The potential of AI: generating answers for multiple-choice questions using ChatGPT

Laura Kuusemets[1]

[1]University of Tartu, Faculty of Science and Technology, Tartu, Estonia; laura.kuusemets.1@ut.ee

Computer-based multiple-choice tests reduce teachers' workloads and enable students to receive immediate feedback on their performance, which is why they are widely used in the education system. To create a high-quality multiple-choice test, the author has to develop plausible distractors to be included among the provided answers. Therefore, creating a high-quality test is time-consuming. The purpose of this paper is to determine the suitability of ChatGPT for generating answer options for arithmetic and textual multiple-choice questions using Estonian and English prompts as examples, based on typical mistakes made by students and their similarities and differences with the answer options created by a human expert. This paper uses ChatGPT3.5 to test if it is possible to generate answers for multiple-choice questions that are based on given parameters. The results show that it is possible to generate answer options for multiple-choice questions using ChatGPT.

Keywords: Multiple-choice questions, artificial intelligence, mathematics education, distractors.

Introduction

The capacity of computers to execute cognitive tasks commonly associated with human intelligence, especially in learning and problem-solving, is defined as artificial intelligence (AI). AI-based systems are used to support teachers, reduce their workload, and automate assessment (Baker & Smith, 2019). In addition, AI-based assessment systems make the assessment process easier and faster for teachers (Kersting et al., 2014). Chat Generative Pre-Trained Transformer (ChatGPT) is an AI technology that generates conversational interactions based on user prompts (OpenAI et al., 2023). Large language models (LLM), such as ChatGPT have been pre-trained on huge volumes of textual data and are therefore able to answer questions with high accuracy, generate text and perform other language-related tasks (Kasneci et al., 2023). The ChatGPT model has proven its potential in various domains, including education (Liu et al., 2023).

Automated assessment and multiple-choice questions (MCQs) tools have been used for a long time. Because a trained language model can answer questions, it can be used to generate multiple-choice answers for tests. However, experience shows that the generation of a high-quality MCQ depends on the quality of the prompts (Kıyak, 2023). While it should not be overlooked, a study conducted by Liu et al., (2023) showed that ChatGPT has rather limited mathematical ability.

In the case of MCQs, the content of the question should be as concise, clear, precise, and unambiguous as possible, (Kelly, 1916); it should refer to the substance of the problem and the statement followed by the question to be answered. One answer option is always correct – it is called ’the key’ – and one or more answer options are always false, known as ‘decoy responses’ or ‘distractors’ (Gierl et al., 2017; McNichols et al., 2023). The answer options must not be partially true/false (Kelly, 1916). Distractors must be plausible and linked to the mistakes made by students and they should be misleading for students but not entirely false, which would make them easy to eliminate (Gierl et al., 2017; McNichols et al., 2023). If many distractors need be generated, this becomes burdensome for the test writer (Gierl et al., 2017).

Method

The main objective in using computer-based environments and MCQs, is to reduce teachers' workload through automated testing. Automated tests with multiple-choice answers are not widespread in the Estonian education practice, as preparing tasks and multiple-choice answers is time-consuming. Due to the aging teaching staff in Estonia, mathematics teachers are reluctant to use computer-based programs because they lack of digital competencies, and their command of English is relatively weak. If ChatGPT could generate answer options for MCQs in Estonian, the teacher would only have to enter the questions and answers for the test. As a result, teachers' workload would be significantly reduced, and teachers would have more time to support the students who need additional tutoring.

The purpose of this paper is to determine the suitability of ChatGPT for generating answer options for arithmetic and textual MCQs using Estonian and English prompts as examples, based on typical mistakes made by students and their similarities and differences with the answer options created by s human expert. In this study, both arithmetical and textual tasks are discussed, both arithmetic skills and knowledge of rules and theorems are checked in Estonian school mathematics. The topics and questions for the study have been taken from the textbook "Testid koolimatemaatikas I" (“Tests in school mathematics I”) by Lea Lepmann, a mathematics didactician at the University of Tartu. Following the textbook, we compare arithmetic and textual answer options generated by ChatGPT with those created by a human expert, a mathematics didactician, in terms of their accuracy and the quality of distractors and keys. When creating the prompt for ChatGPT, the recommendations for creating multiple-choice answers were taken into account, based on typical mistakes made by students. The ChatGPT prompt was entered in two languages – Estonian and English. To answer the research question, the answer options generated by ChatGPT and their correctness and the comparability of the false answers were analyzed and compared with the answer options created by the mathematics didactician. ChatGPT3.5 has been used to carry out the study, and three research questions have been formuled:

1) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in an arithmetic task compared to human expert-generated answer options?

