Annual Review of Applied Linguistics (2025), **45**, 274–293
doi:10.1017/S0267190525000030

RESEARCH ARTICLE

Exploring the dual impact of AI in post-entry language assessment: Potentials and pitfalls

Tiancheng Zhang[1,2], Rosemary Erlam[3] and Morena Botelho de Magalhães[2]

[1]Faculty of International Studies, Southwestern University of Finance and Economics, Chengdu, China; [2]DELNA, The University of Auckland, Auckland, New Zealand and [3]Faculty of Arts and Education, The University of Auckland, Auckland, New Zealand
Corresponding author: Tiancheng Zhang; Email: tzha305@aucklanduni.ac.nz

Abstract

This paper explores the complex dynamics of using AI, particularly generative artificial intelligence (GenAI), in post-entry language assessment (PELA) at the tertiary level. Empirical data from trials with Diagnostic English Language Needs Assessment (DELNA), the University of Auckland's PELA, are presented.

The first study examines the capability of GenAI to generate reading text and assessment items that might be suitable for use in DELNA. A trial of this GenAI-generated academic reading assessment on a group of target participants ($n = 132$) further evaluates its suitability. The second study investigates the use of a fine-tuned GPT-4o model for rating DELNA writing tasks, assessing whether automated writing evaluation (AWE) provides feedback of comparable quality to human raters. Findings indicate that while GenAI shows promise in generating content for reading assessments, expert evaluations reveal a need for refinement in question complexity and targeting specific subskills. In AWE, the fine-tuned GPT-4o model aligns closely with human raters in overall scoring but requires improvement in delivering detailed and actionable feedback.

A Strengths, Weaknesses, Opportunities, and Threats analysis highlights AI's potential to enhance PELA by increasing efficiency, adaptability, and personalization. AI could extend PELA's scope to areas such as oral skills and dynamic assessment. However, challenges such as academic integrity and data privacy remain critical concerns. The paper proposes a collaborative model integrating human expertise and AI in PELA, emphasizing the irreplaceable value of human judgment. We also emphasize the need to establish clear guidelines for a human-centered AI approach within PELA to maintain ethical standards and uphold assessment integrity.

Keywords: Generative Artificial Intelligence (GenAI); Post-Entry Language Assessment; Diagnostic English Language Needs Assessment (DELNA); Automated Writing Evaluation; Automated Item Generation; Human-Centered AI

Introduction

In today's globalized academic environment, universities in English-speaking countries are increasingly navigating the complexities introduced by students' diverse linguistic backgrounds. This diversity is largely driven by significant immigration, by sustained recruitment drives targeting international students, and by national policies aimed at widening access to higher education for underrepresented groups, including ethnic minorities and individuals from low-income backgrounds (Murray, 2016; Read, 2016). Consequently, identifying and providing initiatives for students in need of additional academic English language support has become a priority for these institutions (Dunworth, 2009), not least at Waipapa Taumata Rau/the University of Auckland (WTR/UoA), the context for this study, where the importance of ensuring equitable experiences for students and of enhancing student retention and success is underlined by the university's strategic plan (The University of Auckland, 2020).

One mechanism for identifying students at risk due to limited academic language proficiency and guiding them to appropriate language support is the implementation of a post-entry language assessment (PELA) program (Doe, 2014; Read, 2015). WTR/UoA offers one of the most comprehensive PELAs in Australasia, a low-stakes diagnostic tool known as the Diagnostic English Language Needs Assessment (DELNA) (Elder & Erlam, 2001).

All first-year undergraduate students at WTR/UoA are strongly encouraged to engage with the DELNA process, and for some faculties compliance is mandated. All doctoral candidates, regardless of faculty or language background, are also required to complete DELNA, and sub-doctoral postgraduate students are invited to take the assessment, too. DELNA comprises up to three stages (see Figure 1). The initial phase, known as the Screening, is a computer-based assessment which differentiates proficient academic English users from those with lower proficiency levels (Read, 2015). It includes a vocabulary task and a timed cloze-elide activity (samples of the two tasks can be accessed from https://www.delnatask.com/tasks/practice/vocab.php). Students who do not meet the minimum satisfactory standard are classified in the "Diagnosis required" category, directing them to the second phase of the DELNA process (Read, 2008).

The Diagnosis is a comprehensive, 2-hour evaluation that provides a deeper assessment of students' academic English skills, focusing on reading, listening, and writing. It can be completed on paper or online. The *DELNA Handbook* specifies that the listening component requires students to listen to a short lecture on a topic that does not require specialized knowledge. The reading section includes two passages, also on general-interest topics, totaling approximately 1,200 words. A range of response types assess comprehension, including cloze exercises, summarization, matching of ideas, information transfer, multiple-choice, true/false, and short-answer questions.

The writing tasks are differentiated. The majority of students complete a short writing task in which they interpret information presented in a table or diagram, crafting a 200- to 250-word commentary. Doctoral candidates complete a more extensive writing task, divided into two parts. In Task 1, they summarize key points from two contrasting texts, and in Task 2, they write an essay on a topic related to these texts. For more

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

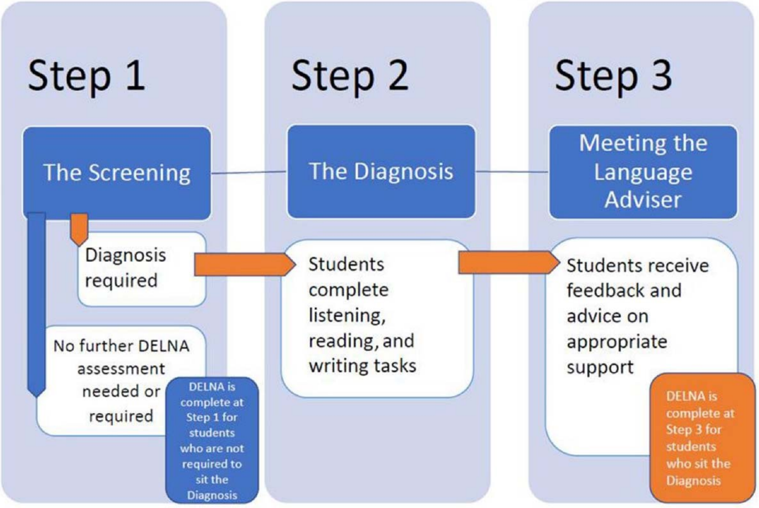

Step 1
The Screening
Diagnosis required
No further DELNA assessment needed or required
DELNA is complete at Step 1 for students who are not required to sit the Diagnosis
Step 2
The Diagnosis
Students complete listening, reading, and writing tasks
Step 3
Meeting the Language Adviser
Students receive feedback and advice on appropriate support
DELNA is complete at Step 3 for students who sit the Diagnosis

Figure 1. The three stages of the DELNA process.

detailed information and task examples, refer to the *DELNA Handbook* (The University of Auckland, 2024).

