Evaluation of Technical Description Writing: An Assessment for ESP Learners in Engineering Programs

SAMIA NAQVI
Center for Foundation Studies, Middle East College, Oman
Author email: snaqvi@mec.edu.om

Article information

Article history:
Received: 17 Feb 2025
Last revised: 23 Dec 2025
Accepted: 28 Dec 2025
Available online: 30 Dec 2025

Keywords:
Test validation
English for specific purposes
Technical description
Validity
Reliability
Practicality
Authenticity
Interactiveness
Impact

Abstract

This paper reports an empirical evaluation of a CBT (Closed Book Test) designed to assess technical description writing skills among first-year engineering students enrolled in an English for Specific Purposes (ESP) module. Grounded in Bachman and Palmer's (1996) test usefulness framework, the study examines the assessment in terms of its validity, reliability, practicality, authenticity, interactiveness, and impact. The CBT required students to produce a written description of an electronic object, using appropriate terminology, critically evaluating the product, and suggesting improvements. Test development involved content expert validation, internal and external moderation, and alignment with ESP module outcomes. Data were collected through test scripts from the entire student cohort (N = 34), expert CVI ratings, post-test survey responses (Likert-scale and open-ended items), and moderators' comments. Analysis included blind marking of all test scripts by two examiners using a standardised analytic rubric, paired samples t-test for inter-rater reliability (p = 0.163), and exploratory factor analysis for construct validity. The mixed-methods approach combined quantitative analysis (survey ratings, statistical tests) with qualitative analysis of open-ended survey responses and moderator feedback. The post-test student survey across all six usefulness dimensions yielded consistently high mean scores (4.1–4.5). The evaluation confirmed the CBT's overall test usefulness across all six dimensions through multiple validation methods, with 85% of students affirming its effectiveness in improving their technical writing skills. Limitations include the small sample size, single-institution context, and potential response bias. Future research should focus on scaling the CBT model across institutions and disciplines, implementing hybrid automated scoring systems, refining rubric analytics, and conducting longitudinal studies to examine skill transfer to professional contexts.

INTRODUCTION

English for Specific Purposes (ESP) is a significant branch of English language instruction that equips learners with language skills tailored to their specific professional, academic, or technical contexts. Unlike General English courses, ESP addresses learners' immediate language needs, often shaped by their current or future professional environments. Dudley-Evans and St. John

(1998) describe ESP as goal-oriented language instruction designed to enable learners to use English effectively within a particular field, such as engineering, medicine, or business.

Clear and precise communication is crucial in technical fields such as engineering. Engineers are required to write technical reports, manuals, product specifications, and other documents that demand accuracy, clarity, and the correct use of technical terminology. Hyland (2006) emphasises that technical communication involves not only linguistic competence but also mastery of the genres and conventions of professional writing. Therefore, ESP courses in technical disciplines must focus on developing these genre-specific writing skills.

Malmström et al.'s (2018) study emphasises the importance of tailoring vocabulary instruction in ESP courses to meet the specific needs of students across various disciplines. Their findings contrast the receptive and productive academic vocabulary needs of university students, highlighting the importance of discipline-specific language instruction in fields such as engineering.

Despite a growing body of literature on ESP testing frameworks, there remains a notable lack of empirical evaluation of how these principles are implemented in authentic classroom contexts, particularly in technical disciplines like engineering. While several studies discuss the theoretical underpinnings of effective language testing (e.g., Bachman & Palmer, 1996; Douglas, 2000; Hyland, 2006), few provide a practical lens through which real-world assessments are analysed. This knowledge gap is filled by this empirical study, which adopts a mixed-methods evaluative design in ascertaining the effectiveness of the CBT utilised in an ESP module for engineering students.

Guided by Bachman and Palmer's (1996) six-component test usefulness framework—validity, reliability, authenticity, interactiveness, impact, and practicality—the study draws on a combination of student performance data, classroom observations, institutional assessment procedures, and student survey responses. By triangulating these data sources, the study provides a comprehensive understanding of how theoretical assessment constructs are implemented in an applied ESP context, offering insights for curriculum design, assessment practice, and future research in technical education.

Research question

In light of the increasing emphasis on authentic and effective assessment in ESP contexts, this study is guided by the following research question:

To what extent does the Closed Book Test (CBT) implemented in an English for Specific Purposes (ESP) module for engineering students demonstrate the six components of test usefulness proposed by Bachman and Palmer (1996)—validity, reliability, authenticity, interactiveness, impact (washback), and practicality?

LITERATURE REVIEW

This section begins by exploring the increasing significance of ESP in undergraduate curricula and technical writing, then shifts to focus on language testing in ESP and the validation of ESP assessments, with particular emphasis on Bachman and Palmer's (1996) framework.

Technical writing in ESP

The necessity for developing language proficiency relevant to specific academic and professional contexts has led to the rapid growth of ESP as a discipline in English language instruction. ESP is distinguished from regular English language courses by its emphasis on task authenticity, in which students complete practical exercises that mirror the types of communication they would encounter in their future professional settings (Dudley-Evans & St. John, 1998). In his thorough review of contemporary ESP theory and practice, Anthony (2018) highlights the significance of needs analysis and genre-based approaches in ESP curriculum design. This supports Hyland's (2006) assertion that ESP is growing in significance in a globalised society where English is frequently used for professional and technical communication. ESP courses are designed to prepare learners for discipline-specific communication tasks, such as writing technical reports, giving presentations, or participating in professional meetings.

Technical writing is one of the most critical skills that students in technical fields, such as engineering, must develop. Adams (2014) and Dobrin (2019) emphasise that technical writing requires the ability to describe objects, processes, and systems with precision and clarity. In engineering, technical writing skills are crucial for documenting how machinery works, writing product specifications, and providing instructions for assembly or repair. Hyland and Jiang (2017) note that effective technical writing necessitates not only linguistic accuracy but also a thorough understanding of the conventions and genres of professional communication. They also highlight the requirement of a high degree of precision and clarity in technical reports prepared by engineers.

In addition to precise language, visual aids play an essential role in enhancing technical writing comprehension and accuracy. The use of visuals such as diagrams and charts can help clarify complex descriptions and make technical writing more accessible to readers (Bobek & Tversky, 2016; Carifio & Perla, 2009). Hence, students should be provided with images of objects to facilitate their writing. This helps students not only to conceptualise but also to describe objects with greater accuracy and relevance.

A key challenge in technical writing lies in maintaining the right balance between clarity and conciseness. Bazerman (1988) supports this view, arguing that engineering and technical documents should be clear enough for non-expert readers while still preserving the precise terminology and discipline-specific language required for professional communication. Furthermore, the task of evaluating the product underscores the shift in ESP writing assessment towards analytical writing. Analytical tasks require students to apply their linguistic as well as critical thinking skills in a professional context, which mirrors real-world engineering practices. Critical thinking is a crucial skill that all academic disciplines must incorporate; however, teaching individuals to become critical thinkers can be challenging. It can be achieved through the application of effective strategies and classroom techniques (Rashtchi & Khoshnevisan, 2020). As ESP students prepare for careers where they will need to make design recommendations, integrating evaluation tasks within assessments like this CBT aligns the learning process with practical, critical thinking and industry-relevant skills.

