Comparative Study of GenAI (ChatGPT) vs. Human in Generating Multiple Choice Questions Based on the PIRLS Reading Assessment Framework

Lam, Yu Yan Hong Kong Metropolitan University, Hong Kong | yuylam@hkmu.edu.hk

Chu, Samuel Kai Wah Hong Kong Metropolitan University, Hong Kong | skwchu@hkmu.edu.hk

Ong, Elsie Li Chen Hong Kong Metropolitan University, Hong Kong | eong@hkmu.edu.hk

Suen, Winnie Wing Lam Hong Kong Metropolitan University, Hong Kong | wlsuen@hkmu.edu.hk

Xu, Lingran Hong Kong Metropolitan University, Hong Kong| lixu@hkmu.edu.hk

Lam, Lavender Chin Lui University of Hong Kong, Hong Kong | lavlcl@connect.hku.hk

Wong, Scarlett Man Yu University of Hong Kong, Hong Kong | scar1022@connect.hku.hk

ABSTRACT

Human-generated multiple-choice questions (MCQs) are commonly used to ensure objective evaluation in education. However, generating high-quality questions is difficult and time-consuming. Generative artificial intelligence (GenAI) has emerged as an automated approach for question generation, but challenges remain in terms of biases and diversity in training data. This study aims to compare the quality of GenAI-generated MCQs with humans-created ones. In Part 1 of this study, 16 MCQs were created by humans and GenAI individually with alignment to the Progress in International Reading Literacy Study (PIRLS) assessment framework. In Part 2, the quality of MCQs generated was assessed based on the clarity, appropriateness, suitability, and alignment to PIRLS by four assessors. Wilcoxon rank sum tests were conducted to compare GenAI versus humans generated MCQs. The findings highlight GenAI's potential as it was difficult to differentiate from human created questions and offer recommendations for integrating AI technology for the future.

KEYWORDS

Reading; PIRLS; GenAI; question creation; question assessment.

INTRODUCTION

The PIRLS is a widely recognized assessment framework used to evaluate reading comprehension skills among students worldwide. It was designed to measure and monitor children's reading literacy levels across different countries (Mullis et al., 2017). The framework covers 4 key aspects of reading literacy: information retrieval (fact finding), making inferences, interpreting texts, integrating ideas, and evaluating information. For many decades, creating multiple choice questions (MCQs) based on the PIRLS framework has required careful consideration of the assessment objectives, reading comprehension skills, and cognitive processes targeted by the assessment. Human-generated multiple MCQs are one of the common assessment methods that benefit from the expertise and knowledge of educators and assessment specialists. This is because writers can tailor questions to specific learning objectives, ensure clarity and alignment with curriculum standards, and incorporate nuances that reflect real-world reading comprehension challenges faced by students. However, the biggest limitation is that high-quality questions are difficult and time-consuming to create because educators need a comprehensive understanding of the PIRLS assessment framework, including its objectives, content domains, and proficiency levels. For example, MCQs should assess students' ability to comprehend and analyze various types of texts, including narratives, expository texts, and informational graphics. Additionally, questions should target higher-order thinking skills, such as inference-making and critical evaluation.

In recent years, the use of GenAI has emerged as a novel approach to automate the process of question generation for educational assessments (Cheung et al., 2023; Doughty et al., 2024; Laupichler et al., 2024). GenAI systems, such as ChatGPT, have demonstrated remarkable capabilities in natural language understanding and generation. These systems leverage large language models (LLM) trained on vast amounts of text data to generate human-like responses and content. In the context of educational assessments, GenAI shows potential for automating question generation tasks, including MCQs.

Recent studies compared the effectiveness of GenAI versus human-generated MCQs in educational assessments. For example, Cheung et al. (2023) highlighted the potential of AI in reducing the workload of medical staff and improving the quality of medical education assessments. Consistent with this, they reported that ChatGPT generated questions faster than humans, but human-generated questions scored slightly higher in relevance. AI, such as

ChatGPT, has been used in education to improve learning outcomes and can generate quality MCQs for medical education. AI-generated questions outperformed human ones in some assessment domains, showing potential for high-quality question creation.