An ongoing challenge for DELNA is the creation of new assessment materials, driven by the need for assessment security, given the large number of students involved in the process. Another challenge, more pertinent during peak enrolment periods, is the demand for timely and accurate feedback on students' writing tasks. Meanwhile, rapid advancements in generative artificial intelligence (GenAI), particularly breakthroughs in large language model (LLM) architectures, such as OpenAI's ChatGPT, could offer solutions while at the same time introducing new challenges for PELAs such as DELNA.

Unlike traditional AI, which typically relies on rule-based algorithms and structured datasets, GenAI harnesses self-supervised learning architectures to generate novel content through probabilistic pattern recognition in unstructured data ecosystems. GenAI has impacted a range of industries, reshaping both the professional and personal domains of everyday life (Grossmann et al., 2023). In language learning, teaching, and assessment, GenAI offers potential for the dynamic creation of content, personalized learning experiences, and automated evaluation, creating new possibilities for scalable and adaptive educational tools (Hao et al., 2024; Kohnke et al., 2023).

Although there has been a recent surge of research on GenAI in language learning and teaching (Hockly, 2023; Yang & Li, 2024), studies focusing on its application in language assessment, particularly in PELA contexts, remain limited. To address this gap, our study conducted two empirical investigations, in the context of DELNA, exploring the use of GenAI in automated item generation (AIG) and automated writing evaluation (AWE). While these experiments specifically targeted text-generation capabilities of LLMs, the subsequent Strengths, Weaknesses, Opportunities, and Threats

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

(SWOT) analysis encompasses broader GenAI implementations – including multimodal content generation (audio, video, and interactive media) – relevant to PELA stakeholders.

Literature review

AI-assisted AIG in reading tasks

As explained earlier, AIG introduces new opportunities for test item generation applying natural language processing. This technology can substantially reduce the time and resources required for item development, making assessments more cost-effective and accessible (Cardwell et al., 2024; Zirar, 2023).

Bezirhan and Von Davier (2023) used GPT-3 to generate reading passages based on materials released from the Progress in International Reading Literacy Study. The features of the discourse of these AI-generated passages were similar to those of the input materials but covered different topics. Human evaluators assessed the coherence, appropriateness, and readability of the passages for fourth-grade readers, finding that the GPT-3-generated texts were comparable to those written by humans. The study highlighted that combining GenAI capabilities with well-crafted prompts and human editing was an effective and efficient approach for generating reading passages.

Shin and Lee (2023) extracted five reading passages and corresponding multiple-choice questions from South Korea's College Scholastic Ability Test (CSAT) English section. They then used ChatGPT to generate an alternative set of readings and items in the same format. In a survey of 50 teachers, the AI-generated passages were rated similarly to the CSAT passages in terms of natural flow and expression. However, the CSAT items were deemed to have more appealing multiple-choice options and were more fully developed.

Lin and Chen (2024) evaluated ChatGPT's ability to generate multiple-choice reading comprehension items. Benchmarking against Item Response Theory models and human evaluation, they found that ChatGPT-generated items were comparable to human-authored items. The study concluded that ChatGPT has potential as a tool for test development and as an aid for teaching and learning reading comprehension.

While AIG enables automated item development, human oversight remains crucial to maintain the quality and fairness of educational assessments. Hao et al. (2024), building on Attali et al. (2022), describe the automated content generation process used in the Duolingo English Test, which incorporates a human-in-the-loop approach (see Figure 2). Here, human experts remain involved throughout the processes of construct definition, task design, item generation, and refinement, ensuring that items meet standards for quality, fairness, and bias reduction. This human-in-the-loop model aligns with the concept of Human-Centered AI proposed by Shneiderman (2022). Human-Centered AI advocates for designing AI systems that augment human capabilities and deliver a positive social impact. It prioritizes human values, ensuring accessibility, usability, and the enhancement, rather than replacement, of human involvement.

The possibility of using AI to generate reading texts and assessment items for the DELNA diagnosis is an attractive option, particularly in terms of alleviating workload pressures. However, it would necessitate investigation into text suitability,

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

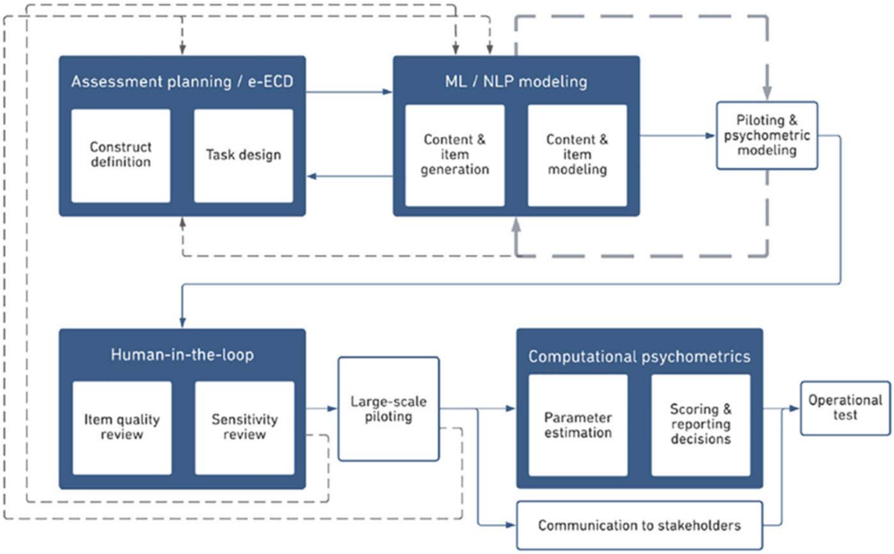


Figure 2. Scalable content creation using human-in-the-loop AI in the Duolingo English Test (Hao et al., 2024, p. 3).

item quality, and the potential impact on the reliability and validity of the assessment (Lee, 2015).

AI-assisted AWE

AWE refers to the use of technology to assess and score written responses, applying standardized measurement processes to open-ended or structured responses (Ifenthaler, 2023). While research on AWE dates back to the 1960s (Bermouth, 1970), recent advancements in computing, data analysis, and LLM fine-tuning have improved the effectiveness, objectivity, reliability, and validity of AWE systems for the evaluation of written texts (Ifenthaler, 2023; Mizumoto & Eguchi, 2023).

In a study that analyzed 12,100 TOEFL writing samples from the Educational Testing Service (ETS) Non-Native Written Corpus, Mizumoto and Eguchi (2023) observed that GPT-generated scores aligned closely with the typical TOEFL scoring patterns. The authors compared the performance of multiple regression models using a model comparison approach to arrive at such conclusion. This finding indicates that LLMs, such as ChatGPT, have potential for practical applications in AWE, offering valuable insights for both research and real-world assessment. However, the TOEFL writing samples in the study were not scored using the original 0–5 scale. Instead, they were grouped into three categories: high, medium, and low. Consequently, the study lacked ground-truth scores – official ratings from trained human evaluators – considered the benchmark for assessing AWE reliability in comparison to human raters (Powers et al., 2015). The model also had not undergone fine-tuning, that is, a process of

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

additional training on a specific dataset or task (Peters et al., 2019). LLMs are generally pretrained on diverse tasks rather than writing evaluation specifically, so fine-tuning for AWE can significantly enhance performance (Wang & Gayed, 2024).

