Generating multiple-choice items for a B2 English reading test with GPT-4: targeting higher-order cognitive processing

Olena Rossi[1], Josep Maria Montcada Escubairó[2]

[1] Independent Researcher, Italy
[2] Department of Education and Vocational Training, Government of Catalonia

Corresponding author: Olena Rossi, olena.rossi@itemwriting.co

Abstract

This study investigated the potential of generative AI to produce multiple-choice reading comprehension items for B2-level English assessment, with a focus on higher-order cognitive processing. Using GPT-4 configured within a custom environment, 164 items were generated from six authentic texts aligned with official test specifications of the Escoles Oficials d'Idiomes (Catalonia). Items underwent expert review and were trialled with 775 test-takers. A triangulated analysis combined linguistic analysis, expert judgements, psychometric modelling, and test-taker feedback. Findings showed that GPT-4 frequently attempted to target higher-order cognitive processing, but the resulting items were often misclassified and suffered from flaws such as implausible distractors and text misinterpretation. An item generation log revealed unstable model behaviour across rounds. Linguistic analysis of item stems highlighted formulaic structures and GPT-4's confusion regarding the cognitive processing required for item completion. Expert reviewers confirmed that most items required substantial revision, with distractor plausibility and construct alignment as recurrent concerns. Psychometric indices indicated that the items exhibited acceptable model fit and discrimination but were generally easy for the trial group. The study concludes that GenAI can replicate surface features of items

targeting higher-order cognitive processing but rarely provides substantive coverage of complex reading processes.

Keywords:
Generative AI; Reading assessment; Multiple-choice items; Automated item generation; Language testing

1. Introduction

Reading comprehension is an important skill in second language (L2) learning. By the B2 level of the Common European Framework of Reference for Languages (CEFR), L2 learners are expected not only to retrieve explicitly stated information but also to engage in higher-order cognitive processing by making inferences, integrating ideas, and evaluating meaning (Council of Europe, 2020). These abilities, therefore, form a central focus of standardised reading assessments at higher proficiency levels. Multiple-choice (MC) items are widely used in assessment contexts because they target various types of cognitive processing and maintain efficiency in administration and scoring. However, they are notoriously difficult to write because effective MC items must elicit intended cognitive processes while avoiding construct-irrelevant clues (Haladyna et al., 2002).

The rapid spread of generative artificial intelligence (GenAI) has prompted growing interest in its potential to support item writing. Large language models (LLMs) can generate fluent text quickly and at scale, raising hopes that they might assist in or even partially automate test development. Recent studies suggest that GenAI can produce reading comprehension items that are linguistically coherent and psychometrically adequate, yet concerns remain about their construct validity, particularly when items are intended to elicit higher-order processing (Lin & Chen, 2024; Wen et al., 2025; Zhang et al., 2025). Against this backdrop, the present study examined whether GPT-4-generated MC items for B2 reading assessment meaningfully targeted

higher-order cognitive processing, evaluating item quality through a triangulated design that incorporated linguistic analysis, expert review, psychometric trialling, and test-taker feedback.

2. Literature review

2.1. Assessing reading comprehension at higher proficiency levels

As second language (L2) reading proficiency develops, learners progress from understanding concrete, local information to making inferences, evaluating implicit meaning, and comprehending the text as a whole (Council of Europe, 2020). Empirical research (e.g., Long & Chong, 2001; Oakhill et al., 2005) shows that less-skilled readers struggle to make inferences and build coherent textual models. Skilled readers, in contrast, generate inferences and connect ideas, enabling them to move beyond literal comprehension (Alderson, 2000; Grabe, 2009). This progression - from decoding local textual information to inferential, evaluative, and global comprehension - is reflected in contemporary models of reading comprehension. The cognitive model proposed by Khalifa and Weir (2009) presents a hierarchy of reading comprehension processes governed by metacognitive control. The model distinguishes between lower-level processes (such as word recognition and syntactic parsing) and higher-level processes (such as inference generation and constructing a situation model of the text), which interact during comprehension. The reading descriptors in the CEFR reflect similar distinctions, highlighting the progression from understanding local details to recognizing implicit meaning and grasping the overall message of texts (Council of Europe, 2020).

By the B2 level, learners typically decode efficiently at the sentence level, freeing cognitive resources for meaning-making beyond the sentence (Khalifa & Weir, 2009). Accordingly, reading comprehension tests at this level emphasise higher-order cognitive processing such as inferencing, meaning evaluation, and global comprehension. For example, many questions in the Cambridge B2 First examination require learners to “report not on information contained in the

text but upon what that information entails" (Khalifa & Weir, 2009, p. 75). MC formats are particularly effective at assessing higher-order cognitive processing because distractors can represent plausible but incorrect interpretations (Haladyna et al., 2002). For example, IELTS Academic Reading uses MC items to identify main ideas and writer purpose (Cullen et al., 2025). Similarly, TOEFL iBT includes MC questions that test higher-order cognitive processing such as inferencing and rhetorical purpose (Sawaki, 2017). However, MC items are challenging to produce, which is why the advent of GenAI - promising rapid generation of varied texts, including assessment items - has been greeted with considerable enthusiasm in language assessment. The next section reviews recent empirical research on this topic.

2.2. Automated generation of multiple-choice reading comprehension items to assess higher-order cognitive processing

Recent studies (e.g., Alshehri & Alharbi, 2025; Shin et al., 2025) have explored the use of GenAI to produce reading test items without distinguishing between types of cognitive processing. Alshehri and Alharbi (2025) used GPT-4 to generate texts and MC items for B1 learners, which were expert-reviewed and trialled with 150 students. Although they showed acceptable psychometric qualities, they were generally too easy, a result the authors interpreted as a GenAI bias toward lower-order information retrieval.

Several studies have examined whether LLMs can generate items targeting specific types of cognitive processing. Lin and Chen (2024) instructed GPT-3.5 to produce items targeting explicit detail, word meaning, inference, main ideas, and sentiment. Five experts evaluated item quality and classified each item by cognitive processing type. Judgements largely matched the model's labels, but many "explicit detail" items simply copied text, reducing the processing required to lexical matching. The model also overproduced "explicit detail" and underproduced higher-order items (e.g., inference or sentiment), which the authors linked to GPT-3.5's weak abstract reasoning. Wen and Chu (2025) also used GPT-3.5, applying the PIRLS framework, which

distinguishes four processes: (1) retrieving explicitly stated information, (2) making straightforward inferences, (3) interpreting and integrating ideas, and (4) evaluating and critiquing content. Two experts judged ChatGPT 3.5 to have excelled at generating factual retrieval questions but performed significantly worse on inference and evaluation. Other recurring issues were content oversimplification (complex ideas reduced to recall), and difficulty in handling nuanced meaning. The authors concluded that GPT-3.5 lacked the deeper reasoning required to produce items targeting higher-order comprehension.