Test validation in ESP

Assessment plays a vital role in developing ESP writing skills by measuring individual progress, identifying strengths and weaknesses, and guiding corrective actions (Alqurashi, 2022; Rachmawati & Hastari, 2022;). However, designing language tests for ESP learners presents unique challenges, as these tests must assess both general language proficiency and field-specific communication skills. This dual focus is essential because ESP learners are expected to function in professional or academic settings where both everyday communicative competence and discipline-specific language are required (Douglas, 2000; Dudley-Evans & St John, 1998). For instance, an engineering student must be proficient in general writing skills as well as in describing technical processes clearly and accurately. To ensure such assessments are fit for purpose, ESP teachers often design tests tailored to their learners' language needs, professional goals, and content-area knowledge, enhancing the overall utility of the test (Korolyova, 2017).

Sèna (2022) identifies two primary challenges: developing assessment instruments that measure critical aspects of ESP writing (appropriate language, content understanding, idea organization, and genre usage) and ensuring objective, consistent assessment to avoid personal bias. These challenges require educators to develop valid and reliable assessment instruments while considering specific contexts and adopting student-centered learning approaches.

Bachman and Palmer's test usefulness framework

One comprehensive approach to tackling these issues is Bachman and Palmer's (1996) Test Usefulness Framework, which has become central to evaluating language assessments in ESP. Their model evaluated test quality across six dimensions—validity, reliability, authenticity, interactiveness, impact, and practicality—offering a theoretical foundation to guide both the development and evaluation of effective ESP assessments (see Figure 1).

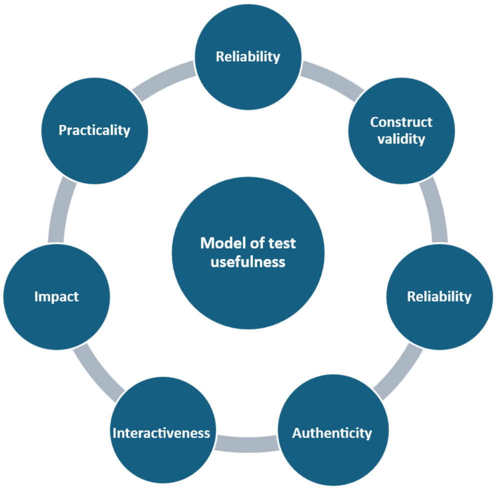


Figure 1 Bachman and Palmer's model of test usefulness framework

Building on Bachman and Palmer's (1996) emphasis on validity, other scholars such as Fulcher and Davidson (2007) and Hughes (2020) also highlight its central role in language testing. Messick's (1989) unified validity framework advances this understanding by integrating content, construct, and consequential validity, moving beyond the treatment of these components as isolated domains. Supporting Bachman and Palmer's view that authenticity enhances test usefulness, Knoch and Macqueen (2020) stress the importance of task authenticity in professional language assessments. Similarly, Douglas (2000) and Dudley-Evans and St. John (1998) argue that ESP assessments must reflect real-world workplace communication needs, a principle applied in the CBT's product description task. Reliability, another essential aspect of test usefulness, presents challenges in ESP writing due to subjective judgment (Weigle, 2002). Adding to this, Winke and Lim (2017) examine how test preparation influences performance, with their findings reinforcing the importance of designing ESP assessments that consider washback effects on student learning strategies.

Reliability remains a critical concern in ESP writing assessments, particularly due to the subjectivity involved in evaluating both linguistic and content-specific features. In technical writing tasks—such as describing a product—raters must assess accuracy in both language use and technical content, increasing the potential for inter-rater variation (Knoch, 2009; Weigle, 2002). This dual demand complicates scoring and highlights the need for standardised evaluation practices. To enhance reliability, researchers advocate the use of analytic rubrics, double marking, and rater training (Knoch & Elder, 2010), which help minimise subjectivity and promote consistency. Additionally, statistical tools such as inter-rater correlation and paired samples t-tests are employed to verify scoring agreement and validate assessment outcomes (Naqvi et al., 2023). These measures underscore the importance of structured, transparent evaluation systems in maintaining the credibility and fairness of ESP assessments in high-stakes academic contexts.

In summary, Bachman and Palmer's (1996) framework offers a comprehensive approach to designing useful ESP assessments, while other scholars emphasize the importance of task authenticity, scoring consistency, and the broader effects of assessment on both learning and instruction. However, while theoretical frameworks are well-established, empirical studies that systematically apply these models to real-world ESP assessment tasks—particularly within engineering or technical education contexts—are limited. This study seeks to bridge this gap by evaluating an authentic CBT used in an ESP engineering module, thereby offering practical insights into how theoretical principles manifest in classroom-based assessments.

METHODOLOGY

This section outlines the research design, participants, data collection tools, and procedures used to evaluate the assessment based on Bachman and Palmer's (1996) test usefulness framework.

Research design

This empirical study employs a mixed-methods evaluative design to investigate the effectiveness of a Closed Book Test (CBT) administered within an ESP module for engineering students. Anchored in Bachman and Palmer's (1996) six-component framework of test usefulness, the methodology integrates both qualitative and quantitative data sources. These include institutional assessment records, test artifacts, and student survey responses. By triangulating these sources, the study aims to provide a comprehensive and context-sensitive evaluation of how the CBT under discussion functions as an ESP assessment tool in a real-world academic setting.

Participants

The participants in this evaluation were first-year engineering students enrolled in an ESP course at a university college in the Sultanate of Oman. The cohort included 34 students (20 males and 14 females). Approximately 90% of the learners were Omani Arabs, while the remaining 10% were from other nationalities. Most students had completed a one-year English Foundation course comprising three levels: Pre-Elementary (Level 1), Elementary (Level 2), and Intermediate (Level 3). Placement into these levels was based on the institutional placement test scores: 0–24 for Level 1, 25–44 for Level 2, and 45–59 for Level 3. Students scoring 60 and above were either granted direct entry into the academic semester or placed in Semester 1 if they achieved an IELTS score of 5.5 or its equivalent. Some students entered directly into their degree programmes by meeting one of these criteria. The students were enrolled in Mechanical, Civil, Electrical, and Electronics engineering programmes.

ESP module and Technical Description (TD) writing practice

The ESP module aims to develop students' discipline-specific communicative competence and technical writing skills. It was designed after a thorough needs analysis and input from subject specialists. They were also consulted for material design and assessments. The resource content and respective activities align with the module outcomes: 1) Prepare students to communicate effectively in written and oral forms; 2) Encourage critical thinking and reasoning skills; 3) Train students to deal with their professional environment.

The unit on Technical Description (TD) spans three weeks. The students receive intensive practice in writing TDs during the first three weeks, before the CBT in week four. To prepare the students to write the TD, they are oriented to the specific jargon and language used in English for Science and Technology (EST) (Swales, 1990). Although the sentence structure in EST is not very different from general English, there is a tendency to favour specific forms, for instance, present simple, passive voice and nominal compounds (Ewer & Latorre, 1969). Therefore, prior to writing full descriptions, the students are given practice through worksheets on shapes, dimensions, materials used in making devices, use of passive voice, nominal compounds and phrasal verbs. Once the students are given practice in the required target language through pre-, in, and post-class activities, the structure of TD is introduced.

