LANGUAGECERT®

Exploring LLM Simulations for Pre-Operational Prediction of Reading Test Item Difficulty

David Coniam

Michael Milanovic

Leda Lampropoulou

March 2026

Abstract

This paper reports on an exploratory simulation-based investigation into the potential use of a large language model (LLM) to predict item facility values for an objective reading test. The study used a 30-item multiple-choice reading test from the LANGUAGECERT Academic (LCA) suite, for which robust empirical data from 671 test takers were available. These item statistics provided a baseline against which the LLM's predictive performance could be evaluated.

Four progressively constrained simulation scenarios were designed. In each scenario, the LLM was instructed to respond to the test items as if it were test takers at specified CEFR proficiency levels, generating multiple responses per level. The resulting datasets were analysed using classical test statistics and compared with empirical item facility values derived from live administrations. Later scenarios incorporated increasingly realistic CEFR-level distributions and, in the most constrained condition, a detailed CEFR-aligned syllabus.

Under constrained conditions, the model produced test-level statistics that closely approximated empirical results, with a mean facility discrepancy of 0.02 in the most aligned scenario. However, variation remained at the individual item level, with a small number of items displaying notable divergence from observed values. These findings suggest that while simulation-based LLM approaches may approximate aggregate test behaviour, consistent item-level precision remains challenging.

The study highlights both the potential and the limitations of simulation-based psychometric prediction. Further refinement of prompting strategies and hybrid modelling approaches may enhance precision, while complementary data-driven methods grounded in calibrated item banks may offer a more stable basis for operational implementation.

Suggested citation

Coniam, D., Milanovic, M., & Lampropoulou, L. (2025). *Exploring LLM Simulations for Pre-Operational Prediction of Reading Test Item Difficulty*. LANGUAGECERT.

Introduction

The rapid emergence of Large Language Models (LLMs) such as ChatGPT-4 has generated much interest in their potential applications to many fields. The English language assessment team at LANGUAGECERT has been exploring such applications since 2022. There has been research on areas related to test development, such as automated item generation (Dourda & Jones, 2026; Jones & Dourda, 2026) or the development of preparation and practice materials (Milanovic & Jones, 2026) as well as areas related to test security such as deep fake technology, item harvesting and autogenerated responses. While LLM tools have demonstrated remarkable capabilities in generating human-like responses, completing complex linguistic tasks, and even passing standardised examinations, their use in the technical domain of psychometric prediction remains comparatively unexplored.

At LANGUAGECERT, attention has recently focused on a programme of work investigating the extent to which LLMs might be used to predict item statistics normally accessible only after pretesting or live administration. This paper presents findings from an initial exploratory study that considered the extent to which an LLM (GPT-4) could meaningfully approximate item performance statistics for language assessment, specifically focusing on item facility (IF) values for a reading comprehension test.

The motivation for this research stemmed from a range of practical challenges in language test development. These include the need to pretest items before operational deployment, to generate vast quantities of test materials to mitigate item harvesting and malpractice, and to maintain stable test difficulty while scaling production. Traditional pretesting is resource-intensive, requiring appropriate test-taker samples, pilot test administrations, and post-hoc statistical analyses. If LLM-based approaches could provide reasonably accurate preliminary estimates of item performance, they might serve as an additional screening tool, helping test developers identify potentially problematic items before investing in full-scale pretesting or they may even contribute to the generation of items with preliminary predicted statistical values prior to empirical validation.

The study reported here used a simulation-based methodology, asking an LLM (GPT-4) to respond to test items as if it were test takers at different Common European Framework of Reference (CEFR) levels. By generating multiple responses at each proficiency level and analysing the resulting response patterns, the study examined whether predicted item facility values approximate empirically observed data from actual test administrations. Four progressively refined scenarios were explored, each introducing additional constraints and contextual information intended to improve prediction accuracy.

