Operational Validation Evidence for the LANGUAGECERT Automated Writing Scoring System: Phase 1

David Coniam
Leda Lampropoulou
Vlasis Megaritis

December 2025

Abstract

This paper reports Phase 1 of a multi-stage validation programme for the LANGUAGECERT Automated Writing Scoring System, focusing on its large-scale operational behaviour on the LANGUAGECERT Academic Writing Test (LCAWT). The purpose of this phase is to evaluate whether the automated scoring system operates coherently and reliably under authentic testing conditions, and whether its behaviour aligns sufficiently with trained human examiners to support progression to subsequent validation phases.

The study draws on a dataset of 2,394 test takers, each completing two writing tasks assessed against four analytic criteria: Task Fulfilment, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, and Organisation and Coherence. Automated scores were compared with scores produced by a composite of trained and quality-assured human markers, consistent with LANGUAGECERT's operational marking procedures. Analyses included descriptive statistics, correlational analyses, and Many-Facet Rasch Measurement modelling to examine score correspondence, marker behaviour, and facet stability across tasks, criteria, and prompts.

Results indicate strong agreement between automated and human scores at the total-score level (Spearman's ρ = 0.87 for both tasks), with similarly high correspondence across most analytic criteria. Rasch analyses showed acceptable fit for marker, criterion, and question facets, and discrepancy rates relative to operational quality-assurance thresholds were low (approximately 1.5% per task), substantially below typical human–human discrepancy levels.

Taken together, the findings demonstrate that the automated scoring system exhibits stable, interpretable, and construct-aligned behaviour under operational conditions. In line with its intended evidential role, Phase 1 establishes a robust operational foundation for subsequent precision- and fairness-focused phases of the validation programme, rather than constituting a standalone or definitive validation of automated scoring performance.

Keywords: LANGUAGECERT Academic Writing Test, Automated Writing Assessment, Human–Machine Score Agreement, Operational Validation

Suggested citation

Coniam, D., Lampropoulou, L., & Megaritis, V. (2025). *Operational validation evidence for the LANGUAGECERT automated writing scoring system: Phase 1.* LANGUAGECERT.

Introduction

Automated scoring systems are increasingly used in high-stakes language assessment, where their deployment must be supported by a coherent body of validation evidence. Such evidence is rarely provided by a single study; rather, it typically accumulates across multiple investigations that address different aspects of system performance under varying conditions.

Within this context, the present paper forms part of a broader validation programme for the LANGUAGECERT Automated Writing Scoring System. The work reported here constitutes Phase 1 of this programme and focuses on the system's large-scale operational behaviour on the LANGUAGECERT Academic Writing Test (LCAWT), using authentic test data and established analytic approaches.

whether, under realistic operational conditions, the automated scoring system operates within an acceptable performance range when compared with trained and quality-assured human markers. The focus is on large-scale score behaviour, including stability, consistency, and alignment with the intended writing construct, rather than on fine-grained estimates of human rater variability or subgroup-level effects. As such, the findings are intended to provide foundational operational evidence that informs subsequent stages of validation, rather than a comprehensive evaluation of automated scoring performance in isolation.

The next section outlines the validation framework within which this study is positioned and clarifies the evidential role of the present analyses.

Validation Framework and Study Positioning

The validation of automated scoring systems intended for use in high-stakes language assessment is typically supported by evidence accumulated across multiple, complementary investigations. Different stages of validation address distinct questions, ranging from large-scale operational behaviour to more tightly controlled examinations of scoring precision and fairness.

Within this broader context, the validation of the LANGUAGECERT Automated Writing Scoring System is being undertaken through a staged programme of related studies, each contributing a different form of evidence to the overall validity argument. The programme currently comprises three phases at different stages of development.

Phase 1, reported in the present paper, constitutes a foundational operational study and is now completed. Its purpose was to examine the large-scale behaviour of the automated scoring system when applied to authentic test data and compared with trained and quality-assured human markers. The emphasis at this stage is on score stability, consistency, and construct-aligned behaviour across tasks, criteria, and prompts under realistic scoring conditions.

Phase 2 is a multi-rater study, for which a detailed study plan has been developed. This phase is designed to examine agreement between the automated scoring system and multiple independent human raters, situating automated–human agreement within the broader distribution of human–human variability. Evidence from Phase 1 informed the design parameters of this work, but did not constitute its sole motivation.

Phase 3 extends the validation programme to questions of subgroup behaviour and fairness. This phase examines automated scoring performance consistency across relevant test-taker subgroups and contributes fairness-related evidence required for responsible use in high-stakes assessment contexts.

The present paper (Phase 1 study) should therefore be interpreted in relation to its specific evidential role within this broader programme. It provides foundational operational evidence that supports, but does not replace, subsequent precision- and fairness-focused validation work, and should not be read as a standalone or definitive evaluation of automated scoring performance.

