Leveraging Generative AI to Scale Item Generation Amid Rising Assessment Demands in Education

Jonah Kenei

Kenya National Examination Council

jkenei@knec.ac.ke

https://doi.org/10.62049/jkncu.v5i2.331

Abstract

In the last few years assessment bodies have witnessed larger student numbers, reduced resources and increasing costs of exam administration which have led to the emergence of e-assessment as efficient and cost-effective method of assessment in education. As educational institutions shift towards the complete adoption of e-assessments for high-stake examinations, there are associated challenges that accompany this development. The transition from traditional paper-based examinations to digital formats present both opportunities and challenges for educational systems worldwide. A significant challenge is to develop many test items to support e-assessment. Leveraging technology, such as automated item generation (AIG), can scale up the number of items required in e-assessment but with its fair share of challenges. This paper investigates the computational techniques of scaling up item development to support e-assessment and facilitating a seamless transition to e-assessments while maintaining quality and fairness of assessment. Through a review of literature, the paper explores the challenge of traditional item writing in transitioning to digital e-assessment platforms. Additionally, the paper examined the use of automated item generation tools that are currently in use and their contribution in assessment practice. From the literature, it was observed that template-based AIG has demonstrated great success with many practical use cases. However, despite the success of template-based AIG over traditional item writing, it still requires subject matter experts to develop cognitive models and item models. Finally, the potential of generative artificial intelligence (AI) in addressing this problem was examined. Finally, we explored ChatGPT generative AI model in generating test items, with the objective of finding out its capabilities. By using ChatGPT, it is possible to generate different types of items that are comparable to those created by human authors and this novel development is expected to help assessment providers seeking to navigate the complexities of transitioning to digital high-stakes examinations.

.

Keywords: *Assessment, Test Item, Test Development, Generative AI, Test Bank*

Introduction

In the last few years, there has been significant advancement in the field of artificial intelligence, and it has significantly impacted different sectors, including education. AI technologies, like generative AI, have the potential to play a significant role in educational assessment [1]. In today's digital age, generative AI is making inroads into a wide range of applications, unlocking new possibilities, and driving innovations in unprecedented ways. Generative AI is an emerging field of artificial intelligence that uses algorithms to automatically generate content such as text, images, audio, and video [2]. It involves training models on large datasets and using them to generate new content that resembles the patterns and characteristics of the training data [3]. Manually developing test items is a time-consuming and expensive process [4],[5]. Being able to automatically generate test items by making use of generative AI might present a solution for this problem. Automating test item development is an evolving concept of leveraging computer algorithms to generate test items automatically. Automated test item development is a promising application of generative AI with the potential of scaling item development process, while producing high-quality items quickly and efficiently in the assessment process. The growing popularity of online assessments has resulted in an increased demand for the quick development of high-quality test items [6]. Automated item development using generative AI has the potential to reduce reliance on human subject experts in item writing. Despite the potential of generative AI in item writing, there is still no existing cases to encourage a general application of generative AI in item writing. It is therefore important to evaluate generative AI regarding its feasibility, validity, and item quality. In this paper we explored the potential application of generative AI in the domain of educational assessments, specifically in the context of automated item writing. This study examines test development and the potential of generative Artificial Intelligence (AI) in item writing and prospects to integrate it into education assessment.

Background

Assessment is a crucial component of the education system, serving as an instrument that provides evidence of learning, measures students' progress, and demonstrates their understanding of the curriculum [7]. Generating test items is a crucial task in test development, demanding substantial time and human effort. The advent of automatic item generation introduces a novel way for test item generation. Test items are fundamental components of the assessment process and serve as how the knowledge, abilities, and competencies of students are assessed. The process of developing items is referred to as item writing which refers to the process of creating assessment items or questions for use in educational assessments. Traditionally, item writing has predominantly been a manual process, relying on the expertise and effort of human item writers. However, manual item development is expensive, time-consuming, and labor-intensive process. It relies heavily on human subject experts. In this approach, each item writer develops an item, which is then reviewed, edited, and revised by a team of experts to make sure it satisfies predefined quality control requirements. This cooperative approach helps in ensuring the reliability, validity, and quality of the assessment items [8]. However, manual item development by item writers is time-consuming and costly [4], [5], [9]. Another issue is the shortage of experienced and well-trained item writers who can develop high-quality items [8] as it requires training, experience, and resources. The detrimental effect of this is inability to develop enough test items for assessments and hindering new advances like adaptive assessment that require a large pool of items [10].