Recent work has tested newer GPT models. Ma et al. (2025) prompted GPT-4 to generate inference questions from a taxonomy of inference subtypes. Three experts evaluated item quality and construct alignment: 93.8% of items were acceptable, but only 46.1% matched the intended inference subtype. Zhang et al. (2025) applied a framework of academic reading subskills to generate varied item types, rated by five experts and trialled with 132 university students. Many items were found too easy for the test-taker sample. Experts commented that vocabulary items lacked nuance, while overall the items allowed answers without deep engagement with the text.

Overall, findings are consistent. GenAI produces fluent MC items that function reasonably well psychometrically but are generally easier than human-written ones (Alshehri & Alharbi, 2025; Zhang et al., 2025). More critically, GenAI struggles to target higher-order cognitive processes such as inference, integration, and evaluation. Even advanced prompting, including chain-of-thought or iterative techniques, has not ensured construct fidelity (Lin & Chen, 2024; Ma et al., 2025; Wen & Chu, 2025), and newer LLMs have not considerably outperformed earlier ones.

Building on this evidence, important gaps remain. First, most prior studies relied on broad expert judgements without systematically cataloguing item flaws or quantifying how these flaws vary by type of cognitive processing. Second, no fine-grained linguistic analysis of generated items has been conducted, an analysis that would provide objective evidence linking surface item features to the ability to target higher-order cognitive processes. Third, studies typically

describe prompting approaches but do not document the generation process itself so model behaviour over time remains opaque. Finally, many reports rely exclusively on expert review without item trial data, and virtually none incorporate the test-taker perspective. The present study sought to address these gaps by investigating the quality of multiple-choice items generated by GPT-4 for a B2-level English reading comprehension test:

RQ1: To what extent, and with what frequency, do recurrent flaws appear in the generated items, and how do these vary across cognitive processing types?

RQ2: What linguistic properties characterise items intended to target higher-order cognitive processing?

RQ3: How do prompting techniques and workflow choices affect item quality?

RQ4: To what extent do the generated items meet quality standards when judged through a triangulated lens of expert review, psychometric analysis of item performance, and test-taker feedback?

3. Study Context

The *Escoles Oficials d'Idiomes* (EOI) are a network of publicly funded adult education centres governed by the *Departament d'Educació i Formació Professional* of the *Generalitat de Catalunya*. Operated under the *Servei de Llengües Estrangeres i d'Origen*, and specifically the *Àrea d'Ensenyament d'Idiomes EOI*, led by the second author of this paper, this network provides structured language instruction to both enrolled learners and external test-takers. EOIs also administer official language certification exams aligned with the CEFR, covering levels B1, B2, C1, and C2.

Certification exams are offered in several modern languages, including English, and are accessible to both current students and the wider public. The B2-level English exam is a

particularly high-demand qualification, often used to meet university entry requirements or for professional advancement. The English B2 certification, similar to exams at the other CEFR levels, consists of five separate components - reading, writing, listening, speaking, and mediation - designed to reflect real-world, communicative language use. The reading test uses authentic texts and follows a selected-response format, including MC, true/false, and multiple matching items. The reading test includes three or four tasks, each based on a different text type (e.g., newspaper articles, blog posts, essays). Test-takers answer a total of 25 to 30 questions, with a time allowance of 60 minutes (Departament d'Educació i Formació Professional, n.d.).

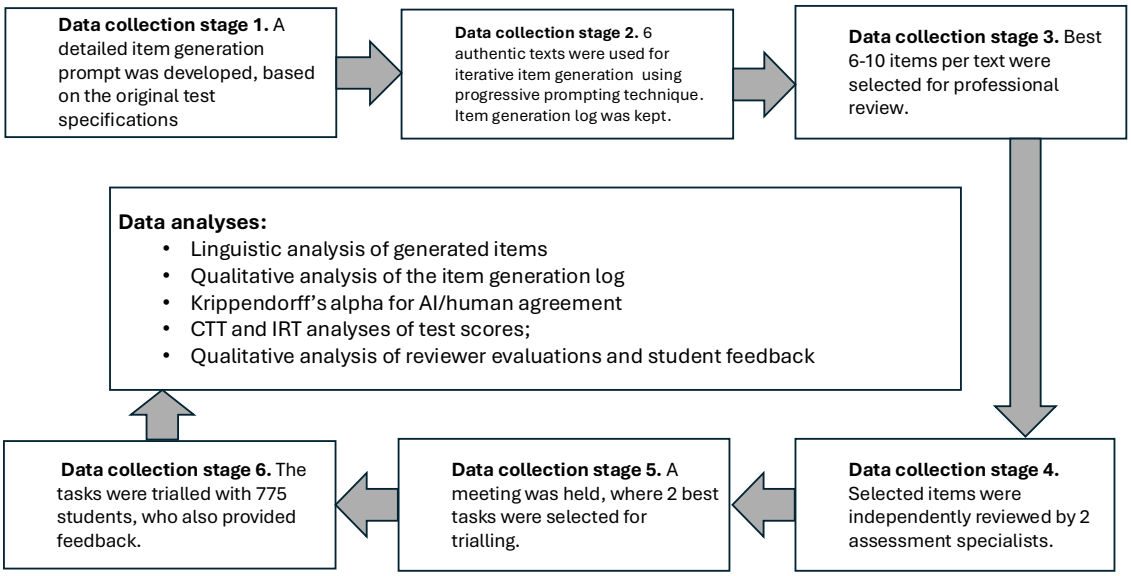
Within the *Àrea d'Ensenyament d'Idiomes EOI*, a structured and collaborative test development model is employed. Experienced language teachers from the EOI network are trained as item writers and work in small teams to develop test materials against detailed specifications. Each draft undergoes peer review within the writing group, followed by a professional review by trained assessment specialists within the department. Once reviewed, the items are trialled. Only those tasks that meet psychometric quality standards based on statistical analysis of trial data and qualitative feedback from students and teachers are selected for official test use. While this system has traditionally relied on expert-led human item development, the increasing scale of test delivery and the rise of GenAI have prompted the organisation to explore new avenues for item creation. The present study forms part of this broader innovation initiative.