Background

The use of computers to score student essays dates back to the 1960s with the work of Page (1966), whose Project Essay Grade *(PEG)* modelled relationships between surface linguistic features (e.g., sentence length, word count, and punctuation) and human-assigned scores using regression analysis. Although primitive by contemporary standards, PEG represented a critical conceptual leap: that aspects of writing quality could be quantified and predicted algorithmically.

From the 1980s onward, advances in computational capacity and the availability of larger linguistic corpora contributed to increased research activity in computer-based testing and assessment (Chapelle & Douglas, 2006). During this period, the educational potential of computational tools was increasingly recognised within broader Computer-Assisted Language Learning (CALL) frameworks. Warschauer and Healey (1998), for example, anticipated the emergence of an Intelligent CALL phase in which technology-supported systems would provide adaptive guidance and personalised feedback. Developments in natural language processing (NLP) and artificial intelligence (AI) have since enabled aspects of this vision to be realised within automated writing assessment.

Evolution of Analytic Techniques in Automated Scoring

Throughout the 2000s and 2010s, research and commercial applications of automated writing assessment (AWA) systems expanded substantially. Ramineni and Williamson (2013) documented a range of AWA systems, each employing different combinations of linguistic analysis and statistical modelling. These approaches included rule-based feature modelling (e.g., later iterations of PEG; Page, 2003), semantic vector methods such as Latent Semantic Analysis (e.g., the Intelligent Essay Assessor; Foltz, Laham, &

Landauer, 1998), and NLP-driven feature extraction approaches used in systems such as e-rater and IntelliMetric (Burstein, 2003).

Validation of AWA systems has typically been framed around multiple dimensions reflecting both psychometric and practical considerations. These dimensions commonly include statistical agreement with human scores, fairness across test-taker groups, relationships with external performance indicators, and the broader consequences of automated scoring use. Table 1 summarises these dimensions as articulated by Ramineni and Williamson (2013), which continue to inform contemporary validation practice.

Table 1: *Criteria for assessing AWA system performance (adapted from Ramineni & Williamson, 2013)*

Dimension	Description
Statistical agreement between computer- and human-derived scores	The degree of correspondence between machine-generated and human-assigned scores, typically measured through correlations, kappa coefficients.
Fairness	The extent to which scores are consistent and unbiased across different groups of test takers, such as those differing by gender or first language.
Relationship with external variables	The strength of association between automated scores and other relevant indicators of performance, such as results on related test components or English class grades.
Consequential validity	The practical and ethical implications of using automated scores in place of human scores.

By the mid-2010s, NLP techniques had become integral to the design of nearly all AWA systems. The growing sophistication of language models enabled systems to move beyond surface feature counting to capture aspects of coherence, cohesion, and rhetorical organisation. Systems such as ETS’s E-rater (Burstein, 2003) were used operationally in large-scale assessments such as the TOEFL, functioning either autonomously or within hybrid human–machine marking models. (Attali & Burstein, 2006; Ramineni & Williamson, 2013).

Integration of AI-Based Scoring in Large-Scale English Language Assessment

Over the past decade, the use of AI-based automarking has transitioned from experimental to operational in a number of large-scale English language assessments. Adoption across the sector, however, has been neither uniform nor unconditional.

Testing organisations differ markedly in the extent to which automated scoring is integrated into operational decision-making, reflecting varying institutional assessments of risk, accountability, and evidential sufficiency.

In high-stakes contexts, automated scoring is most commonly implemented within hybrid or human-in-the-loop frameworks, in which automated scores are complemented by human oversight and review. Such models are designed to combine the efficiency and consistency of algorithmic scoring with the interpretive judgement, contextual sensitivity, and ethical accountability of trained human examiners.

While advances in AI-based scoring have produced increasingly strong alignment with human ratings under controlled conditions, fully automated and unsupervised deployment remains comparatively limited in high-stakes writing assessment. This reflects ongoing concerns relating to construct representation, edge-case behaviour, and the handling of atypical or severely deficient responses, areas in which human judgement continues to play a critical role. As a result, sector-wide practice tends to emphasise progressive integration rather than wholesale replacement of human scoring.

Importantly, a number of high-stakes English language assessments continue to rely exclusively on human marking for writing, reflecting differing institutional positions on automation and risk management. Across the sector as a whole, however, a consistent emphasis can be observed on reliability, fairness, transparency, and human oversight wherever automated scoring is used. Automated scoring is thus increasingly positioned as a complement to expert human judgement rather than a wholesale replacement.

From NLP to Large Language Models

The emergence of transformer-based Large Language Models (LLMs) such as ChatGPT has marked a transformative phase in automated writing assessment. Whereas earlier systems relied on manually engineered linguistic features and regression modelling, LLMs can infer discourse-level, pragmatic, and rhetorical relationships directly from vast text corpora. This allows for more nuanced evaluation of content relevance, coherence, and stylistic appropriateness.