The investigation used a 30-item reading test from the LANGUAGECERT Academic (LCA) suite, for which robust empirical data from 671 test takers were available. The availability of confirmed item statistics provided a solid baseline against which the LLM's predictive performance could be evaluated. Through systematic manipulation of instructions, sample distributions, and background information provided to the model, the study examined whether an LLM can predict item difficulty, as well as what types of scaffolding and constraints might be necessary to optimise its performance in this task.

Literature Review: AI in English language assessment

While earlier generations of "artificial intelligence" (AI) have been used in educational assessment since the late 1960s, interest in AI applications intensified in the early 2020s, particularly following the emergence of Large Language Models (LLMs). In language assessment, this has accelerated research and operational experimentation across multiple parts of the assessment lifecycle, from item development to scoring and delivery.

One of the most established applications of AI in language assessment is automated marking and scoring, particularly for productive skills such as writing and speaking. Automated essay scoring systems have been studied extensively in relation to reliability, construct representation, and operational deployment (Attali & Burstein, 2006; Coniam et al., 2025; Williamson et al., 2012). Similarly, research on automated spoken-response scoring has demonstrated how speech recognition, acoustic features, and supervised learning models can support large-scale speaking assessment, while emphasising the need for careful validation and monitoring (Zechner et al., 2009; Bernstein et al., 2010). Across both writing and speaking contexts, the assessment literature consistently frames automated scoring not as a replacement for psychometric principles, but as a system that must be evaluated within established measurement frameworks.

Another important strand of technology-enabled assessment concerns computerised adaptive testing (CAT) and related models of dynamic test assembly. In CAT, item selection occurs during test administration based on real-time ability estimates derived from item response theory, allowing the test to adapt to the test taker's performance (Wainer, 2000; Van der Linden & Glas, 2010). Similarly, linear-on-the-fly test assembly (LOFT) approaches construct equivalent test forms dynamically from large, calibrated item banks using optimisation algorithms (Van der Linden, 2005). These models demonstrate that "on-the-fly" delivery is feasible and operationally robust when grounded in empirically estimated item parameters and well-defined measurement models. Crucially, both adaptive selection and dynamic assembly rely on stable, pre-calibrated item characteristics derived from field data.

A developing strand of AI in language assessment is automated test item generation (AIG). Traditionally, AIG has relied on structured item and cognitive models to generate large item banks while maintaining construct alignment and psychometric control (Gierl & Lai, 2012, 2016). With the emergence of LLMs, research has increasingly explored their use across pre-generation, generation, and post-generation stages of the AIG process (Dourda & Jones, 2026; Tan et al., 2025). A recent review of 60 empirical studies found that models such as T5, BERT, and GPT variants are widely used to generate items across domains and languages; however, evidence regarding key measurement properties, such as difficulty, discrimination, and reliability remains limited (Tan et al., 2025). This distinction between linguistic generation and psychometric validation remains central in high-stakes assessment contexts.

Taken together, these strands of research illustrate both the potential and the constraints of AI integration in language assessment. Despite advances in automated scoring, item generation, and adaptive delivery, these applications share a common feature: they rely on empirically derived data and established measurement frameworks. Automated scoring models are trained on large corpora of human-rated responses, and adaptive testing depends on pre-calibrated item parameters estimated from field data. By contrast, comparatively little research has examined whether LLMs can approximate psychometric item properties prior to pretesting. While simulation-based approaches may offer preliminary insights, the extent to which they can produce stable and operationally reliable item-level estimates remains an open empirical question. Given the operational importance of pretesting in large-scale language assessment, further empirical investigation of this issue is warranted.

Research Question

The broad research question being pursued involved the extent to which an LLM such as GPT-4 would be able to predict item facility values for objective reading test items in the context of test takers at different CEFR levels. Specifically, the study sought to determine whether a simulation-based approach, in which the model is instructed to respond as test takers at different proficiency levels, can produce item facility estimates that are sufficiently close to empirically observed values to be of practical use in test development.