4. Methodology

The methodology adopted for this study followed a multi-stage data collection process designed to develop, review, and trial B2-level reading comprehension items. An overview of the study's design is presented in *Figure 1* below. Each stage is described in detail in the sections that follow.

Figure 1

Overview of the Study Methodology



4.1. Prompt development and GPT setup

A detailed item generation prompt was developed based on the official test specifications provided by the *Àrea d'Ensenyament d'Idiomes EOI,* reflecting the cognitive demands and stylistic requirements of the B2 reading exam. The prompt instructed GPT-4, operating in ChatGPT Plus environment, to produce ten three-option MC items based on the text provided, targeting various types of cognitive processing: identification of specific details, understanding main ideas, global comprehension, lexical inference and inference of meaning, understanding attitudes and moods, identification of text function and type, and grasping implicit/allusive/connotative use. The prompt did not specify how many items should be assigned to each type of cognitive processing. The prompt also included detailed guidance on item structure, stem clarity, option plausibility, and language level appropriateness. The prompt also required GPT-4 to provide metadata for each item: the correct answer, the targeted cognitive process, and the relevant information in the text. The full version of the prompt is available through the Open Science Framework (OSF) https://osf.io/djeqh/overview?view_only=05c0fc44b4084859bfbbcec735506b86

A dedicated custom GPT was created in the 'My GPTs' section of ChatGPT Plus and assigned the persona of an experienced item writer, a common prompt-engineering technique also used in prior research (e.g., Lin & Chen, 2024). Moreover, the GPT was supplied with supporting documentation, including the official test specifications and sample tasks provided by the department, as well as guidelines on producing MC items.

4.2. Item generation

B2 reading comprehension items were generated by the first author using six texts selected from a pool of retired tasks that had been used in live certification exams five to ten years prior to the study. All texts were authentic and drawn from a variety of public sources, including magazines, newspapers, and online blogs. Some texts were slightly edited for length when originally used, while maintaining their original structure and tone.

Item generation was conducted using GPT-4 within the 'My GPTs' environment. The initial prompt was used for the first three texts and then iteratively refined to address recurring issues. For example, for Text 4 revisions included additional guidance to improve inferencing items, avoid lexical overlap with the source text, and ensure coverage of the cognitive processes identified by the model.

For most texts, three rounds of item generation were conducted, each producing ten items. After each round, items were briefly screened for quality, and notes were made on issues identified in individual items. Better quality items were copied to a task template. The first author maintained a three-day item generation log recording qualitative observations and recurring issues (see Table 1).

After the item generation, all selected items were reviewed to remove overlaps, and final sets of six to ten items – matching the number of items in the original retired tasks - were compiled for review. In selecting the final items, the primary considerations were quality and

independence, essential principles in reading item design. Item independence requires that no two items target the same content in the text, and that no item provides clues that could assist in answering another. Some items that had initially been marked as usable were later excluded for violating this principle, due to overlap between items generated in different rounds. There was no requirement to distribute items evenly across different cognitive processing types within each set.

Table 1

Excerpt from the Item Generation Log

Text 1 Round 1	*Generated items exceeded my expectations – main idea items did focus on main ideas rather than individual sentences in the text, there was almost no lexical overlap with the text. The inference items do target inference rather than specific details. Taken individually, most items seem usable. However, taken as a set, items overlap and give clues to each other, therefore I had to carefully compare them to select items for the final set.*
Text 1 Round 2	*This round was less encouraging: lexical inference items focus on reasonable phrases, but the distractors are so clearly incorrect that the answer is obvious and doesn't require inferencing. Implausible distractors, which make the item too easy to answer, was the main problem in this round of item generation. Other problems included lexical overlap (e.g., the item asks about the writer's profession, uses "freelance journalist" in the key which is the exact phrase used in the text, while the two distractors are never mentioned in the text), and targeting information that GPT has misinterpreted from the text...*

4.3. Item review

The compiled item sets were submitted for independent review by two assessment specialists from the *Departament d'Educació i Formació Professional*. Each reviewer completed a review checklist for every task (i.e., the set of items for a particular text) evaluated. The reviewers identified cognitive processes required to complete the item (e.g., global comprehension, inference of meaning, etc.) and assessed each item across multiple quality criteria. These included clarity and conciseness of the stem and options, structural consistency across options, correctness of the key, plausibility of distractors, and overall appropriateness for the B2 proficiency level. Reviewers also had the opportunity to provide written comments. Importantly,

reviewers did not make any edits to the items because the study focus was to evaluate GenAI's capability in producing high-quality test items. The full version of the item review checklist is available in OSF https://osf.io/djeqh/overview?view_only=05c0fc44b4084859bfbbcec735506b86

Reviewers were already familiar with the original test specifications through their routine work in test development. A follow-up meeting was held between the first author and the two reviewers. The purpose of the meeting was to jointly select the two most promising tasks (out of six) to advance to the trialling stage.

4.4. Item trialling

Following the selection of two tasks, a digital test form was created and prepared for online administration. The form included both reading tasks, each followed by a brief questionnaire designed to collect feedback on item quality. The questionnaire asked test-takers to rate the overall difficulty of the task, to identify specific items they found particularly difficult or easy, and to indicate whether any items were unclear or seemed unusual. The full version of the feedback questionnaire is available in OSF https://osf.io/djeqh/overview?view_only=05c0fc44b4084859bfbbcec735506b86

The test was administered online under controlled, exam-like conditions in 30 schools of the *Xarxa d'Escoles Oficials d'Idiomes de Catalunya* (Network of Official Schools of Languages in Catalonia), with 775 students participating. The trialling took place in October 2024 and involved learners who had completed a B2 English course the previous summer and had just started a C1 course. Trialling the tasks on students slightly above the B2 level was deliberate. The only alternative - using students at the start of their B2 course - would have introduced confounding factors related to insufficient proficiency. Administering the test to students well below the target level could have made it unclear whether incorrect responses reflected item flaws or limited ability. Although trialling with students slightly above B2 level may have made the

tasks being relatively easy, it allowed irregularities in response patterns to be more confidently attributed to item quality rather than test-taker ability.

4.5. Data analyses

Test-taker responses to the trialled items were analysed using both Item Response Theory (IRT) and Classical Test Theory (CTT) approaches. IRT analysis was conducted in Jamovi using the dichotomous Rasch model. An expected score curve was generated for each item. In parallel, CTT analysis was conducted in SPSS (version 30), where item discrimination indices were calculated.