Laupichler et al., (2024) conducted a comparative study between AI-generated questions and human questions in the medical preparatory exam. They found that there was no statistically significant difference in the difficulty of questions and there is a better outcome in evaluation on high- from low-performing students in LLM questions. The findings revealed that LLMs including ChatGPT could be successfully employed to create questions for formative exams in medical schools.

Despite the potential of GenAI for question generation, several challenges remain. These include the need to ensure the diversity and representativeness of training data, address biases inherent in AI models, and enhance the interpretability and explainability of generated content. Additionally, human-generated MCQs offer advantages in terms of pedagogical expertise, contextual understanding, and domain-specific knowledge. Therefore, the current study aims to explore and compare GenAI and human-generated MCQs based on the PIRLS reading assessment framework.

METHOD

The current research composed of two parts: Part 1 Questions Creation by human versus by GenAI; and Part 2 Questions Assessment. This bifurcated approach was developed to comprehensively evaluate the capabilities of humans versus GenAI, specifically ChatGPT-4, in generating MCQs aligned with the PIRLS reading assessment framework.

Data Sources

In this study, two children's books were extracted from eFunReading.com and Reading Battle 2.0 online platform. They were chosen for their appeal to the target age group of 8-12 and their diversity in genre. These selections aimed to assess the comparative effectiveness of humans and GenAI (ChatGPT) in generating MCQs aligned with the PIRLS reading literacy framework. The task of selecting MCQs and developing specific prompts for ChatGPT was undertaken by the same individual. This dual role facilitated a nuanced approach to generating 16 MCQs that accurately reflected the comprehensive reading comprehension levels defined by PIRLS, ensuring a cohesive and direct comparison between human and AI-generated content.

Part 1: Question Creation

Group 1 focused on drawing upon the human expertise to select questions available from the data source while Group 2 emphasized generation of the MCQs by ChatGPT with the alignment with the PIRLS assessment framework.

Group 1 (Human Expert): One human expert was selected as they have extensive experience in question creation and a profound understanding of the PIRLS framework. 16 MCQs were selected from the data source which included eight questions from each of the two books' MCQs dataset, ensuring a balanced representation across the PIRLS assessment levels.

Group 2 (GenAI): Few-Shot Learning (FSL) was applied to GenAI models, where ChatGPT was given explicit MCQ examples for each PIRLS level alongside the book excerpts. Few-Shot Learning (FSL) refers to the ability of GenAI models to generalize few labeled training examples (Parnami & Lee, 2022). This approach aimed to refine the AI's question generation process further.

The main study aim was to compare between human expertise and GenAI's (ChatGPT) capability, in generating MCQs aligned with the PIRLS reading assessment framework. The human-led process (Group 1) involved four educators in selecting MCQs that meet the nuanced requirements of the PIRLS levels. Group 2 utilizes ChatGPT's application of Few-Shot Learning to generate 16 MCQs based on the content of two books. In the designated prompt, ChatGPT received the excerpts from the books together with sample book content, and corresponding explicit examples of MCQs tailored to each PIRLS level as Few-Shot Learning.

Prompt for ChatGPT-4.0 Few-Shot Learning Task with Real Examples

*"Based on the following excerpt from a children's book: [insert book excerpt here], create 8 multiple-choice questions suitable for 5th graders. Ensure each pair of questions aligns with a different level of the PIRLS reading assessment framework to test a range of comprehension skills. While creating these questions, incorporate the comprehension skills as detailed in the ****Guidance on PIRLS Levels**** and ****A Sample passage and the real examples for 4 PIRLS Levels****. Drawing inspiration from the examples provided, generate 8 MCQs suitable for 9-year-olds from the book excerpt. Each question should be created to align with a distinct level of the PIRLS reading assessment framework, with two questions on each level. Provide four options (A, B, C, D) for each question, clearly identifying the correct answer and offering a brief explanation that*

connects the correct choice to the excerpt's key ideas or themes.