Test design and administration

The CBT evaluated in this study was developed as part of the ESP module and was designed to assess students' ability to compose technical descriptions, use technical terminology and evaluate the object independently. It is a summative achievement assessment that falls under the category of the achievement test. The students were trained in writing technical descriptions during the first three weeks of the fourteen-week semester, and the CBT was conducted in the fourth week. The CBT is worth 20% of the overall course grade and requires students to describe an electronic object in 350–400 words— the Philips Hair Straightener HP469—in this instance. The test is structured into two parts:

Part 1: Technical Description – Students were asked to describe the physical details, including materials of construction, shape, colour, texture, and dimensions, and functions of the Philips Hair Straightener.

Part 2: Product Evaluation – Students were required to evaluate the design and performance of the product and provide suggestions for improvement, focusing on user-friendliness, durability, and potential enhancements.

The descriptions were typed by the students on computers in the computer lab in a proctored environment. To assist students in visualising the object and accurately describing its features, visual prompts, including images of the Philips Hair Straightener, were provided. The time allotted was 90 minutes. To ensure security, the CBT was configured on the exam server of the institutional Moodle-based online learning platform called MECLearn. The students used their institutional log-in IDs and passwords to access the computers. Additional passwords were provided to access the CBT. The CBT was graded manually using a standardised grading rubric that was uploaded to MECLearn.

Student survey

A structured questionnaire was administered to all 34 participants within one week of completing the CBT to capture student perceptions of test usefulness. The survey was designed to align with Bachman and Palmer's six-component framework, including questions about validity (whether the test measured intended skills), reliability (test fairness and consistency), authenticity (relevance to real engineering tasks), interactiveness (level of engagement required), impact (influence on learning), and practicality (test administration and accessibility). The survey included both Likert scale items and open-ended questions to gather quantitative and qualitative feedback on student experiences.

Descriptive statistics were calculated for student survey responses across the six dimensions of Bachman and Palmer's framework. Mean scores and standard deviations were computed for each dimension to summarize students' perceptions of the CBT's usefulness. Frequency distributions were also generated to show student responses on the Likert scale items, providing a clear overview of how students evaluated each aspect of the assessment.

Evaluation procedure using Bachman and Palmer's framework

Validity evaluation

To obtain quantified content validity evidence, five subject matter experts in ESP and technical writing independently rated each test component on a 4-point relevance scale (1 = not relevant to 4 = highly relevant). Content Validity Index (CVI) was calculated as the proportion of experts rating items as relevant (scores of 3–4), with acceptable thresholds set at ≥ 0.78 for individual items and ≥ 0.80 for the overall scale. The student response on the validity of the CBT via survey was also considered.

A formal moderation process involving internal and external reviewers through the automated Content Management System (CMS) was undertaken. Moderators were provided with structured guidelines to assess the task on five key aspects: clarity of task instructions, alignment with students' proficiency level, appropriateness of content and language, authenticity of the task, and alignment with module learning outcomes. Moderators submitted written feedback on these criteria, and the question paper (QP) and answer key (AK) were revised where needed.

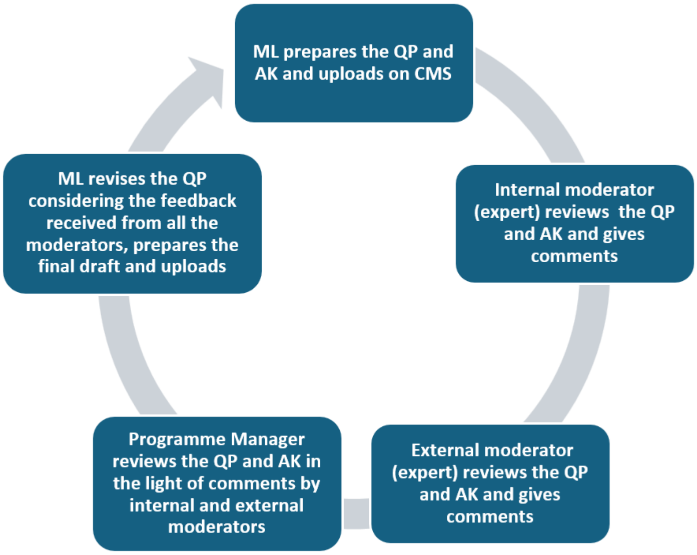

ML prepares the QP and AK and uploads on CMS
Internal moderator (expert) reviews the QP and AK and gives comments
External moderator (expert) reviews the QP and AK and gives comments
Programme Manager reviews the QP and AK in the light of comments by internal and external moderators
ML revises the QP considering the feedback received from all the moderators, prepares the final draft and uploads

Figure 2 Moderation of question paper and answer key

Construct validity was examined through exploratory factor analysis despite the small sample size (*n* = 34), following procedures outlined by Tabachnick et al. (2019).

Additionally, construct validity was assessed by mapping the CBT components against the learning outcomes and instructional content of the ESP module. These outcomes focused on students' ability to use discipline-specific vocabulary, demonstrate grammatical accuracy, describe technical features, and think critically—all of which were embedded in the CBT design. This process helped ensure that the test measured the intended skills without construct under-representation or irrelevant content.

Reliability evaluation

Following institutional assessment and moderation policy, a minimum of 12 scripts or 12% of submissions per cohort—whichever is higher—are typically second-marked. If discrepancies are found, the entire bundle is reassessed by the original marker. However, as this was part of a research study focused on evaluating inter-rater reliability, all 34 scripts were independently marked by two experienced examiners. Both examiners used the same standardised analytic rubric, which included criteria for evaluating technical descriptions, vocabulary use, structure, and clarity. Discrepancies were operationally defined as > 5 marks difference. None exceeded this threshold; hence, no re-marking or arbitration was required. A paired samples *t*-test was performed to compare inter-rater agreement. Student perceptions of test fairness were captured via student survey.

Practicality evaluation

Practicality was evaluated through systematic documentation of test administration, marking workflows, and platform functionality. Administration procedures were monitored to record time requirements for test setup, duration, and invigilation needs during the 90-minute testing session. Marking workflows were documented to capture time per script and total marking duration, including moderation process efficiency. Resource requirements such as staffing needs, facility booking, and associated costs were tracked throughout the assessment process. Academic integrity measures were observed and documented, including security protocols, invigilation procedures, and system access controls. MECLearn platform functionality was monitored to assess system usability, submission processes, and technical performance. Student survey data were collected to capture perceptions of platform ease of access and test administration experience.

Authenticity evaluation

Authenticity was assessed by comparing the CBT task to real-world technical documentation practices. The use of a real consumer product and accompanying visuals was reviewed in relation to genre conventions. Inputs from subject specialists from the engineering discipline during the module design process were also considered to assess alignment with discipline-specific communicative needs and professional writing requirements in engineering contexts. Student perceptions of task relevance to their respective fields were captured through survey responses.

Interactiveness evaluation

In Bachman and Palmer’s (1996) framework, interactiveness refers to the extent to which a test engages participants’ knowledge, communicative strategies, and interest in the task. In ESP assessments, this concept is particularly significant because students are expected to apply discipline-specific language in professional contexts. The CBT required students to integrate language skills and technical content knowledge to describe and evaluate a real product, thus promoting high interactiveness. This combination of cognitive and linguistic engagement supports construct validity and mirrors real-world communicative demands (Douglas, 2000; Dudley-Evans & St. John, 1998). To evaluate this dimension, a student survey was administered to assess how far the task required them to apply both content understanding and language use. Tasks that promote such active engagement contribute to test usefulness by making assessments more relevant, engaging, and authentic for learners (Flowerdew, 2016).