Assessment is a key component in any education system, and the prevalence of e-assessments has experienced continued popularity in the last few years. As a result, there is a pressing need to generate many assessment items quickly and efficiently [11]. In literature, one of the most common challenges is coming up with sufficient high-quality assessment items for regular assessment [12]. On the other hand, the negative consequences of poor-quality items are widely acknowledged in literature such as in [13] and [14]. The shift to competency-based education as a progressive model of education has increased significantly over the last few years. In competency-based education, formative assessment is used in the learning process, to get regular feedback and monitor learners' progress [15]. E-assessment provides many benefits that can greatly support formative assessment practices.

The Problem of Scaling Up Item Development to Support E-Assessment

The shift from paper-based evaluation to e-assessment in education is laying the groundwork for the widespread use of technology-based solutions in assessment practice. However, this has also opened formidable new challenges, particularly in item development. Since assessments are administered to learners on a regular basis, many high-quality, diverse test items are required. This implies that a large pool of test items is required to develop item banks to support electronic delivery of formative assessment [16]. Traditional approaches for item development are facing challenges in today's era of continuous assessment. Much research works in literature have attempted to address the problems related to development of large test items using techniques generally referred to as automatic item generation (AIG) techniques. It is a relatively new but rapidly evolving research area where computer algorithms are used to automatically generate test items for assessment purposes. These techniques employ computational techniques to generate many items quickly and efficiently [16] as compared to traditional manual item writing by human subject experts. These techniques emerged as a solution to the challenges experienced by test developers in developing many quality test items [10]. A considerable amount of literature has been published on automatic item generation as a means of creating test items. Most works in literature use template-based methods to generate test items [17].

Automated Item Generation (AIG) and Emerging E-Assessment

In the last few years, e-assessments have experienced increased popularity, which may be attributed to advancements in technology and the increasing availability of digital platforms [9]. With the increasing demand for assessments, especially online assessments, the ability to generate items efficiently and at higher scale becomes crucial. With the shift to competence-based curriculum, assessment is a continuous process that takes place throughout the learning cycle. Therefore, the development of many test items becomes necessary. The demand for a large pool of test items, especially in e-assessments, has led to the development of Automatic item generation (AIG) tools. Automatic item generation (AIG) is a new area of assessment research where a set of test items are generated using computational techniques. It generates test items by combining computational techniques with cognitive and psychometric modeling techniques [18]. The motivating factor behind research on automatic item generation is generating unlimited test items for this assessment. Automatic item generation is not a new concept; it was first described in the 1960's John Bormuth's work which laid the foundation for the development of automated techniques for generating test items. Though, its actualization and widespread acceptance wasn't achieved until later years [19]. The advancements in computing technology in more recent years have made AIG gain significant attention and

an active research area in educational assessment. Most AIG techniques in use today are based on the pioneering works of John Bormuth. The AIG frameworks and approaches in use today can be classified into two categories: template-based AIG and non-template-based AIG. Template-based AIG uses predefined item templates to generate test items. The function of a template is to provide a structure for the items including the format, content, and response options. These templates are usually created by subject matter experts (SMEs) to guide the process of item generation. Non-template-based AIG leverages more advanced techniques, such as natural language processing (NLP) techniques to generate item features from a corpus of data that is relevant to the subject domain in which the items are to be generated [20]. Both these approaches each have their advantages and are used in different scenarios based on the specific needs and requirements of the assessment. Template-based AIG techniques are mostly used, which entails subject matter specialists developing cognitive and item models to generate several distinct assessment items using software. Template-based AIG has demonstrated great success with many practical use cases. However, despite the success of template-based AIG over traditional item writing, it still requires subject matter experts to develop cognitive models and item models [21].

Research Gap

The lack of a comprehensive method that can be used to automatically generate test items, independent of human subject matter experts, is what constitutes the research gap in this area. While there are many state-of-the-art approaches available, they still rely on human subject experts. This creates an urgent need for a more robust alternative solution that can be used independent of subject matter experts. Human experts can be a limited resource, and assessment bodies often face challenges in continuously renewing and expanding their item banks to sustain assessments. The process of item development can be time-consuming, expensive, and resource intensive. This is where the use of algorithms and automated item generation (AIG) techniques can offer potential solutions. By leveraging algorithms and AIG, testing agencies can potentially overcome some of the limitations associated with traditional item development processes. Closing this research gap would provide educators and exam administration bodies with more accessible and cost-effective approach to item generation that could be used across a range of disciplines and educational settings.