Part 2: Question Assessment

Author assessors

Four author assessors (i.e. Expert Review Panel) were selected based on some inclusion criteria. They must have experience in question creation, teaching experience with children and familiarity with the PIRLS framework. These assessors were either a teacher, researcher, or an experienced question creator from the Reading Battle platform evaluated the MCQs based on four domains in an assessment tool. Prior to that, they have received sufficient guidance on the meaning of each domain before conducting the assessments.

Procedure

Data sources were extracted based on 2 children's books and the same as the ones used in Study 1. All 32 MCQs (16 MCQs from each group of participants, with 8 MCQs created by AI from book 1; 8 MCQs created by humans from book 1; 8 MCQs created by AI from book 2; and 8 MCQs created by human from book 2) were individually numbered and randomized using a computer-generated sequence. The assessment tool categorized into four domains were adapted from the measures used in Cheung et al. (2023):

1. **Alignment with the specified PIRLS levels**: determine if the question is set at the level stated (1-Information Retrieval, 2- Inferences, 3- Interpretation, 4- Evaluation);

2. **Clarity and Specificity**: determine if the question is clear and specific without ambiguity, its answerability and without being under- or over-informative;

3. **Appropriateness**: determine if the question is correct, appropriately constructed with appropriate length and well-formed;

4. **Suitability for Specific Age Group**: determine whether the question is suitable for assessing reading comprehension skills in the target age group.

Each question was assessed on the reviewer's agreeableness to the statement on a numeric scale from 0–10, with "0" being as extremely disagree to "10" as extremely agree. The assessors were also asked to determine if the questions were constructed by AI or by humans, which they were blinded by the total number of questions created.

RESULTS

Source	Correctly Identified	Incorrectly identified	Total	Percentage Correctly Identified
AI-Generated	21	43	64	32.81%
Human-Generated	36	28	64	56.25%

Table 1: Identification Accuracy of AI-Generated vs. Human-Generated MCQs

As the data did not meet the assumptions required for the parametric tests partly due to small sampling number, three Wilcoxon signed-rank tests were conducted and indicated that there were no significant differences between Human and GenAI in their ratings of (1) Clarity and Specificity, (2) Appropriateness, and (3) Suitability for Specific Age Group ($p > .05$). However, MCQs generated by GenAI were significantly higher in the ratings of PIRLS assessment framework alignment (7.98) than those generated by humans (6.61), $p = .035$.

DISCUSSION

Recent studies such as Cheung et al. (2023) and Doughty et al. (2024) have already established that Generative AI has helped tremendously in the process of generating MCQs for educational assessment in Higher educational settings. The current study attempted to explore the power of this GenAI tool further with alignment to the PIRLS assessment framework and in a younger age group. Findings indicated that the MCQs generated by GenAI were comparable to humans on the Clarity and Specificity, Appropriateness, and Suitability for Specific Age Groups. Based on the ratings of the four assessors, both humans and GenAI created clear and accurate questions efficiently. They were effective in generating a diverse range of questions that covered various levels in the PIRLS framework. The percentage of correctly identified MCQs in Table 1 shows that assessors were more accurate in identifying human-generated questions (56.25%) compared to AI-generated questions (32.81%). This suggests that characteristics or qualities of human-generated content might be more easily recognizable to assessors reflecting on the current state and potential areas for improvement in AI-generated educational content, in order to generate

human-like questions. Overall, the low percentage figures indicated that it was very hard to differentiate whether questions were created by humans or GenAI although more data will be needed to confirm this further.

One of the primary challenges in categorizing questions within the PIRLS framework is the ambiguity in question phrasing. Questions may contain multiple components or require question creators to make inferences based on implicit information, making it difficult to assign them to a specific PIRLS level. This ambiguity can lead to inconsistencies in categorization and affect the overall reliability of the assessment.