Recent studies have shown that LLM-based approaches can achieve levels of agreement with human raters comparable to, and in some cases exceeding, earlier NLP-based systems, particularly when guided by explicit rubrics and exemplar material (e.g., Mizumoto & Eguchi, 2023; Xiao et al., 2024). Related work across multiple languages suggests that such models may generalise scoring behaviour beyond English, although concerns regarding bias, construct representation, explainability, and governance remain active areas of research (Kostić et al., 2024; Kwon et al., 2023).

The Role of AI Feedback and Future Directions

Beyond summative assessment, AI-based systems have increasingly been explored for *formative feedback generation*. Prior research has examined the use of automated systems to provide targeted guidance on vocabulary, grammar, and discourse structure (Ramineni & Williamson, 2013; Mansour et al., 2024; Shi & Aryadoust, 2024). This capacity aligns with the long-standing pedagogical vision of *Intelligent CALL* outlined by Warschauer & Healey (1998). While feedback generation lies outside the scope of the present study, it represents a potential extension of automated scoring systems in future work.

LANGUAGECERT Automated Writing Scoring: Analytic and Modelling Framework

The LANGUAGECERT Automated Writing Scoring System adopts a hybrid modelling approach that integrates linguistically motivated features with contemporary deep learning techniques. The design reflects a balance between construct-relevant linguistic analysis and contextual language representation, with the aim of supporting reliable operational scoring while enabling transparent validation. The process is described briefly below.

Scripts identified as unsuitable for automated analysis are excluded at this stage. These include responses that are excessively short (fewer than 85 unique words), blank or score-zero submissions, and texts that exhibit characteristics of non-language output, such as a very low proportion of valid English words or unusually high repetition. The remaining scripts are then partitioned into training and evaluation sets. This filtering step reflects standard operational quality-control practice and is distinct from the validation analyses reported in this study.

A range of readability, lexical, and syntactic metrics is subsequently computed. Readability measures include indices such as Flesch Reading Ease and SMOG, while lexical and syntactic features encompass average sentence length, indicators of vocabulary diversity and complexity, and part-of-speech (POS) n-gram distributions (uni- to tri-gram). These features are selected to align with aspects of writing performance explicitly referenced in the LCAWT rating scale and to support construct-aligned score modelling.

In parallel, contextual representations are derived using Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). A pre-trained BERT model is fine-tuned on scored writing scripts to produce dense embeddings that capture semantic and stylistic characteristics not readily represented by surface linguistic features alone.

Score prediction is performed using a regression-based learning framework (XGBoost; Chen & Guestrin, 2016). For each analytic criterion, the modelling process allows for different combinations of available feature sets, including contextual embeddings, handcrafted linguistic features, and POS-based representations. In practice, most deployed models incorporate all feature types, though the framework does not require a fixed fusion of components for all criteria. Model performance during development is evaluated using standard predictive indices, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Quadratic Weighted Kappa (QWK), which provide complementary information about average prediction error, sensitivity to large deviations, and agreement with human scores.

The system underwent two major development iterations. Version 1 was trained on data collected up to August 2024, while Version 2 was retrained and re-evaluated using additional data collected between September and December 2024. Table 2 summarises the training and testing configurations for each version. All analyses reported in the present study are based on Version 2 of the model.

Table 2: *Automarker training and testing parameters*

Automarker version	Time frame	Training scripts	Testing scripts
Version 1	Up to Aug 2024	1,838	460
Version 2	Sept-Dec 2024	2,232	558

A hybrid design was selected over an end-to-end generative language model for three primary reasons. First, the inclusion of explicit linguistic features provides interpretable signals aligned with the construct definitions of the LCAWT rating scale, supporting transparent validation work, even where BERT-derived representations typically carry greater predictive weight. Second, the use of contextual embeddings enables modelling of semantic and stylistic nuance while remaining comparatively stable and data-efficient, compared with large generative models. Third, the XGBoost regression-based stage offers predictable behaviour under controlled conditions, reducing the risk of drift, hallucination, or unexplainable variance associated with generative LLM scoring. Taken together, this architecture balances modern contextual modelling with the operational reliability and controlled interpretability required in high-stakes assessment.

The Current Study

To contextualise the analyses that follow, this section outlines the assessment instrument, dataset, and analytic design used to compare automated and human-generated scores.

The LANGUAGECERT Academic Writing Test

The present study draws on scripts from the LANGUAGECERT Academic Writing Test (LCAWT), a high-stakes, multi-level test designed to assess the ability to understand, use, and communicate effectively in written English for academic purposes. The LCAWT lasts approximately 50 minutes and comprises two tasks: (1) a structured response to visual input (e.g., a graph or table) of 150-200 words, and (2) an argumentative academic essay of approximately 250 words.