Research comparing AWE scores with human rater evaluation is limited and inconclusive in terms of findings. Yavuz et al. (2024) examined the validity and reliability of ChatGPT and Google's Bard in grading higher education essays based on an analytical grading rubric. Fifteen experienced English as a foreign language instructor, and two GenAIs assessed three essays of varying quality. Inter-rater reliability, measured by intraclass correlation (ICC), suggested that LLM scores were similar to those of human raters, though human ratings tended to offer more nuanced feedback. A limitation was that their sample size was small and lacked details on fine-tuning. Another study compared ChatGPT's scoring of 200 essays with that of human raters (Bui & Barrot, 2024). Correlation results showed weak to moderate alignment with an experienced human rater, and low ICC values indicated inconsistent scores across multiple rounds of scoring. A limitation was that the study did not include fine-tuning of GPT-3.5, and only one human rater was involved.

Another limitation of the studies referred to above is that they did not set the LLM's temperature to zero, a parameter that controls output randomness. According to OpenAI's documentation, setting the temperature to zero ensures the model provides the same response to identical prompts each time, which is especially important for AWE applications, where consistency in scoring is vital for fairness and reliability (see https://platform.openai.com/docs/api-reference). One exception was Wang and Gayed (2024) who did explicitly address this issue and developed an AWE system to score argumentative essays by fine-tuning the GPT model. They compared the fine-tuned model's effectiveness to non-fine-tuned GPT-3.5 and GPT-4 models using zero-shot prompting. The dataset, comprising 480 argumentative essays from ETS's TOEFL Public Writing Dataset, included ground-truth scores under two prompts. Findings suggest that task-specific fine-tuning enhances AWE performance, and fine-tuning does not require an extensive variety of prompts.

Academic writing assessment is a crucial component of PELA, as assessment outcomes correlate with university students' academic performance (Read, 2016). To accurately evaluate student writing and provide targeted feedback, DELNA currently uses a traditional rating system which is time- and resource-intensive, with the added limitation of potential for subjective bias. Each essay is double-rated; and if the two raters' scores differ significantly, a third rater is consulted. Past PELA research has focused on rater training to improve consistency and accuracy (e.g., Elder et al., 2007; Erlam et al., 2013). Meanwhile, recent advancements in GenAI, particularly in LLMs, mean that AWE systems present new possibilities for reducing PELA raters' workload and supporting them with timely, reliable scoring.

In summary, rapid advancements in AI technology have significant potential for more efficient and accurate language assessment in PELA. However, current studies on GenAI-assisted AIG and AWE remain few in number and primarily focus on generating reading passages and multiple-choice questions; other item types are largely unexplored. Moreover, most studies on GenAI-assisted AWE lack fine-tuning and zero-temperature adjustments.

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

This research study aims to address some of these gaps with the following questions:

1) How effective is GenAI-assisted AIG in designing a reading assessment within the context of a PELA?
2) Can GenAI-assisted AWE effectively support human evaluation of student essays?
3) What additional opportunities and challenges does AI introduce for PELA?

Methods

To address the research questions, two empirical studies were conducted. Study 1 includes two parts: Part 1 examines whether language advisors/teachers can distinguish between GenAI-generated and human-generated test items. Part 2 aims to (a) understand how language advisors/teachers evaluate a reading text, and items generated entirely by GenAI and (b) conduct a trial of this reading assessment. Study 2 explores the effectiveness of the fine-tuned GPT-4o-2024-08-06 model in the automated grading of DELNA essays, using the DELNA rubric.

Participants

Study 1

Invitation emails were sent to 11 experienced professionals, all of whom were either DELNA language advisors or WTR/UoA language teachers, with 5 agreeing to participate. Of these five, two were DELNA language advisors, two were language teachers at WTR/UoA, and one was a former academic advisor of DELNA, who is a seasoned expert in language assessment.

Study 2

The dataset consisted of 348 short essays from the DELNA Writing Dataset, provided by the DELNA Office. These essays were written by first-year undergraduate and sub-doctoral postgraduate students at WTR/UoA as part of the DELNA Diagnosis between March 2021 and October 2022; all were written in response to the same prompt. Each essay was evaluated by at least two human raters, using the DELNA rubric, with a third rater involved if there was a significant discrepancy between scores. Final scores were assigned on a 4–9 scale, with increments of 1. Upon publication, DELNA will retire this writing version; the prompt is provided in Appendix A, while the rubric remains confidential.

Instruments

Study 1

For Study 1, Microsoft Copilot, previously known as “Bing Chat Enterprise” and available exclusively to business customers with Microsoft 365, was utilized. Copilot, which is powered by ChatGPT-4 and DALL-E 3, was accessed through a WTR/UoA account to ensure data privacy and intellectual property protection (refer to https://teachwell.auckland.ac.nz/resources/generative-ai/gen-ai-usage-standard/).

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

Study 2

The GenAI model for AWE was selected based on three criteria: (1) a pretrained base model, not linked to a specific web interface; (2) the ability to support custom fine-tuning; and (3) demonstrated effectiveness in AWE tasks as suggested by Wang and Gayed (2024). The GPT-4o-2024-08-06 model was chosen, therefore, offering enhanced capabilities for language understanding and text generation, outperforming previous GPT-4 versions in terms of coherence and contextual appropriateness and making it especially suitable for the precise requirements of AWE. At the time of data analysis, it represented the latest GPT model.

Research design

Study 1

Part 1

Participants were asked to distinguish between human and GenAI-generated questions. They received a document with a DELNA human-generated reading text, followed by 20 questions of which 10 were GenAI-generated based on a series of prompts. They were asked to decide whether each item was: “Human-Generated,” “GenAI-Generated,” or “Not Sure” (included to reduce random guessing). Lastly, participants were invited to outline the criteria used to make their decisions in writing.

Part 2

Using the *DELNA Handbook*’s descriptions of reading materials and question types (The University of Auckland, 2024), GenAI was provided with refined prompts and asked to generate an entirely new reading text and associated items. As prompt accuracy significantly influences GenAI output quality (Giray, 2023; Knoth et al., 2024), a “chain-of-thought” prompting approach (Wei et al., 2022) was employed, generating the reading text first and then creating questions to assess specific subskills identified by Liu (2018). Only formatting changes were made to the GenAI output; the content remained unchanged. Study 1 experts, advised of the subskills each item targeted, evaluated the quality of the GenAI-generated items and the text. Subsequently, 132 students completed the GenAI-generated reading assessment to trial it. These students ranged from undergraduate to doctoral level and spoke more than 20 different first languages. Students’ performance on the reading assessment was analyzed using Cronbach’s Alpha and the Rasch model to evaluate reliability and item difficulty. Among these participants, 49 students had taken the DELNA Diagnosis within 6 months prior to the trial. Their trial results were converted to Bands A through D, consistent with the DELNA Diagnosis scoring format; and the differences between their trial and Diagnosis scores were analyzed using the Wilcoxon signed-rank test. Cronbach’s Alpha and the Wilcoxon signed-rank test were conducted using IBM SPSS Statistics version 26, while Rasch analysis was performed using Winsteps version 3.72.3.