In order to an answer to the research question, four different types of arithmetic tasks were chosen for one subtopic of Lepmann’s (1991) textbook – addition fractions with same denominators, subtraction of fraction with mixed numbers, addition and subtraction of fractions with different denominator.

2) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in a textual task compared to human expert-generated answer options?

In order to answer the research question, two theory-based textual tasks were chosen from Lepmann's (1991) textbook – natural and prime numbers. The aim was to check whether ChatGPT understood the prompt and wether the output was formulated using correct mathematical vocabulary.

3) What are the differences in answer options between the English and Estonian prompts in ChatGPT?

ChatGPT prompt can also be in Estonian and can be used for understanding and constructing textual tasks in the local language, Estonian. To evaluate if ChatGPT can provide didactically correct answer options in the Estonian language in arithmetic and textual tasks, research questions 1 and 2 must be analyzed.

Results

The results of each prompt are presented as a table, where each answer option is accompanied by its type: key (k), distractor (d), or incorrect option (i).

1) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in an arithmetic task compared to human expert-generated answer options?

Prompts 1.1-1.4: Prepare four answer options for the actions ________, where one answer option is correct, and three are incorrect. When preparing incorrect answer options, rely on typical mistakes made by students.

Prompt 1.1 – After receiving prompts in English and Estonian, ChatGPT prepared four answer options, one correct and three false (Table 1). There was a correct answer, in both instance and the generated distractor responses were plausible. The prompt given in Estonian, resulted in a non-reduced answer ($\frac{8}{34}$). For some of the answer options created by Lepmann, a partial calculation step has been added. There is also a non-reduced fraction ($\frac{8}{34}$).

Table 1. Four asnwers proposed by ChatGPT for action $\frac{5}{17}+\frac{3}{17}$

Input	**Output**		**Lepmann, 1991**
	In Estonian	**In English**	
$\frac{5}{17}+\frac{3}{17}$	$\frac{8}{17}$ (k) $\frac{15}{34}$ (d) $\frac{1}{10}$ (d) $\frac{8}{34}$ (d)	$\frac{8}{17}$ (k) $\frac{15}{17}$ (d) $\frac{2}{17}$ (d) $\frac{35}{17}$ (d)	$\frac{8}{17\cdot17}=\frac{8}{289}$ (d) $\frac{5\cdot17+3\cdot17}{17}=\frac{136}{17}$ (d) $\frac{8}{17}$ (d) $\frac{8}{34}$ (k)

Prompt 1.2 – After receiving a prompt in Estonian, ChatGPT generated four answer options, one correct and three false (Table 2), i.e., they match the given parameters. Two of the generated distractor answers were negative, and one option is in a non-reduced form. ChatGPT generated four answer options for the English prompt, all of which were false. One incorrect answer option was marked as

the key. The wrong answer options were all in a non-reduced form. If the same prompt was given a second time, ChatGPT created new answer options and indicated the correct answer, but it was in a non-reduced form ($1\frac{2}{6}$). One option was also in a non-reduced form ($2\frac{2}{6}$), two other answer options were not simplified correctly and were not reduced ($1\frac{8}{6}$ and $2\frac{8}{6}$). The answer options created by Lepmann show a partial calculation step. One of the answer options is in a non-reduced form ($2\frac{4}{6}$).

Table 2. Four answers proposed by ChatGPT for action $4\frac{1}{6} - 2\frac{5}{6}$

Input	Output			Lepmann, 1991
	In Estonian	In English		
$4\frac{1}{6} - 2\frac{5}{6}$	$1\frac{1}{3}$ (k)	$1\frac{2}{3}$ (k)	$1\frac{2}{6}$ (d)	$(4-2)+\left(\frac{1}{6}-\frac{5}{6}\right)=2-\frac{4}{6}=1\frac{1}{3}$ (k)
	$-1\frac{1}{3}$ (i)	$1\frac{4}{6}$ (d)	$2\frac{2}{6}$ (d)	$(4-2)+\left(\frac{5}{6}-\frac{1}{6}\right)=2+\frac{4}{6}=2\frac{2}{3}$ (d)
	$-\frac{4}{3}$ (i)	$1\frac{4}{36}$ (d)	$1\frac{8}{6}$ (d)	$(4-2)-\left(\frac{1}{6}-\frac{5}{6}\right)=2-\frac{1-5}{6}=2\frac{4}{6}$ (d)
	$3\frac{2}{3}$ (d)	$3\frac{3}{12}$ (d)	$2\frac{8}{6}$ (d)	$(4-2)-\left(\frac{1}{6}+\frac{5}{6}\right)=2-\frac{6}{6}=1$ (d)