Potential of Generative AI Aid in Item Writing

Generative AI large language models (LLMs) have garnered considerable attention in recent years as artificial intelligence (AI) tools that generate content such as text at a level comparable to that of humans. It leverages deep learning models to generate human-like content [22]. One major benefit of LLMs is their accuracy in understanding the context of an input and generating the appropriate output [23]. ChatGPT is one such LLM that has gained significant attention since its launch in 2022 [24]. ChatGPT is a generative AI model that processes and generates text in natural language using deep learning techniques. It has demonstrated the potential of generating test items [25]. It can generate questions by simply taking the input of a subject or topic area. Many papers are currently devoted to analyzing the potential of these generative AI models in item writing. In literature, use of generative AI for automatic item generation has received little attention. As mentioned earlier, developing test items is a time-consuming process, and any application capable of automating this process could be highly valuable in test development. The emergence of OpenAI's state-of-the-art large language models (LLMs) such as GPT-3.5 and GPT-4 [26, 27] offers

some potential for addressing this challenge. This is due to their human-like text understanding and generation, and this could be pivotal in automating the creation of assessment items. In this section, we discuss the potential of generative AI in item writing. We provide an overview of generative AI along with its pros and cons.

Methodology

In this paper we introduce ChatGPT generative AI model developed by OpenAI as a new approach to generate test items. We adopted experimental research method for our study using ChatGPT to generate test items. We created ChatGPT account which is available freely on the internet at https://chat.openai.com. The experimental setup for our research had the following components:

- Selection of Test Item Topics: Selecting a subject area to generate test items from.
- Creation of Prompts: Developing a set of prompts that cover a range of concepts or skills within the chosen subject area. These will be used as inputs to ChatGPT to generate the corresponding test items.
- Chosoing ChatGPT version: Selecting specific version of ChatGPT that will be used.
- Generation of Test Items: Use ChatGPT to generate test items based on the prompts created above. Specify the type of test items (e.g., multiple-choice questions, short answer questions, etc.).
- Baseline Test Item Generation: Selecting existing test items that were developed using traditional manual processes.
- Evaluation: Evaluating the quality and suitability of the test items generated by ChatGPT and the baseline method relevance, accuracy, clarity, and alignment with the learning outcomes.

Results

ChatGPT was able to generate test items on different subjects and their corresponding subject areas. ChatGPT can be used to generate test items or questions by inputting a prompt into the model and then running it to generate an item based on the input. The input can be provided in the form of a sentence or a paragraph, depending on the degree of difficulty associated with the item.

The table below shows sample test items generated by using different prompts.

Prompt	**Mathematics multiple choice Question on place value of a number with thousands and three decimal places**
Question	What is the place value of digit 8 in the number 5,432.789? A) Ones B) Tens C) Thousands D) Thousandths
Prompt	Multiple-choice question converting fraction number into percentage
Question	What is 3/5 expressed as a percentage? A) 25% B) 40% C) 50% D) 60%

1st Prompt	multiple-choice question similar to A teacher shared pencils among his learners. The girls received 2/3 of the pencils and the boys received 5/6 of the remainder. The teacher remained with 8 pencils. How many pencils did the teacher have before sharing?
2nd Prompt	Another scenario
Question	A teacher shared pencils among his learners. The girls received 3/5 of the pencils, and the boys received 1/4 of the remainder. The teacher remained with 12 pencils. How many pencils did the teacher have before sharing? A) 30 B) 40 C) 48 D) 60

Based on the experimental results obtained, ChatGPT can be used to generate questions quickly and easily with appropriate prompts. By leveraging ChatGPT prompts one can successfully create different types of assessment items. The results of the experiments carried out produced items that are comparable to those generated by human authors. The total time required to formulate the initial prompt and generate an item is relatively short compared to manual item writing.