The Study

As outlined above, the study adopted a simulation-based methodology, asking GPT-4 (a large language model developed by OpenAI) to respond to a set of objective reading test items as if it were test takers at different CEFR proficiency levels. By generating multiple responses at each level and analysing the resulting response patterns using classical test statistics, the study examined whether AI-generated data could yield item facility values comparable to those derived from live test administrations.

The investigation reported in this paper was conducted in mid-2025. It used a 30-item multiple-choice reading test from the LCA suite, for which robust empirical data from 671 test takers were available. These empirically derived item statistics provided a stable baseline against which the model’s predictions could be evaluated. The test is level-agnostic, covering CEFR levels B1–C2, with scoring aligned to both CEFR and the

LANGUAGECERT Global Scale.

To explore the feasibility of simulation-based prediction under different conditions, four scenarios were designed, each introducing progressively tighter constraints intended to improve the plausibility of the simulated data. Across all scenarios, the LLM was instructed to respond to each item multiple times as if it were a test taker at specified CEFR levels. The resulting response sets were analysed using classical test statistics, in line with standard LANGUAGECERT procedures for reading tests.

The four simulation scenarios are summarised in Table 1, which outlines the constraints applied, sample sizes, and CEFR levels represented in each condition.

In scenarios 1 and 2, the model was asked to generate responses with minimal constraints. In Scenario 1, it responded as if it were test takers at CEFR levels A1 through C2, producing 50 responses per level. Scenario 2 restricted responses to CEFR levels A2 through C1, again with equal numbers of responses per level. These scenarios were intended to establish baseline behaviour and to observe how the model distributed responses across proficiency levels in the absence of realistic population constraints.

Scenario 3 introduced a more realistic test-taker distribution. In this condition, the composition of the requested response sets was designed to resemble the CEFR distribution observed in the baseline live test. The LLM was instructed to generate responses reflecting the actual CEFR distribution observed in the live test sample (below-B1 to C2), with response counts matched to the empirical dataset. This scenario aimed to examine whether aligning the simulated sample more closely with real test-taker populations would improve the correspondence between predicted and observed item statistics.

Scenario 4 built on Scenario 3 by adding a further layer of constraint. Its sample was specified as for Scenario 3. In addition to using the empirical test-taker distribution, the LLM was provided with a CEFR-aligned syllabus detailing relevant vocabulary, grammar, syntax, topics, and language functions, available to potential test takers from the LANGUAGECERT website. The intention was to determine whether supplying explicit linguistic and proficiency-level information would enable the model to produce response patterns more closely aligned with empirical data.

Table 1: *Four GPT analysis scenarios*

Scenario	Constraints	Sample size	CEFR levels to respond to
1	No constraints. The model should respond to each item as if it were an A1, A2, B1, B2, C1, C2 test taker.	300 50 responses to each level	A1 to C2
2	No constraints. The model should respond to each item as if it were an A2, B1, B2, C1 test taker.	200 50 responses to each level	A2 to C1
3	The model should respond to each item as if it were a below-B1, B1, B2, C1, C2 test taker.	671 below-B1:45 B1:199 B2:234 C1:146 C2:47	below-B1, B1, B2, C1, C2
4	Syllabus containing vocabulary, topics, functions, grammar pertinent to specific CEFR levels uploaded for GPT to digest The model should respond to each item as if it were a below-B1, B1, B2, C1, C2 test taker.	671 below-B1:45 B1:199 B2:234 C1:146 C2:47	below-B1, B1, B2, C1, C2

For all scenarios, the model was provided with the full set of test items; however, it was not given access to the item keys or any information about correct answers. Responses were recorded in structured Excel templates corresponding to each CEFR level, and item-level and test-level statistics were subsequently calculated and compared with the live test data.

The detailed instructions provided to the model for Scenario 3 are illustrated in Figure 1. The instructions were broadly similar across all scenarios, with variations primarily in the specified sample sizes. In Scenario 4, the LLM was additionally informed that a CEFR-aligned syllabus would be uploaded and that it should digest this information before generating revised predictions.

Figure 1: *Instructions for GPT*

I want to get input as to how English language reading test items work.