Study 2

In Study 2, the GPT-4o-2024-08-06 model for AWE was fine-tuned based on Wang and Gayed (2024). Using API calls in Python, the model was trained to understand

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

how experienced human raters apply rubric criteria. A set of 80 essays, each with corresponding scores, was used for training. Following the approach of Wang and Gayed (2024), a 5:1 training/validation split was applied, resulting in 16 essays being allocated for validation. The remaining 252 essays in the dataset were used for testing. The primary criteria for testing were consistency, accuracy, and reliability. Consistency involved verifying that fine-tuned GPT processed input accurately and produced identical scores across repeated evaluations, as suggested by Wang and Gayed (2024). The temperature was set to zero to ensure consistent scoring.

Accuracy was assessed using root mean square error (RMSE) between GPT and ground-truth scores, a standard AWE metric (Klebanov & Madnani, 2022; Yuan et al., 2020). Reliability was measured through percentage agreement and quadratic weighted kappa (QWK), comparing GPT and human rater scores. An absolute agreement of 85% or higher is considered a high alignment between system and human scores. QWK provided further insight into agreement levels in AWE. Lower QWK values indicate lower consistency between machine and human scores, while higher values suggest stronger agreement (Doewes et al., 2023; Ramesh & Sanampudi, 2021).

Scores were compared across fluency, content, form, and total score, as aligned with the DELNA rubric. The R packages Metrics (Hamner & Frasco, 2018) and irr (Gamer et al., 2019) were employed to calculate the RMSE and QWK, respectively, while percentage agreement was determined using Microsoft Excel.

Study 2 prompts are provided in Appendix B and C, as are Part 2 of Study 1 prompts, that is, the GenAI-generated reading text and items, along with the expert questionnaire. The reading text and items used in Part 1 are not disclosed because they are taken from a current DELNA assessment task.

Results

Study 1

Part 1

The overall success rate for correctly identifying GenAI-generated items was low at 50% or below. Only one expert identified 5 of the 10 GenAI-generated test items; the remaining four identified from between 1 and 4 items. While the experts admitted that they relied on intuition to a significant extent, judgments were based on three main criteria:

Item type. Experts were more likely to classify certain item types as AI-generated or human-generated based on the complexity and specificity of the tasks. Items requiring synonym choices and paragraph summaries were frequently identified as AI-generated, likely due to the relatively formulaic or structured nature of these tasks, which GenAI models are well-suited to handle. In contrast, items such as true/false questions, fill-in-the-blanks, and those involving diagrams and tables were often assumed to be human-generated. This assumption stemmed from the perception that these item types require detailed content analysis or specialized contextual understanding, which may go beyond the typical capabilities of AI.

Item quality. The quality of the items also played a critical role in influencing expert judgments. Items with alternative answers that appeared implausible, unrealistic, or poorly aligned with the main question were more likely to be flagged as AI-generated. Similarly, questions that were perceived as overly simplistic or poorly constructed – lacking depth or clarity – were attributed to AI, reflecting a belief that human-generated items are generally more thoughtful and refined.

Language cues. Linguistic details were a significant factor in determining whether an item was AI- or human-generated. Experts closely examined aspects such as word choice, syntax, and punctuation. Well-crafted syntax was more likely to be attributed to human authorship, reflecting natural fluency and attention to detail.

Part 2

Experts evaluated each GenAI-generated item according to the statement, "This is a good test item," using a rating scale: 1 (Strongly Disagree) to 5 (Strongly Agree). Opinions on item quality varied widely, with no item universally rated as either high or low quality (a summary of the results can be found in Appendix D). Items 3, 4, 6, and 12 received particularly low ratings. For instance, experts criticized Item 3 (focused on word meaning analysis) for failing to effectively explain "aesthetic." Items 4 and 6 were deemed overly simple; some experts noted that students might answer Item 6 without reading the passage. Item 12, which aimed to assess multiple subskills (understanding relationships between texts and drawing conclusions), was criticized for not adequately testing both subskills.

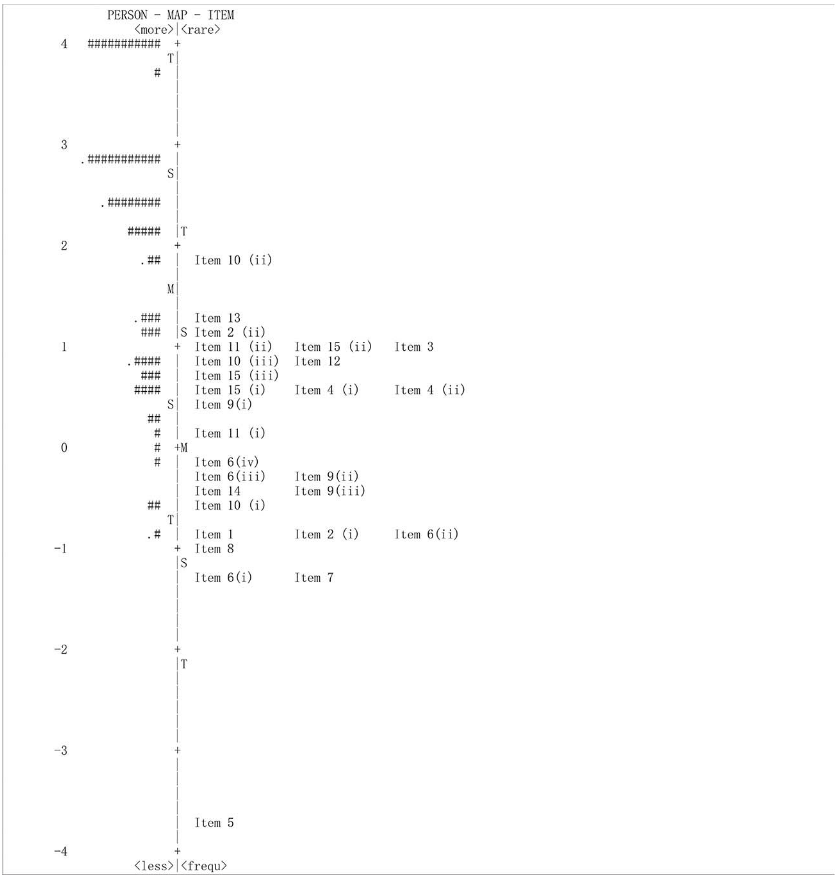
In contrast, the experts rated the GenAI-generated text highly, noting that it met traditional reading test criteria: it was well-structured and neutral, presenting factual information and viewpoints on nontechnical topics.