Agreement between the GPT-4-generated labels and human reviewer classifications of the cognitive process targeted by each item was evaluated using Krippendorff's alpha. Quantitative data from item review checklists and test-taker feedback questionnaires were analysed using frequency counts and descriptive statistics. Optional textual comments – though limited in number - from both sources were analysed qualitatively by noting individual observations and identifying common themes.

Item generation data were organised in a spreadsheet to track item output, selection, and categorisation by cognitive processing type, based on metadata produced by GPT-4. For higher-order cognitive processing categories, item stems were collated and examined for question type and recurring structural and analytical patterns. Comments on item quality made during the initial screening phase were compiled and frequency counts were produced to quantify the occurrence of specific flaws (e.g., lexical overlap with the source text, ambiguous distractors, misalignment with the intended cognitive process). Finally, entries in the item generation log were thematically coded independently by both authors, yielding an inter-coder reliability of 91%. Any discrepancies were subsequently resolved through discussion.

All sources of data were triangulated in the final analysis to produce a comprehensive evaluation of the quality of GPT-4-generated items.

5. Findings

5.1. Findings from the item generation

A total of 164 items were generated across the six texts, of which 46 (approximately 28%) advanced to the review stage following initial screening. For each item, GPT-4 specified the target cognitive processing type; their distribution is shown in Table 2. Notably, GPT-4 did not limit itself to lower-order processing; on the contrary, it frequently produced items aimed at higher-order cognitive processes. The most frequently targeted cognitive process was *inferencing of meaning (n=49)*, followed by *specific details* (n=28) and *main ideas (n=24)*. In contrast, few items were generated for categories such as *identification of text function/type* (n=1) and *grasping implicit/connotative use (n=4)*. The third column of Table 2 displays the percentage of items in each category that were selected for further review. These percentages represent category-level proportions, not a total sum across all items. They indicate that items targeting higher-order cognitive processing were as likely to advance to review as those targeting lower-order processing.

Linguistic analysis of item stems for each higher-order cognitive process as labelled by GPT-4 revealed that, in most categories, items tended to follow consistent patterns. The analysis focused on those higher-order cognitive categories that contained more than ten items, as this quantity was sufficient to identify patterns.

Lexical inferencing

Nearly all lexical inferencing questions followed a uniform format - *What does X mean/imply/suggest?* or *What is implied by X?* - where *X* referred to idioms, metaphors, figurative expressions, or emotionally charged phrases drawn from the source text. For example:

*What **is implied** by John's muscles "starting to burn" during the swim?*

Table 2

Cognitive Processing Types: Item Counts and Percentage Forwarded for Review

Cognitive processing type (as stated by GPT-4)	N° items generated	% sent for review (within category)
Lower-order cognitive processing		
Specific details	28	35.7
Main ideas	24	16.7
Higher-order cognitive processing		
Lexical inferencing	20	30.0
Inferencing of meaning	49	24.5
Understanding attitudes/moods	16	31.2
Global comprehension	22	40.9
Identification of text function/type	1	0
Grasping Implicit/connotative use	4	0

Inferencing of meaning

Items in this category followed three broad analytical patterns:

1) Implied meaning in the text, using verbs *imply* and *infer*, for example:

*What **is implied** about the current job market for graduates?*

2) Authorial attitude or viewpoint, for example:

*What **can be inferred** about the author's view of being single?*

3) Implied mental state, mood, feelings or motivations of individuals other than the writer, for example:

*What **can be inferred** about Trimble's views on the criticism she received?*

Across this category, the question structure was highly repetitive: 43 out of 49 questions began with *What*, and 19 began with *What can be inferred*. Additionally, nine items asked about the meaning of specific expressions (e.g., *What does the phrase "all hell broke loose" imply about the narrator's experience upon arriving at the inn?*), indicating that GPT-4 did not consistently

differentiate between lexical inference and inference of meaning, despite labelling them as different.

Understanding attitudes/moods

The questions focussed on emotional reactions, tone, and affective stance. Common lexical features included nouns *mood, attitude*, and *feeling*, and the verb *feel*. Typical question structures included:

What can be inferred about …?

What is the tone/attitude/mood …?

How does the author/person feel about …?

Several items fitting this description also appeared in the *Inference of Meaning* category, reinforcing that the model did not clearly differentiate between various cognitive processes related to implied meaning.

Global comprehension

This category showed more structural variation than others. Semantically, stems addressed emotional state (e.g., *What was the narrator's emotional state as they drove to the inn?*), motivation or reason (e.g., *What is the primary reason the narrator accepts the invitation to the Cedar Inn and Spa?*), cause-effect relationships (e.g., *What ultimately led to the family being rescued?*), and overall message or significance of the text. However, only the last group aligns with the construct of global comprehension, as it requires synthesis of information across the text or a substantial portion of it. Among the 22 items labelled as *Global Comprehension*, only five met this criterion, for example:

What is the significance of the title "An Everyday Hero"?

What is the author's view on aid in general, according to the entire text?

The analysis of comments made during the initial screening of items identified eight recurrent item flaws:

1. Lexical overlap (LO): The item stem and/or key (i.e., correct answer) repeated a substantial portion of the corresponding text verbatim, allowing test-takers to respond through lexical matching rather than engaging the intended cognitive process.
2. Double key (DK): The item contained more than one correct answer.
3. Wrong key (WK): The response option labelled by GPT-4 as correct did not contain the information necessary to answer the question accurately.
4. Longer key (LK): The correct answer was noticeably longer than the distractors, providing a clue to test-wise test-takers.
5. Implausible distractor (ID): At least one distractor was clearly incorrect, making the item solvable by eliminating implausible options rather than applying the intended type of cognitive processing.
6. Obvious answer (OA): The correct answer was evident for reasons unrelated to length or plausibility of distractors (i.e., not attributable to #4 or #5).
7. Misinterpretation of the text (MT): The item misrepresented the meaning of the relevant part of the text.
8. Wrong type of cognitive processing (WCP): Items could be answered without engaging the type of cognitive processing indicated by the GPT-4 label. This misalignment stemmed from the underlying item design and was not attributable to others flaws such as lexical matching or the use of implausible distractors.

The classification of flaw types was developed by the authors, drawing on professional item writing experience and informed by established principles of multiple-choice item construction (e.g., Haladyna et al., 2002). The frequency of each flaw across items targeting different types of cognitive processing is summarised in Table 3.