I am going to upload to you a 30-item Reading Test, and an Excel Response Template containing a grid for responses.

In the Excel file there are 5 worksheets. Each worksheet contains space for responses to all 30 items by test takers as CEFR levels below-B1 (b-B1), B1, B2, C1, C2 respectively.

Record your answers to each item on the appropriate Excel worksheet.

Can you respond to each of the 30 Reading Test items the number of times below according to the level of below-B1, B1, B2, C1, C2 test takers.

Level	N
Below-B1 (b-B1)	45
B1	199
B2	234
C1	146
C2	47

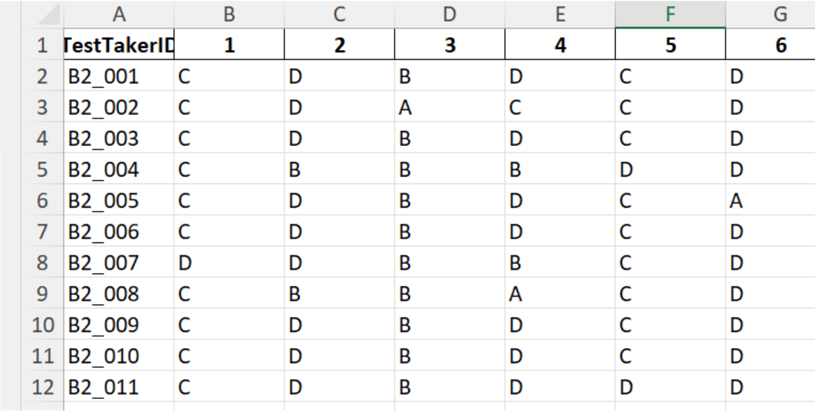
When you respond to the test items, follow the restrictions below

Test part	Items	Permissible responses
1A	1-6	A-D
1B	7-11	A-C
2	12-17	A-H
3	18-24	A-D
4	25-30	A-D

Make sure you give thought to each item. Do not just answer randomly.

Following completion of each simulation condition, the model produced response datasets that were analysed using classical test statistics to evaluate alignment with empirically observed item behaviour. An example of the response dataset generated for one CEFR level is shown in Table 2, illustrating the structure of the simulated data used for subsequent statistical analysis.

Table 2: *GPT's predicted responses*



	A	B	C	D	E	F	G
1	TestTakerID	1	2	3	4	5	6
2	B2_001	C	D	B	D	C	D
3	B2_002	C	D	A	C	C	D
4	B2_003	C	D	B	D	C	D
5	B2_004	C	B	B	B	D	D
6	B2_005	C	D	B	D	C	A
7	B2_006	C	D	B	D	C	D
8	B2_007	D	D	B	B	C	D
9	B2_008	C	B	B	A	C	D
10	B2_009	C	D	B	D	C	D
11	B2_010	C	D	B	D	C	D
12	B2_011	C	D	B	D	D	D

Analysis

Following the completion of GPT's predictions for each scenario, all output underwent classical test statistics (CTS) analysis. The analysis focused first on test-level statistics, followed by a more detailed examination of item-level performance, with item facility values used as the primary basis for comparison.

Overall test statistics

This section reports test-level statistics for the live test alongside the four GPT-generated scenarios. Table 3 presents test and group mean scores, standard deviation (SD), and reliability (alpha). The live test statistics are shown in the second column in red font and serve as the empirical baseline. The LANGUAGECERT Academic Reading test targets CEFR levels B1 to C2, although some test takers obtain scores below B1; these are referred to here as "below-B1".

