DOI: https://doi.org/10.4038/jipe.v2i2.28

RESEARCH ARTICLE *JIPE, 2026*

Vol. 02, No. 02, (pp. 39-80)

Mindful Reading: A Moodle-Based Approach to Metacognitive Strategy Use in English for Specific Purposes

Ashmini Kalika Karunarathne*

Abstract

This study examined how Moodle-based continuous assessment tasks associate with undergraduates' metacognitive engagement in English for Specific Purposes (ESP) reading in a first-year engineering course at a Sri Lankan public university. It used a single-case study design, treating a public university's ESP reading module as the bounded case. Quantitative (CA1: N = 997; CA2: N = 973) and qualitative (n = 51) evidence were drawn from records of two compulsory continuous assessments and volunteer subset post-task reflections, respectively. The research questions focused on performance patterns across comprehension, summary writing, multiple-choice, and inference tasks; relationships among task components; comparative difficulty and ceiling effects across CA1 and CA2; reported strategy use and online assessment experience; and implications for task design. The objectives were to: (1) analyse student performance patterns across task types in CA1 and CA2; (2) examine relationships among assessment components; (3) evaluate difficulty, discrimination, and ceiling effects; (4) explore learner reflections on strategy use, confidence, and online assessment experience; and (5) propose assessment design implications for capturing metacognitive engagement. Quantitative analysis, using descriptive statistics, Pearson correlations, score distributions, and Moodle quiz statistics, examined difficulty and discrimination. Qualitative analysis, using thematic analysis, identified reported strategies, challenges, and suggestions. Results showed consistently high achievement ($\bar{x}$ > 97/100) and strong ceiling effects, limiting discrimination among learners. Summary writing functions as a distinct component, showing minimal association with recognition-based items and eliciting evidence of planning, monitoring, and evaluation in reflections. Learners valued summarising and note-taking, reported weaker planning and monitoring, and noted tensions between high scores and low confidence, alongside technical issues during Moodle delivery. Findings suggest rebalancing ESP reading assessment towards more demanding tasks, strengthening productive components, and incorporating brief reflection prompts to capture metacognitive engagement in large-cohort digital contexts.

Keywords: Metacognition, English for Specific Purposes, academic reading, Moodle, assessment design, engineering undergraduates

**Contact: ashminik@uom.lk*

ORCID: https://orcid.org/0009-0003-5202-8152

(Received 24 September 2025; Revised 16 July 2026; Accepted 18 July 2026) © OUSL

Introduction

Background and Rationale

Academic reading is central to success in higher education, especially in English for Specific Purposes (ESP) where learners must work with dense disciplinary texts, extract key ideas, and use information for study tasks. For engineering undergraduates, reading is not only about understanding sentences; it often involves identifying claims and evidence, following technical explanations, and selecting what is relevant for problem solving and report writing. These demands make reading a strategic activity, not a passive one.

Metacognition in Academic Reading

Metacognition refers to how learners plan, monitor, and evaluate their learning processes (Flavell, 1979). In reading, metacognition includes decisions made before reading (for example, setting a purpose and previewing structure), checks during reading (for example, noticing confusion and rereading), and judgments after reading (for example, evaluating whether the main ideas were understood). Research in L2 reading has shown that metacognitive strategy use is strongly linked to successful comprehension and academic performance (Anderson, 2003; Mokhtari & Reichard, 2002). In ESP contexts, metacognitive regulation becomes even more important because texts often contain unfamiliar concepts, specialised vocabulary, and discipline-specific ways of reasoning.

Assessment Design as a Driver of Strategic Reading

Assessment does more than measure learning. It signals what counts as reading ability and shapes how learners approach texts. Recognition-based formats such as multiple-choice and cloze items are efficient and can sample comprehension quickly, but they may not consistently elicit planning, monitoring, or synthesis. In contrast, productive tasks such as summary writing require learners to select, organise, and re-express meaning. This makes summaries a stronger site for observing metacognitive regulation and higher-order reading processes. However, in large-cohort university contexts, productive tasks are often given less weight because they require manual marking and moderation.

This creates a validity issue. If assessments are dominated by low-difficulty recognition items, high scores may reflect ease of items rather than deep strategic reading. In such cases, ceiling effects may compress score variation and reduce the assessment's ability to discriminate between learners, which also weakens interpretation of correlations among components.

Digital Assessment in Large Cohorts

Learning management systems such as Moodle are increasingly used for assessment because they support scale, speed, and administrative efficiency. They also provide analytics that can inform item quality. Yet online assessment introduces contextual pressures including connectivity problems, platform glitches, time pressure, and test anxiety. These factors can shape learner experience and may distort what scores appear to show. In settings where large cohorts are common, including Sri Lankan public universities, these practical realities are part of the

assessment context and should be treated as part of the evidence base when making claims about learning and metacognition.

Purpose and Objectives of the Study

This study examines how two Moodle-based continuous assessments in an ESP reading module functioned in practice, both as measures of performance and as sites that may reveal metacognitive strategy use.

The objectives of the study were to:

1. analyse student performance patterns across task types in CA1 and CA2
2. examine relationships among assessment components to identify whether tasks measured overlapping or distinct dimensions of reading
3. evaluate difficulty, discrimination, and ceiling effects using Moodle item statistics and score distributions
4. explore learner reflections on strategy use, confidence, and the experience of Moodle-based assessment
5. propose assessment design implications for capturing metacognitive engagement more validly in large-cohort digital ESP contexts

Research Questions

The study was guided by the following research questions:

1. What patterns of performance are evident in CA1 and CA2 across comprehension, summary writing, and inference tasks?

2. How do relationships among task components reflect different dimensions of reading and metacognitive engagement?
3. What commonalities and differences emerge when comparing CA1 and CA2 in terms of difficulty, discrimination, and ceiling effects?
4. What insights do student reflections provide about strategy use, confidence, and the experience of online assessment?
5. How can these findings inform the design of ESP reading assessments to better capture metacognitive strategy use and learner variability?

Contribution of the Study

This study contributes in three ways. First, it offers empirical evidence on how different task types within ESP reading assessment relate to metacognitive engagement, with specific attention to summary writing as a productive task. Second, it strengthens assessment design discussions by showing how ceiling effects and item ease can limit discrimination and weaken the interpretability of results in large cohorts. Third, it foregrounds learner experience of Moodle-based assessment, showing how performance outcomes and perceived difficulty can coexist in tension.

Review of Literature

Metacognition as a Framework for Academic Reading

Metacognition is commonly defined as awareness and regulation of one's cognitive processes (Flavell, 1979). In reading, this regulation is typically described through three core processes:

planning (pre-reading decisions and goal setting), monitoring (checking comprehension during reading), and evaluation (judging understanding and strategy effectiveness after reading). This framework is widely used in education and applied linguistics because it explains not only reading outcomes, but also the processes that lead to them.

In second language (L2) academic reading, metacognition supports purposeful engagement with texts through actions such as previewing, predicting, adjusting reading rate, rereading, and summarising. Anderson (2003) argues that skilled readers do not rely on one strategy. They select and adapt strategies in response to text difficulty and task demands. This adaptive control is central to metacognitive regulation. It is also relevant to digital contexts, where reading and assessment conditions can influence attention, time management, and monitoring behaviour.

Measuring Metacognitive Reading Strategies

A major challenge is that metacognition is not always visible in product scores. For this reason, researchers often use self-report instruments to capture perceived strategy use. Mokhtari and Reichard (2002) developed the Metacognitive Awareness of Reading Strategies Inventory (MARSI), which organises strategies into global strategies, problem-solving strategies, and support strategies. The revised version (Mokhtari et al., 2018) reports improved structure and validity. Research using MARSI-based measures generally shows that learners who report stronger metacognitive awareness also tend to perform better in academic reading tasks (Zhang & Wu, 2009). At the same time, self-report data may not fully match what learners do under

timed assessment conditions. This has led researchers to combine performance data with reflections or interviews to better interpret metacognitive engagement.

