criteria that reflected A2 proficiency. Thus, when the criteria do not match the proficiency level exhibited by Level 2 learners, the scoring inference is either weak or invalid because the scores no longer accurately represent the target construct, i.e. A2 proficiency. It helped make local assessments more valid (Kane, 2000). These changes made the rubric easier to use, more relevant to teaching, and more transparent about how students are performing. The new rubric makes it easier for different people to agree on the meaning of the band descriptors and connects them directly to the course material. Additionally, it evaluates student writing more fairly.

Despite these strengths, assessors highlighted some room for refinement in several areas of the revised rubric. Nearly all participants were concerned about the potential for grade inflation, especially in the organisation and vocabulary sections of the rubric. It suggests that the modified band descriptors may inadvertently raise scores when a corresponding improvement in students' writing results might not be noticeable. Ghanbari and Barati (2020) expressed similar concerns in their rubric validation study, where they observed inflationary scoring patterns caused by imprecise descriptors in analytic rubrics. They emphasised the inclusion of clear examples and rater calibration to address this concern. Additionally, participants in this study believed the grammar and vocabulary bands should be more specifically defined to encompass all the learning outcomes taught at this level and earlier levels. It might explain the lack of a unified writing syllabus in all UTAS branches, or leaving the assessment of the language element in writing Level 2 relatively open to teachers' subjective perception of students' language ability in Level 2.

In summary, although all the assessors praised the revised rubric for its closer alignment with actual classroom instruction and delivery plan and recommended its application in level exit examinations, they underscored the need for continuous structured rating moderation sessions and student orientation to raise the awareness of both stakeholders. These observations highlight the vital role of professional development in establishing the reliable application of such analytic rubrics (Brookhart, 2018).

The current study also reveals that a significant challenge for UTAS Level 2 EFL teachers seeking to utilise rubrics effectively is the lack of unified and task-specific teaching resources. Because teachers must interpret vague or overly broad descriptors independently, this misalignment makes it challenging to apply

assessment criteria consistently and confidently. Studies have consistently shown that teachers struggle to use rubrics effectively if they are not based on a clear syllabus and supported by well-organised instructional materials (Brown & Harris, 2014; Davison & Leung, 2009). The new rubrics created in this study addressed these problems by using clearer language and being closely aligned with the writing syllabus. This integration helps teachers understand and clarify the rubric features more closely and makes testing more consistent. Brookhart and Chen (2015) support this claim that clear rubrics, which are aligned with the lesson plan, are more likely to be used correctly. The new rubric not only helps teachers feel more confident about their work, but it also compensates for the lack of a consistent syllabus, leading to fairer and more relevant assessments.

Conclusion

The results of this study have several important implications for writing assessment and pedagogy. First, they emphasise the crucial role of rubric design in influencing both teachers' scoring practices and student learning outcomes. The fact that the biggest improvements were shown in Grammar and Vocabulary shows how important it is to periodically review and improve rubric criteria to make sure they appropriately reflect the aims of the lesson and cover all the students' skills. Second, the study shows that even small adjustments to assessment methods can make a big difference in how fair and good student evaluations are. It is especially important in situations where writing tests are very important for getting a degree or for continuous studies.

Furthermore, the results support calls in the literature for increased teacher involvement in rubric development and revision (Brookhart, 2018; Panadero & Jonsson, 2013). Teachers who participate in the design and refinement of assessment tools are better equipped to interpret and apply criteria consistently, resulting in more reliable and valid assessments. It, in turn, benefits students by providing clearer guidance on how to improve their writing and achieve higher levels of performance.

Although the findings can enhance the assessment reliability and validity in the UTAS writing assessment system, future research should include more diverse branches and larger samples. Longitudinal studies examining how rubric modifications influence student progress over time and the integration of rubric use in teacher training programmes would also provide deeper insights into sustainable assessment improvements. Moreover, the future studies can include writing Task 1 in Level 2 or assess writing rubrics in levels 1, 3 and 4. Many teachers from other UTAS branches can be incorporated.

The findings offer a readily transferable, low-cost model for any multilingual similar programs worldwide that uses centrally designed rubrics with pre-intermediate learners. Institutions anywhere can replicate the proven three-session professional-development cycle (think-aloud scoring, collaborative identification and redrafting of problematic descriptors, followed by moderated re-scoring) to align institutional rubrics with local instructional realities and learner

proficiency levels. This project is being adopted in countries namely, Iran, India and Pakistan.

The authors confirm that no generative artificial intelligence (AI) tools or services were used in the drafting, writing, analysis, or revision of any part of this manuscript. The authors used Grammarly to check the language of the manuscript.