Prompt 1.3 – In response to a prompt in Estonian, ChatGPT generated three answer options, one correct and three false, i.e., matching the given parameters (Table 3). If the prompt was given in English, ChatGPT did not produce any answer options that matched the parameters. A wrong answer was marked as the key, and there was no correct answer option. When the same prompt was given a second time, the results were partially different; again, a wrong answer was marked as the key, and a correct answer option was missing. In both cases, there was no correct answer among the outputs. If ChatGPT was prompted to solve an action, it returned the correct answer with solution steps. Only then, it generated answer options corresponding to the parameters, and the correct answer option was in a reduced form. Three of the four answer options created by Lepmann show a partial calculation step. Only one out of the four answer options was in a non-reduced form ($\frac{7}{21}$).

Table 3. Four answers proposed by ChatGPT for action $\frac{2}{3} + \frac{5}{7}$

Input	Output				Lepmann, 1991
	In Estonian	In English			
$\frac{2}{3} + \frac{5}{7}$	$\frac{29}{21}$ (k)	$\frac{31}{21}$ (d)	$\frac{31}{21}$ (k)	$1\frac{8}{21}$ (k)	$\frac{2\cdot7+3\cdot5}{3+7}=\frac{29}{10}$ (d)
	$\frac{7}{10}$ (d)	$\frac{1}{7}$ (d)	$\frac{7}{10}$ (d)	$\frac{7}{10}$ (d)	$\frac{2\cdot7+3\cdot5}{21}=\frac{29}{21}$ (k)
	$\frac{10}{21}$ (d)	$\frac{10}{21}$ (k)	$\frac{15}{21}$ (d)	$\frac{29}{21}$ (d)	$\frac{2+5}{21}=\frac{7}{21}$ (d)
	$\frac{34}{21}$ (d)	$\frac{7}{10}$ (d)	$\frac{2}{10}$ (d)	$\frac{5}{7}$ (d)	$\frac{7}{10}$ (d)

Prompt 1.4 – Having receiveda prompt in Estonian, ChatGPT suggested four answer options, one correct and three false (Table 4). The output matched the parameters. The distractors are all in a reduced form. When providing a prompt in English, the output was incorrect the first time. The output was similar to the output given in Estonian and to the answer options produced by Lepmann.

However, a wrong answer option was marked as correct and it was in a non-reduced form. In response to a prompt requiring ChatGPT to solve the given operation first and then generate the answer options, it produced the correct answer option in a reduced form. Two of the generated false answers were equal – one in a reduced and the other in a non-reduced form ($\frac{2}{6}$ and $\frac{1}{3}$). Another distractor was also in a non-reduce form ($\frac{3}{12}$). All of the answer options created by Lepmann include calculation steps.

Table 4. Four answers proposed by ChatGPT for action $\frac{2}{3}-\frac{1}{6}$

Input	**Output**			**Lepmann, 1991**
	In Estonian	**In English**		
$\frac{2}{3}-\frac{1}{6}$	$\frac{1}{2}$ (k)	$\frac{1}{6}$ (d)	$\frac{1}{2}$ (k)	$\frac{2-1}{6}=\frac{1}{6}$ (d)
	$\frac{1}{3}$ (d)	$\frac{1}{3}$ (k)	$\frac{2}{6}$ (d)	$\frac{4-1}{3}=\frac{3}{3}=1$ (d)
	$\frac{1}{6}$ (d)	$\frac{3}{6}$ (d)	$\frac{3}{12}$ (d)	$\frac{4-1}{6}=\frac{3}{6}=\frac{1}{2}$ (k)
	$\frac{3}{4}$ (d)	$\frac{2}{9}$ (d)	$\frac{1}{3}$ (d)	$\frac{12-3}{6}=\frac{9}{6}=\frac{3}{2}$ (d)

2) How appropriate are ChatGPT-generated answer options, based on students' typical mistakes, for multiple-choice questions in a textual task compared to human expert-generated answer options?

Prompts 2.1-2.4: Prepare four answer options for the assertion "______", where one is correct and three are incorrect. When preparing incorrect answer options, rely on typical mistakes made by students.