Impact (Washback)

To evaluate the washback effect, the study examined instructional materials, pre-assessment in-class activities and mock tests. The focus was on identifying the extent to which the CBT influenced teaching methods, classroom focus, and student engagement. The study also considered the alignment between assessment requirements and the content delivered through instructional materials and preparatory exercises. Informal teacher observations and a student survey aligned with Bachman and Palmer’s (1996) framework were used to capture student perceptions regarding the test’s impact on their confidence, learning strategies, and perceived relevance of the task to future professional use.

FINDINGS AND DISCUSSION

This section discusses the evaluation of the assessment considering the six aspects that Bachman and Palmer (1996) consider as the cornerstones of test usefulness: validity, reliability, practicality, authenticity, interactiveness and impact.

Validity

Let us first examine the validity since it is considered the central concept in testing and assessment (Fulcher & Davidson, 2007).

Content validity

Five subject matter experts in ESP and technical writing independently evaluated the CBT components on a 4-point relevance scale (1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, 4 = highly relevant), following established CVI procedures (Polit & Beck, 2006). The expert panel comprised professionals with extensive experience in ESP curriculum development and assessment design.

Table 1

Content validity index results for CBT components

CBT Component	Expert Ratings	I-CVI	Interpretation
Part 1: General Description Task	4, 4, 3, 4, 4	1.00	Excellent
Technical Specifications Analysis	4, 3, 4, 4, 3	0.80	Acceptable
Part 2: Component Analysis Task	4, 4, 4, 3, 4	1.00	Excellent
Three-Part Structure Requirement	4, 3, 4, 4, 4	1.00	Excellent
Technical Vocabulary Assessment	4, 4, 3, 4, 4	1.00	Excellent
Part 3: Evaluation and Suggestions	3, 4, 4, 3, 4	0.80	Acceptable
Critical Analysis Component	4, 3, 4, 4, 3	0.80	Acceptable
Word Count Requirements (100–150)	3, 4, 4, 3, 4	0.80	Acceptable
Word Count Requirements (400–500)	3, 3, 4, 4, 3	0.60	Needs Revision
Word Count Requirements (75–100)	3, 3, 3, 4, 3	0.20	Poor

Note: Scale-Level CVI = 0.82 (Above acceptable threshold of 0.80; Polit & Beck, 2006)

The content validity analysis revealed strong expert consensus on most CBT components, consistent with established CVI interpretation guidelines (Davis, 1992). The overall scale-level CVI of 0.82 exceeded the acceptable threshold, indicating good content validity (Polit & Beck, 2006). Individual tasks demonstrated excellent validity: the general description task (I-CVI = 1.00), component analysis task (I-CVI = 1.00), and technical vocabulary assessment (I-CVI = 1.00). However, word count restrictions, particularly for Parts 2 and 3, received lower ratings, with experts noting that rigid word limits may not reflect authentic professional writing contexts where content completeness takes precedence over arbitrary length constraints (Hyland, 2019).

Subject specialist input analysis

Three engineering faculty members and two ESP specialists provided systematic feedback during the module design phase regarding CBT authenticity and professional relevance. Their input validated the appropriateness of technical terminology, task complexity, and alignment with real-world engineering communication practices. Key contributions included confirmation of the suitability of the technical object for description, validation of the evaluation criteria as reflective of engineering design thinking, and endorsement of the vocabulary and structural requirements as professionally relevant.

Moderation and alignment analysis

As represented by Figure 2, the moderation process of the QP and AK ensures that the test content corresponds with the module learning outcomes and is appropriate in terms of readability, formatting, language used, suitability with students' proficiency level, relevance to the learning context, and usefulness. The formal moderation process involved structured assessment across these five dimensions by both internal and external reviewers through the institutional CMS.

Table 2 provides a summary of the moderators' feedback in light of five key aspects, including clarity of instructions, level appropriateness, language appropriateness, task authenticity and constructive alignment. Both internal and external moderators provided positive feedback as well as suggestions for improvement. The final version incorporated their suggestions before CBT was administered.

Table 2

Moderation feedback summary

Criterion	Internal Moderator Feedback	External Moderator Feedback	Action Taken
Clarity of Instructions	"Instructions are clear and well-structured. Minor suggestion: add time allocation for each part."	"Task instructions are comprehensible. Consider providing word count guidance for each section."	Added suggested time allocation (60 min Part 1, 30 min Part 2)
Proficiency Level Alignment	"Appropriate for first-year engineering students. Vocabulary level matches foundation course outcomes."	"Task complexity aligns well with intermediate-level learners. Good scaffolding is evident."	No changes required
Content Appropriateness	"Content is relevant and authentic. Good choice of everyday technical object."	"Excellent selection of product for description. Terminology is accessible yet technical."	No changes required
Task Authenticity	"Mirrors real-world engineering communication needs effectively."	"Strong alignment with professional writing requirements. Evaluation component has added value."	No changes required
Learning Outcome Alignment	"Clear mapping to module outcomes visible. Covers all required skills."	"Good constructive alignment. Suggestion: explicitly state assessment criteria in rubric."	Enhanced rubric descriptors for clarity

Construct validity

Factor analysis results

Construct validity was examined through exploratory factor analysis (EFA) to explore the underlying structure of the test components and assess alignment with the theoretical constructs of the CBT. Despite the relatively small sample size (*n* = 34), the analysis was conducted following established procedures outlined by Tabachnick et al. (2019). While this sample size is below the recommended minimum of 100 participants for stable factor solutions (Comrey & Lee, 2013; Hair et al., 2019), the statistical outputs provided preliminary evidence supporting the test's intended structure. These findings offer an initial indication of construct alignment, though further validation using a larger and more diverse sample is recommended to ensure robustness and generalisability.

Table 3

Factor analysis summary statistics

Measure	Value	Interpretation	Reference
Kaiser-Meyer-Olkin (KMO)	0.62	Marginally adequate	Kaiser (1974)
Bartlett's Test of Sphericity	$\chi^2 = 156.78$, $p < 0.05$	Significant	Bartlett (1954)
Sample Adequacy	n = 34	Below recommended minimum	Hair et al. (2019)

Table 4

Three-factor solution with varimax rotation

Factor	Eigenvalue	% Variance	Cumulative %	Factor Name
1	2.18	36.3%	36.3%	Technical Description Skills
2	1.82	30.3%	66.6%	Language Accuracy
3	1.01	16.8%	83.4%	Critical Evaluation

Table 5

Factor loadings for CBT components

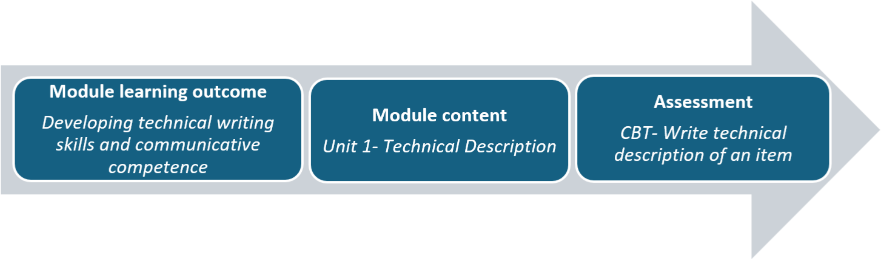
Test Component	Factor 1	Factor 2	Factor 3
General Description Task	**0.74**	0.32	0.18
Component Analysis	**0.68**	0.41	0.25
Technical Vocabulary Use	**0.71**	0.38	0.22
Specification Interpretation	**0.66**	0.29	0.31
Grammar Accuracy	0.31	**0.78**	0.19
Sentence Structure	0.28	**0.72**	0.34
Coherence and Cohesion	0.35	**0.69**	0.28
Register Appropriateness	0.42	**0.64**	0.21
Evaluation Quality	0.24	0.31	**0.73**
Suggestion Feasibility	0.19	0.28	**0.68**
Critical Thinking	0.33	0.25	**0.65**