Table 3

Distribution of Item Flaws by Cognitive Processing Type

Item flaw	Specific details	Main ideas	Global compre-hension	Lexical inferencing	Inferencing of meaning	Under-standing attitudes/ moods	Total #	Total %
LO	9	4	1	1	2	1	18	13.7
DK	0	1	0	2	5	1	9	6.9
WK	0	0	0	0	2	0	2	1.5
LK	2	3	3	1	8	1	18	13.7
ID	3	2	4	3	8	2	22	16.8
OA	6	3	3	4	10	2	28	21.4
MT	0	0	1	0	2	1	4	3.1
WCP	0	0	17	3	7	3	30	22.9
Normalised flaw rates per 10 items	7.14	5.42	13.2	7.0	9.0	6.9		

Among the flaw categories, the most frequent was the incorrect attribution of items to a cognitive processing type (*n*=30, 22.9%), followed by items with an obvious answer (*n*=28, 21.4%). In contrast, flaws involving an incorrect key or misinterpretation of the text were relatively rare. To enable comparisons across cognitive processing types, flaw rates were normalised to account for differences in the number of items generated per category. When adjusted for item count, *global comprehension* items exhibited the highest flaw rate (13.2 flaws per 10 items). This was primarily due to misclassification: in 17 cases, GPT-4 mislabelled items as requiring global comprehension. Items targeting *inferencing of meaning* were also associated with a high number of flaws, although these were more evenly distributed across flaw types. Common issues included obvious answers, visibly longer keys, implausible distractors, and occasional misclassification of the cognitive process. By contrast, items focused on lower-order processing, such as *specific details* and *main ideas*, tended to exhibit flaws related to lexical overlap with the source text, obvious answers, implausible distractors, and key length. Notably, these items did not exhibit problems related to misclassification, incorrect keys, or textual misinterpretation, issues that were concentrated exclusively in items targeting higher-order cognitive processes.

Analysis of the item generation log identified six central themes, reflecting patterns in GPT-4's performance, its response to prompts and prompt modifications, item flaws, and GPT-4's ability to target different cognitive processing types. The themes, their frequencies, and distribution across texts are shown in Table 4.

Table 4

Thematic Analysis of Item generation Log

Theme	Frequency	Text where the theme occurred
GPT-4 adaptation to instructions	8	3, 5, 6
GPT-4 behaviour	13	2, 3, 4, 5, 6
Item flaws	16	1, 2, 3, 4, 6
Prompt modifications	9	1, 3, 4, 5, 6
Targeting different cognitive processing types	14	1, 2, 3, 4, 5, 6
Variations in output quality	10	1, 2, 3, 4

One of the most consistent observations concerned the fluctuating quality of generated items across rounds and texts. The quality of items sometimes improved midway through the process, only to decline again. As noted in the log, *"each item batch has one or two prevailing problems… as if ChatGPT performed at will,"* suggesting that output quality might be influenced by unpredictable factors. Recurring flaws, such as implausible distractors, longer keys, or lexical overlap with the text, were frequently documented. Another persistent issue was item overlap, where one item gave away the answer to another or multiple items targeted the same information, thus violating the rule of item independence. Prompts were iteratively adjusted in response to such problems, and some modifications yielded improvements. For example, in Round 3 of one session, a list of rules was prepared *"worded very strictly",* reflecting earlier issues. This resulted in better items, free of lexical overlap and targeting inferencing more accurately. However, not all changes were effective, and some improvements were short-lived.

The log reflects GPT-4's uneven ability to target specific cognitive processes. While some items, particularly those addressing main ideas and specific details, were well targeted, others

were misclassified. Items labelled as global comprehension, for example, often targeted specific details or even individual words in the text. A frequent confusion was between paraphrase and inference: *"ChatGPT doesn't know the difference between paraphrase and inference… something that human item writers have problems with too."*

GPT-4's behaviour emerged as a distinct theme. The model was described as unpredictable, sometimes reverting to previous flaws after showing signs of progress. Additionally, it was observed that GPT-4 *"tends to target some parts of a text a lot, while overlooking some other parts,"* and that item generation sometimes followed a loop: *"It goes on to generate items in order, but when it finds itself having exhausted info it comes back to the beginning."* Although adaptation was partial and inconsistent, there was evidence of responsiveness to explicit instruction. For example, inference items improved after clarifying the difference between inference and paraphrase. Likewise, the quality of attitude items improved following repeated correction: *"I've told it a couple of times that these are a type of inference items, and it seems to have taken it on board."* Nonetheless, the model remained inconsistent in following instructions, with the log noting that GPT-4 *"is selective in what it conforms to."*

5.2. Findings from the item review

Six item sets, comprising 46 items in total, were independently reviewed by two professional reviewers. Agreement between GPT-4 and the two reviewers regarding the cognitive processes targeted by the items was examined using Krippendorff's α. Agreement was highest between GPT-4 and Reviewer 1 (α=0.802, 95% CI=0.661–0.915), indicating substantial consistency. Agreement between GPT-4 and Reviewer 2 was lower (α=0.589, 95% CI=0.404–0.758), as was agreement between the two human reviewers themselves (α=0.538, 95% CI= 0.352–0.694). When all three coders (GPT-4, Reviewer 1, and Reviewer 2) were considered together, agreement was α=0.644 (95% CI=0.523–0.763).

The reviewers also identified flaws in individual items; table 5 summarises their counts. The most frequent issue noted by both reviewers was distractor implausibility (Reviewer 1: *n*= 14; Reviewer 2: *n*=16). Additionally, Reviewer 1 judged that distractors were possible answers in 15 items, whereas Reviewer 2 expressed concern about the appropriateness of the language level in 14 items. Notably, the two reviewers differed in the extent to which they observed these issues. Several flaw types were rarely noted; for example, both reviewers recorded low counts for unclear stems, unclear options, and an incorrect key.