Strategy Instruction in ESP and STEM Contexts

ESP learners, especially in engineering and technology programmes, work with texts that include specialised vocabulary, dense information, and discipline-specific reasoning. These features increase cognitive load and can reduce comprehension if learners lack strategies to manage reading demands. Strategy instruction aims to reduce this burden by teaching learners how to plan reading, monitor understanding, and extract and restructure meaning through note-taking and summarising.

Research in ESP and English for Science and Technology contexts suggests that explicit training in metacognitive strategies can improve reading performance, including among engineering undergraduates (Khellab, 2022). Broader learning strategy frameworks also support the view that metacognitive strategies are teachable and benefit from explicit modelling, guided practice, and reflection (O'Malley & Chamot, 1990). However, even when instruction is provided, whether learners apply strategies during assessment depends heavily on task design and test conditions.

Assessment Design, Validity, and Discrimination

Assessment tasks shape the type of reading that learners perform. Recognition-based tasks such as multiple-choice, cloze, or short-answer comprehension questions can efficiently sample understanding, but they often reward recognition rather than synthesis. In contrast, productive

tasks such as summary writing require selection, organisation, and reformulation of meaning. These processes align more closely with metacognitive regulation because learners must decide what matters, monitor coherence, and evaluate whether the summary represents the text accurately.

From a measurement perspective, assessments must also discriminate across learners. Item difficulty and discrimination influence what an assessment can reveal. Moodle analytics commonly report facility indices (percent correct) and indicators linked to discrimination and variance contribution (MoodleDocs, n.d.). Very easy items often produce ceiling effects, compressing score ranges and limiting correlation patterns. When this happens, high scores may reflect task ease rather than strong underlying ability, which weakens the validity of inferences about reading and strategy use.

Washback is also important in ESP assessment. When assessment formats strongly favour recognition items, teaching and learning may shift toward test-taking and surface comprehension instead of strategic reading and synthesis. Language testing literature argues that assessment design should therefore balance practicality with construct validity so that the assessment rewards the kinds of learning the course intends to develop (Kuang, 2020).

Digital Assessment and Learner Experience

Digital assessment platforms offer scalability, immediate feedback, and administrative convenience, particularly in large cohorts. However, e-assessment research identifies concerns

about technical reliability, unequal access, unfamiliarity with formats, and increased anxiety (Jordan, 2016). These challenges may influence learner's performance and perceptions even when scores remain high.

This issue became more visible during and after the COVID-19 period, when studies reported that students experienced connectivity problems, time pressure, and stress during online assessments (Rameez et al., 2020; Subashini et al., 2022). These constraints are not separate from assessment validity because they shape test conditions and can alter how learners apply strategies. Moodle also creates a practical limitation: platform-generated reliability and item statistics often focus on auto-graded items and may exclude manually scored tasks such as summaries, even when those tasks represent the construct more authentically (MoodleDocs, n.d.).

ESP Assessment in the Sri Lankan Higher Education Context

Sri Lankan universities commonly implement English benchmarks such as the University Test of English Language (UTEL) (University Grants Commission, 2017). At the same time, ESP programmes in public universities often manage very large cohorts, making efficiency a key concern. Moodle-based assessment offers a workable solution, but it can encourage heavy reliance on auto-gradable items, which risks narrowing construct coverage. Local studies on e-learning adoption and online assessment experiences have also documented uneven access, technical barriers, and varied learner confidence, even in cases where academic outcomes appear satisfactory (Rameez et al., 2020; Senaratne & Samarasinghe, 2019; Subashini et al., 2022).

Synthesis and Research Gap

Across literature, metacognition is recognised as central to academic reading, especially in ESP settings where texts are demanding and strategy use is essential (Anderson, 2003; Flavell, 1979; Mokhtari & Reichard, 2002). At the same time, assessment design strongly shapes what is measured and what learners do during reading tasks, particularly in digital environments (Jordan, 2016; Kuang, 2020). However, there is limited empirical work in Sri Lankan ESP higher education that combines large-scale Moodle assessment analytics with learner reflections to examine how task types (recognition versus production) reveal or mask metacognitive engagement. This study addresses that gap by analysing two Moodle-based ESP reading assessments alongside student reflections to evaluate performance patterns, task functioning, and strategy-related experience in a large-cohort context.

Theoretical Framework

This study is guided by a metacognitive framework of academic reading grounded in Flavell's (1979) conception of metacognition as the awareness and regulation of cognition. In reading, metacognitive regulation is commonly operationalised through three interrelated processes: planning, monitoring, and evaluation. These processes offer a clear lens for examining how learners approach academic texts and how assessment tasks can either encourage or obscure strategic engagement.

Planning

Planning refers to the decisions learners make before or at the start of reading. In academic reading tasks, planning may include setting a purpose, previewing text structure, predicting content, identifying what information is needed for the task, and allocating time across sub-tasks (for example, reading, answering comprehension items, and writing a summary). In assessment contexts, planning is influenced by task format and time limits. Tasks that require synthesis, such as summarisation, tend to demand stronger planning because learners must decide what to prioritise and how to organise information.

Monitoring

Monitoring involves checking comprehension while reading and adjusting strategies when understanding breaks down. This can include rereading difficult sections, slowing down, verifying meanings of key terms through context, tracking main ideas across paragraphs, and ensuring that interpretations remain consistent with the text. Monitoring is especially important in ESP contexts where technical vocabulary and dense information can overwhelm working memory. In digital assessments, monitoring may be shaped by screen-based reading and by test navigation features that affect pacing and attention.

Evaluation

Evaluation refers to reflecting on outcomes after reading, judging whether comprehension goals were achieved, and checking whether responses meet task requirements. For summary writing, evaluation includes checking whether the summary captures key ideas accurately, whether it is

concise and coherent, and whether it avoids copying. In metacognitive terms, evaluation is closely linked to self-regulation because it informs future strategy choices and helps learners refine their approach to similar tasks.

Linking the Framework to Assessment Design

This framework is relevant because different assessment tasks are likely to activate different levels of metacognitive regulation. Recognition-based tasks such as multiple-choice or short comprehension items can be completed with limited planning or evaluation, particularly if items are easy. In contrast, productive tasks such as summarisation require selection, organisation, and reformulation of meaning, which typically involves planning, monitoring, and evaluation. Therefore, the framework supports an important assumption in this study: task type and difficulty influence how visible metacognitive engagement becomes in assessment evidence.

How the Framework Guided the Study

The framework informed the study in three ways:

1. The reflection prompts and analytic categories were aligned to planning, monitoring, and evaluation, enabling interpretation of learner-reported strategies through a clear theoretical lens.

2. Component relationships and score distributions were interpreted in relation to whether tasks were likely to elicit metacognitive regulation. For example, weak associations between summary writing and recognition-based items were examined as evidence that summary writing may capture a different dimension of reading linked more strongly to metacognitive control.

3. The case-based interpretation examined whether performance patterns (such as ceiling effects) aligned with learner experience and strategy reports, supporting a more valid account of metacognitive engagement than scores alone.

In short, Flavell's metacognitive regulation model provides a coherent theoretical basis for analysing how Moodle-based ESP reading assessments represent learner engagement, and for evaluating how assessment design can support or limit the measurement of metacognitive strategy use.

Methodology

Research Design

The study adopted a single-case study design. The bounded case was the Moodle-based ESP reading assessment practice in the EL1030 Language Skills Enhancement module at one Sri Lankan public university. A case study design was suitable because the study examined one institutional assessment context in depth and interpreted the findings within its real-life teaching, learning, and digital delivery conditions (Yin, 2018). Quantitative and qualitative evidence were used as complementary data sources within the case study and not as a separate research design label. The assessment records showed performance patterns and item functioning, while the student reflections provided evidence of perceived strategy use, confidence, and Moodle-based assessment experience. The two data sources were analysed according to their nature and then interpreted together to build a fuller case-based explanation of assessment performance and metacognitive engagement.