Conclusion

Harnessing the power of generative AI models, such as ChatGPT, to automatically generate assessment items makes the process of item writing fast and more efficient. ChatGPT for example has potential in generating assessment items for Kenyan educational system. By focusing on the various types of items and using the appropriate prompts, it is possible to generate different types of items such as multiple-choice questions, short answer questions, essays that are comparable to those created by human authors. There is need for further research to ensure that the items generated by generative AI models have acceptable psychometric characteristics. Generative AI has the potential in revolutionizing assessments by automating item generation thus making it possible unlimited number of items for use in assessments. Having an unlimited number of items will support e-assessment and other features of e-assessment such as adaptive testing. However, ethical considerations, biases, and user perceptions are critical factors that require careful attention in the integration of generative AI in assessments. Further research and development are needed to address these challenges and leverage the full potential of generative AI in educational assessments. It's important to note that while generative AI can be able to generate test items, human expertise and judgment are still recommended to involve subject matter experts and assessment specialists throughout the process to ensure the quality and validity of the assessment. As generative AI continues to evolve, embracing innovative tools like ChatGPT can be a game-changer in automating test items generation.

References

[1] Abunaseer, H. The Use of Generative AI in Education: Applications, and Impact. *Technology and the Curriculum: Summer 2023*.

[2] Kalota, F. (2024). A Primer on Generative Artificial Intelligence. *Education Sciences*, *14*(2), 172.

[3] Wu, T., & Zhang, S. H. (2023, August). Applications and Implication of Generative AI in Non-STEM Disciplines in Higher Education. In *International Conference on AI-generated Content* (pp. 341-349). Singapore: Springer Nature Singapore.

[4] Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., & Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Advances in Health Sciences Education*, *28*(5), 1441-1465.

[5] Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A. P., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and learning in medicine*, *28*(2), 166-173.

[6] Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, *5*, 100161.

[7] Oldfield, A., Broadfoot, P., Sutherland, R., & Timmis, S. (2012). Assessment in a digital age: a research review, Graduate School of Education, University of Bristol.

[8] Haladyna, T.M., & Rodriguez, M.C. (2013). Developing and Validating Test Items (1st ed.). Routledge. https://doi.org/10.4324/9780203850381

[9] Rossi, O. (2023). Using technology to write language test items. Talk delivered at IATEFL TEASIG online conference *Developing Assessment Tasks for the Classroom*. September 2023.

[10] Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*, 121-204.

[11] Circi, R., Hicks, J., & Sikali, E. (2023, May). Automatic item generation: foundations and machine learning-based approaches for assessments. In *Frontiers in Education* (Vol. 8, p. 858273). Frontiers.

[12] Karthikeyan, S., O'Connor, E., & Hu, W. (2019). Barriers and facilitators to writing quality items for medical school assessments–a scoping review. *BMC Medical Education*, *19*, 1-11.

[13] Areekkuzhiyil, Santhosh. (2021). Issues and Concerns in Classroom Assessment Practices. Edutracks, 20(8), pp 20-23

[14] Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC medical education*, *16*, 1-10.

[15] Thangaraj, P. (2021). Concept of Formative assessment and Strategies for its effective implementation under Competency-Based Medical Education: A Review. *National Journal of Research in Community Medicine*, *10*(1), 016-024.

[16] Gierl, M. J., & Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied psychological measurement*, *42*(1), 42-57.

[17] Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. Educational Measurement: Issues and Practice, 32, 3650.

[18] Bulathwela, S., Muse, H., & Yilmaz, E. (2023, June). Scalable Educational Question Generation with Pre-trained Language Models. In *International Conference on Artificial Intelligence in Education* (pp. 327-339). Cham: Springer Nature Switzerland.

[19] Bormuth, J. (1969). *On a Theory of Achievement Test Items*. Chicago, IL: University of Chicago Press.

[20] Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). Automatic item generation: Theory and practice. Routledge.

[21] Kıyak, Y. S. (2023). A ChatGPT prompt for writing case-based multiple-choice questions. *Revista Española de Educación Médica*, *4*(3).

[22] Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13*(9), 856.

[23] Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

[24] OpenAI, C. (2023). Optimizing language models for dialogue, 2022. *URL: https://openai. com/blog/chatgpt*.

[25] Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European journal of clinical pharmacology*, 1-7.

[26] Kalyan, K. S. (2023). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 100048.

[27] Nunes, D., Primi, R., Pires, R., Lotufo, R., & Nogueira, R. (2023). Evaluating GPT-3.5 and GPT-4 models on Brazilian university admission exams. *arXiv preprint arXiv:2303.17003*.