Note: Bold values indicate primary factor loadings ≥ 0.60 (Stevens, 2012)

The exploratory factor analysis suggested a three-factor structure (see Table 3) explaining 83.4% of total variance, aligning with the theoretical framework underlying the CBT and consistent with multivariate analysis principles (Hair et al., 2019). Factor 1 (TD Skills) captured students' ability to describe physical features and interpret specifications. Factor 2 (Language Accuracy) reflected grammatical competence and appropriate register use. Factor 3 (Critical Evaluation) represented analytical thinking and suggestion formulation. However, given the small sample size ($n = 34$), these results should be interpreted cautiously, and validation with a larger sample ($n \geq 100$) is recommended for reliable construct validation (Comrey & Lee, 2013; MacCallum et al., 1999).

Constructive alignment analysis

Regarding the test construct, the CBT assesses the required macro and micro skills covered in the unit on technical description, which includes describing physical and quantifiable details, grammatical competence, lexical range, item evaluation and suggestions for improvement, thus establishing the constructive alignment between the module outcomes, aims and assessment (see Figure 3). This also confirms that there is no threat to validity from construct under-representation or construct irrelevance (Messick, 1989).


Module learning outcome
Developing technical writing skills and communicative competence

Module content
Unit 1- Technical Description

Assessment
CBT- Write technical description of an item

Figure 3 Constructive alignment

The CBT avoids construct under-representation by comprehensively assessing both macro skills (organizing and structuring technical descriptions) and micro skills (precise vocabulary use, grammar, and detailed evaluation). All test components directly relate to technical description writing within an engineering context, ensuring no construct irrelevance threatens validity (Messick, 1989).

However, construct validity could be improved by incorporating additional sub-tasks, such as the explanation of technical processes (e.g., how the hair dryer converts electrical energy into heat), to more fully capture the range of writing skills required in engineering communication (Hyland, 2006). Additionally, while the product evaluation component required students to think critically about the design of the product, further opportunities to engage in problem-solving (e.g., proposing design alternatives) would enhance the task's alignment with real-world engineering scenarios (Artemeva, 2009).

Student perceptions of validity

Student survey responses provided additional validity evidence from the learner perspective, supporting a comprehensive validity argument (Kane, 2013; Messick, 1989). Students demonstrated high agreement (Mean = 4.3, SD = 0.49) that the test was aligned with module content and learning outcomes, with 85% agreeing or strongly agreeing with validity statements.

Table 6

Student perceptions of test validity (N = 34)

Response	Frequency	Percentage
Strongly Disagree	0	0%
Disagree	1	3%
Neutral	4	12%
Agree	17	50%
Strongly Agree	12	35%

Students agreed that the CBT measured the intended skill—technical description writing—and was connected to the course content, vocabulary, and writing structure. Representative student comments included: "*The CBT tested the skills taught in the ESP module*" and "*The test questions matched what we learned in class.*"

This careful alignment ensures there is no construct irrelevance, which could otherwise compromise the validity of the assessment (Messick, 1989). The convergent evidence from expert ratings, factor analysis, alignment documentation, and student perceptions provides a comprehensive validity argument for the CBT (Kane, 2013), though areas for improvement have been identified, particularly regarding word count constraints and the need for larger-sample construct validation.

Reliability

Test reliability refers to the consistency and trustworthiness of assessment results, minimizing potential bias in scoring. According to Badjadi (2013), "…examiner bias is the most prevailing type of reliability akin to essay testing and is referred to as inter-rater reliability" (p. 8). Since this was a subjective test, minimizing rating bias in the CBT evaluated here was important.

Double blind marking and paired samples t-test analysis

A standardised analytic rubric was used to ensure high inter-rater reliability by providing consistent scoring guidelines. However, subjectivity can still arise, particularly during the assessment of more creative or analytical responses (e.g., suggestions for product improvement). Therefore, to ensure reliability, all 34 scripts were blind-marked by two examiners, and a paired samples *t*-test revealed no statistically significant difference (*p* = 0.163) between their marks (see Table 7).

Table 7

Paired samples test results

		Paired Differences					T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Marker_1 - Marker_2	.863	4.731	.611	-.361	2.087	1.412	59	.163

Note: Alpha significance (α = 0.05)

Student perceptions of reliability

Student perceptions supported reliability (Mean = 4.1, SD = 0.58), with 76% agreeing the assessment was fair and consistent as shown in Table 8. They also found the rubric helpful in understanding expectations.

"*I believe all students were assessed using the same criteria.*"

Table 8

Frequency of responses for reliability item (N = 34)

Response	Frequency	Percentage
Strongly Disagree	0	0%
Disagree	2	6%
Neutral	6	18%
Agree	16	47%
Strongly Agree	10	29%

Although the standardised rubric supported in maintaining consistency and fairness, an area of improvement is the need for a more detailed analytic rubric to provide clearer feedback (Jönsson et al., 2021; Tomas et al., 2019). Interestingly, Barkaoui (2008) found that holistic scales sometimes led to greater inter-rater agreement than analytic rubrics. In this case, however, the consistency observed between the two markers could be attributed to their extensive experience teaching this module over the past decade, a factor that may not hold with less experienced markers. The moderation process, which involved both internal and external examiners, helped to mitigate discrepancies in scoring, but the implementation of automated scoring systems (e.g., tools that check for grammatical accuracy and coherence) could further improve the reliability of the assessment, particularly in terms of reducing marker bias. Automated scoring would also save time for teachers, who are often overwhelmed by the task of marking a large number of papers. However, ethical concerns can arise regarding transparency and fairness in marking, as algorithms may result in biases (Williamson & Breyer, 2012).

Recent advancements in automated essay scoring (AES) demonstrate significant improvements in reliability through the integration of large language models (LLMs) and hybrid evaluation systems. Wang and Gayed (2024) found that fine-tuned GPT models achieved substantial agreement with human raters (QWK = 0.78, 84.72% score agreement), though they emphasized that human oversight remains crucial. Similarly, hybrid frameworks combining deep learning with linguistic features show robust reliability (Faseeh et al., 2024; QWK ≈ 0.94), while GPT-4 applications in academic contexts report strong consistency (Quah et al., 2024; ICC = 0.79–0.86). These studies suggest that while modern AES systems approach human-level accuracy, the most effective strategy remains a human-in-the-loop framework combining automated efficiency with expert oversight. This approach could potentially enhance the reliability of future CBT assessments while maintaining the nuanced evaluation that complex analytical tasks require.

Practicality

The practicality of administration and scoring is a central challenge in language test design. Practicality is associated with the resource requirements of the test against existing institutional resources (Bachman & Palmer, 1996), including time, effort, money, and human resources.