Participants and sampling

The study population was all first-year undergraduates enrolled in EL1030 Language Skills Enhancement, a compulsory ESP module for the Faculty of Engineering at a Sri Lankan public university. The case was selected purposively because it represented a large-cohort ESP reading assessment context delivered through Moodle. Since CA1 and CA2 were compulsory assessments, the quantitative dataset was treated as a complete sample of available cohort records, excluding only cases with missing submissions due to absence or technical disruption (CA1: N = 997; CA2: N = 973).

For the qualitative data source, post-task reflections were collected through a Google Form shared after the assessments. Participation was voluntary. Therefore, the reflection sample was a purposive volunteer subset of students from the same cohort (n = 51).

To make the sampling strategy explicit, the institutional case was selected purposively; the assessment-record dataset used complete/total-population sampling of available compulsory CA1 and CA2 records; and the reflection dataset used purposive sampling with voluntary self-selection. This approach is appropriate for a single-case study because different sources of evidence can be used to explain the same bounded case.

Table 1

Participant dataset by data source and sampling approach

Data source	Source	Sampling approach	Sample size
Assessment records	Moodle gradebook exports (CA1, CA2)	Complete/total-population sampling of available cohort records	CA1: 997; CA2: 973
Student reflections	Post-task reflections (Google Form)	Purposive sampling with voluntary self-selection	51

Data Sources and Measures

Continuous Assessments (CA1 and CA2)

Two Moodle-based continuous assessments served as the primary quantitative instruments. Each assessment required students to read an academic passage and complete four task types:

- Comprehension questions (recognition-based)
- Summary writing (productive, open-ended)
- Multiple-choice questions (recognition-based)
- Inference questions (higher-order comprehension)

CA1 used a general academic text, while CA2 used a discipline-related text on artificial intelligence in engineering. Both assessments were delivered through Moodle. Auto-graded items (comprehension, inference, MCQs) were scored automatically, while summaries were manually graded using analytic rubrics. Moodle generated item-level statistics for auto-graded items, including facility index and effective weight. Summary tasks were marked using a rubric within Moodle; auto-graded items produced Moodle quiz statistics.

Post-Task Reflection Instrument

A structured reflection questionnaire was administered after the assessments, using Google Forms. It included:

1. Likert-scale items (1–4) measuring planning, monitoring, and evaluation strategies
2. Perception items on task clarity, timing, and Moodle delivery
3. Open-ended questions on helpful strategies, challenges, and suggestions for improvement

The instrument was adapted from established metacognitive strategy frameworks (Mokhtari & Reichard, 2002) and contextualised for Moodle-based ESP assessment.

Data Collection

Quantitative data were exported from Moodle gradebooks and quiz analytics reports into Excel for analysis. CA1 data included 997 students and CA2 data included 973 students. Moodle's "Quiz statistics" reports provided facility indices and effective weight values for auto-graded items. Qualitative data were collected from 51 complete student reflection responses. Participation in reflections was voluntary and independent of course grading.

Data Analysis

Quantitative Analysis

The following analyses were conducted:

- Descriptive statistics - mean, median, standard deviation, minimum, and maximum for each component and total score

- Correlation analysis - Pearson's r to examine relationships among task components
- Distributional analysis - histograms and scatterplots to identify ceiling effects and component relationships
- Item analysis - interpretation of Moodle facility indices and effective weights for auto-graded items

These analyses were used to evaluate task difficulty, discrimination, and construct representation across assessment components.

Quantitative analyses were conducted in Microsoft Excel (descriptive statistics and correlations) and Moodle's built-in Quiz Statistics Report (facility index and effective weight).

Qualitative Analysis

The reflection dataset initially contained 52 submissions. One non-student test submission was removed before analysis, leaving 51 valid responses. Closed-response items were summarised using frequencies, percentages, means, and standard deviations. For the five-point items, 1 represented the most positive response and 5 the most negative response. The three open-ended questions were analysed using a hybrid deductive-inductive form of thematic analysis (Braun & Clarke, 2006). Each complete written response was treated as a unit of analysis. Deductive codes were informed by the metacognitive framework and covered planning, monitoring, and evaluation. Inductive codes were developed from recurring content, including summarising, note-taking, time management, text difficulty, Moodle instability, navigation between pages, and suggestions for task redesign.

Codes were compared across the three open-ended questions and grouped into four themes: productive strategies as the most visible form of engagement, uneven use of planning and monitoring, confidence and perceived difficulty, and digital delivery conditions as a constraint on strategic reading. Frequency counts were used only to show how often a pattern appeared in the volunteer subset and not as a substitute for thematic interpretation. The themes were checked against all 51 valid responses, and representative quotations were selected without registration numbers or other identifying details.

Case-Based Synthesis of Evidence

After separate analysis, the two sources of evidence were synthesised at the interpretation stage. Patterns observed in performance data, such as ceiling effects and weak correlations, were examined alongside learner reflections to provide a coherent case-based account of task functioning, strategy use, and learner experience.

Ethical Considerations

Ethical principles were followed in line with institutional guidelines. Student identities were removed from datasets prior to analysis. Participation in reflections was voluntary and did not affect assessment outcomes. Data are reported in aggregate form, and reflection evidence is presented without identifying information.

Results

This section presents findings from CA1 (N = 997), CA2 (N = 973), and post-task reflections (N = 51). Results are organised according to the five research questions to improve clarity and alignment between evidence and claims.

RQ1: What patterns of performance are evident in CA1 and CA2 across comprehension, summary writing, and inference tasks?

Performance Patterns

Across both continuous assessments, students demonstrated extremely high achievement. In CA1, the mean total score was 97.24 (SD = 4.47), while in CA2, the mean total score rose to 99.57 (SD = 2.87). In both assessments, most students achieved near-perfect or perfect scores across most task components.

Recognition-based tasks (comprehension, multiple-choice, and inference items) showed very limited variability, with most learners achieving full marks. Summary writing displayed slightly greater variability but still showed heavy clustering near the maximum score.

These patterns indicate a strong ceiling effect, particularly in CA2, where almost all learners achieved the highest possible scores. This suggests that while learners could complete the tasks successfully, the assessments provided limited scope to differentiate between levels of ability.

Figure 1

Distribution of Total Scores in CA1

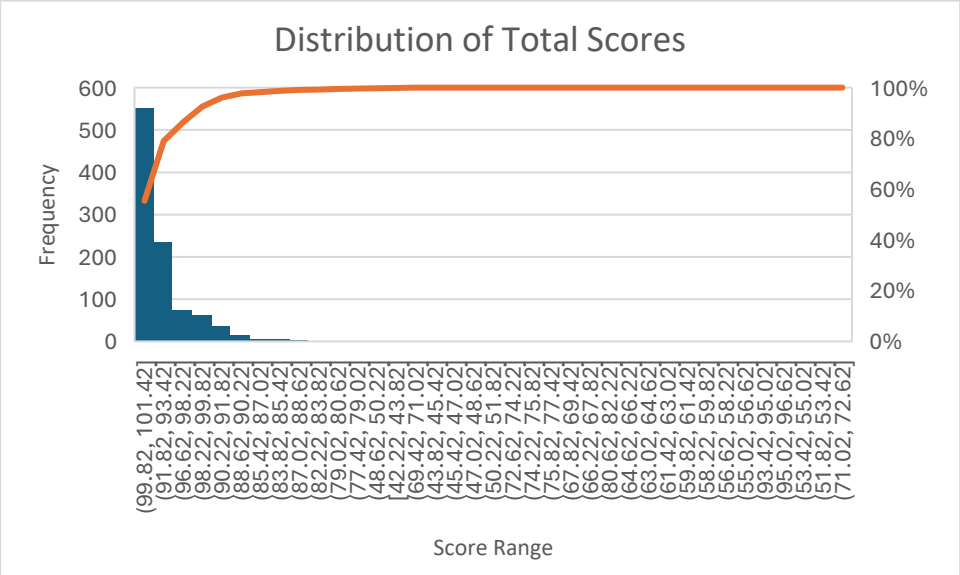


Figure 1 shows clustering of scores near 100, with limited spread below 90.