References

Abderrahmane, D., & Mebitil, N. (2025). AI and the pedagogical shift: Adaptability strategies for Algerian EFL teachers. *Logos Universality Mentality Education Novelty Social Sciences*, *14*(1), 1-17. https://doi.org/10.18662/lumenss/14.1/112

Alghizzi, T. M., & Alshahrani, T. M. (2024). Effects of grading rubrics on EFL learners' writing in an EMI setting. *Heliyon*, *10*(18), e36394. https://doi.org/10.1016/j.heliyon.2024.e36394

Alshakhi, A. (2019). Revisiting the writing assessment process at a Saudi English language institute: Problems and solutions. *English Language Teaching*, *12*(1), 176–185. https://doi.org/10.5539/elt.v12n1p176

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* Continuum.

Al-Saadi, Z., Khalil, H., Yousef, A. M. F. (2025). Exploring Omani EFL student teachers' perceptions on fostering critical thinking through ethical use of AI. *Education Process: International Journal, 17*(1), 12–25. https://doi.org/10.22521/edupij.2025.17.319

Andrade, H. L. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27-31. https://doi.org/10.3200/CTCH.53.1.27-31

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research, and Evaluation*, *10*(3), 1–11.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86–107. https://doi.org/10.1016/j.esp.2005.09.002

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (2018). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review, 67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565

Brown, G. T., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline learning research, 2*(1), 22–30. https://doi.org/10.14786/flr.v2i1.24

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43-56.

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*(3), 393–415. https://doi.org/10.1002/j.1545-7249.2009.tb00242.x

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Ghanbari, N., & Barati, H. (2020). Development and validation of a rating scale for Iranian academic writing assessment: A mixed-methods study. *Language Testing in Asia, 10*(1), 1–22. https://doi.org/10.1186/s40468-020-00112-3

Goodwin, R. (2019). Opportunities and questions: A short report on rubric assessments in Asia and the Middle East. *Arab World English Journal (AWEJ),10*. https://dx.doi.org/10.24093/awej/vol10no3.2

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Routledge.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternatives. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 487–504). Springer.

Hubias, A., & Muftahu, M. (2022). Internationalization of curriculum in Omani higher education: Perceptions of academic staff in UTAS. *International Journal of Higher Education, 11*(5), 134-144. https://doi.org/10.5430/ijhe.v11n5p134

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. T. (2000). Validity as an argument in educational assessment. *Measurement, 2*(2-3), 31–34.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4). https://doi.org/10.1177/0265532217710049

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*(S1), 127–159.

Le, X. M., Phuong, H. Y., Phan, T., Thao, L. T. (2024). Impact of using analytic rubrics for peer assessment on EFL students' writing performance: An experimental study. *Multicultural Education, 10*(3), 41–53. https://doi.org/10.5281/zenodo.7750831

Lee, I. (2009). Ten mismatches between teachers' beliefs and written feedback practice. *ELT Journal, 63*(1), 13-22. https://doi.org/10.1093/elt/ccn010

Jin, H. (2025). When AI meets source use: Exploring ChatGPT's potential in L2 summary writing assessment. *System, 133*. https://doi.org/10.1016/j.system.2025.103126

Li, J., & Lindsey, P. (2015). Understanding variations between student and teacher application of rubrics. *Assessing Writing, 26*(5), 67-79. https://doi.org/10.1016/j.asw.2015.07.003

Lim, J., & Sudweeks, R. (2020). Rubric revision and its impact on rater reliability and student performance. *Assessing Writing, 44*, 100450.

Mosquera, L. (2017). The impact of analytic rubrics on students' writing. *Profile Issues in Teachers' Professional Development, 19*(1), 149–159.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nurhayati, A. (2020). The implementation of formative assessment in EFL writing: A case study at a secondary school in Indonesia. *8* (2), 126. https://doi.org/10.32332/pedagogy.v8i2.2263

Pallant, J. (2020). *SPSS survival manual: A step-by-step guide to data analysis using IBM SPSS*. Routledge.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129-144.