Writing performance is assessed using analytic marking criteria aligned with the descriptors of the Common European Framework of Reference for Languages (CEFR). The four criteria are Task Fulfilment (TF), Accuracy and Range of Grammar (GRA), Accuracy and Range of Vocabulary (VRA), and Organisation and Coherence (OC). Performance on each criterion is rated on a nine-point scale (0–8), yielding a maximum composite score of 32. Each task is double-marked by trained human examiners, and the final task score represents the mean of the two ratings. While the LCAWT reports results across CEFR levels A1 to C2, its primary purpose is essentially to indicate readiness for tertiary-level study; hence, most test-taker scores fall between B1 and C1 levels.

Sample

The dataset comprises 2,394 LCAWT test takers, each contributing one script for each of the two writing tasks. Across the dataset, each task was administered using 11 distinct prompts, yielding a total of 22 unique writing prompts. Minor variation in the number of scripts per task reflects routine data-cleaning procedures applied prior to analysis. Five scripts from Task 1 were excluded because they did not meet minimal analysable criteria (e.g. truncated or empty responses). No candidates were removed in full, and the total number of test takers therefore remains unchanged.

Markers

Two sources of scores were compared in the study. Marker 1 represents the finalised human scores produced through LANGUAGECERT's routine operational marking process. These scores were generated by multiple trained and accredited human markers, all of whom marked in accordance with standard procedures. Where applicable, scores were subject to routine moderation and sign-off processes before being finalised. The resulting human scores therefore reflect operationally valid outcomes rather than individual examiner judgements.

Marker 2 refers to the automated writing scoring system described above, which generated criterion-level and total scores for the same scripts. Comparisons in the present study are thus between operationally finalised human scores and automated scores produced independently by the system.

Research Questions

The broad research hypothesis of the study is that the LANGUAGECERT automarker would demonstrate statistical qualities comparable to those reported among human markers, reflecting the intended evidential role of Phase 1. Specifically:

1. Correlations between automarker and human total scores (out of 32 for the four criteria) will be above 0.80.
2. Rasch fit statistics for marker, criterion and question facets will fall within the range 0.50–1.50.
3. Discrepancies between automarker and human ratings will remain below 10%.

Analysis

This section analyses test-taker writing performance as scored by two sources: the operationally finalised human scores (Marker 1) and the automated scoring system (Marker 2). Analyses are presented in stages, beginning with descriptive statistics, followed by correlational analyses and Many-Facet Rasch Analysis (MFRA).

Descriptive statistics are first reported for both score sources at the criterion level and for total composite scores. Inferential analyses then examine the degree of association between human and automated scores. Given the ordinal nature of the 0–8 rating scales and the focus on association rather than mean differences, Spearman's rank correlation coefficient (ρ) is used as the primary measure of correspondence (Lumley, 2002).

In addition, MFRA is employed to examine score behaviour across multiple facets simultaneously. MFRA has been widely used in performance assessments such as writing to model variability associated with test takers, raters, tasks, and rating criteria (Engelhard, 1992; McNamara, 1996). In the present study, MFRA is used to examine marker behaviour and score consistency across test-taker, task, and criterion facets within a unified measurement framework.

Descriptives

Table 3 presents descriptive statistics for two writing tasks by score source. The maximum possible score for each task is 32.

Table 3: *Descriptive statistics for human and automarker scores*

Task	Task 1			Task 2	
	Human marker	Automarker		Human marker	Automarker
Number	2,389	2,389		2,394	2,394
Mean	18.10	17.65		18.03	17.66
SD	6.24	5.37		6.17	5.35
Minimum	1	4		2	4
Maximum	32	28		32	27

Across both tasks, mean scores produced by the human and automated scoring sources differed by less than one point. Human mean scores were highly consistent across tasks (18.10 and 18.03), while the automarker's mean scores were virtually identical (17.65 and 17.66), indicating stable scoring behaviour across task types.

Correlations

Table 4 presents Spearman's ρ correlations between human and automated scores at the total-score level and for each analytic criterion. Following Hatch and Lazaraton's (1991) descriptors for inter-rater reliability, correlations of 0.80 or above are interpreted as strong, 0.50–0.79 as moderate to strong, and around 0.50 as moderate.

Table 4: *Correlations (Spearman's ρ) between human and automarker scores by task and criterion*

Task	Total	TF	GRA	VRA	O&C
1	0.87	0.81	0.84	0.83	0.75
2	0.87	0.81	0.84	0.84	0.76

At the total-score level, correlations between human and automated scores were strong for both tasks (ρ=0.87). Criterion-level correlations were similarly high, with strong correlations (ρ ≥ 0.80) across three of the four criteria. The slightly lower, though still substantial, correlations on Organisation and Coherence (ρ ≈ 0.75) are consistent with prior findings that discourse-level constructs present greater challenges for automated scoring models.