Based on the trial results, the reading assessment generated by GenAI had a Cronbach's Alpha of 0.863, indicating a high level of internal consistency and reliability. This suggests that the GenAI-generated reading assessment items effectively measure the same underlying reading ability or construct, demonstrating strong inter-item correlations and providing stable and consistent results. However, the Wright map (see Figure 3) showed that this version of the reading assessment was not well-aligned with the ability levels of the test-taker sample, with several items, particularly items 5–8, being generally too easy for the participants, which aligned with the conclusions of the experts. The Wilcoxon signed-rank test further revealed a significant difference between participants' trial results and Diagnosis results ($n = 49$, standardized test statistic $= 4.872$, $p < 0.001$). Specifically, 31 participants scored higher on the trial results compared to their Diagnosis results, 16 showed consistent scores across both assessments, and only 2 participants scored lower in the trial results.

Study 2

Results in Table 1 show the fine-tuned model's accuracy; lower RMSE values reflect smaller discrepancies, signifying higher accuracy (Chai & Draxler, 2014). All RMSE values were below 0.5, showing that the model provides good prediction accuracy across both overall and individual scoring metrics, approaching human evaluators' accuracy.

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press



```
                PERSON - MAP - ITEM
                    <more>|<rare>
    4   ###########  +
                    T|
                   # |
                     |
                     |
                     |
                     |
    3                +
       .###########  |
                    S|
                     |
          .########  |
                     |
              #####  |T
    2                +
                .##  |  Item 10 (ii)
                     |
                    M|
                     |
               .###  |  Item 13
                ###  |S Item 2 (ii)
    1                +  Item 11 (ii)   Item 15 (ii)   Item 3
              .####  |  Item 10 (iii)  Item 12
                ###  |  Item 15 (iii)
               ####  |  Item 15 (i)    Item 4 (i)     Item 4 (ii)
                    S|  Item 9(i)
                 ##  |
                  #  |  Item 11 (i)
    0             #  +M
                  #  |  Item 6(iv)
                     |  Item 6(iii)    Item 9(ii)
                     |  Item 14        Item 9(iii)
                 ##  |  Item 10 (i)
                    T|
                 .#  |  Item 1         Item 2 (i)     Item 6(ii)
   -1                +  Item 8
                     |S
                     |  Item 6(i)      Item 7
                     |
                     |
                     |
                     |
   -2                +
                     |T
                     |
                     |
                     |
                     |
                     |
   -3                +
                     |
                     |
                     |
                     |
                     |  Item 5
                     |
   -4                +
                    <less>|<frequ>
```

Figure 3. Wright map for the reading trial of the GenAI-generated assessment.

Table 1. Performance of the fine-tuned LLM-based AWE model in writing assessment

	RMSE	Percentage of identical scores	QWK
Total score	.36	86.90%	.73
Subscore			
Fluency	.45	79.76%	.61
Content	.47	78.17%	.60
Form	.45	80.16%	.61

The total score absolute percentage agreement reached 86.9%, while agreement across the sub-scores was approximately 80%, indicating strong consistency. The adjacent agreement of the total scores (i.e., where scores differed by only one point) was 100%. This indicates that, even when discrepancies occurred between GPT and human ratings, the differences were minimal, with no score differing by more than one point. The high level of adjacent agreement may be attributed to the concentration of DELNA scores around bands 5 and 6. This result underscores the model's reliability in closely aligning with human scoring standards.

The third metric, QWK, consistently exceeded 0.6. In general, QWK values between 0.6 and 0.8 could be interpreted as reflecting the model's capability to provide scores in close alignment with human judgment (Doewes et al., 2023). However, ETS sets a standard QWK value of at least 0.70 for its e-Rater system (Williamson et al., 2012, as cited in Wang & Gayed, 2024); in this study, only the total score met this benchmark, with the three sub-scores slightly below. This indicates that, while the model's performance aligns well with human raters on total scores, there is room for improvement in its sub-score predictions.

DELNA provides students with an overall academic writing score along with sub-scores in three specific categories. Human-generated comments for each of the categories inform a language advisor on areas needing improvement. While raters' comments may be generic at times, they are usually accompanied by examples from the student's text. The advisor then collaborates with the student in reviewing their performance, offering detailed feedback and tailored suggestions for improvement. Raters' comments for one of the essays in the sample are below.

> **Score**
> fluency: 5
> No visual paragraphs. Hard to follow throughout. Insufficient use of/ repetition of cohesive devices: however, such as. Little understanding of academic style, informal and colloquial: use of brackets, "thing" (2X), personal pronouns.
> content: 5
> Data partially described in a very general manner. Only dates and trends, no figures. Interpretation brief with only one reason - "increasing the living area". Discussion absent (incomplete sentence).
> form: 5
> Limited control of sentence structures. Most sentences are convoluted/ incomplete, word order issues. A range of significant errors affect expression of ideas: use of prepositions (regarding to, reasons of), articles, verb, singular/ plural. Range and use of vocabulary inadequate. Numerous errors in word use/ form (expresses, late years, ecosystems methods). Spelling mistakes (Amazone, goverment).

In contrast, although the AWE model developed in this study is capable of generating explanations for scores and providing learning recommendations for essays in the dataset, the feedback was often general and lacked concrete examples. To illustrate this, below are machine-generated feedback and a summary of recommendations for the same essay as in the human-written comments above.

> **Feedback**:
> fluency: 5
> The essay lacks clear organization and logical flow. Paragraphs are fragmented, and the message is difficult to follow. There is a lack of cohesion, with abrupt transitions between ideas. The style is informal and lacks an academic tone.
> content: 5

> The description of the data is incomplete and lacks clarity. The interpretation of trends is brief and lacks elaboration. The discussion of policies is missing, making the content insufficient and unclear.
> form: 5
> The essay contains numerous grammatical errors and awkward sentence structures. Vocabulary is limited and there are several spelling mistakes. The overall expression is unclear and lacks sophistication.
>
> **Summary of recommendations**:
>
> 1. Practice organizing your essays with clear introductions, body paragraphs, and conclusions.
> 2. Ensure you address all parts of the prompt and provide detailed explanations and examples.
> 3. Work on improving your grammar and vocabulary through practice and reading.
> 4. Consider seeking feedback from teachers or peers to help identify areas for improvement.

Discussion

GenAI performance in AIG and AWE

This section addresses the first two research questions. In Part 1 of Study 1, experts were tasked with distinguishing between GenAI-generated and human-generated test items. The findings indicate that, although experts could recognize some patterns typical of GenAI-generated items, the language produced by AI is becoming increasingly sophisticated. This advancement is beginning to blur the lines between AI-generated language and what we traditionally consider as "human-like" language. The findings also reveal that GenAI's ability to generate test items surpassed expert expectations, indicating that its capacity is not limited to simple question types but extends to more complex tasks that require in-depth analysis. This suggests that GenAI capabilities will likely continue to evolve, potentially rivaling human-generated content in more complex assessments, further challenging experts in distinguishing AI-generated from human-authored items.