Figure 2

Distribution of Total Scores in CA2

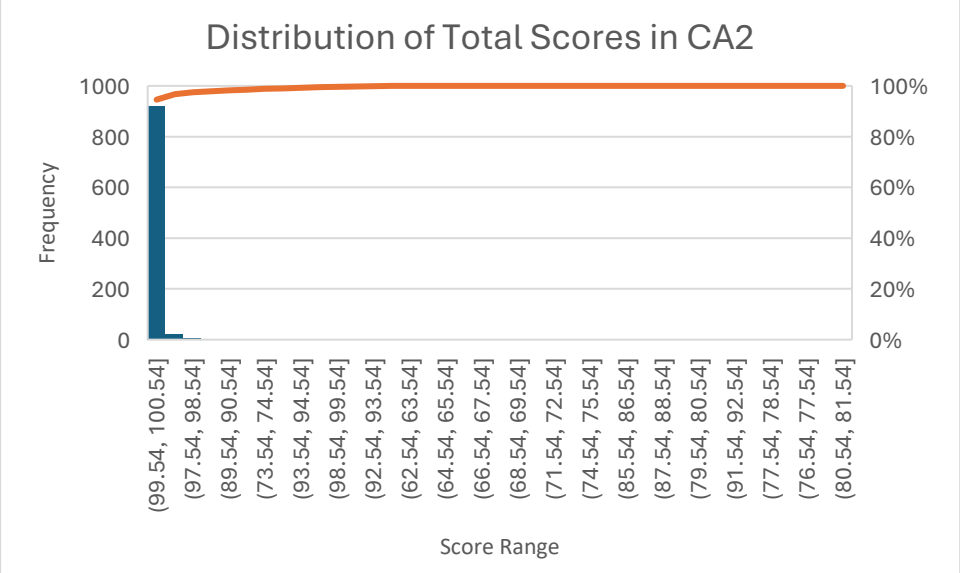


Figure 2 demonstrates extreme clustering at 100, confirming a stronger ceiling effect than CA1.

RQ2: How do relationships among task components reflect different dimensions of reading and metacognitive engagement?

Correlation analysis revealed clear differences between task types. In CA1, comprehension strongly predicted total score (r = 0.77), while summary writing showed weak or no correlation with comprehension, inference, or MCQ performance. In CA2, correlations were compressed due to ceiling effects, but the same pattern was visible: summary writing functioned independently of recognition-based components.

This pattern indicates that summary writing measured a distinct dimension of reading performance. Unlike recognition tasks, which rely on surface-level identification of correct information, summary writing required selection, organisation, and reformulation of meaning. These processes align more closely with metacognitive regulation than with recognition-based comprehension.

Figure 3

Scatter plot: Q1 (Comprehension) vs Q2 (Summary Writing) – CA1

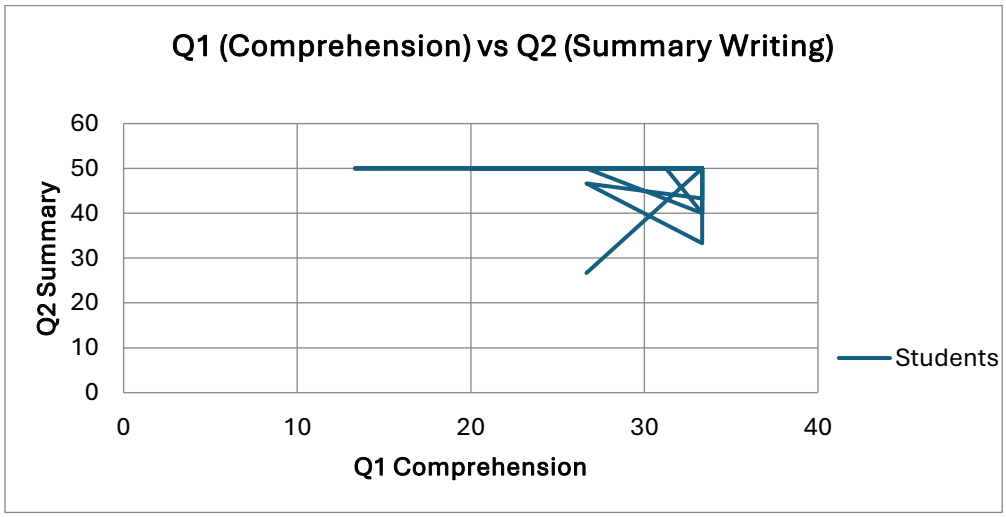


Figure 3 shows no systematic relationship between comprehension and summary performance.

Figure 4

Scatterplot: Q1 (Comprehension) vs Q2 (Summary Writing) – CA2

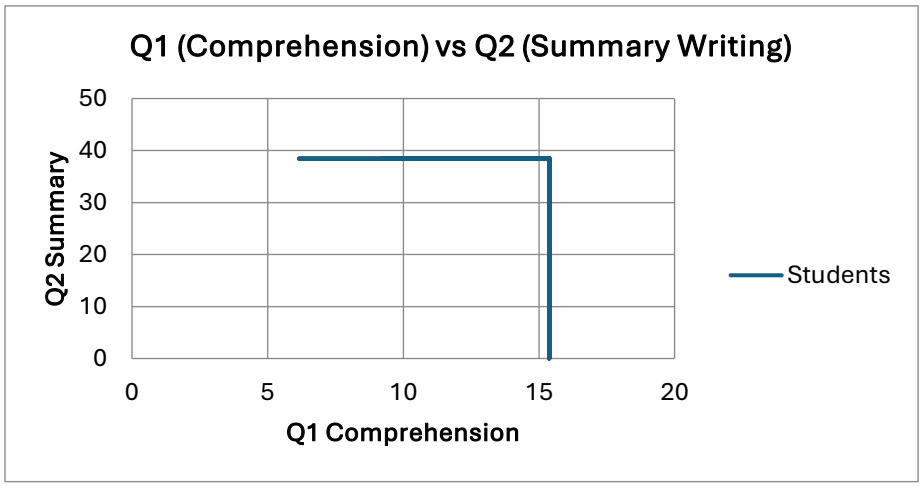


Figure 4 shows compressed clustering due to ceiling effects, masking correlation patterns.