Interestingly, Task Fulfilment, despite its reliance on prompt-specific information, also demonstrated correlations above 0.80, indicating that the automarker was able to infer topic relevance with reasonable accuracy. Overall, the descriptive and correlational results indicate that the automarker performs comparably to an experienced human marker in terms of scores awarded, range of scores, and correlations between criteria.

Organisation and Coherence correlations (ρ≈0.75) fall below the 0.80 threshold, consistent with prior research on discourse-level constructs. For Phase 1, correlations above 0.70 are acceptable given strong total-score correspondence. However, this relatively weaker performance will be looked at in Phase 2 through multiple independent raters.

Many Facet Rasch Analysis

In performance assessments such as writing tests Many-Facet Rasch Analysis (MFRA) has become a widely accepted approach for modelling multiple sources of variability (e.g., facets for test takers, raters, tasks, and criteria) (Engelhard, 1992; Weigle, 1998, 2002). In MFRA, the measurement scale derived by application of a unified metric such as the Rasch model means that various phenomena – marker severity, question difficulty etc – can be modelled and compensated for (McNamara, 1996; Weir, 2005). In MFRA, the measurement scale derives from the probability of observed ratings given facet locations; thus, situational factors -here, test-taker ability, question difficulty, and marker severity- are explicitly modelled in constructing the overall measurement picture. The present study specified a four-facet design: markers, test takers, task prompts (labelled 'Questions' in FACETS output), and marking criteria. The human marker facet represents an operationally finalised composite of multiple trained examiners rather than an individual rater, reflecting standard LANGUAGECERT marking and moderation procedures. Analyses were conducted with FACETS v4.1.8 (Linacre, 2024).

MFRA offers advantages over purely classical statistics by calibrating all facets onto a single unidimensional latent scale, enabling direct comparison across marker severity, task difficulty, and criterion behaviour (Eckes, 2015). In line with previous LANGUAGECERT work (e.g., Coniam et al., 2021a; Papargyris & Yan, 2022), fit is reported as a key diagnostic. Fit relates to how well obtained values match expected values and is divisible into related, if slightly different, categories. The most widely used is the infit mean square (MnSq) statistic. Infit values around 1.0 indicate expected variation, with values between approximately 0.5-1.5 commonly taken as acceptable in operational assessment contexts (Lunz & Stahl, 1990). High infit suggests excess, unsystematic variation (misfit); very low infit indicates over-predictability (overfit).

Data preparation

For MFRA purposes, scripts from both tasks were combined into a single dataset and subjected to routine preprocessing prior to calibration. The resulting FACETS analysis comprised 39,240 criterion-level rating observations, distributed across four analytic criteria, two scoring sources, 22 task prompts, and 2,394 test takers. This dataset represents the complete set of valid rater–response interactions retained for multifaceted calibration.

Model Fit and Facet Maps

Overall model fit was examined prior to interpretation of individual facets. Following Linacre (2002), satisfactory fit is indicated when no more than approximately 5% of standardised residuals exceed ±2 and no more than 1% exceed ±3. In the present analysis, 39,240 valid responses contributed to parameter estimation. Of these, 21 responses (0.05%) exceeded ±2, and 79 responses (0.20%) exceeded ±3, well within acceptable limits.

As some raw scores included decimal points, all scores were multiplied by 10 to meet FACETS input requirements. This transformation affects the scale of reported scores, as in some cases (total score, observed score, fair average), the output appears 10 times larger than it actually is, however, fit and measure are not affected.

Figure 1 presents the Wright map displaying the relative locations of all facets. Stricter markers are located higher up the ruler; more lenient markers down the ruler. More able test takers are located higher up the scale, less able test takers further down the scale. Marking criteria shown higher are harder (i.e., test takers are awarded lower scores on these criteria); criteria which are lower are easier.