Administration and security

From a practical perspective, CBT was relatively easy to administer, requiring no specialized equipment beyond computer access and visual prompts. The analysis, as depicted in Table 9, indicates that the assessment was feasible within institutional constraints, with no additional costs incurred beyond regular faculty duties. The closed-book format ensured students depended solely on their knowledge and skills, which was vital for assessing independent technical writing competence.

Academic integrity was maintained through strict invigilation by two supervisors following institutional guidelines. To ensure security, the question paper and answer key were uploaded to the CMS, which provides secure, role-based access limited to authorized personnel. No cases of academic integrity violation were observed during administration.

Table 9

Resource requirements and time analysis

Resource Component	Requirement	Duration/Cost	Comments
Test Administration	Computer lab booking	90 minutes	Standard computer lab booking using the online room reservation system
Invigilation	2 invigilators	90 minutes	As per institutional policy
Marking Time	Per script	8–12 minutes	Varies with response quality
Total Marking Time	34 scripts	6–8 hours	For experienced markers
Moderation Process	Internal/External review	2 hours	Automated system reduces time
Feedback Provision	Per student	5 minutes	Standardised rubric-based comments
Additional Costs	None	00	Part of regular faculty duties

Marking and moderation efficiency

The CBT preparation by the module leader is time-consuming; however, automation of the moderation process saves time through automated email reminders to moderators. Teachers are allocated two weeks to complete marking and moderation, which proved feasible within the institutional framework.

While generic marking criteria might be time-efficient, they can be insufficient for confident assessment, particularly for novice teachers. For larger cohorts, the essay-type responses create a tedious marking burden. The test required significant time and effort in grading due to the need for detailed feedback on student compositions.

Student perceptions of reliability

Students found the CBT easy to access and complete through the MECLearn platform (Mean = 4.4, SD = 0.52), appreciating the simplicity of task submission and platform familiarity.

"*The MECLearn system was easy to use for accessing the test.*"

Table 10

Frequency of responses for practicality item (N = 34)

Response	Frequency	Percentage
Strongly Disagree	0	0%
Disagree	0	0%
Neutral	3	9%
Agree	17	50%
Strongly Agree	14	41%

To improve practicality, future implementations should consider hybrid assessment approaches combining automated grammar and mechanics checking with human evaluation of content and critical thinking components (Ramineni & Williamson, 2018). The introduction of automated scoring tools for basic language mechanics, including spelling, grammar, and punctuation, may significantly reduce the burden on markers and allow them to focus more on content quality and analysis. Additionally, the development of more detailed marking criteria would support less experienced teachers in conducting confident assessments.

Authenticity

Context relevance

The context relevance of a test is of utmost importance since the usefulness of the test changes according to the context. According to Messick (1989), the evidential basis of test use is also construct validity, but with specific reference to the context for which the test is designed or used. The learners who take the CBT discussed here are future engineers and, therefore, training in writing technical descriptions is useful for their future jobs. This adds to the authenticity of the test, which is described as the relationship between the characteristics of the task given in the test and the characteristics of tasks in the real world.

The test closely mirrored the kinds of communication tasks that engineers frequently perform in their professional roles (Douglas, 2000; Dudley-Evans & St. John, 1998). By using the Philips Hair Straightener as the object of description, the students were given an accessible yet sufficiently complex task to demonstrate their understanding of both technical terminology and product evaluation techniques. Knoch and Macqueen (2020) emphasise the importance of task authenticity in ESP assessment. In light of their research, this CBT's use of a real-world product (the Philips Hair Straightener HP469) as the subject of description and evaluation aligns well with current best practices in ESP assessment.

Student perceptions of authenticity

The CBT task was rated as authentic by most participants (Mean = 4.2, SD = 0.47). The scenario involving an engineering product and a picture-based description mirrored real-world communication expected in professional settings.

"The task reflected real-world engineering communication."

Table 11

Frequency of responses for authenticity Iitem (N = 34)

Response	Frequency	Percentage
Strongly Disagree	0	0%
Disagree	1	3%
Neutral	5	15%
Agree	19	56%
Strongly Agree	9	26%

To further enhance the authenticity of the task, diverse engineering-specific scenarios can be included. However, this can be counterargued that this assessment is a CBT and mainly focuses on the basic physical description and evaluation of the item and does not delve deeper into the discipline.

Interactiveness

Interactiveness in test design

The CBT was purposefully designed to engage students both cognitively and linguistically by requiring the integration of content knowledge and language use. This approach aligns with Flowerdew's (2016) view that learner engagement is enhanced when tasks reflect real-world professional demands. By asking students to describe and evaluate a real product using technical vocabulary, the test promoted the application of discipline-specific communication skills in an authentic context—reflecting the kind of task integration that Bachman and Palmer (1996) consider central to interactiveness.

Student perceptions of interactiveness

Student responses confirmed the effectiveness of this design. With a mean score of 4.3 (SD = 0.45), 87% of the students agreed or strongly agreed that the test encouraged them to apply both content and language knowledge. This suggests they were actively synthesising information rather than relying on memorisation. The absence of disagreement and the strong positive ratings indicate a shared perception of the test as relevant and engaging. These findings support the test's intended function of simulating communicative demands in technical contexts, enhancing both learning depth and assessment authenticity (Douglas, 2000; Flowerdew, 2016).

"*The test encouraged me to use both content knowledge and language.*"

Table 12

Frequency of responses for interactiveness item (N = 34)

Response	Frequency	Percentage
Strongly Disagree	0	0%
Disagree	0	0%
Neutral	4	13%
Agree	18	54%
Strongly Agree	12	33%

Impact

Washback on teaching and learning

The CBT demonstrated a strong positive washback effect, substantively influencing both teaching strategies and student learning behaviours. This aligns with contemporary understanding of washback as the effect of testing on pedagogical practices (Cheng, 2004; Rahman et al., 2021). Students received structured in-class instruction, followed by post-class writing tasks and a mock test, which helped reduce anxiety and build familiarity with technical description writing (Ahmadjavaheri & Zeraatpishe, 2020). These iterative cycles—practice, feedback, revision—embodied formative washback (Rahman et al., 2021), guiding learners toward greater clarity, precision, and fluency in technical writing.

Observations of student engagement and perceived relevance

As the module instructor, the author observed sustained student motivation and enhanced engagement throughout the technical description tasks, which students perceived as both relevant and transferable to their academic and professional contexts. The closed-book format encouraged recall and deeper cognitive engagement. Although some might view repeated practice as "teaching to the test," in this context, it is defensible and beneficial due to the authenticity and transferability of the task to professional settings where engineers routinely write technical descriptions and product evaluations. The object descriptions that students practiced in class directly mirror authentic workplace communication requirements, making the washback effect genuinely beneficial for their future careers.

Student survey results support these observations, with 94% of students agreeing that the CBT enhanced their technical writing confidence and use of discipline-specific vocabulary (M = 4.5, SD = 0.41).

Student perceptions of impact

"The CBT helped me improve my technical writing skills."