Table 3: *Test-level statistics*

Test type	Live test	GPT			
Scenario		1	2	3	4
Levels assessed	Below-B1 to C2	A1 to C2	A2 to C1	below-B1 to C2	below-B1 to C2
Sample size		50 per level	50 per level	as per live test	as per live test
Other					with Syllabus
Item total	30	30	30	30	30
Sample total	671	300	200	671	671
Test mean	17.03	17.59	21.09	19.39	16.63
Test mean%	56.77	58.63	70.3	64.62	55.43
SD	5.91	11	7.17	6.27	5.96
SD%	19.70	36.67	23.90	20.90	19.87
Cronbach's α	0.84	0.97	0.92	0.85	0.85

Scenarios 1 and 2 were demonstrably out of scope and therefore offer no value for operational test development. In scenario 1, the overall mean appeared superficially close to that of the live test; however, this was effectively due to the inclusion of extreme A1 and C2 groups. Scenario 2, which excluded these extreme groups, yielded a substantially higher mean score, reflecting the absence of lower-ability test takers.

Scenarios 3 and 4 showed closer alignment with the live test at the test level. In particular, Scenario 4, where GPT was provided with a CEFR-aligned syllabus, produced a mean score (55.43%) that is the closest to the live test mean (56.77%).

Reliability estimates (Cronbach's alpha) were extremely high for Scenarios 1 and 2. This is hardly surprising given the artificial stratification with equal sample sizes across A1 to C2. The standard deviations (SD) in these scenarios are also markedly inflated (e.g., SD= 11 in Scenario 1 compared with 5.91 in the live test), primarily due to the inclusion of extreme proficiency groups in equal proportions. In contrast, alphas and SDs for

Scenarios 3 and 4 were more comparable to those of the live test, suggesting that, at an aggregate level, the simulated data can approximate global test characteristics when more realistic distributions are used.

To further exemplify mean scores, detail on group (i.e., CEFR level) performance is presented in Table 4.

Table 4: *Group distributions*

	Live test		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Group	N	Mean %	N	Mean %	N	Mean %	N	Mean %	N	Mean %
A1	–	–	50	4.8	–	–	–	–	–	–
A2	45	21.34	50	20.87	50	33.93	45	25.85	45	21.85
B1	199	39.10	50	51.07	50	68.13	199	46.21	199	38.64
B2	234	57.38	50	81	50	86.53	234	68.21	234	55.71
C1	146	76.65	50	94.73	50	92.6	146	86.03	146	76.96
C2	47	92.75	50	99.27	–	–	47	95.39	47	90.43
Means	671	56.77	300	58.63	200	70.3	671	64.62	671	55.43

Drawing on the overall group means for the live Reading Test, the whole group mean for all 671 test takers was 56.77%.

Under Scenario 1, A1 test takers scored almost zero and C2 test takers virtually 100%, an unlikely distribution in operational contexts. In Scenario 2, removing A1 and C2 test takers produced a more plausible distribution, though performance remained inflated at higher levels.

Scenario 3's group mean scores were reasonably close to the live test whole group although ability appeared to be overestimated somewhat at the C1 level.

Scenario 4 produced group mean scores that were reasonably close to those observed in the live test, with differences generally within a few percentage points. This suggests that increased constraint and contextual information can improve test-level and group-level approximation.

Item statistics

While test-level results provided an initial indication of plausibility, item-level behaviour was the critical focus of this study. Item facility (IF) values were therefore examined in detail across all scenarios. Item discrimination indices were above the accepted benchmark of 0.3 (Falvey et al., 1994) in most scenarios; however, for clarity and relevance, the discussion below focuses primarily on IF values. Table 5 presents IF values for the live test alongside those generated under the four ChatGPT scenarios.