Table 5

Reviewer judgements: Item flaw counts

Type of item flaw	Reviewer 1	Reviewer 2
Stem is unclear	5	1
Stem is not concise	5	1
Options are unclear	4	1
Options are not concise	4	1
Options are dissimilar in appearance	0	5
Key is incorrect	4	1
Distractors are implausible	14	16
Distractors are possible correct answers	15	3
Language level is inappropriate	0	14

The optional qualitative comments corroborated the quantitative findings for the three most frequent flaw types. For implausible distractors, both reviewers noted that some distractors were "*not supported with the text*," "*poor choices*," or "*too far away from the correct option*" to be credible (Reviewer 1: 5 comments; Reviewer 2: 8 comments). Two main causes of implausibility emerged: valence imbalance, often in attitude and mood items where the key and distractor differed in polarity, and distractors that were the direct opposite of the correct answer, making them unrealistic. For distractors judged to be possible correct answers (Reviewer 1: 4; Reviewer 2: 2) reviewers observed that some were partially or wholly true. Regarding inappropriate language level, Reviewer 2 cited several words above B2 level (e.g., *resilient, disheartened, aftermath*), while Reviewer 1 did not view them as problematic. Few additional comments concerned other flaw types listed in Table 5 (≤ 2 per reviewer).

5.3. Findings from the item trialling

Responses from 775 test-takers were analysed using both IRT and CTT. Jamovi (snowIRT module) was used to obtain proportion correct, Rasch difficulty estimates (in logits) and fit indices (Table 6). Proportion correct values ranged from .54 to .98, indicating that most items were relatively easy for the test-taker sample. Consistently, all Rasch difficulty values are negative (−3.95 to −0.17 logits), suggesting that the items were located below the mean ability of the group.

Table 6

Item statistics from Rasch (IRT) and Classical Test Theory (CTT) analyses

Item	Proportion Correct	Rasch Measure (logits)	SE	Infit	Outfit	CTT Discrimination
Task1Item1	0.974	-3.894	0.229	0.958	0.838	0.05
Task1Item2	0.696	-0.918	0.082	1.073	1.116	0.26
Task1Item3	0.975	-3.947	0.235	0.958	0.801	0.04
Task1Item4	0.964	-3.537	0.196	0.961	0.788	0.07
Task1Item5	0.811	-1.609	0.096	1.079	1.166	0.08
Task1Item6	0.938	-2.949	0.152	0.977	0.928	0.11
Task1Item7	0.641	-0.645	0.079	1.015	1.014	0.41
Task1Item8	0.784	-1.425	0.091	0.995	0.995	0.37
Task1Item9	0.543	-0.188	0.076	1.062	1.073	0.35
Task2Item1	0.807	-1.582	0.095	0.97	0.952	0.33
Task2Item2	0.538	-0.165	0.076	1.001	1.001	0.45
Task2Item3	0.631	-0.595	0.079	1.01	1.017	0.37
Task2Item4	0.744	-1.181	0.086	0.985	0.975	0.38
Task2Item5	0.808	-1.591	0.095	0.948	0.895	0.39
Task2Item6	0.887	-2.263	0.118	0.957	0.905	0.24
Task2Item7	0.922	-2.697	0.138	0.948	0.88	0.16

Item fit statistics (Table 6) showed a strong alignment between model expectations and observed data: Infit and Outfit mean squares were consistently close to 1.0 across all items, confirming that the items functioned as intended within the Rasch framework. Item characteristic curves are available in OSF https://osf.io/djeqh/overview?view_only=05c0fc44b4084859bfbbcec735506b86

Because Rasch models constrain item discrimination to be equal, CTT item–total correlations (point-biserials) were also calculated to provide additional evidence of item quality. These ranged from .04 to .45, indicating that while some items contributed relatively little to distinguishing between stronger and weaker test-takers, others showed good discriminatory power. Importantly, none of the items displayed negative discrimination, indicating that weaker test-takers were not outperforming stronger ones on any item, a situation that would point to a flawed item.

Test-taker feedback on the two trialled tasks is summarised in Table 7. Overall, few respondents found items unclear, unusual, or extremely difficult or easy. By cognitive processing type, items requiring *inferencing of meaning* and *global comprehension* drew the most 'unclear' responses (up to 10%) and were more often rated as 'unusual' or 'extremely difficult' than those targeting other cognitive processes.

Table 7

Percentage of test-takers reporting items as unclear, unusual, or extremely difficult/easy

Item #	Cognitive processing (as stated by GPT-4)	Unclear (%)	Unusual (%)	Extremely difficult (%)	Extremely easy (%)
Task1Item1	Lexical inferencing	2.4	0.1	0.1	0.7
Task1Item2	Inferencing of meaning	10.0	2.9	3.7	0.1
Task1Item3	Global comprehension	0.7	0.2	0.1	0.7
Task1Item4	Specific details	0.9	0.3	0.3	2.4
Task1Item5	Inferencing of meaning	2.7	0.5	1.3	0.3
Task1Item6	Specific details	1.4	0.3	0.1	1.4
Task1Item7	Inferencing of meaning	5.5	1.9	2.8	0.5
Task1Item8	Global comprehension	6.6	2.8	2.2	0.2
Task1Item9	Global comprehension	7.5	4.6	3.3	0.1
Task2Item1	Inferencing of meaning	3.8	1.1	2.3	0.2
Task2Item2	Understanding attitudes	4.2	0.7	2.5	0.3
Task2Item3	Inferencing of meaning	6.8	1.5	4.5	0.2
Task2Item4	Global comprehension	9.5	1.4	4.6	0.0
Task2Item5	Specific details	1.6	0.2	1.2	1.6
Task2Item6	Main ideas	3.6	1.5	1.8	0.6
Task2Item7	Global comprehension	2.9	0.6	1.4	0.5

Ten comments concerned 'unclear' items, all targeting higher-order cognitive processing, describing them as "*open*" or interpretive and stating, "*you cannot find the exact/specific answer in the text*." Eleven comments on 'unusual' and eight on 'extremely difficult'

items also referred to higher-order cognitive processing, citing that such questions were interpretive, *"unfair for neurodivergent people,"* and not directly text-based. Respondents most often linked difficulty to unfamiliar vocabulary (e.g., *"some word was essential and I wasn't sure of the meaning'*) but also mentioned ambiguous options *("none of the answers were completely true"*) and referential uncertainty *("the use of "their" for me wasn't clear—singular or plural?"*). The latter is notable, because both reviewers independently flagged the singular gender-neutral 'they' as problematic.

Items targeting specific details and main ideas (lower-order processing) were seldom rated as 'unclear' or 'unusual' and rarely as 'extremely difficult.' Reports of 'extremely easy' items were uncommon across all cognitive processing types but slightly higher for specific details items, and unaccompanied by qualitative comments. Overall, the convergence between the quantitative survey patterns and the limited qualitative feedback indicates that concerns about clarity, familiarity, and difficulty, where they did arise, were concentrated in items targeting higher-order cognitive processing. It is noteworthy, however, that psychometric analyses showed that all items were objectively easy for the test-taker sample.