The PIRLS assessment has been administered in multiple countries and languages, each with its own cultural and linguistic nuances. It has been argued that questions created by humans are straightforward for students in one context but may be challenging for students in another context due to differences in language proficiency, background knowledge, age, and cultural references (Ibrahim et al., 2020). In the current study, having a completely different nature of participants (GenAI) to create questions aimed at young children is a good example of these linguistic and perceptual differences. This may partly explain why there was a significant difference in how well the questions created by humans compared to GenAI adopted the PIRLS framework. The lower rating for human created questions reflected that categorizing questions into the four levels of the PIRLS framework requires subjective judgment by the creators. Different question creators may interpret the complexity of questions differently, leading to inconsistencies in scoring and potentially impacting the validity of the assessment results. Moreover, different levels of understanding the categorization and question difficulties may result in variations that would not be present in the situation if GenAI were the question creator.

The current findings have significant implications for the future of educational assessment design because by demonstrating that GenAI can effectively create MCQ questions based on topic, difficulty level, and question format, the study highlights the potential for educators to utilize AI tools to customize assessments to better fit the unique needs of students and their curriculum. This flexibility allows educators to quickly produce tailored assessment materials that align with specific learning objectives and student capabilities, thereby enhancing the relevance and effectiveness of the evaluation process. Additionally, the efficiency and scalability of AI-generated questions can reduce the workload for educators, freeing up more time for personalized teaching and student engagement. As educational technologies continue to evolve, integrating GenAI into the question creation process promises to revolutionize how educators approach testing, potentially leading to more dynamic, responsive, and individualized learning environments.

FUTURE DIRECTIONS

Some questions often contain elements of multiple complexity levels, making it challenging to assign them to a specific PIRLS level. Categorizing such questions accurately within the PIRLS framework can be a complex and nuanced task. This would be an issue for both GenAI and human questions creators to resolve in future categorization tasks. In light of the subjectivity of the MCQ creation and categorization exercises, the provision of comprehensive training for assessors and GenAI on both the categorization criteria and guidelines within the PIRLS framework can help improve consistency in scoring practices. For curriculum developers and educational policymakers, it is crucial to recognize the dual implications of using GenAI in educational technology given the findings that MCQs generated by AI scored significantly higher in alignment with the PIRLS assessment framework compared to those created by humans. Despite the enlightening finding of this research in the contribution of AI, as a foresight to the future, we must however remind ourselves that over- reliance on AI for generating MCQs may risk losing the deeper educational objectives.

GENERATIVE AI USE

In this study, OpenAI's ChatGPT has been utilized to generate MCQs for a comparative analysis between GenAI and human expertise in creating educational content, which aligned with the PIRLS assessment framework.

The authors take full responsibility for the content of this submission, ensuring that the integration of GenAI-generated content adheres to academic standards and contributes meaningful insights into the potential of AI in educational settings.

REFERENCES

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PloS one, 18(8), e0290691. https://doi.org/10.1371/journal.pone.0290691

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024). A Comparative Study of AI-Generated (GPT-4) and Human-crafted MCQs in Programming Education. In Proceedings of the 26th Australasian Computing Education Conference (ACE 2024). ACM. https://doi.org/10.1145/3636243.3636256

Ibrahim, A., Alhosani, N., & Vaughan, T. (2020). Impact of language and curriculum on student international exam performances in the United Arab Emirates. Cogent Education, 7(1). https://doi.org/10.1080/2331186X.2020.1808284

Laupichler, M. C., Rother, J. F., Kadow, I. C. G., Ahmadi, S., & Raupach, T. (2024). Large Language models in Medical Education: comparing ChatGPT-to Human-generated exam questions. Academic Medicine, 99(5), 508-512.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (Eds.). (2017). PIRLS 2016 assessment framework. International Association for the Evaluation of Educational Achievement (IEA).

Parnami, A., & Lee, M. (2022) Learning from Few Examples: A Summary of Approaches to Few-Shot Learning