Table 5: *Item analyses*

	Live test		Scenario 1	Scenario 2	Scenario 3	Scenario 4
			A1-C2	A1-C2	below B1-C2	below B1-C2
			50 per level	50 per level	distribution as per live test	distribution as per live test / with syllabus
Item	IF		IF	IF	IF	IF
R1	0.86		0.79	0.86	0.55	0.86
R2	0.84		0.74	0.90	0.55	0.86
R3	0.72		0.72	0.86	0.54	0.83
R4	0.74		0.68	0.85	0.57	0.83
R5	0.56		0.69	0.86	0.55	0.85
R6	0.38		0.69	0.82	0.57	0.84
R7	0.63		0.66	0.82	0.57	0.72
R8	0.68		0.69	0.82	0.56	0.70
R9	0.51		0.65	0.80	0.50	0.72
R10	0.89		0.66	0.81	0.52	0.73
R11	0.82		0.61	0.79	0.59	0.70
R12	0.40		0.68	0.73	0.57	0.71
R13	0.49		0.62	0.77	0.56	0.45
R14	0.54		0.58	0.72	0.54	0.51
R15	0.27		0.61	0.72	0.54	0.48
R16	0.55		0.59	0.71	0.54	0.52
R17	0.31		0.58	0.66	0.50	0.54
R18	0.50		0.58	0.68	0.54	0.51
R19	0.56		0.50	0.64	0.53	0.53
R20	0.60		0.54	0.63	0.57	0.36
R21	0.62		0.52	0.64	0.52	0.33
R22	0.44		0.52	0.67	0.57	0.31
R23	0.40		0.52	0.62	0.55	0.32
R24	0.57		0.53	0.52	0.55	0.37
R25	0.57		0.46	0.57	0.53	0.35
R26	0.28		0.40	0.56	0.57	0.31
R27	0.79		0.44	0.51	0.55	0.34
R28	0.53		0.44	0.57	0.54	0.37
R29	0.51		0.46	0.50	0.58	0.33
R30	0.47		0.45	0.46	0.55	0.35
MEAN	0.57		0.59	0.70	0.55	0.55

In Scenarios 1 and 2, GPT's predictions imposed a clear and artificial facility gradient across the test, with items becoming progressively more difficult as the test progresses. While such a pattern may appear intuitively reasonable, it was not reflected in the empirical data, where item difficulty was distributed more irregularly. Scenario 3 produced a different but still implausible pattern, with item facilities clustering closely together across items. This lack of variation again contrasted with the live test data and suggests that aligning simulated sample distributions alone is insufficient to capture realistic item behaviour.

Scenario 4 showed greater variation and, at face value, appeared to approximate the empirical pattern more closely. Table 6 compares IFs for the live test and Scenario 4 directly. The final column provides the difference between each set of IFs.

Table 6: *IFs for the Live Test and for Scenario 4.*

	Live Test	GPT Scenario 4	Difference
Item	IF	IF	(Live – GPT)
R1	0.86	0.86	0.00
R2	0.84	0.86	-0.02
R3	0.72	0.83	-0.11
R4	0.74	0.83	-0.09
R5	0.56	0.85	-0.29
R6	0.38	0.84	-0.46
R7	0.63	0.72	-0.09
R8	0.68	0.70	-0.02
R9	0.51	0.72	-0.21
R10	0.89	0.73	0.16
R11	0.82	0.70	0.12
R12	0.40	0.71	-0.31
R13	0.49	0.45	0.04
R14	0.54	0.51	0.03
R15	0.27	0.48	-0.21
R16	0.55	0.52	0.03
R17	0.31	0.54	-0.23
R18	0.50	0.51	-0.01
R19	0.56	0.53	0.03
R20	0.60	0.36	0.24
R21	0.62	0.33	0.29
R22	0.44	0.31	0.13
R23	0.40	0.32	0.08
R24	0.57	0.37	0.20
R25	0.57	0.35	0.22
R26	0.28	0.31	-0.03
R27	0.79	0.34	0.45
R28	0.53	0.37	0.16
R29	0.51	0.33	0.18
R30	0.47	0.35	0.12
MEAN	0.57	0.55	0.02

While some items showed relatively small differences, only two items exhibited discrepancies exceeding 0.4. The overall mean difference between live and simulated IF values was 0.02, indicating strong aggregate alignment (0.57 vs 0.55). However, variation at the individual item level remained evident.

Across the test as a whole, item-by-item correspondence varied, with some items closely aligned and others displaying moderate divergence. In particular, the model tended to overestimate facility for earlier items and underestimate facility for later items, again reflecting an imposed progression that was not supported by empirical evidence. Although the aggregate mean IF for Scenario 4 (0.55) is close to that of the live test (0.57), this overall similarity masked substantial item-level inaccuracies.