Figure 5

Scatter plot: Total Score vs Q2 (Summary Writing) – CA1

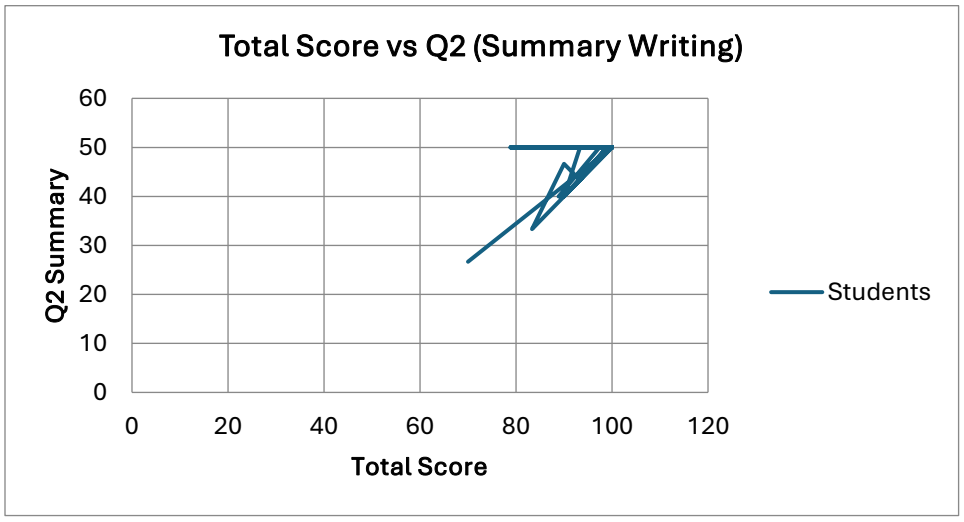
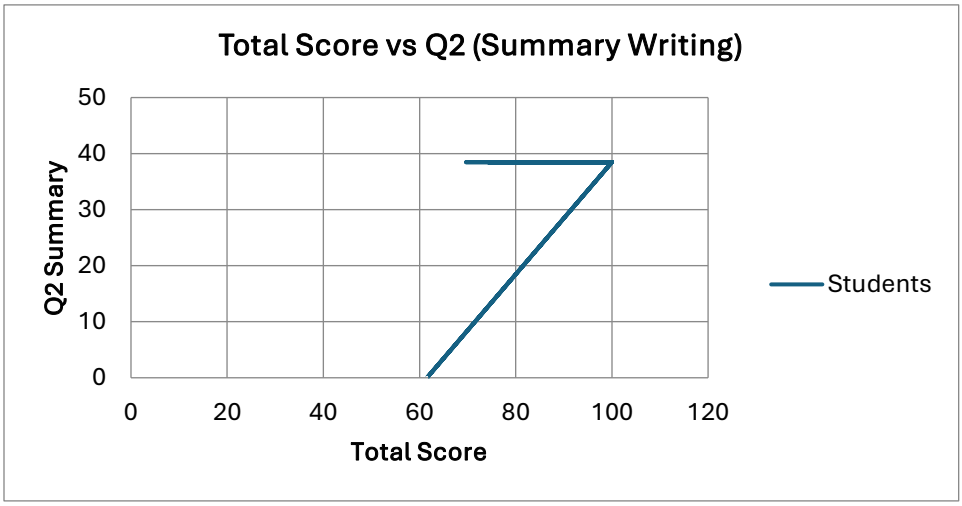


Figure 5 shows weak association driven mainly by score weighting rather than skill overlap.

Figure 6

Scatter plot: Total Score vs Q2 (Summary Writing) – CA2



RQ3: What commonalities and differences emerge when comparing CA1 and CA2 in terms of difficulty, discrimination, and ceiling effects?

Both CA1 and CA2 demonstrated high achievement, but CA2 exhibited a substantially stronger ceiling effect. CA1 showed limited variation, particularly in comprehension, whereas CA2 produced near-universal perfect scores across most components.

Moodle item statistics confirmed that many recognition-based items in CA2 were extremely easy (facility indices above 95 percent), resulting in minimal discrimination. In both assessments, summary writing remained the most authentic task but was under-represented in automated item analysis because it required manual grading.

These findings indicate that task difficulty and weighting structures reduced the assessments' ability to discriminate between learners and limited their construct coverage of metacognitive engagement.

Figure 7

Scatter plot: Total Score vs Q1 (Comprehension) – CA1

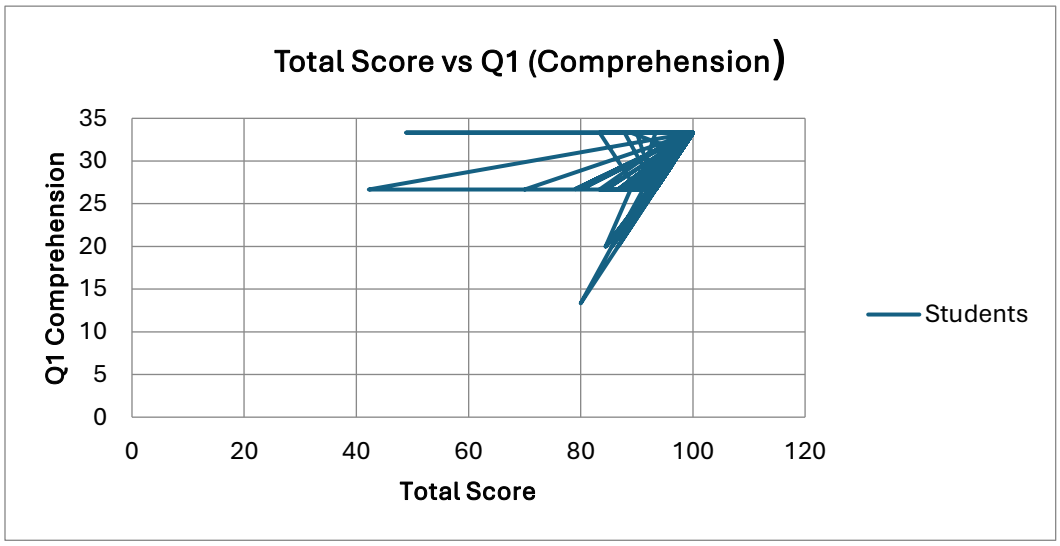


Figure 8

Scatter plot: Total Score vs Q1 (Comprehension) – CA2

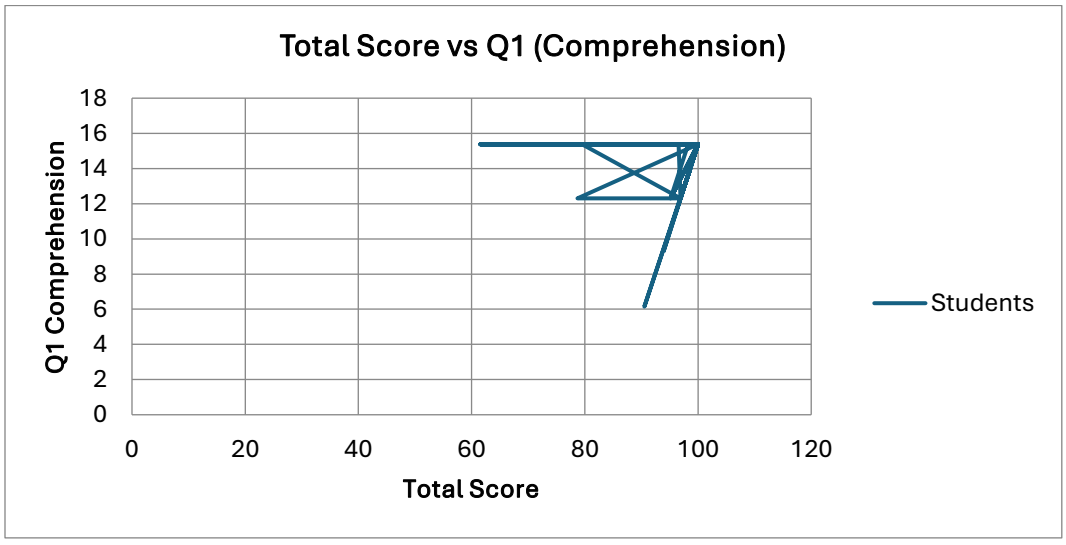
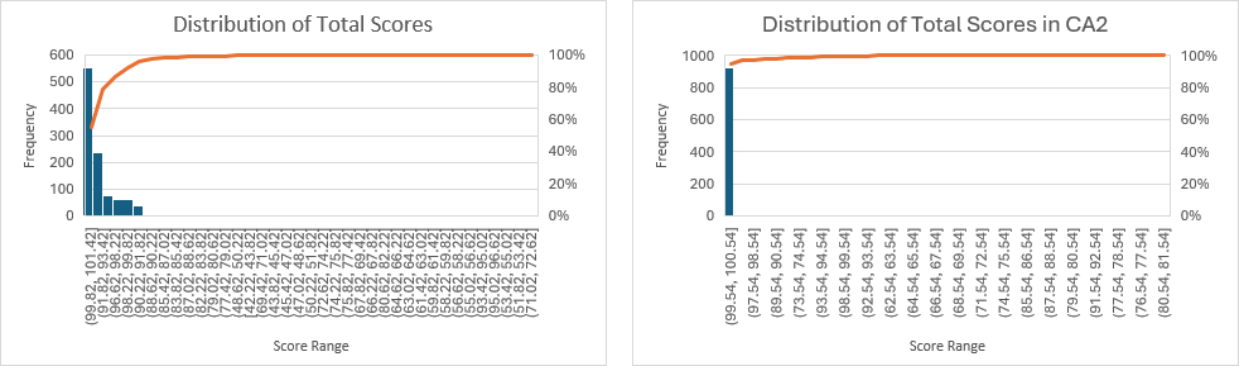


Figure 9

Comparative Distribution of Total Scores in CA1 and CA2



The paired distributions show substantially stronger clustering at the maximum score in CA2 than in CA1.