In Part 2 of Study 1, experts evaluated a reading assessment entirely generated by GenAI. The results revealed significant variability in their opinions regarding the quality of the GenAI-generated test items. This finding underscores the difficulty of establishing consistent standards for assessing the quality of GenAI-generated items and points to a need for more rigorous, universally accepted standards, especially for items designed to assess multiple skills. Experts noted that some GenAI-generated questions failed to meet DELNA's quality requirements, particularly vocabulary questions, which lacked nuance in word meaning. Others were deemed too simple, allowing students to answer without engaging deeply with the text, indicating potential gaps in the ability of GenAI to set appropriate difficulty levels. Such feedback reinforces other research advocating for the continued importance of human oversight in monitoring item quality, fairness, and bias (e.g., Hao et al., 2024; Shin & Lee,

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

2023). Experts observed that GenAI struggled to integrate and assess multiple sub-skills in the same item, limiting its effectiveness as a holistic assessment tool. Enhancing GenAI to develop integrated testing methods capable of assessing multiple skills while maintaining high item quality could lead to more comprehensive assessments. This advancement would also support innovations like Cognitive Diagnostic Assessment and Cognitive Diagnostic Computerized Adaptive Testing, enabling precise, tailored evaluations that efficiently diagnose examinees' specific knowledge structures and skill levels.

Analysis of the trial of the GenAI-generated items showed that while these met standards for reliability, some items did not align well with the ability level of the assessment participants. These items were often not at the appropriate level of difficulty, which aligned with the experts' conclusions, as mentioned above, that they were too easy. The experts rated the quality of the GenAI-generated text highly, noting that GenAI has potential for generating texts that meet traditional standards for reading comprehension tests in terms of structure and topic suitability. This further reinforces the potential for GenAI in test development and suggests that, while content creation capabilities are promising, further refinement is needed in item design and skill assessment.

The fine-tuned GPT model demonstrated high accuracy in AWE scoring, particularly for total scores. However, performance in sub-scores was not as strong, and feedback and subsequent learning recommendations were often general and lacking in specificity, as observed by Zhai and Ma (2021, 2022). This indicates that there is room for improvement in the fine-tuning process. Future fine-tuning efforts could integrate human raters' detailed scoring criteria to enhance consistency in evaluating finer features of writing. This approach would not only help the model emulate human raters' thought processes more accurately but also provide more personalized learning guidance for students.

Building an AWE system that genuinely understands reasoning and critical thinking in human writing remains a significant challenge. This study analyzed short writing tasks completed by sub-doctoral students, which, though requiring the expression of personal opinion, primarily focused on data description. The AWE model's effectiveness for writing tasks which demand a greater level of critical thinking, such as DELNA's argumentative essays for PhD students, requires further research.

Another issue to consider is the social aspect of writing. The National Council of Teachers of English (2013) argues that machine scoring, when there is no engagement with human readers, could imply that writing lacks value. A solution to some of these concerns would be to use a fine-tuned GenAI-assisted AWE system to support, rather than replace, human raters. In such a model, a human rater could be paired with the AWE system, and a second human rater involved in cases of discrepancy. This would have the benefit of ensuring that feedback to students remained mediated by language advisors. It would seem that, regardless of AI advancements, the irreplaceable value of human interaction remains essential.

Opportunities and challenges posed by AI for PELA

This section addresses the third research question. The potential impact of AI on language assessment, in the specific context of a PELA such as DELNA, was analyzed

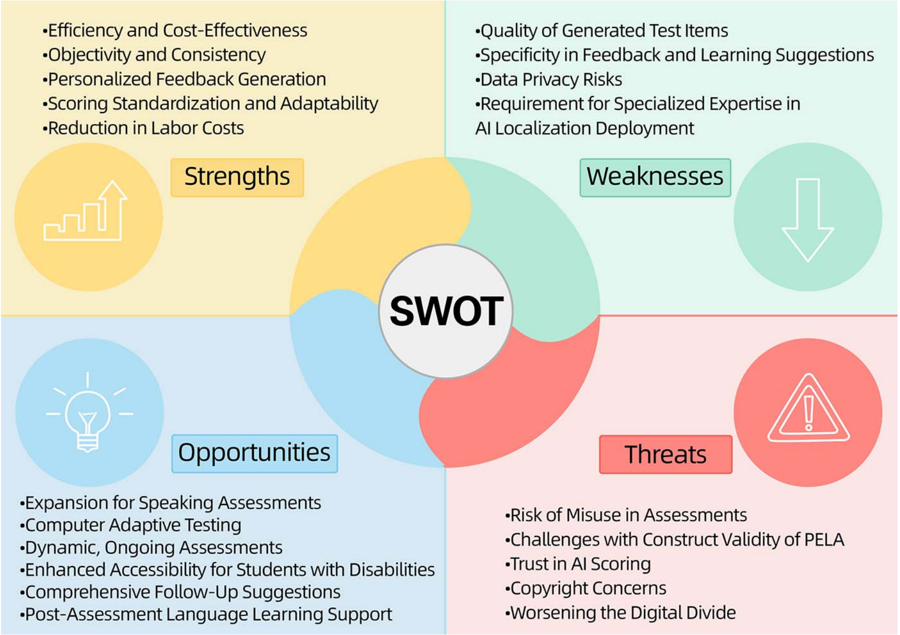

•Efficiency and Cost-Effectiveness
•Objectivity and Consistency
•Personalized Feedback Generation
•Scoring Standardization and Adaptability
•Reduction in Labor Costs
Strengths
•Quality of Generated Test Items
•Specificity in Feedback and Learning Suggestions
•Data Privacy Risks
•Requirement for Specialized Expertise in AI Localization Deployment
Weaknesses
SWOT
Opportunities
•Expansion for Speaking Assessments
•Computer Adaptive Testing
•Dynamic, Ongoing Assessments
•Enhanced Accessibility for Students with Disabilities
•Comprehensive Follow-Up Suggestions
•Post-Assessment Language Learning Support
Threats
•Risk of Misuse in Assessments
•Challenges with Construct Validity of PELA
•Trust in AI Scoring
•Copyright Concerns
•Worsening the Digital Divide

Figure 4. SWOT analysis of AI in PELA.

using the SWOT framework – a tool frequently applied in educational contexts (Farrokhnia et al., 2024), as inspired by Liu (2024).