Figure 1: *Facet maps*

```
+-----------------------------------------------------------------------------------------+
|Measr|-Markers          |+Test takers|-Questions|-Criteria                         |WRITE|
|-----+------------------+------------+----------+----------------------------------+-----|
|   2 +                  +            +          +                                  + (8) |
|     |                  |            |          |                                  | --- |
|     |                  |            |          |                                  |     |
|     |                  | .          |          |                                  |     |
|     |                  |            |          |                                  |  7  |
|     |                  | .          |          |                                  |     |
|     |                  | .          |          |                                  | --- |
|     |                  | .          |          |                                  |     |
|     |                  | *.         |          |                                  |     |
|     |                  | *.         |          |                                  | --- |
|   1 +                  + **.        +          +                                  +     |
|     |                  | ****       |          |                                  |  6  |
|     |                  | ****.      |          |                                  | --- |
|     |                  | ****.      |          |                                  |     |
|     |                  | *******.   |          |                                  | --- |
|     |                  | *******.   |          |                                  |  5  |
|     |                  | ********.  |          |                                  | --- |
|     |                  | *********. |          |                                  |     |
|     |                  | *******.   |          |                                  |  4  |
|     |                  | ******.    | *****.   | Coherence                        | --- |
*   0 * Human Automarker * *****.     * *.       * Grammar                          *     *
|     |                  | ****.      | *.       | Task Fulfilment  Vocabulary      | --- |
|     |                  | ***.       | **       |                                  |  3  |
|     |                  | **.        |          |                                  | --- |
|     |                  | *.         |          |                                  |     |
|     |                  | *.         |          |                                  | --- |
|     |                  | .          |          |                                  |     |
|     |                  | .          |          |                                  |  2  |
|     |                  | .          |          |                                  | --- |
|     |                  | .          |          |                                  |     |
|  -1 +                  + .          +          +                                  +     |
|     |                  | .          |          |                                  |     |
|     |                  | .          |          |                                  |     |
|     |                  |            |          |                                  | --- |
|     |                  |            |          |                                  |     |
|     |                  | .          |          |                                  |     |
|     |                  |            |          |                                  |     |
|     |                  | .          |          |                                  |     |
|     |                  | .          |          |                                  |  1  |
|     |                  | .          |          |                                  |     |
|  -2 +                  +            +          +                                  +     |
|     |                  |            |          |                                  |     |
|     |                  |            |          |                                  |     |
|     |                  | .          |          |                                  |     |
|     |                  |            |          |                                  | --- |
|     |                  |            |          |                                  |     |
|     |                  |            |          |                                  |     |
|     |                  |            |          |                                  |     |
|     |                  |            |          |                                  |     |
|     |                  |            |          |                                  |     |
|  -3 +                  +            +          +                                  + (0) |
|-----+------------------+------------+----------+----------------------------------+-----|
|Measr|-Markers          | * = 15     | * = 2    |-Rating_Scales                    |WRITE|
+-----------------------------------------------------------------------------------------+
```

As can be seen from the map, test takers show a spread of approximately three logits, while marker, criterion, and task facets cluster closely around the zero-logit line, indicating broadly comparable behaviour across these facets. Task prompt difficulty spans a narrow range (approximately −0.21 to +0.15 logits), with infit values close to 1.0, suggesting minimal variation in prompt difficulty.

Marker Facet

Table 5 presents the MFRA statistics for the two scoring sources.

Table 5: *Marker Measurement Report*

```
+-----------------------------------------------------------------------------------------------+
| Total   Obsvd  Fair(M)|   Model       |    Infit      Outfit          |                       |
| Count  Average Average| Measure  S.E.|  MnSq ZStd  MnSq ZStd |Discrm| PtBis   | N Markers    |
|-------------------------------+------------+----------+------------------------------+------|
|19344      44.04   43.86 |     .01    .00 |   .82  -9.0    .80 -9.0 | 1.34 |     .40   | 2 Automarker|
|19896      44.00   44.31 |    -.01    .00 | 1.18    9.0  1.15   9.0 |   .67 |     .39   | 1 Human      |
|-------------------------------+------------+------------------------------+------------|
|19620.0    44.02   44.08 |     .00    .00 | 1.00     .0    .97    .0 |       |     .39   |              |
|  390.3      .03     .32 |     .01    .00 |   .25   12.7    .24 12.7 |       |     .01   |              |
+-----------------------------------------------------------------------------------------------+
|Model, Sample: RMSE .00  Adj (True) S.D. .01  Separation 3.18  Strata 4.58  Reliability .91 |
|Model, Fixed (all same) chi-squared: 11.1  d.f.: 1  significance (probability): .00          |
-----------------------------------------------------------------------------------------------+
```

Marker reliability was high (0.91), indicating consistent internal ranking of scripts. Both the human and automated markers were located close to the zero-logit line, suggesting comparable overall severity. Infit statistics for both sources fell comfortably within accepted operational thresholds.

Task Prompt Facet

In previous MFRA analyses, Tasks 1 and 2 showed very comparable statistics: both were located around the zero-logit line indicating that, as facets, they were operating very similarly. Consequently, the analysis below presents a composite analysis of the questions that constitute both Tasks 1 and 2. Table 6 reports Task Prompt Measurement statistics (labelled as "Questions" in FACETS output).