Discussion

This paper presented an exploratory, simulation-based analysis of whether an LLM can predict item facility values for an objective reading test. Four progressively constrained scenarios were examined, with findings that reveal both the potential and the persistent limitations of this approach.

The most constrained scenario, which used realistic test-taker distributions and a detailed CEFR-aligned syllabus, produced test-level statistics closely aligned with empirical values (mean IFs of 0.55 vs 0.57 for the live test), with an overall mean discrepancy of 0.02 across items. Yet aggregate correspondence masked considerable item-level variability: two items showed discrepancies exceeding 0.4, and several others displayed moderate divergence.

A recurring pattern in the less constrained scenarios was the model's tendency to impose an artificial facility gradient, predicting items as increasingly harder across the test, a pattern absent from the empirical data, where difficulty was distributed more irregularly. Even with extensive scaffolding – vocabulary lists, grammatical and syntactic specifications, and topic descriptors for each CEFR level – the model could not fully capture the complex interaction between item characteristics and heterogeneous test-taker behaviour that underlies observed facility values. Strong linguistic competence, it seems, does not straightforwardly translate into stable psychometric modelling of population-level performance.

Two directions appear most promising for future work. The first concerns methodological refinement: improvements to prompting strategies, conditioning mechanisms, or hybrid modelling approaches may enhance item-level precision. The second shifts orientation more substantially. Rather than relying on simulated test-taker responses, data-driven methodologies trained on large corpora of calibrated items with confirmed psychometric properties could offer a more stable empirical foundation A system of this kind might also yield interpretable diagnostic output, identifying which item features drive predicted difficulty or discrimination values, and thereby supporting the attribution of reliable statistical characteristics to items produced through automated authoring.

These findings underscore the importance of rigorous methodological evaluation when integrating AI tools into language assessment. Simulation-based approaches show meaningful promise at the aggregate level, but require substantial refinement before they can inform operational decisions at the item level.

Conclusion

This study investigated whether LLM-based simulation of CEFR-level test-taker behaviour could yield reliable predictions of item facility values for an objective reading test. Under the most constrained conditions, aggregate alignment was strong, with a mean discrepancy of just 0.02 between predicted and observed values. Item-level precision, however, remained inconsistent: while most predictions fell within a reasonable range of empirical estimates, a small number of items showed notable divergence — a degree of variability that falls short of the stability required for high-stakes operational use.

The broader implication is that direct simulation of test-taker behaviour through LLMs is a methodologically demanding undertaking, one where current capabilities are better suited to approximating test-level tendencies than to modelling individual item behaviour with the reliability that assessment contexts demand. Progress will likely depend on advances in constraint design and prompting, alongside closer integration with empirically validated item banks and established psychometric frameworks. As AI capabilities continue to develop, their effective integration into language assessment will depend on sustained alignment with established principles of validity, reliability, and empirical calibration.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1–30. https://ejournals.bc.edu/index.php/jtla/article/view/1650

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355-377. https://doi.org/10.1177/0265532210364404

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

Dourda, C., & Jones, C. (2026). *AI-assisted item generation in high-stakes language assessment: A Design-Based Study of Human-AI Item Development*. LANGUAGECERT.

Falvey, P., Holbrook, J., & Coniam, D. (1994). *Assessing students*. Hong Kong: Longman.

Jones, C., & Dourda, C. (2026). *Making the invisible visible: how human-AI collaboration transforms item development practices in language testing.* LANGUAGECERT.

Milanovic, M., & Jones, C. (2026). *Mobile-based English language learning through bite-sized formative CEFR-aligned assessment*. LANGUAGECERT.

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education, 12*(2), 317–340. https://doi.org/10.21449/ijate.1602294

Van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer. https://doi.org/10.1007/0-387-29054-0

Van der Linden, W. J., & Glas, C. A. (Eds.). (2010). *Elements of adaptive testing* (Vol. 10). Springer.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.

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

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of nonnative spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009

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