Table 13

Frequency of responses for impact item (N = 34)

Response	Frequency	Percentage
Strongly Disagree	0	0%
Disagree	0	0%
Neutral	2	6%
Agree	16	47%
Strongly Agree	16	47%

The CBT's closed-book format appears to foster active recall and cognitive processing, supporting students in internalizing technical language without reliance on external aids. This aligns with findings from Agarwal et al. (2008), who showed that test formats requiring recall can significantly enhance learners' autonomy and strategic learning behaviors. Additionally, the structured combination of in-class practice, mock assessments, and criterion-based feedback facilitated iterative learning cycles: students rehearsed, received feedback, reflected on performance, and improved—an embodiment of formative washback (Rahman et al., 2021).

Together, these elements reinforce that the CBT served not only as an evaluative tool but as a pedagogical agent, guiding learners towards deeper understanding and professional communication competencies. The alignment between instructional design, student perceptions, and observed performance trends illustrates how well-constructed assessments can positively shape both teaching and learning outcomes.

The washback effect of the CBT on teaching and learning was generally positive. The author also teaches the module and has observed that students, in general, are interested in writing these descriptions and consider them useful. The closed-book format also encouraged students to develop their recall abilities with respect to the specific jargon, which is critical in professional environments where engineers may not have immediate access to reference materials. Through in-class practice sessions and the provision of feedback on students' technical writing and product evaluation skills, students are better prepared for the CBT and develop a deeper understanding of the course content.

Overall student perceptions: Summary of survey results

In addition to the dimension-wise analysis, a consolidated view of student responses to the post-assessment survey is presented here. Tables 14 and 15 present the distribution of responses, descriptive statistics across each dimension, and student responses to open-ended questions, providing a comprehensive understanding of student feedback on the CBT experience.

Table 14 displays the frequency distribution of student responses across each of the six dimensions of the Bachman and Palmer's (1996) framework. A majority of students agreed or strongly agreed with the statements in all categories, indicating generally positive perceptions. Practicality and Impact were the most strongly endorsed dimensions, with 91% and 94% of students, respectively, selecting 'Agree' or 'Strongly Agree.' No students expressed strong disagreement across any dimension, suggesting that the CBT was well-received.

Table 14

Frequency distribution of student responses (%) by dimension (N = 34)

Dimension	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Validity	0%	3%	12%	50%	35%
Reliability	0%	6%	18%	47%	29%
Practicality	0%	0%	9%	50%	41%
Authenticity	0%	3%	15%	56%	26%
Interactiveness	0%	0%	13%	54%	33%
Impact	0%	0%	6%	47%	47%

Table 15 presents the descriptive statistics for each survey dimension, including the mean score, standard deviation, and the highest-rated item. The highest overall mean was recorded for the Impact dimension (M = 4.5), with students particularly valuing how the CBT improved their technical writing skills. All dimensions scored above 4.0 on average, further supporting the overall positive reception of the assessment.

Table 15

Descriptive statistics of student survey responses across six dimensions (N = 34)

Dimension	Mean	SD	Highest Rated Item
Validity	4.3	0.49	The CBT tested the skills taught in the ESP module.
Reliability	4.1	0.58	I believe all students were assessed using the same criteria.
Practicality	4.4	0.52	The MECLearn system was easy to use for accessing the test.
Authenticity	4.2	0.47	The task reflected real-world engineering communication.
Interactiveness	4.3	0.45	The test encouraged me to use both content knowledge and language.
Impact	4.5	0.41	The CBT helped me improve my technical writing skills.

Table 16

Open-ended questions and student responses summary

Question	Students' Comments
What part of the CBT was most helpful or engaging?	- "The product evaluation helped me think critically." - "The time to complete was a bit short to think deeply." - "Using pictures made it easier to visualize the object."
What challenges did you face during the CBT?	- "Finding the right technical words was challenging but useful." - "I struggled with time management." - "Some instructions need more clarity." - "Understanding the technical terms needs practice."
How did the feedback help you improve your writing?	- "The rubric showed exactly where I needed to improve." - "I received detailed feedback for each part." - "I need clear feedback." - "Add a checklist for reviewing before submission." - "I learned to structure my descriptions better."
What suggestions do you have to improve the CBT?	- "More practice tests would be helpful." - "Test duration should be increased." - "Two weeks after the unit completion should be there before the actual CBT". - "Including more engineering-specific items would add interest." - "Some technical terms need clearer explanation." - "Incorporate peer-review sessions before the test." - "Provide more samples for reference." - "Provide more detailed comments on grammar and vocabulary."

Overall, the findings from both quantitative survey data and qualitative student feedback indicate that the CBT was well-received and effective in meeting its intended objectives. Students generally perceived the assessment as valid, reliable, practical, authentic, interactive, and impactful for their learning. While minor areas for improvement were noted, the results support the continued use and further refinement of the CBT within the ESP module. The student suggestions will be considered during the next cycle.

CONCLUSION

The evaluation of the CBT for ESP engineering students, grounded in Bachman and Palmer's (1996) framework, revealed several strengths. The assessment was closely aligned with the module's learning outcomes, particularly in evaluating technical description writing, vocabulary use, and product evaluation (Hyland, 2006). The test demonstrated construct validity, reliability, and authenticity, supported by expert CVI ratings (Polit & Beck, 2006) and factor analysis findings (Tabachnick et al., 2019).

Students also perceived the assessment as valid and beneficial, and 85% agreed/strongly agreed that the CBT was instrumental in developing their writing competencies (Kane, 2013). The test's closed-book format fostered deeper learning and recall of discipline-specific terminology—a valuable skill in professional settings (Quah et al., 2024).

More importantly, the CBT model presented provides useful insights into the ESP curriculum development, especially for technical disciplines. The authentic task design simulates real writing situations (Artemeva, 2009), thus enhancing the constructive alignment among assessment, learning outcomes and prospective workplace communication requirements.

LIMITATIONS

This study, however, has a few limitations. The results from the study are not generalizable because of the relatively small sample size (N = 34). Additionally, the study's factor analysis results, while informative, represent preliminary evidence (Comrey & Lee, 2013; Hair et al., 2019). Despite this limitation, the findings lay a foundation that can be further substantiated through replication with larger and more diverse cohorts. While the standardised assessment rubric supported reliability, the analytic descriptors, as noted by Jönsson et al. (2021) and Tomas et al. (2019), could have been more precise, limiting inter-rater agreement. Future studies with expanded sample sizes will likely be more effective in providing strong and nuanced construct validation.

RECOMMENDATIONS AND DIRECTIONS FOR FURTHER RESEARCH

Building on the findings of this study, several directions for practice and future research are proposed. First, while the current CBT rubric is standardised and broadly effective, it can be improved on in future iterations by increasing the analytic granularity. Separating combined criteria, such as grammar, lexis and mechanics, into distinct subcomponents may enhance scoring precision, support more targeted feedback, and further strengthen inter-rater reliability (Tomas et al., 2019). A clearer distinction between content, organization, and language features would also better align the assessment with best practices in ESP writing evaluation.

Second, the adoption of hybrid automated scoring tools, particularly those leveraging large language models (Faseeh et al., 2024; Wang & Gayed, 2024), offers potential for improving scoring consistency, reducing marking loads, and minimizing bias. However, implementation must address issues of transparency, ethical considerations, and the need for human oversight (Williamson & Breyer, 2012). Third, the CBT framework should be adapted and integrated into ESP programmes, particularly in technical disciplines. Task types such as technical specifications, product evaluations, and visual analysis, as emphasized by Hyland (2006), have proven effective and should be embedded into both formative and summative assessments across modules. Fourth, future research should focus on scaling the CBT model across institutions and disciplines, including both STEM and non-STEM contexts (Comrey & Lee, 2013), and conducting longitudinal studies to examine how CBT-acquired skills transfer to capstone projects or professional workplace writing.