Table 6: *Task Prompt Measurement Report*

```
+---------------------------------------------------------------+
|  Fair(M)|  Model          |  Infit          Outfit  |           |
|  Average|  Measure   S.E. | MnSq  ZStd      MnSq ZStd | Question  |
|--------------------------------+--------------+----------------
|   42.28 |    .06    .01   | 1.21    6.7   1.17    5.3 |  57408    |
|   43.59 |    .02    .01   | 1.17    5.3   1.13    4.0 |  59651    |
|   42.78 |    .04    .01   | 1.16    4.9   1.11    3.3 |  59075    |
|   41.74 |    .08    .01   | 1.14    4.7   1.11    3.6 |  58195    |
|   41.23 |    .09    .01   | 1.12    3.6   1.08    2.5 |  74112    |
|   46.95 |   -.09    .01   | 1.09    2.3   1.10    2.6 |  67259    |
|   42.47 |    .05    .01   | 1.04    1.2    .98    -.5 |  57595    |
|   40.81 |    .11    .01   | 1.03     .9    .99    -.1 |  58140    |
|   49.79 |   -.19    .01   |   .99   -.2    .99    -.2 |  70394    |
|   50.32 |   -.21    .01   |   .98   -.5   1.00     .0 |  66924    |
|   40.34 |    .12    .01   |   .99   -.1    .96   -1.3 |  71701    |
|   43.63 |    .01    .01   |   .98   -.8    .95   -1.8 |  72344    |
|   42.13 |    .06    .01   |   .98   -.7    .94   -1.8 |  59565    |
|   40.02 |    .13    .01   |   .92  -2.7    .91   -3.2 |  58194    |
|   41.90 |    .07    .01   |   .94  -2.2    .91   -3.2 |  59782    |
|   47.48 |   -.11    .01   |   .90  -2.9    .89   -3.0 |  69580    |
|   49.24 |   -.17    .01   |   .89  -4.4    .90   -3.9 |  70134    |
|   47.65 |   -.12    .01   |   .89  -2.9    .88   -3.1 |  67370    |
|   39.54 |    .15    .01   |   .89  -3.9    .87   -4.5 |  59776    |
|   42.24 |    .06    .01   |   .84  -5.7    .81   -6.9 |  72343    |
|   48.66 |   -.15    .01   |   .80  -5.5    .79   -5.7 |  67356    |
|--------------------------------+--------------+----------------
| 44.04   |    .00    .01   | 1.00   -.1    .97    -.9 |           |
|  3.51   |    .12    .00   |   .12   3.7    .11    3.4 |           |
+------------------------------------------------------------------------------------
Model, Sample: RMSE .01  S.D. .12  Separation 11.18  Strata 15.23  Reliability .99
Model, Fixed (all same) chi-squared: 2376.7 d.f.: 20 significance (probability): .00
------------------------------------------------------------------------------------
```

Task prompt reliability was high (0.99), and all prompts demonstrated infit and outfit values within the 0.5–1.5 range, indicating good fit to the model. Severity ranges from about –0.21 to +0.15 logits, i.e., < 0.5 logit spread. The narrow spread of prompt difficulty supports the interpretation that prompts functioned equivalently across the dataset.

Discrepancy Analysis

Discrepancy analysis examined cases in which differences between human and automated scores exceeded LANGUAGECERT's routine quality-assurance threshold, triggering review by a Chief Examiner. Under normal operational conditions, approximately 10% of scripts marked by two human raters require such review.

Table 7 below presents the figures for Tasks 1 and 2 where the differences between human and automarker scores met the criteria for third-marker review under this policy.

Table 7: *Task discrepancy figures between the two markers*

Task	Raw figure	Percentage discrepancy
1	37/2,389	1.55%
2	38/2,394	1.59%

In the present dataset, discrepancies meeting the review threshold occurred in approximately 1.5% of scripts for each task. These figures indicate that automated scores aligned closely with operationally finalised human scores and generated substantially fewer review-triggering cases than are typically observed in human–human marking.

Taken together, the results suggest that, under operational conditions, the automated scoring system exhibits stable and coherent scoring behaviour, consistent with its intended role in Phase 1 of the validation programme. However, it important to emphasise that Marker 1 represents quality-assured operational scores after LANGUAGECERT moderation procedures. Approximately 10% of scripts underwent Chief Examiner adjudication; others represent concordant double-marking. The 1.5% automarker discrepancy rate therefore reflects disagreement with moderated outcomes, not pre-moderation human-human variance. Pre-moderation variability is not analysed in Phase 1. It is however a key issue and will be examined in Phase 2's independent multiple-rater design.

Limitations

The scope of the present study is intentionally bounded by its role within a staged validation programme. While the automated scores are compared against operationally finalised human scores produced through LANGUAGECERT's standard marking and moderation processes, the design does not incorporate multiple independent human ratings of each script. As a result, the study does not estimate the full distribution of human inter-rater variability or define a human benchmark beyond the operational composite.

This reflects the specific evidential purpose of Phase 1, which is to examine large-scale operational behaviour, score stability, and construct-aligned performance under realistic scoring conditions. Questions that require independent rater replication—such as fine-grained estimates of rater severity differences or direct comparison of automated–human agreement against the full range of human–human variability—lie outside the scope of this phase.