Phuong, H. Y., Phan, Q. T., Thao, L. T. (2023). The effects of using analytical rubrics in peer and self-assessment on EFL students' writing proficiency: A Vietnamese contextual study. *Language Testing in Asia, 13*(1). https://doi.org/10.1186/s40468-023-00256-y

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into practice, 48*(1), 4–11. https://doi.org/10.1080/00405840802577536

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Reddy, Y. M., & Andrade, H. (2010). *A review of rubric use in higher education*. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment, Research, and Evaluation, 15*(1). https://doi.org/10.7275/psk5-mf68

Riddle, E. J., Smith, M., & Frankforter, S. A. (2016). A rubric for evaluating student analyses of business cases. *Journal of Management Education*, *40*(5), 595-618. https://doi.org/10.1177/1052562916644283

Sabermoghaddam Roudsari, S., Azabdaftari, B., & Seifoori, Z. (2024). The Intervention of Criteria-Referenced Self-Assessment in Developing the Accuracy, Lexical Resource, and Coherence of Advanced Iranian EFL Learners' Writing: Shared vs. Independent Tasks. *Journal of Applied Linguistics and Applied Literature: Dynamics and Advances*, *12*(2), 31-58. https://doi.org/10.22049/jalda.2024.28336.1523

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159-179. https://doi.org/10.1080/02602930801956059

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Pearson Education.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Yang, C., Zhang, L. J. (2023). *Think-aloud protocols in second language writing: a mixed-methods study of their reactivity and veridicality*. Springer.

Appendices

Appendix I

The Original Writing Rubric

Mark	TASK RESPONSE	ORGANISATION	GRAMMAR	VOCABULARY
5	Fully addresses all parts of the task. Presents a fully developed response with well supported ideas.	Logically organizes information and ideas into individual paragraphs. Uses a range of discourse markers appropriately. There is a clear progression of thought.	Uses the full range of structures required for the task with no grammatical or punctuation errors.	Correct choice of vocabulary and no spelling errors. The register or style consistently matches the nature of the task.
4	Sufficiently addresses all parts of the task. Presents a developed response with supported ideas.	Arranges information and ideas coherently and there is an overall progression. Uses discourse markers though there may be some incorrect use. Paragraphing may not be logical.	Uses most of the structures required for the task with only rare grammatical or punctuation errors which do not impede communication.	Only rare spelling errors or incorrect choice of vocabulary which do not impede communication. The register or style may in occasional instances not match the nature of the task.
3	Addresses most parts of the task. Presents ideas which may not be fully developed and/or lack focus.	Presents information with some organization but there may be a lack of overall progression. Uses only a limited number of discourse markers and there may be frequent incorrect use. Does not divide ideas into individual paragraphs.	Uses only some of the structures required for the task with some grammatical or punctuation errors which may impede communication.	Some spelling errors and incorrect choice of vocabulary which may impede communication. The register or style frequently does not match the nature of the task.

Mark	TASK RESPONSE	ORGANISATION	GRAMMAR	VOCABULARY
2	Addresses some parts of the task. Presents ideas which may be inadequately developed.	Ideas are not arranged coherently and there is no clear progression in the response. Uses very few discourse markers and their use is usually inaccurate or repetitive.	Uses very few of the structures required for the task with frequent grammatical or punctuation errors which usually impede communication.	Frequent spelling errors and incorrect choice of vocabulary which usually impede communication. The register or style usually does not match the nature of the task.
1	Addresses the task only minimally.	Has very little control of organizational features. Makes no attempt to use discourse markers.	The range of structures used is extremely limited and grammatical and punctuation errors are so prevalent that hardly any communication takes place.	Spelling errors and incorrect choice of vocabulary occurs to such an extent that hardly any communication takes place. The register or style does not match the nature of the task at all.

1. Give a zero if the task is not attempted, if the answer is totally incomprehensible or if the answer is plagiarized and there is evidence for this.
2. If the answer is totally irrelevant and unrelated to the task in any way, award a zero for the Task Response, and don't give more than 2 marks for Organization, 2 for Grammar and 1 for Vocabulary.
3. Candidates are penalized for writing fewer than the required number of words. The reduction is from Task Response as follows:
    - 60 words = -3; 61–100 words = -2; 101–140 = -1
4. If the word count is less than 50% do not award more than 3 marks for the other three criteria.
5. To receive the mark allocated for a criterion, all positive features mentioned in the descriptors should be achieved.