Figure 10

Comparative Component Means in CA1 and CA2

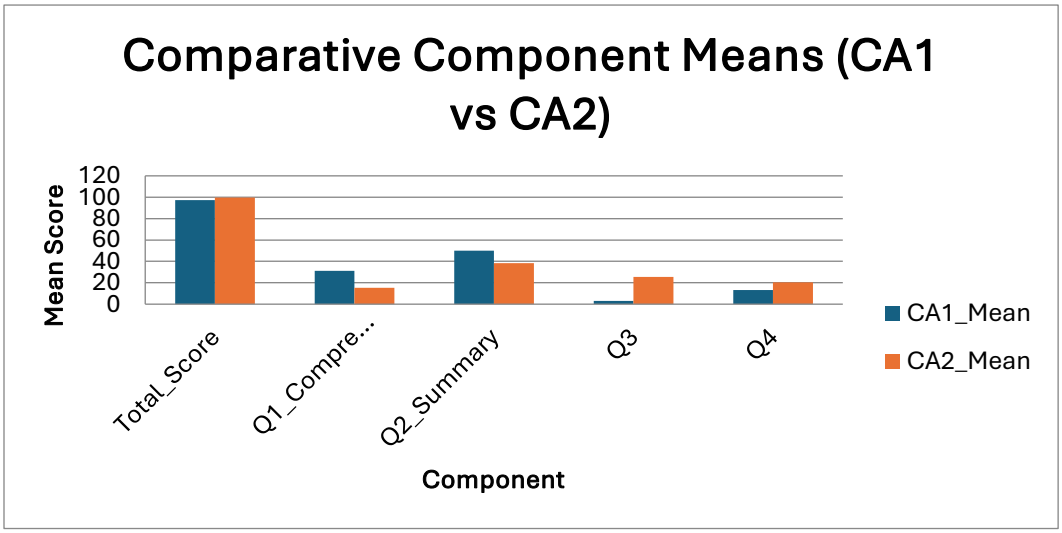
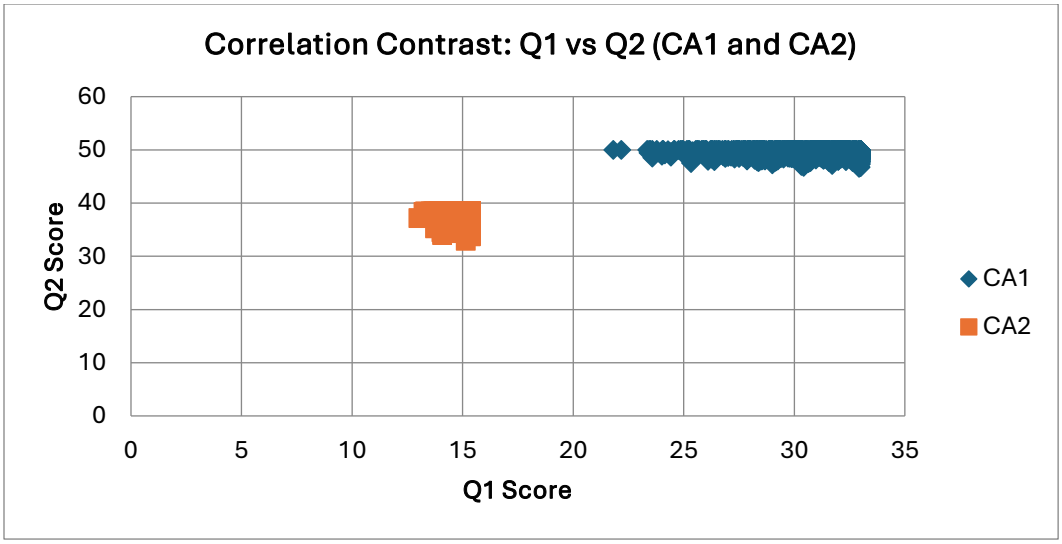


Figure 11

Correlation Contrast: Q1 vs Q2 in CA1 and CA2



RQ4: What insights do student reflections provide about strategy use, confidence, and the experience of online assessment?

Qualitative Findings from Open-Ended Reflections

The 51 valid reflections contained both closed-response ratings and short written comments. The closed responses provided an overview of reported strategy use, while the open responses explained how students experienced the task. Table 2 integrates the principal numerical patterns with the four themes and representative comments.

Table 2

Integrated findings from the post-task reflections

Theme	Evidence from the 51 reflections	Interpretation
1. Productive strategies were most visible	40/51 (78.4%) took brief or detailed notes; 40/51 (78.4%) used notes for the summary; 22/35 helpful-part responses mentioned summarising and 5 mentioned note-taking. Representative comment: "The summarizing part helped me the most. It improved my ability to identify key points and express them clearly."	Learners most readily recognised strategies tied to selecting, condensing, retaining, and producing information.
2. Planning was less consistent than monitoring	32/51 (62.7%) reported having a plan, but only 22/51 (43.1%) reported prediction. Monitoring responses were stronger: pausing to check understanding, 38/51 (74.5%); adjusting reading speed, 41/51 (80.4%); checking the summary, 36/51 (70.6%).	Metacognitive activity was more visible during and after reading than before reading, and time pressure constrained deliberate monitoring.
3. Confidence was positive but incomplete	36/51 (70.6%) reported confidence in the summary, while 15/51 (29.4%) were neutral or negative. Comments referred to difficult vocabulary, rereading, word-limit demands, and rushing the summary.	Near-perfect scores did not represent all aspects of perceived difficulty, effort, or control.
4. Digital conditions shaped task engagement	39/51 (76.5%) rated Moodle positively and 37/51 (72.5%) found the timing and instructions clear. Yet 19/41 challenge responses mentioned time, 12 mentioned technical or network problems, and 4 mentioned page navigation.	Overall acceptance of Moodle coexisted with delivery problems that affected pacing, concentration, and comprehension monitoring.

Theme 1: Productive strategies were the most visible form of engagement

Productive strategies were prominent in both parts of the reflection. Forty students (78.4%) reported using notes to write the summary, and 40 students (78.4%) reported taking either brief

or detailed notes while reading. In the open question on the most helpful part of the task, 22 of the 35 substantive responses referred directly to summarising, while five mentioned note-taking. One learner explained that short notes made it easier to remember important points and focus on the main ideas when answering questions and writing the summary. Another stated that summarising improved the ability to identify key points and express them clearly. These responses support the quantitative finding that summary writing involved selection and reformulation processes that were less visible in recognition-based items.

Theme 2: Planning was less consistent than monitoring

The closed responses showed a clearer distinction between planning and monitoring. Although 32 students (62.7%) agreed that they had a plan for completing the reading and summary task, only 22 (43.1%) agreed that they had tried to predict the content before reading. Monitoring was more strongly reported: 38 students (74.5%) agreed that they paused to check comprehension, 41 (80.4%) adjusted their reading speed when the text became difficult, and 36 (70.6%) checked whether the summary was clear, focused, and accurate. The written comments nevertheless showed that these strategies were difficult to sustain under time pressure. One learner described the tension directly: reading quickly reduced understanding but slowing down created fear of not finishing. Thus, monitoring was reported more often than advance planning, but its use was constrained by the assessment conditions.