Finally, investigating teacher and student perceptions of AI-based scoring, particularly in high-stakes contexts, is essential to ensure responsible and widely accepted implementation (Quahet al., 2024). Advancing ESP assessments along these lines will strengthen their validity, practicality, and alignment with the evolving academic and professional needs of learners.

THE AUTHOR

Samia Naqvi, PhD, is an Associate Professor and the Head of the Center for Foundation Studies at Middle East College, Oman. She holds a PhD in English Language Teaching and is a Senior Fellow of Advance HE (UK). Dr. Naqvi has also published extensively in renowned journals and edited books. Her research interests include ESP, academic writing and innovation in language learning.
snaqvi@mec.edu.om

REFERENCES

Adams, D. J. (2014). *Clarity, organisation, precision, economy: A technical writing guide for engineers.* University of New Haven, Civil Engineering Faculty Book Series. https://digitalcommons.newhaven.edu/cgi/viewcontent.cgi?article=1000&context=civilengineering-books

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L. III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861–876. https://doi.org/10.1002/acp.1391

Ahmadjavaheri, Z., & Zeraatpishe, M. (2020). The impact of construct-irrelevant factors on the validity of reading comprehension tests. *International Journal of Language Testing, 10*(1), 1–10. https://www.ijlt.ir/article_114277_2cf1a67513f425e6dea47b4b181aae97.pdf

Alqurashi, F. (2022). ESP writing teachers' beliefs and practices on WCF: Do they really meet? *Journal of Language and Linguistic Studies, 18*, 569–593. http://www.jlls.org

Anthony, L. (2018). *Introducing English for specific purposes*. Routledge.

Artemeva, N. (2009). Stories of becoming: A study of novice engineers learning genres of their profession. In C. Bazerman, A. Bonini, & D. Figuieredo (Eds.), *Genre in a changing world: Perspectives on writing* (pp. 158–178). The WAC Clearinghouse and Parlor Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford University Press.

Badjadi, N. E. I. (2013). *Conceptualising essay tests' reliability and validity: From research to theory.* ERIC. https://files.eric.ed.gov/fulltext/ED542099.pdf

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* [Doctoral dissertation, University of Toronto]. TSpace. https://utoronto.scholaris.ca/items/5a310a23-8dd2-473d-bf47-17d5103612a7

Bartlett, M. S. (1954). A note on the multiplying factors for various χ2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological), 16*(2), 296–298. https://doi.org/10.1111/j.2517-6161.1954.tb00174.x

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science.* University of Wisconsin Press.

Bobek, E., & Tversky, B. (2016). Creating visual explanations improves learning. *Cognitive Research: Principles and Implications, 1*, 1–14. https://doi.org/10.1186/s41235-016-0031-6

Carifio, J., & Perla, R. J. (2009). A critique of the theoretical and empirical literature of the use of diagrams, graphs, and other visual aids in the learning of scientific-technical content from expository texts and instruction. *Interchange, 40*(4), 403–436. https://doi.org/10.1007/s10780-009-9102-7

Cheng, L. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing* (pp. 25–40). Routledge. https://doi.org/10.4324/9781410609731-9

Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*. Psychology Press.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194 –197. https://doi.org/10.1016/S0897-1897(05)80008-4

Dobrin, D. N. (2019). What's technical about technical writing? In P. V. Anderson, R. J. Brockmann, & C. R. Miller (Eds.), *New essays in technical and scientific communication* (pp. 227–250). Routledge.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for specific purposes*. Cambridge University Press.

Ewer, J. R., & Latorre, G. (1969). *A course in basic scientific English.* Longman.

Faseeh, M., Nadeem, M., & Arif, M. (2024). Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics, 12*(21), Article 3416. https://doi.org/10.3390/math12213416

Flowerdew, J. (2016). English for specific academic purposes (ESAP) writing: Making the case. *Writing & Pedagogy, 8*(1), 5–32. https://doi.org/10.1558/wap.v8i1.30051

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment.* Routledge.

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

Hughes, A. (2020). *Testing for language teachers*. Cambridge University Press.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Hyland, K. (2019). *Second language writing.* Cambridge University Press.

Hyland, K., & Jiang, F. K. (2017). Is academic writing becoming more informal? *English for Specific Purposes, 45*, 40–51. https://doi.org/10.1016/j.esp.2016.09.001

Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice, 28*(3), 212–227.https://doi.org/10.1080/0969594X.2021.1884041

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *System, 38*(1), 63–74. https://doi.org/10.1016/J.system.2009.12.006

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304. https://doi.org/10.1177/0265532208101008

Knoch, U., & Macqueen, S. (2020). *Assessing English for professional purposes*. Routledge.

Korolyova, L. Y. (2017). The discursive approach to developing tests for ESP assessment in higher educational institutions. *Educational Studies Moscow, 2017*(4), 167–172. https://doi.org/10.17277/voprosy.2017.04

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Malmström, H., Pecorari, D., & Shaw, P. (2018). Words for what? Contrasting university students' receptive and productive academic vocabulary needs. *English for Specific Purposes, 50*, 28–39. https://doi.org/10.1016/j.esp.2017.11.002

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Naqvi, S., Srivastava, R., Al Damen, T., Al Aufi, A., Al Amri, A., & Al Adawi, S. (2023). Establishing reliability and validity of an online placement test in an Omani higher education institution. *Languages, 8*(1), Article 61. https://doi.org/10.3390/languages8010061

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W. & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education, 24*, Article 962. https://doi.org/10.1186/s12909-024-05881-6

Rachmawati, D. L., & Hastari, S. (2022). Formative assessment as an innovative strategy to develop ESP students' writing skills. *Voices of English Language Education Society, 6*(1), 78–90. https://doi.org/10.29408/veles.v6i1.5174

Rahman, K. A., Seraj, P. M. I., Hasan, M. K., & Rahman, M. M. (2021). Washback of assessment on English teaching-learning practice at secondary schools. *Language Testing in Asia, 11*(1), Article 12. https://doi.org/10.1186/s40468-021-00129-2

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series, 2018*(1), 1–31. https://doi.org/10.1002/ets2.12192

Rashtchi, M., & Khoshnevisan, B. (2020). Lessons from critical thinking: How to promote thinking skills in EFL writing classes. *European Journal of Foreign Language Teaching, 5*(1), Article 3153. https://doi.org/10.46827/ejfl.v5i1.3153

Sèna, U. O. (2022). Appraising the challenges related to the teaching of ESP to advanced learners in Beninese higher education. *Journal of English Language and Literature, 9*(1), 118–136. https://doi.org/10.54513/JOELL.2022.9114

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Taylor & Francis. https://doi.org/10.4324/9780203843130

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. *Frontiers in Education, 4*, Article 89. https://doi.org/10.3389/feduc.2019.00089

Wang, Q., & Gayed, J. M. (2024). Effectiveness of large language models in automated evaluation of argumentative essays: Finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2024.2371395

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

Williamson, D. M., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Winke, P., & Lim, H. (2017). The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly, 14*(4), 380–397. https://doi.org/10.1080/15434303.2017.1399396