Appendix II

The Modified Writing Rubric

Mark	TASK RESPONSE	ORGANIZATION	GRAMMAR	VOCABULARY
5	Fully addresses all parts of the task. Presents a fully developed response with well supported ideas.	Logically organizes information and ideas into individual paragraphs. **The topic sentence and the supporting sentences form logically- and coherently-developed paragraph.** Uses a range of discourse markers listed in the syllabus (and, but, so, because) appropriately. There is a clear progression of thought.	Uses the full range of structures **required for the task (present simple, present continuous, adverbs of frequency)** with no grammatical or punctuation (dot, comma, capitalization) errors.	Correct choice of vocabulary **(the ones taught for this writing task)** and no spelling errors. The register or style consistently matches the nature of the task.
4	Sufficiently addresses all parts of the task. Presents a developed response with supported ideas.	Arranges information and ideas coherently and there is overall progression. **The topic sentence and the supporting sentences are well-developed.** Uses discourse markers listed in the syllabus (and, but, so, because) though there may be some incorrect use. Paragraphing may not be logical.	Uses most of the structures **required for the task (present simple, present continuous, adverbs of frequency)** with only rare grammatical or punctuation (dot, comma, capitalization) errors which do not impede communication.	Only rare spelling errors or incorrect choice of vocabulary **(the ones taught for this writing task)** which do not impede communication. The register or style may in occasional instances not match the nature of the task.
3	Addresses most parts of the task. Presents ideas which may not be fully developed and/or lack focus.	Presents information with some organization but there may be a lack of overall progression. **The topic sentence and supporting sentences are partially developed but do not represent a well-developed paragraph.** Uses only a limited number of discourse markers listed in the syllabus (and, but, so, because) and there may be frequent incorrect use. Does not divide ideas into individual paragraphs.	Uses only some of the **structures required for the task (present simple, present continuous, adverbs of frequency)** with some grammatical or punctuation (dot, comma, capitalization) errors which may impede communication.	Some spelling errors and incorrect choice of vocabulary **(the ones taught for this writing task)** which may impede communication. The register or style frequently does not match the nature of the task.

Mark	TASK RESPONSE	ORGANIZATION	GRAMMAR	VOCABULARY
2	Addresses some parts of the task. Presents ideas which may be inadequately developed.	Ideas are not arranged coherently and there is no clear progression in the response. **The topic sentence or/and supporting sentences are not coherently developed.** Uses very few discourse markers listed in the syllabus (and, but, so, because) and their use is usually inaccurate or repetitive.	Uses very few of the structures required for the task **(present simple, present continuous, adverbs of frequency)** with frequent grammatical or punctuation (dot, comma, capitalization) errors which usually impede communication because of lack of grammatical elements.	Frequent spelling errors and incorrect choice of vocabulary **(the ones taught for this writing task)** which usually impede communication. The register or style usually does not match the nature of the task.
1	Addresses the task only minimally.	Has very little control of organizational features **to make the topic sentence and develop the related supporting ideas.** Makes no attempt to use discourse markers listed in the syllabus (and, but, so, because).	The range of structures (present simple, **present continuous, adverbs of frequency)** used is extremely limited and grammatical and punctuation (dot, comma, capitalization) errors are so prevalent that hardly any communication takes place.	Spelling errors and incorrect choice of vocabulary **(the ones taught for this writing task)** occur to such an extent that hardly any communication takes place. The register or style does not match the nature of the task at all.

1. Give a zero if the task is not attempted, if the answer is totally incomprehensible or if the answer is plagiarized and there is evidence for this.
2. If the answer is totally irrelevant and unrelated to the task in any way, award a zero for the Task Response, and don't give more than 2 marks for Organization, 2 for Grammar and 1 for Vocabulary.
3. Candidates are penalized for writing fewer than the required number of words. The reduction is from Task Response as follows:
    - 60 words = -3;
    - 61–100 words = -2;
    - 101–140 = -1
4. If the word count is less than 50% do not award more than 3 marks for the other three criteria.
5. To receive the mark allocated for a criterion, all positive features mentioned in the descriptors should be achieved.

- Task Response: The writing is not off topic and covers all areas of the question content.

Authors' Biography

Zahra Zargaran holds a Ph.D. in TEFL from Azad University, Science and Research Branch, Tehran, Iran. She is a lecturer at University of Technology and Applied Sciences, Shinas, Oman and is also the coordinator of Academic Advising Committee. She is a certified trainer by the British Council, IDP Australia, and Cambridge and has actively trained hundreds of teachers in both public and private sectors nationally and internationally since 2011. With 19 years of experience in teaching and teacher training, Dr. Zargaran specializes in language education, curriculum design, materials development, and professional skills training. She has conducted many research training workshops and courses and is an active researcher in the areas of language teaching/learning, educational assessment, teacher education, learning strategies, teacher cognition, AI in education and IELTS teaching.

Mohsen Ghorbanpoor received his B.A. and M.A. degrees in English Language Teaching from Kharazmi University and Tarbiat Modares University, Tehran, respectively. He began his teaching career in 2010 and has taught English in both public education and private institutes in Iran. In 2020, he joined the University of Technology and Applied Sciences in Al Mussanah, Oman, where he currently serves as an ELT lecturer and the coordinator for professional development and research. His areas of interest include teacher education, materials development, listening, writing and digital pedagogy. He has also conducted in-house training programs and coordinated institutional research activities.