Prompt 2.1 – After receiving a prompt in Estonian, ChatGPT produced an output partially corresponding to the parameters (Table 5). The marked correct answer is partially true because the set of natural numbers is closed under addition and multiplication but not under subtraction and division. For the English prompt, ChatGPT gave an output that partially matches the parameters, indicating addition as the correct answer while noting subtraction and division as incorrect.

Table 5. Options proposed for the assertion "The set of natural numbers is closed …"

Input	**Output**		**Lepmann, 1991**
	In Estonian	**In English**	
"Naturaalarvude hulk on kinnine ..." *„The set of natural numbers is closed ..."*	• Liitmise ja lahutamise suhtes (*Under addition and subtraction); (k)* • Jagamise suhtes (*Under division); (d)* • Murdude suhtes (*Under fractions); (d)* • Korrutamise suhtes (*Under multiplication). (d)*	• Under addition; (k) • Under subtraction; (d) • Under multiplication; (d) • Under division. (d)	• Ainult liitmise suhtes (*Only under addition);* • Ainult liitmise, lahutamise ja korrutamise suhtes (*Only under addition, subtraction); (d)* • Ainult liitmise ja korrutamise suhtes (*Only under addition, multiplication); (k)* • Kõigi nelja aritmeetilise tehte suhtes (*Under all four arithmetical operation). (d)*

Prompt 2.2 – In response to a prompt in Estonian, ChatGPT generated an output corresponding to the parameters (Table 6). For English prompts, however, ChatGPT solved the task completely differently

by generating true/false statements. Each answer option explains why the statement is true/false and the reasoning is relevant.

Table 6. Options proposed for the assertion "Of the following numbers, 1, 2, 3, 4, 5, 6, 7, 8, 9 are prime numbers"

Input	**Output**		**Lepmann, 1991**
	In Estonian	**In English**	
"Järgmistest arvudest 1, 2, 3, 4, 5, 6, 7, 8, 9 on algarvud" *"Of the following numbers 1, 2, 3, 4, 5, 6, 7, 8, 9 are prime numbers"*	• 2, 3, 5 ja 7; (k) • 1, 2, 3, 5, ja 7; (d) • 2, 4, 6, ja 8; (d) • 2, 3, 4, ja 7. (d)	• False. 1, 4, 6, 8 and 9 are not prime numbers (k) • True. All the numbers are prime; (d) • False. Only 2, 3, 5, 7 are prime numbers; (d) • True. Any single-digit number is a prime number. (d)	• Ainult 1, 3, 5, 7 (*only 1, 3,5, 7); (d)* • Ainult 2, 3, 5, 7 (*only 2, 3, 5, 7); (k)* • Ainult 1, 3, 5, 7, 9 (*only 1, 3, 5, 7, 9); (d)* • Kõik nimetatud arvud (*All given numbers). (d)*

Discussion

The study results showed that when constructing multiple answers to given operations, the wording of the parameters in the prompt requires careful consideration. ChatGPT generated multiple-choice answers based on the parameters.

From the answer options generated by ChatGPT it is evident that they still require human verification. The generated multiple-choice answers included many mistakes, such as marking a wrong answer option as the correct one, especially with the English prompt. With the Estonian prompt, ChatGPT generated the correct answer option and three false options for all the given tasks on the first run. Most of the false responses were also plausible, which means that ChatGPT can generate suitable answer options in response to an Estonian prompt. Comparing the answer options generated by ChatGPT with the options generated by the mathematic didactician Lea Lepmann, there are some similarities for some tasks but no complete overlaps. For some tasks, no lure responses were the same for the Estonian prompt, the English prompt, and the response options of the human expert (Table 2). However, for some tasks, there were response options that occurred in all three (Table 3, 4). With an English prompt, ChatGPT marked an incorrect answer option as the key 75% of the time on the first run. Most of the false responses were also plausible, which means that in the case of an English prompt, the answer options generated by ChatGPT may not be suitable and need to be checked by teacher. ChatGPT provides suitable answer options if it first solves the action. An important difference is that the answer options generated by the human expert are structurally different – they include calculation steps that represent the mistake made in the calculation of fraction. None of the answer options generated by ChatGPT included calculation steps.

Care must be taken to ensure that the answer options and the task statement are consistent with each other and the curriculum. In the topic of common fractions, the problem statement should indicate the expected form of the answer (reduced or non-reduced). Estonian school mathematics requires the answer to be given in a reduced form. In this case, this should also be described in the stem. In Estonian school mathematics, fractions are taught in the 6th grade, while calculation with negative numbers is taught in the 7th grade. Therefore, two answer options can be eliminated from the Estonian output for Prompt 1.2. If an answer option contains an answer in a form that the students have not yet learned, they are likely to treat this option as incorrect. According to Girel et al., (2017) an answer option must not be completely wrong because in this case, students will be able to eliminate the answer immediately.