Subsequent phases of the validation programme address these complementary questions through designs that incorporate multiple independent human raters and more controlled rating conditions. The present findings should therefore be interpreted as foundational operational evidence that supports, but does not replace, later precision- and fairness-focused validation work.

Conclusion

This study reported Phase 1 of a staged validation programme for the LANGUAGECERT Automated Writing Scoring System, focusing on its operational behaviour on the LANGUAGECERT Academic Writing Test (LCAWT). The analyses drew on a large, operationally representative dataset comprising 2,394 test takers, each completing two writing tasks. Automated scores were compared with operationally finalised human scores produced through LANGUAGECERT's standard marking and moderation processes, reflecting authentic scoring conditions in high-stakes assessment.

Across both tasks, automated and human scores demonstrated strong correspondence at the total-score level, with correlations of 0.87. Many-Facet Rasch Analysis indicated acceptable fit for marker, criterion, and task prompt facets, with stable behaviour across scoring dimensions. Discrepancy rates relative to routine quality-assurance thresholds were low, and substantially below those typically observed in human–human marking under comparable conditions.

Taken together, the findings indicate that the automated scoring system operates coherently within the intended assessment framework and produces score patterns that are interpretable alongside human judgements under operational conditions. In line with the evidential purpose of Phase 1, these results provide foundational operational evidence regarding score stability, construct alignment, and system behaviour at scale. They do not constitute a complete validation of automated scoring performance, nor do they address questions requiring independent human rater replication or subgroup-level fairness analysis.

These complementary questions are addressed through subsequent phases of the validation programme, including controlled multi-rater precision studies and analyses of subgroup behaviour. The present study establishes an appropriate empirical foundation for this continued work by demonstrating that the automated scoring system behaves predictably and proportionately within a human-in-charge assessment model. In this way, the phased approach supports a cumulative and methodologically coherent validation strategy, with the current study serving as a necessary first stage rather than a definitive endpoint.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650/1492

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021). Validating the LANGUAGECERT Test of English scale: the adaptive test. London, UK: LANGUAGECERT.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*. (pp. 4171-4186).

Eckes, T. (2015). Introduction to many-facet Rasch measurement. Frankfurt am Main: Peter Lang.

Foltz, P.W., Laham, D., and Landauer, T.K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), http://imej.wfu.edu/articles/1999/2/04/index.asp

Hatch, E. & Lazaraton, A. (1991). The Research Manual. Heinle and Heinle: Boston, MA.

Kostic, M., Witschel, H. F., Hinkelmann, K., & Spahic-Bogdanovic, M. (2024). LLMs in Automated Essay Evaluation: A Case Study. *Proceedings of the AAAI Symposium Series*, *3*(1), 143-147. https://doi.org/10.1609/aaaiss.v3i1.31193

Kwon, S. Y., Bhatia, G., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Beyond English: Evaluating LLMs for Arabic grammatical error correction. arXiv preprint arXiv:2312.08400.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. Journal of Applied Measurement, 3(1), 85-106.

Linacre, J. M. (2024) FACETS computer program for many-facet Rasch measurement. Beaverton, Oregon: Winsteps.com.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19(3), 246–276. https://doi.org/10.1191/0265532202lt231oa

Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. Evaluation and the Health Profession, 13, 425-444.

Mansour, W., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can Large Language Models Automatically Score Proficiency of Written Essays?. arXiv preprint arXiv:2403.06149.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48* , 238–243.

Page, E. B. (2003). Project essay grade: PEG. In: M. D. Shermis & J. C. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 43–54). Hillsdale, NJ: Erlbaum

Papargyris, Y., & Yan, Z. (2022). Examiner quality and consistency across LANGUAGECERT Writing Tests. International Journal of TESOL Studies, 4(1), 203-212. https://doi.org/10.46451/ijts.2022.01.13

Ramineni, C., & Williamson, D.M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*, 25-39.

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. ReCALL (2024), 36(2), 187–209.

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, *31*(2), 57–71. doi:10.1017/S0261444800012970

Weigle, S. (1998). Using FACETS to model examiner training effects. Language Testing, 15(2), 263-287.

Weigle, S.C. (2002). Assessing Writing. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language Testing and Validation*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Xiao, C., Ma, W., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. arXiv preprint arXiv:2401.06431.

LANGUAGECERT®

by PeopleCert

MILLIONS OF EXAMS DELIVERED WORLDWIDE

LANGUAGECERT is an Awarding Organisation recognised by Ofqual.
It spearheads innovations in language assessment and certfication, providing high-quality services to the global learners' community. It is a UK-based member of PeopleCert Group, a global leader in the certification industry, that delivers millions of exams in over 200 countries.

Learn more about LANGUAGECERT exams at:
www.languagecert.org

LANGUAGECERT is a business name of PeopleCert
Qualifications Ltd, UK company number 09620926

✉ info@languagecert.org   f /LanguageCert.org