6. Discussion

6.1. Item generation

Contrary to earlier studies that found GenAI favoured lower-order cognitive processing (e.g., Lin & Chen, 2024), GPT-4 in this study did not show such a preference. Instead, most generated items were nominally aligned with higher-order processes, particularly inference. Linguistic analysis of their stems revealed that the questions followed uniform repetitive patterns. For example, inference items commonly began with *"What is implied by…?"*. Although such phrasing mimicked higher-order processing demands, the item flaw analysis demonstrated that many items did not require inferential reasoning. Thus, questions beginning with *"What is*

implied by...?" often referred to explicitly stated information. Ma et al. (2025) reported similar mismatches, with fewer than half of their items reflecting the intended inferential process. Likewise, Zhang et al. (2025) observed that AI-generated items seldom required deeper textual analysis. It seems that GPT-4 can successfully imitate the surface-level characteristics of items designed to tap into higher-order processing yet does not consistently generate items that truly demand such engagement. Moreover, this study found that many generated items could be answered through test-wiseness strategies such as eliminating implausible distractors, relying on lexical matching, or selecting the longest option. As a result, even when stems did target implied meaning – as judged by the first author during the initial screening process - flaws in item construction often removed the need for higher-order cognitive processing.

It was also found that GPT-4 did not always distinguish between processes related to implied meaning. Items intended to assess lexical inference were often labelled as *inferencing of meaning*, while items designed to measure meaning inference blended with the ones aimed at *understanding attitudes/moods*. Global comprehension items posed an even greater challenge, as they were frequently recast as questions about emotional states, motivations, or cause–effect relations that did not require understanding the text's overall message. Such patterns mirror difficulties encountered by human item writers (Ma et al., 2025). Automated generation may make these overlaps among cognitive processes more visible, pointing to underlying construct ambiguities.

This study also found that items targeting higher-order cognitive processing sometimes misinterpreted information in the source text. Similar difficulties were noted by Wen and Chu (2025), who also generated items from given texts and reported content misalignment and oversimplified interpretations. Such issues may partly stem from the types of texts used for generation. Previous studies asked GenAI to produce both texts and items (Alshehri & Alharbi, 2025; Zhang et al., 2025), thereby reducing the likelihood of GenAI misinterpreting the source text.

By contrast, our study, as well as that of Wen and Chu (2025), used authentic texts, which may explain why GenAI occasionally misinterpreted passages or produced items misaligned with their deeper meaning. This is not intended as an argument for generating both texts and items with GenAI. On the contrary, based on our item-writing experience, GenAI-produced texts tend to be factual and low in implied meaning, thereby limiting opportunities for inferential items. Since implied meaning is a key focus of assessment at higher proficiency levels, the use of authentic texts remains preferable.

When using authentic texts, however, additional efforts may be required beyond simply supplying GenAI with a text and detailed instructions. In this study, we invested considerable effort in constructing a comprehensive item-generation prompt, assigning the model a persona, and providing multiple item-writing guides and examples. We also employed iterative prompting. Yet the results were not always satisfactory. This suggests that detailed guidance and sophisticated prompting techniques alone are insufficient to ensure high-quality items, and that novel approaches may be needed. One possibility is to engage GenAI in clarifying a text's implied meaning and overall message prior to item generation. Another is to explore text-mapping approaches. Text mapping has been described as a "more principled basis for arriving at test items" (Urquhart & Weir, 2013, p. 162) and is widely used in human item generation, where panels of writers employ consensus techniques to identify the aspects of text meaning to be targeted in items. If valuable in human practice, it may also be relevant in automated generation. Several adaptations are possible: a human panel could agree on item content before tasking GenAI with item production; AI could be asked to map text meaning in collaboration with a human; or multiple LLMs could form a "panel" mapping the text, with results then verified by human experts.

The item generation log recorded GPT-4's unstable performance. Iterative prompting occasionally improved results, but gains were inconsistent, and the model's adherence to instructions was unreliable. This echoes Wen and Chu's (2025) account of ChatGPT-3.5, which

selectively followed complex instructions and retained information only within the immediate context. Despite being a newer model, GPT-4 showed comparable volatility, underscoring persistent constraints in its ability to sustain stable performance across extended item generation sessions.

6.2. Item review

Although only 28% of generated items, the subset judged better quality in pre-screening, were submitted for expert review, reviewers still identified numerous flaws, most commonly in distractors. GenAI's struggles with producing plausible distractors have been noted previously (Attali et al., 2023). Writing effective distractors requires anticipating the kinds of misjudgements likely among less proficient test-takers, which in turn depends on familiarity with test-taker populations or an intuitive sense of how people misinterpret texts. LLMs, lacking such experience, are poorly positioned to simulate these errors. As Kerner (2024) explains, "an LLM is a neural network trained to predict text sequences based on probability, and there is no evidence to suggest that LLMs are capable of reasoning as humans do". This limitation likely constrains GenAI's ability to generate high-quality distractors.

Several workarounds to producing distractors have emerged. To produce distractors for the Duolingo English Test, multiple versions of a passage are generated, with parallel segments repurposed as distractors in reading tasks (Attali et al., 2023). Sayin and Gierl (2024) used a template-based method to create short texts containing a single irrelevant sentence. Test-takers had to identify the intruder, thereby activating coherence inference. In trialling, incorrect options functioned effectively as distractors. Such strategies suggest that it may be more productive to design tasks that exploit GenAI's strengths than to persist with requesting LLMs to directly generate distractors, perhaps even reconsidering the need for conventional item formats that depend on distractors and shifting toward novel item types better suited to AI generation.

Expert review also revealed disagreements about the cognitive processes targeted. While reviewers generally distinguished between lower- and higher-order processing, they often diverged on the specific cognitive process involved. This echoes Ma et al. (2025), who reported only about 70% inter-rater agreement when classifying inference item types, attributing this to the inherent subjectivity of expert judgment in construct validation. The finding further underscores a broader challenge, shared both by humans and GenAI, of pinning down discrete cognitive categories in reading comprehension.