In Prompt 2.1, one can see that ChatGPT makes mathematical mistakes and seems that ChatGPT is aware of the mistaks but still marks the wrong answer as the key. The testing conducted in this study showed that ChatGPT can understand the content of the text and generate a textual response in both Estonian and English. The answers are plausible in Estonian but require linguistic editing. When comparing the answer options in Estonian and English and the one prepared by the didactician, they are similar, or only one of the answer options is different. The most significant difference was for Prompt 2.2, where ChatGPT produced true/false statements for the English prompt, which is undesirable for MCQs because it intentionally complicates the test.

Nevertheless, there were errors or omissions in compiling the multiple-choice answers. For several of the statements, ChatGPT gave a wrong or incomplete answer. For both English and Estonian prompts, there were errors in the mathematical terms or rules in the output. When generating answer options for arithmetical tasks, ChatGPT performed better with Estonian prompts than with English prompts. There were many errors in the outputs in response to English prompts, and the correct answer option often needed to be corrected. Thus, ChatGPT is good at generating multiple-choice answers based on Estonian prompts. Furthermore, it was found that ChatGPT can understand Estonian prompts and produce verbal responses. Based on the results, Estonian teachers can use ChatGPT to create multiple-choice answers, but it does not necessarily result in efficiency gain for teachers, as all the multiple-choice answers still have to be checked by the teacher.

Conclusion and future work

In this work, ChatGPT was used to see if it could make it easier and faster for teachers to create multiple-choice answers. The most important findings are that, first, the variability of the distractors generated by ChatGPT is smaller than the variability of the distractors generated by mathematical didactician. Second, the keys generated by ChatGPT are incorrect in some cases incorrect. In the case of Estonian prompts, ChatGPT failed to give a correct answer only in a textual task where it marked a partially false statement as the key. With English prompts for arithmetic tasks, ChatGPT marked the correct answer option as the key only 25% of the time in the first attempt. In the case of textual tasks, it indicated the wrong answer option as the correct one in some of the outputs. The third and most important result for Estonian mathematic teachers was that ChatGPT generates more evenly matched responses for Estonian prompts than for English prompts. In some cases, the distractors generated in Estonian are more plausible than those generated in English, because there were more answer-options that were non-reduced fractions.

The present work's limitations lie in using an older version of ChatGPT, 3.5, when the newer version 4.0 is more advanced, has newer data, and is better at processing textual information. In this work, version 3.5 is used because it is available free of charge to everyone, while 4.0 is a paid version. Even though ChatGPT 4.0 is not free of charge, would be advisable to use the newer version.

In future work, ChatGPT should be given more details on the desired learning outcomes for a given topic are on students' common misconceptions in that topic. It should then be investigated whether ChatGPT can generate more appropriate distractor responses based on the given parameters and, for each false response, generate feedback that supports the student based on the type of mistake.

Reference

Baker, T., & Smith, L. (2019). *Educ-AI-tion Rebooted? Exploring the future of artificial intelligence in schools and colleges*. https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, *87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kelly, F. J. (1916). The Kansas Silent Reading Tests. *Journal of Educational Psychology*, *7*(2), 63–80. https://doi.org/10.1037/h0073542

Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated Scoring of Teachers' Open-Ended Responses to Video Prompts: Bringing the Classroom-Video-Analysis Assessment to Scale. *Educational and Psychological Measurement*, *74*(6), 950–974. https://doi.org/10.1177/0013164414521634

Kıyak, Y. S. (2023). A ChatGPT Prompt for Writing Case-Based Multiple-Choice Questions. *Revista Española de Educación Médica*, *4*(3). https://doi.org/10.6018/edumed.587451

Lepmann, L. (1991). *Testid koolimatemaatikast I*. Eesti õppekirjanduse keskus.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

McNichols, H., Feng, W., Lee, J., Scarlatos, A., Smith, D., Woodhead, S., & Lan, A. (2023). *Exploring Automated Distractor and Feedback Generation for Math Multiple-choice Questions via In-context Learning* (arXiv:2308.03234). arXiv. http://arxiv.org/abs/2308.03234

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report*. https://doi.org/10.48550/ARXIV.2303.08774