As shown in Figure 4, AI offers potential to enhance PELA's efficiency, reduce costs, provide personalized learning support, and assist students with practice in areas such as speaking and writing. AI also holds potential to address some of the limitations inherent in current PELA. Some examples are listed below:

- Using a Computer Adaptive Testing system, AI could deliver a personalized assessment where each student faces questions matched to their language level. This approach saves time, improves efficiency, and avoids irrelevant questions, leading to more accurate language proficiency assessments and a fairer testing experience (Ratnayanti, 2023).
- Advancements in multimodal technology suggest AI's potential to assess students' speaking skills, a skill currently not part of DELNA, and often not included in PELA.
- AI-driven multimodal technology could also provide better accessibility for students with disabilities.
- AI could allow for ongoing academic English language assessment throughout a students' academic journey, something some staff at WTR/UoA have advocated for. Such a dynamic model of assessment would allow for continuous monitoring and support of students.

However, AI also presents challenges and threats for PELA. One of these, common to all unsecured assessment procedures, is the misuse of AI by students completing a

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

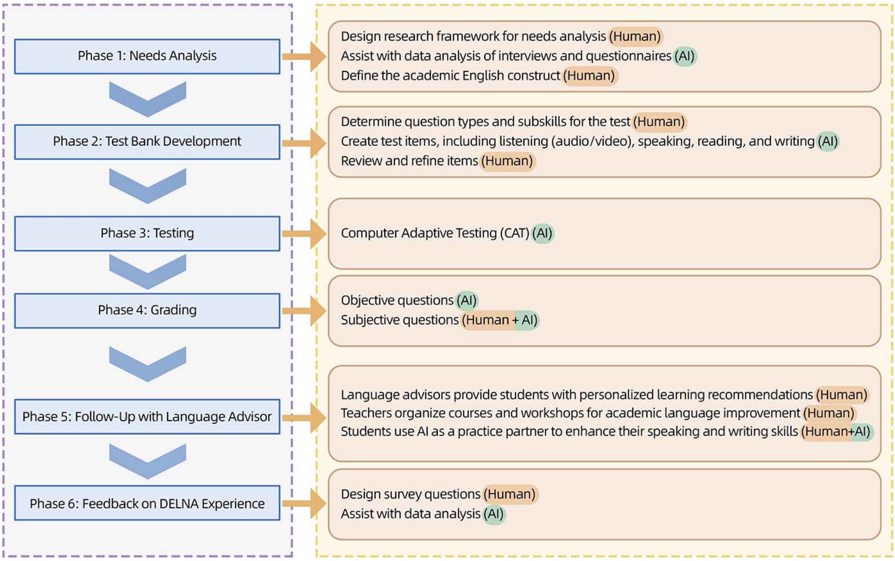

Phase 1: Needs Analysis
Design research framework for needs analysis (Human)
Assist with data analysis of interviews and questionnaires (AI)
Define the academic English construct (Human)
Phase 2: Test Bank Development
Determine question types and subskills for the test (Human)
Create test items, including listening (audio/video), speaking, reading, and writing (AI)
Review and refine items (Human)
Phase 3: Testing
Computer Adaptive Testing (CAT) (AI)
Phase 4: Grading
Objective questions (AI)
Subjective questions (Human + AI)
Phase 5: Follow-Up with Language Advisor
Language advisors provide students with personalized learning recommendations (Human)
Teachers organize courses and workshops for academic language improvement (Human)
Students use AI as a practice partner to enhance their speaking and writing skills (Human+AI)
Phase 6: Feedback on DELNA Experience
Design survey questions (Human)
Assist with data analysis (AI)

Figure 5. A proposed collaborative model between humans and AI within PELA.

PELA, with associated academic misconduct. There are also potential copyright and privacy concerns when using AI-generated items (Hao et al., 2024). The use of AI can also undermine the constructs on which a PELA is based. For instance, some students on social media have argued that since GenAI can handle proofreading and grammar effortlessly, assessment in these areas should be relaxed.

The results of this study suggest that the use of AI does not eliminate the need for human involvement. A more appropriate proposal for the use of AI in PELA might be the implementation of a collaborative model where human involvement is an integral component. Figure 5 sets out how tasks could potentially be apportioned, according to the different stages of the assessment process, in such a model. However, ongoing research is needed to explore the impact of AI and to ensure that PELA such as DELNA remains fit for purpose.

A number of educational bodies aim to create guidelines for responsible AI use; these include the AI Competency Framework for Students (UNESCO, 2024), the Australian Framework for GenAI in Schools (Department of Education, 2023), and ETS's Responsible Use of AI in Assessment (Educational Testing Service, 2023). There is also a need for a comprehensive set of guidelines for the responsible use of AI in language assessment, and in the context of PELA. As Andringa and Godfroid (2020) point out, there is a need for AIG and AWE studies that work with diverse samples and contexts to drive research in the area of language assessment.

Conclusion

This study examined the integration of GenAI in language assessment, particularly its application in automated test item generation and writing evaluation in PELA.

Findings reveal possibilities for the use of GenAI, especially in generating structured reading texts and evaluating writing tasks. While GenAI shows promise in content creation, expert evaluations highlight the need for refinement in question complexity and subskill targeting. In AWE, the fine-tuned GPT model aligns closely with human ratings, particularly in overall scoring, though improvements are needed for the provision of more specific feedback.

A SWOT analysis underscores the potential of GenAI to make PELA more efficient, adaptive, and personalized. AI could expand PELA's reach to areas like oral skills and support continuous student assessment. However, challenges such as academic integrity, privacy, and the need for clear guidelines emphasize the importance of adopting a Human-Centered AI approach to safeguard the integrity of language assessment.

One limitation of this study is that it included only a single AI-generated text and its related items, without incorporating a benchmark comparison. The absence of a human-generated text or pre-established standard for comparison limits the ability to comprehensively evaluate the quality and effectiveness of the AI-generated content. Future research could address this limitation by introducing a comparative analysis with human-generated materials to provide a more robust evaluation of AI-generated texts and items. Another limitation of this study lies in the training of the GenAI for AWE. The training materials used consisted solely of student writing samples and their scores. However, no detailed feedback on specific aspects of each essay, such as issues in word choice, grammar, or logic, was provided. As a result, the feedback generated by the GenAI was relatively general and lacked detailed examples or targeted suggestions. Future studies could enhance the feedback quality by incorporating more granular annotations and detailed feedback in the training process, enabling the GenAI to generate more specific and actionable recommendations for students.

Moving forward, establishing standards for GenAI-generated content and refining AI alignment with human scoring practices are essential. Continued research in diverse contexts will deepen understanding of AI's role in language assessment, fostering a balanced, responsible approach to AI-enhanced educational tools.

References

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. doi:10.1017/S0267190520000033

Anson, C., Filkins, S., Hicks, T., O'Neill, P., Pierce, K. M., & Winn, M. (2013). *NCTE position statement on machine scoring: Machine scoring fails the test*. National Council of Teachers of English. https://ncte.org/statement/machine_scoring/

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, 903077. doi:10.3389/frai.2022.903077

Australian Government, Department of Education. (2023). *Australian Framework for Generative Artificial Intelligence (AI) in Schools*. Retrieved September 14, 2024, from https://www.education.gov.au/schooling/resources/australian-framework-generative-artificial-intelligence-ai-schools

Bermouth, J. R. (1970). *On the theory of achievement test items*. University of Chicago Press.