6.3. Item trialling

Consistent with prior research (Alshehri & Alharbi, 2025; Zhang et al., 2025), virtually all items in our study were objectively easy, with facility values well above .50. None showed negative discrimination, suggesting that they functioned as intended at the technical level. However, the clustering of items below 0 logits meant that the upper end of the ability distribution was poorly targeted. This reflects the trial cohort's composition, students at the cusp of B2/C1, and our deliberate decision to prioritise identifying item flaws over optimizing difficulty. Nevertheless, the lack of harder items has measurement implications: precision in ability estimates for more proficient learners is reduced, echoing reviewers' concerns that many items were "too obvious" or answerable without engaging fully with the text.

This study revealed a divergence between test-takers' subjective perceptions of item difficulty and their objective performance. Questionnaire responses indicated that inference and global comprehension items were perceived as considerably more challenging than literal detail items. Yet IRT analysis showed that all items were objectively easy for the group. The discrepancy likely reflects the cognitive demands of higher-order processing: such items may feel harder because their solution paths are less transparent. However, when items merely appear to target these processes, as was often the case in this study, the correct answer could be easily retrieved from the text.

7. Conclusion

This study evaluated the capacity of GPT-4 to generate MC B2 reading comprehension items intended to target higher-order cognitive processing. The study used a triangulated design that combined expert review, linguistic analysis, trialling with 775 learners, and test-taker feedback. The investigation showed that while AI-generated items can meet basic psychometric standards and replicate surface features of higher-order cognitive processing, they rarely provide substantive coverage of those complex cognitive processing types. It also seems that the shallow alignment between form and substance observed in earlier LLMs persists in more advanced models, though perhaps with greater fluency and polish.

These findings carry important implications for reading assessment practice. By detailing the issues that most commonly affect higher-order items, the study highlights areas that require particular attention in item review. Practitioners considering the use of GenAI to generate reading items should be aware that the cognitive process labels supplied by GenAI are often inaccurate: items may appear to target higher-order cognitive processing while in fact permitting answers through literal retrieval or test-wiseness strategies. Rigorous review is therefore essential, both to verify the type of processing an item is likely to elicit and to detect flaws that make items susceptible to shortcuts that bypass genuine higher-order processing.

The study makes several important contributions to the growing literature on GenAI for item writing in language assessment. It is the first to document the item generation process through a structured log, providing an empirical record of how model behaviour evolved across iterative rounds. It also extends previous work by offering a fine-grained linguistic analysis of item stems, linking surface features to the likelihood of eliciting higher-order cognitive processing. The study also advances validation practice by triangulating evidence from expert review, psychometric trial data, and test-taker feedback, an approach rarely adopted in prior research. Finally, by embedding the investigation within an operational certification context, the study

demonstrates both the practical feasibility and the current limits of using GenAI as a tool in large-scale reading test development.

Several limitations should be acknowledged. First, trialling was conducted with learners slightly above B2 level, which may have affected item difficulty estimates and reduced comparability with the target population. Second, the study drew on a limited set of six texts, with only two advanced to trialling, so the findings may not extend to a broader range of genres and text types. Third, no advanced NLP methods such as LLM fine-tuning or pipeline architectures were applied. However, this "low-tech" approach reflects the realities of most operational test development contexts. Fourth, the reviewer pool was small, comprising only two experts, which restricted the breadth of perspectives on item quality and construct alignment. Finally, the classification of cognitive processes was based on expert judgement. While this is consistent with common practice in reading assessment research, more definitive evidence would require empirical validation through methods such as eye-tracking or think-aloud protocols, which were beyond the scope of the present study.

Future research should address both technical and assessment-oriented aspects of using GenAI for developing reading test items. On the technical side, where expertise is available, efforts could focus on fine-tuning LLMs in API environments or building evaluation-assisted system architectures. On the assessment side, promising directions include integrating text mapping into the item generation process and the development of novel item formats better aligned with GenAI capabilities.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Alshehri, S. M., & Alharbi, M. S. (2025). Evaluating GPT-4 Turbo's ability to design English reading test items for language learners. *English Language Teaching, 18*(7), 48–57. https://doi.org/10.5539/elt.v18n7p48

Attali, Y., LaFlair, G., & Runge, A. (2023, March 31). *A new paradigm for test development* [Webinar]. Duolingo Webinar Series. https://www.youtube.com/watch?v=rRc96oe9bzk

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume*. Council of Europe Publishing.

Cullen, P., French, A., & Jakeman, V. (2014). *The official Cambridge guide to IELTS*. Cambridge University Press & Assessment.

Departament d'Educació i Formació Professional. (n.d.). *Mostres de les proves*. Generalitat de Catalunya. https://educacio.gencat.cat/ca/serveis-tramits/proves/proves-lliures-obtencio-titols/convocat-ordinaria-idiomes/mostres-proves/

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5

Kerner, S. M. (2024, February 8). What is a large language model (LLM)? *TechTarget*. https://www.techtarget.com/whatis/definition/large-language-model-LLM

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System, 123,* 103344. https://doi.org/10.1016/j.system.2024.103344

Long, D. L., & Chong, J. L. (2001). Comprehension skill and global coherence: A paradoxical picture of poor comprehenders' abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1424–1429. https://doi.org/10.1037/0278-7393.27.6.1424

Ma, W. A., Flor, M., & Wang, Z. (2025). Automatic generation of inference-making questions for reading comprehension assessments. *arXiv*. https://doi.org/10.48550/arXiv.2506.08260

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing, 18*(7–9), 657–686. https://doi.org/10.1007/s11145-005-3355-z

Sawaki, Y. (2017). The effects of different levels of performance feedback on TOEFL iBT® reading practice test performance (ETS Research Report No. RR-17-31). *ETS*. https://doi.org/10.1002/ets2.12159

Sayın, A., & Gierl, M. J. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice, 43*(1), 5–18. https://doi.org/10.1111/emip.12590

Shin, D., Lee, J. H., & Kim, K. (2025). An exploratory study on two automated item generators for generating L2 reading test items. *RELC Journal, 0*(0), 1–16. https://doi.org/10.1177/00336882251326284

Urquhart, A. H., & Weir, C. J. (2013). *Reading in a second language: Process, product and practice*. Routledge.

Wen, Z., & Chu, S. K. W. (2025). Using generative AI for reading question creation based on PIRLS 2011 framework. *Cogent Education, 12*(1), 2458653. https://doi.org/10.1080/2331186X.2025.2458653

Zhang, T., Erlam, R., & de Magalhães, M. (2025). Exploring the dual impact of AI in post-entry language assessment: Potentials and pitfalls. *Annual Review of Applied Linguistics, 45,* 1–20. https://doi.org/10.1017/S0267190525000030