Theme 3: Confidence was positive overall but not universal

Thirty-six students (70.6%) agreed that they felt confident about their summaries. However, 11 students (21.6%) selected the midpoint and four students (7.8%) disagreed, meaning that almost one-third did not express clear confidence despite the near-perfect cohort scores. Written responses identified difficulty including all key points within the word limit, uncertainty about vocabulary, and the need to reread difficult passages. One learner reported having to rush the summary after spending too long reading, while another found it difficult to include all key points within the required length. The qualitative evidence therefore moderates the score-based picture by showing that successful performance could still involve uncertainty, effort, and perceived difficulty.

Theme 4: Moodle was generally accepted, but technical and navigational problems affected performance

Closed responses towards Moodle were generally positive. Thirty-nine students (76.5%) rated the Moodle task positively, and 37 (72.5%) agreed that the timing and instructions were clear. The open responses, however, exposed recurring difficulties that the rating items did not fully capture. Of the 41 substantive challenge responses, 19 referred to insufficient time or time management, 12 referred to Moodle, server, loading, or network problems, and four referred specifically to moving between the passage and question pages. Students described Moodle becoming slow or unresponsive, losing concentration while pages loaded, and repeatedly scrolling or moving between pages. Suggestions included placing the passage and questions on one page, allowing additional time for long texts, staggering access to reduce server overload,

and providing practice tasks. These comments show that delivery conditions influenced pacing, concentration, and opportunities to monitor comprehension.

Cross-Source Synthesis

The two evidence sources converged in some areas and qualified each other in others. Learners' strong endorsement of strategy usefulness (43 of 51, 84.3%) and future strategy use (44 of 51, 86.3%) supports the interpretation that note-taking and summarising were meaningful parts of the task. At the same time, weaker prediction, incomplete confidence, time pressure, and technical disruption were not apparent in the near-perfect assessment scores. The reflections therefore explain why score distributions alone provide an incomplete account of metacognitive engagement in this case.

RQ5: How can these findings inform the design of ESP reading assessments to better capture metacognitive strategy use and learner variability?

The combined findings suggest that current assessment designs privilege surface comprehension and recognition over strategic regulation. Summary writing emerged as the strongest indicator of metacognitive engagement, but its interpretive value was constrained by ceiling effects and weighting structures.

To better capture metacognitive strategy use, ESP assessments should:

- include more challenging comprehension and inference tasks to reduce ceiling effects
- increase the weighting and analytic treatment of productive tasks such as summaries

- integrate brief reflective components to capture learner strategy use
- treat delivery conditions and technical reliability as part of assessment validity

These changes would improve construct representation and provide more meaningful evidence of strategic reading in large-cohort digital contexts.

Summary of Results

Across the assessment-record and reflection evidence, learners performed extremely well, but task design limited the visibility of metacognitive engagement. Recognition-based tasks produced ceiling effects and weak discrimination, while summary writing functioned as a distinct productive component. In the reflection subset, 78.4% reported taking notes and the same proportion used notes for summary writing. Monitoring was reported more strongly than prediction, and almost one-third did not express clear confidence in their summaries. Although Moodle was rated positively by most respondents, open comments repeatedly identified time pressure, technical instability, and page navigation as constraints. Together, the findings show that assessment validity depends on task design, weighting, learner regulation, and delivery conditions.

Discussion

This discussion interprets the quantitative findings together with the four qualitative themes identified in the student reflections. The focus is on what the combined evidence means for assessment validity and for designing tasks that represent metacognitive engagement more accurately in large-cohort ESP contexts.

Assessment outcomes and the problem of ceiling effects

A major finding was consistently high performance across both assessments, with mean total scores above 97 percent and medians at 100. Although this appears to signal strong reading ability, the distributions suggest that much of the assessment evidence was compressed at the top end. This ceiling effect was especially pronounced in CA2, where recognition-based components (comprehension, inference, and MCQ items) produced near-universal perfect scores.

From an assessment validity perspective, ceiling effects reduce discrimination and weaken the interpretability of both total scores and statistical relationships among components. When most learners receive maximum marks, it becomes difficult to identify meaningful differences in reading ability or strategy use, even if such differences exist. This aligns with measurement principles that emphasise moderate item difficulty for effective discrimination and meaningful variance (Alderson et al., 1995; MoodleDocs, n.d.). Therefore, the findings suggest that the assessments, particularly CA2, were more effective as confirmation of minimum competence than as diagnostic measures of learner variability.

Summary writing as a distinct construct linked to metacognitive regulation

An important pattern in CA1 was the near-zero relationship between summary writing and comprehension. This suggests that summary writing was not simply an extension of recognition-based comprehension accuracy. Instead, it likely represented a different aspect of academic

reading, one that involves selecting information, organising ideas, and transforming content into a coherent representation. These processes align closely with metacognitive regulation as described in the theoretical framework: planning what to include, monitoring coherence while writing, and evaluating whether the product accurately reflects the text.

This finding strengthens the argument that productive tasks are better suited to capturing metacognitive engagement than recognition-based items. While comprehension questions can confirm that learners identify information, summary writing requires a higher level of control over meaning and structure. This interpretation is consistent with work in L2 reading and strategy research that positions summarisation and note-taking as key strategies associated with skilled reading and metacognitive control (Anderson, 2003; Mokhtari & Reichard, 2002).

At the same time, the study also shows a practical limitation. Despite summary writing being conceptually central, its measurement value was reduced by limited score variability and by the separation between manual scoring and Moodle's automated item analytics. This creates a common "construct versus convenience" tension in digital assessment: the most authentic task is often the least visible in platform-generated reliability indicators.

Limited visibility of planning and monitoring in learner reports

The reflection data showed that planning and monitoring should not be treated as equally visible. A majority reported having a task plan, but prediction was the least strongly endorsed strategy, with fewer than half agreeing that they predicted the text before reading. By contrast, most

students reported pausing to check comprehension and adjusting reading speed when a difficulty arose. The written comments indicate why monitoring remained difficult in practice: students frequently described a trade-off between reading carefully and completing the task within the available time. This suggests that strategy instruction should make pre-reading planning more explicit while also giving learners repeated practice in regulating pace under realistic assessment conditions.

One explanation is that assessment conditions encouraged a performance-focused approach, where learners prioritised completing tasks efficiently rather than engaging in deliberate strategy regulation. Another explanation is that planning and monitoring may be used implicitly but not named or recognised by learners as strategies. Still, the pattern indicates a need for stronger instructional emphasis on planning and monitoring routines, especially in high-density ESP texts where comprehension breakdown is likely. This aligns with claims that metacognitive strategies are teachable and benefit from explicit modelling rather than being assumed as automatic learner behaviours (O'Malley & Chamot, 1990).

The paradox of high scores and low confidence

Confidence was positive overall, but the ceiling-level scores still presented a more secure picture than the reflections. Although 70.6% agreed that they felt confident about their summaries, almost one-third selected a neutral or negative response. Comments about difficult vocabulary, word limits, rereading, and rushing indicate that some high-scoring performances involved considerable uncertainty. This supports the use of reflection evidence alongside assessment

records because the two sources describe different aspects of the same task: scores represent the final product, while reflections provide access to perceived effort, control, and difficulty.

This mismatch supports the value of using reflection data alongside assessment records in case study research. The score data indicated mastery, but reflections exposed hidden struggle. In other words, scores alone risk presenting a "clean" picture that overlooks anxiety, uncertainty, and strategy weakness. This finding also aligns with digital assessment research showing that performance and perception can diverge when learners experience stress, unfamiliarity, or platform constraints (Jordan, 2016).