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

Bezirhan, U., & Von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, *5*, 100161. doi:10.1016/j.caeai.2023.100161

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*, 2041–2058. doi:10.5281/zenodo.8115784

Cardwell, R., LaFlair, G. T., & Settles, B. (2024). *Duolingo English Test: Technical Manual. Duolingo, Inc*. Retrieved November 1, 2024, from https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. doi:10.5194/gmd-7-1247-2014

Doe, C. (2014). Diagnostic English Language Needs Assessment (DELNA). *Language Testing*, *31*(4), 537–543. doi:10.1177/0265532214538225

Doewes, A., Kurdhi, N. A., & Saxena, A. (2023). *Evaluating Quadratic Weighted Kappa as the standard performance metric for automated essay scoring*. 16th International Conference on Educational Data Mining. https://doi.org/10.5281/zenodo.8115784

Dunworth, K. (2009). An investigation into post-entry English language assessment in Australian universities. *Journal of Academic Language and Learning*, *3*(1), 1–13. https://journal.aall.org.au/index.php/jall/article/view/67

Educational Testing Service. (2023). *Responsible use of AI in assessment*. Retrieved October 15, 2024, from https://www.ets.org/Rebrand/pdf/ETS_Convening_executive_summary_for_the_AI_Guidelines.pdf

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37–64. doi:10.1177/0265532207071511

Elder, C., & Erlam, R. (2001). *Development and validation of the Diagnostic English Language Needs Assessment (DELNA): Final report*. The University of Auckland, Department of Applied Language Studies and Linguistics.

Erlam, R., von Randow, J., & Read, J. (2013). Investigating an online rater training program: Product and process. *Papers in Language Testing and Assessment*, *2*(1), 1–29. doi:10.58379/aada5911

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. doi:10.1080/14703297.2023.2195846

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement (Version 0.84.1) [R package]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=irr

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. doi:10.1007/s10439-023-03272-4

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109. doi:10.1126/science.adi1778

Hamner, B., & Frasco, M. (2018). *Metrics: Evaluation Metrics for Machine Learning (Version 0.1.4)*. R package. Comprehensive R Archive Network (CRAN). Retrieved October 10, 2024, from https://CRAN.R-project.org/package=Metrics

Hao, J., Alina, A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement*, *43*(2), 16–29. doi:10.1111/emip.12602

Hockly, N. (2023). Artificial intelligence in English language teaching: The good, the bad and the ugly. *RELC Journal*, *54*(2), 445–451. doi:10.1177/00336882231168504

Ifenthaler, D. (2023). Automated essay scoring systems. In O. Zawacki-Richter & I. Jung (Eds), *Handbook of open, distance and digital education* (pp. 1057–1071). Springer.

Klebanov, B. B., & Madnani, N. (2022). *Automated essay scoring*. Morgan & Claypool Publishers.

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225. doi:10.1016/j.caeai.2024.100225

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550. doi:10.1177/00336882231162868

https://doi.org/10.1017/S0267190525000030 Published online by Cambridge University Press

Lee, Y. W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, *32*(3), 299–316. doi:10.1177/0265532214565387

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

Liu, J. In Chinese. (2024). Language assessment in the era of artificial intelligence: Opportunities and challenges. *Modern Foreign Languages*, *47*(6), 34–49. 10.20071/j.cnki.xdwy.20240824.008.

Liu, X. H. (2018). *Establishing the foundation for a diagnostic assessment of reading in English for academic purposes* [Doctoral thesis, University of Auckland.] UOA Library. http://hdl.handle.net/2292/37053

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. doi:10.1016/j.rmal.2023.100050

Murray, N. (2016). *Standards of English in higher education: Issues, challenges and strategies*. Cambridge University Press.

Peters, M. E., Ruder, S., & Smith, N. A. (2019). To tune or not to tune? Adapting pretrained representations to diverse tasks. *arXiv preprint, arXiv:1903.05987.* doi:10.48550/arXiv.1903.05987

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard. *Applied Measurement in Education*, *28*(2), 130–142. doi:10.1080/08957347.2014.1002920

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. doi:10.1007/s10462-021-10068-2

Ratnayanti, R. (2023). Artificial Intellegence (AI) in association with language assessment. *Journal of Science, Education and Studies*, *2*(3). doi:10.30651/jses.v2i3.20346

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes*, *7*(3), 180–190. doi:10.1016/j.jeap.2008.02.001

Read, J. (2015). Issues in post-entry language assessment in English-medium universities. *Language Teaching*, *48*(2), 217–234. doi:10.1017/s0261444813000190

Read, J., Ed.. (2016). *Post-admission language assessment of university students*. Springer.

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language, Learning and Technology*, *27*(3), 27–40. https://hdl.handle.net/10125/73530

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

UNESCO. (2024). *AI competency framework for students*. Retrieved November 21, 2024, from https://www.unesco.org/en/articles/ai-competency-framework-students

The University of Auckland. (2020). *The University of Auckland vision and strategic plan 2020-2030*. Retrieved November 1, 2024, from https://www.auckland.ac.nz/assets/about-us/the-university/official-publications/strategic-plan/2021-2030/taumata-teitei-vision-2030-and-strategic-plan-2025.pdf

The University of Auckland. (2024). *DELNA handbook*. Retrieved October 17, 2024, from https://www.auckland.ac.nz/assets/delna/delna/delna-handbook-2024.pdf

Wang, Q., & Gayed, J. M. (2024). Effectiveness of large language models in automated evaluation of argumentative essays: Finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*, 1–29. doi:10.1080/09588221.2024.2371395

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave & K. Cho (Eds), *Advances in neural information processing systems* (pp. 24824–24837). Curran Associates, Inc.

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. doi:10.1111/j.1745-3992.2011.00223.x

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System*, *124*, 103351. doi:10.1016/j.system.2024.103351

Yavuz, F., Çelik, Ö., & Çelik, G. Y. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(1), 150–166. doi:10.1111/bjet.13494

Yuan, S., He, T., Huang, H., Hou, R., & Wang, M. (2020). Automated Chinese essay scoring based on deep learning. *Computers, Materials and Continua*, *65*(1), 817–833. doi:10.32604/cmc.2020.010471

Zhai, N., & Ma, X. (2021). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, *35*(9), 2817–2842. doi:10.1080/09588221.2021.1897019

Zhai, N., & Ma, X. (2022). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875–900. doi:10.1177/07356331221127300

Zirar, A. (2023). Exploring the impact of language models, such as ChatGPT, on student learning and assessment. *Review of Education*, *11*(3), e3433. doi:10.1002/rev3.3433

Cite this article: Zhang, T., Erlam, R., & de Magalhães, M. B. (2025). Exploring the dual impact of AI in post-entry language assessment: Potentials and pitfalls. *Annual Review of Applied Linguistics*, *45*, 274–293. https://doi.org/10.1017/S0267190525000030