Digital delivery as part of assessment validity

The Moodle ratings were mainly positive; yet, the open responses showed that general acceptance did not remove specific delivery problems. Time pressure was the most frequent challenge, followed by server, loading, network, and Moodle failures. Students also found it disruptive to move repeatedly between the passage and question pages. These conditions can influence reading speed, concentration, and opportunities to check understanding. In validity terms, they are therefore part of the assessment context rather than merely administrative inconveniences. The findings support practical changes such as placing essential text and questions within an easier navigation structure, allowing timing that reflects passage length and summary demands, and providing contingency arrangements for server instability.

In validity terms, delivery conditions form part of the assessment context. If learners face interruptions or uncertainty due to platform problems, the assessment may measure coping and speed rather than metacognitive reading. Prior research in Sri Lanka has documented similar pressures during online learning and assessment periods, especially where infrastructure is uneven (Rameez et al., 2020; Subashini et al., 2022). Therefore, addressing delivery issues is not only a technical improvement. It strengthens the interpretability of assessment outcomes.

Implications for redesigning ESP reading assessment

The findings suggest several assessment design implications that go beyond general advice and follow directly from observed patterns:

1. **Recalibrate recognition-based items** to reduce ceiling effects, especially in CA2. This may involve including more inference items requiring justification, integrating distractors that test deeper comprehension, or using text segments that require synthesis rather than direct retrieval.
2. **Strengthen the role of productive tasks** by increasing the challenge and discriminative capacity of summary tasks. This can be achieved through tighter rubric descriptors, clearer constraints (word limit, focus question), and moderation procedures that ensure score variation reflects quality.
3. **Embed structured reflection** as a brief, routine component of assessment. This strengthens construct representation by capturing planning, monitoring, and evaluation evidence that is not visible in objective scores.

4. **Improve delivery conditions and safeguards**, including timing buffers, clear navigation guidance, and contingency procedures. These steps support fairness and reduce construct-irrelevant variance caused by technical instability.

Taken together, these implications argue that Moodle-based assessment can support metacognitive development, but only when task design and delivery conditions are treated as part of validity rather than separate logistical concerns.

Conclusion

This study examined how assessment design shapes the visibility of metacognitive engagement in ESP academic reading among first-year engineering undergraduates in one Sri Lankan engineering ESP programme. Using a single-case study design, the study analysed performance data from two Moodle-based continuous assessments (CA1: N = 997; CA2: N = 973) alongside student reflections (N = 51). The purpose was not only to describe achievement, but also to evaluate how different task types and digital delivery conditions influence what assessment evidence can legitimately claim about metacognitive strategy use.

Across both assessments, students achieved exceptionally high scores, with strong clustering at the maximum. These ceiling effects were particularly pronounced in CA2 and reduced the assessments' ability to discriminate between learners. While recognition-based items confirmed surface comprehension, they provided limited evidence of learner variability and limited insight into deeper strategic reading. In contrast, summary writing emerged as the most conceptually

meaningful task for capturing metacognitive regulation. In CA1, summary performance functioned largely independently of comprehension performance, suggesting that it measured a distinct dimension of reading linked to selection, organisation, and evaluation of meaning. Reflection data supported this interpretation by showing that students frequently relied on summarisation and note-taking, while planning and monitoring strategies were less consistently reported.

A further contribution of the study is the qualitative account of learner strategy use and assessment experience. Four themes showed that productive strategies were more visible than planning and monitoring, high scores coexisted with only moderate confidence, and Moodle-related technical conditions shaped pacing and concentration. These findings reinforce the value of examining assessment records together with learner reflections and show that digital delivery conditions should be treated as part of assessment validity and not as external inconveniences.

The study offers practical implications for ESP assessment design in large-cohort digital settings. First, recognition-based items should be recalibrated to reduce ceiling effects and improve discrimination. Second, productive tasks such as summary writing should be strengthened through clearer rubrics, appropriate weighting, and prompt design that increases meaningful variability. Third, brief reflective prompts can be embedded to capture metacognitive evidence that scores alone cannot represent. Finally, technical and procedural safeguards in Moodle delivery are necessary to reduce construct-irrelevant influences on performance and learner confidence.

Two limitations should be noted. Reflection data were obtained from a smaller volunteer subsample and may not represent all learner experiences. In addition, Moodle's automated statistics exclude manually graded components, which limit platform-generated reliability interpretations for the full construct. Future research can address these limitations through larger reflective datasets, triangulation with log data, and redesigned assessments that balance scalability with construct validity.

Overall, the study demonstrates that Moodle-based assessment can support ESP reading evaluation at scale, but the validity of inferences about metacognitive engagement depends on task design, weighting, and delivery conditions. Strengthening these elements can move digital ESP assessment beyond efficient scoring toward richer evidence of the strategic reading processes required for academic success.

About the author:

Ashmini Kalika Karunarathne is a Lecturer in English at the University of Moratuwa, Sri Lanka. She holds a B.Ed. (Hons) focusing on TESL and is pursuing an M.A. in Linguistics. Her research interests include ESP, metacognition, digital assessment, and curriculum innovation in higher education.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Anderson, N. J. (2003). Scrolling, clicking, and reading English: Online reading strategies in a second/foreign language. *The Reading Matrix, 3*(3), 1-33. https://www.readingmatrix.com/articles/anderson/article.pdf

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

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906-911. https://doi.org/10.1037/0003-066X.34.10.906

Jordan, S. (2016). E-assessment: Past, present and future. *New Directions in the Teaching of Natural Sciences, 9*, 87-106. https://doi.org/10.29311/ndtps.v0i9.504

Khellab, F., Demirel, Ö., & Mohammadzadeh, B. (2022). Effect of teaching metacognitive reading strategies on reading comprehension of engineering students. *SAGE Open, 12*(4), Article 21582440221138069. https://doi.org/10.1177/21582440221138069

Kuang, Q. (2020). A review of the washback of English language tests on classroom teaching. *English Language Teaching, 13*(9), 10-18. https://doi.org/10.5539/elt.v13n9p10

Mokhtari, K., & Reichard, C. A. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology, 94*(2), 249-259. https://doi.org/10.1037/0022-0663.94.2.249

Mokhtari, K., Dimitrov, D. M., & Reichard, C. A. (2018). Revising the Metacognitive Awareness of Reading Strategies Inventory (MARSI) and testing for factorial invariance. *Studies in Second Language Learning and Teaching, 8*(2), 219-246. https://doi.org/10.14746/ssllt.2018.8.2.3

MoodleDocs. (n.d.). Quiz statistics report. Retrieved July 9, 2026, from https://docs.moodle.org/502/en/Quiz_statistics_report

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition.* Cambridge University Press.

Rameez, A., Fowsar, M. A. M., & Lumna, N. (2020). Impact of COVID-19 on higher education sectors in Sri Lanka: A study based on South Eastern University of Sri Lanka. *Journal of Educational and Social Research, 10*(6), 341. https://doi.org/10.36941/jesr-2020-0132

Senaratne, S. I., & Samarasinghe, S. M. (2019). Factors affecting the intention to adopt m-learning. *International Business Research, 12*(2), 150-164. https://doi.org/10.5539/ibr.v12n2p150

Subashini, N., Udayanga, L., De Silva, L. H. N., Edirisinghe, J. C., & Nafla, M. N. (2022). Undergraduate perceptions on transitioning into E-learning for continuation of higher education during the COVID pandemic in a developing country: A cross-sectional study from Sri Lanka. *BMC Medical Education, 22*, Article 521. https://doi.org/10.1186/s12909-022-03586-2

University Grants Commission. (2017). *Establishments Circular Letter No. 03/2017: Common undergraduate curriculum and a testing mechanism for English proficiency* (University

Test of English Language - UTEL).

https://www.ugc.ac.lk/attachments/1820_Est.%20Circular%2003_2017.pdf

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.

Zhang, L. J., & Wu, A. (2009). Chinese senior high school EFL students' metacognitive awareness and reading-strategy use. *Reading in a Foreign Language, 21*(1), 37-59. https://files.eric.ed.gov/fulltext/EJ838388.pdf