




[www.ijemst.net](http://www.ijemst.net)

## Assessing AI-generated (GPT-4) Versus Human Created MCQs In Mathematics Education: A Comparative Inquiry into Vector Topics

**Laura Kuusemets**   
University of Tartu, Estonia

**Kristin Parve**   
Tallinn University, Estonia

**Kati Ain**   
University of Tartu, Estonia

**Tiina Kraav**   
University of Tartu, Estonia

### To cite this article:

Kuusemets, L., Parve, K., Ain, K., & Kraav, T. (2024). Assessing AI-generated (GPT-4) versus human created MCQs in mathematics education: A comparative inquiry into vector topics. *International Journal of Education in Mathematics, Science, and Technology (IJEMST)*, 12(6), 1538-1558. <https://doi.org/10.46328/ijemst.4440>

The International Journal of Education in Mathematics, Science, and Technology (IJEMST) is a peer-reviewed scholarly online journal. This article may be used for research, teaching, and private study purposes. Authors alone are responsible for the contents of their articles. The journal owns the copyright of the articles. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of the research material. All authors are requested to disclose any actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations regarding the submitted work.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

## Assessing AI-generated (GPT-4) Versus Human Created MCQs in Mathematics Education: A Comparative Inquiry into Vector Topics

Laura Kuusemets, Kristin Parve, Kati Ain, Tiina Kraav

---

### Article Info

#### Article History

Received:

01 May 2024

Accepted:

03 September 2024

---

#### Keywords

Multiple-choice question

Artificial intelligence

Mathematics education

Distractors

GPT-4

---

### Abstract

Using multiple-choice questions as learning and assessment tools is standard at all levels of education. However, when discussing the positive and negative aspects of their use, the time and complexity involved in producing plausible distractor options emerge as a disadvantage that offsets the time savings in relation to feedback. The article attempts to understand whether, with the AI conquests on the educational landscape, we can now remove this aspect from the list of drawbacks. This paper aims to determine the suitability of GPT-4 for generating questions and answer options for multiple-choice questions using prompts in Estonian on topics related to vectors and their similarities and differences compared to questions and answers created by a human expert. The results show that GPT-4 can generate multiple-choice questions and answer options based on given learning objectives, theory, and sample problems. However, the suggested correct answer option often requires correction and is not yet linguistically at such a level that teachers can use the questions without editing. Verifying the generated tasks still becomes labour-intensive for teachers. Nonetheless, it is more crucial that the AI tool accurately determines the correct answer than that some of the generated distractors are not plausible.

---

### Introduction

Artificial Intelligence (AI) has gained tremendous popularity in various fields in recent years and is increasingly making its presence felt in education. Artificial intelligence tools for education (AIE) have become commonplace in schools worldwide (Baker & Smith, 2019; Zhai et al., 2021). To date, these AI tools still require varying degrees of human involvement to monitor task progress or ensure feedback accuracy (Kaplan & Haenlein, 2019). However, the potential disappearance of the teaching profession is also under discussion, as history provides numerous examples of jobs being eliminated due to the automation of various activities (Lacity & Willcocks, 2017). This situation increasingly necessitates a reevaluation of the teacher's role in conjunction with the continuous development of AI (Fenwick, 2018). Chat Generative Pre-Trained Transformer (ChatGPT) is an AI technology that generates conversational interactions based on user commands (OpenAI et al., 2023). Large language models (LLMs), like ChatGPT, have been pre-trained on vast amounts of textual data and are thus

capable of answering questions with high accuracy, generating text, and performing other language-related tasks (Kasneci et al., 2023). In pre-trained LLMs, the quality of the answers generated depends directly on the instructions and prompts given by the user (Bsharat et al., 2024). Thanks to ongoing development and training, GPT-4 is capable of processing both image and text files to produce textual outputs.

GPT-4 may be inferior to humans in real-life scenarios but demonstrates human-level abilities in many professional and academic fields (OpenAI et al., 2023). It is essential to draft guidelines based on which an LLM can generate high-quality content. Prompts should be brief and devoid of irrelevant information that could be misleading; otherwise, the LLM will provide confusing or irrelevant answers (Bsharat et al., 2024). In the case of large-scale prompts, the language model may make reasoning errors, which may stem from inaccurate interpretation of the input (OpenAI et al., 2023). The tasks must provide appropriate context to aid the model's understanding of the background and scope of the undertaking. Including keywords, domain-specific terminology, or the descriptions of situations helps contextualise the model's responses. Language and structure should be employed to clearly indicate the nature of the task to the model. In the case of a complex task, it is reasonable to specify the desired form or to demonstrate the format and type of the required task (Bsharat et al., 2024). The ChatGPT model has shown its potential across various domains, including education (Liu et al., 2023), but it must be borne in mind that it does not address large-scale and complex problems in the same manner as a human would (OpenAI et al., 2023). In relation to ChatGPT, concerns have been raised about ethical considerations and potential negative impacts on assessment practices, scientific integrity, and students' higher-order thinking skills (Farrokhnia et al., 2023).

Different GPT models are used to construct exam questions. However, the construction of multiple-choice exam questions is more complex than the automatic item generation (AIG) task, as such questions must be relevant to the exam topic, contain a logical question, and provide answer choices, one of which is correct (Mead & Zhou, 2024). von Davier (2019) and Attali et al. (2022) have previously conducted studies on the construction of multiple-choice tests using GPT. von Davier (2019) conducted a study based on GPT-2, and Attali et al. (2022) focused on GPT-3. Both studies demonstrated that GPT can be used to generate selective exam content. However, the usability of the generated test items depends on several factors, such as the type of item being created and the domain. A study by Mead & Zhou (2024) showed that up to 89% of the generated test items were suitable, but most of the items still required some adaptation in terms of wording or answer options, such as multiple correct answers or no correct answer. It is likely that GPT models will be capable of creating high-level exams and tests in the future; however, for the tests to function effectively at present, they need to be revised and validated (Mead & Zhou, 2024).

Testing is a common method for schools to measure educational attainment (William, 2010), and this tradition is likely to continue for some time. Computer-based tests have both positive and negative aspects (Bartram & Hambleton, 2006). Although there are many different environments for test construction, producing high-quality tests can still be difficult and time-consuming, which results in the same questions being reused year after year (Roediger & Marsh, 2005). However, AI-based assessment systems can make the assessment process more accessible and less time-consuming for teachers (Kersting et al., 2014). The use of AI in education is crucial; it

enables students to learn anytime and anywhere, provides various ways of presenting materials, and facilitates differentiation according to students' abilities and levels (Alissa & Hamadneh, 2023). According to several sources, traditional assessment poses a significant time challenge for teachers and lecturers working with large groups of learners (Moreno & Pineda, 2020) and can lead to assessment bias. Computer-based tests also offer the advantage of assessment objectivity (Dong & Zhang, 2016).

Multiple-choice tests (MCQs) have been used for a very long time as both a control and a learning tool. The use of MCQs in schools and higher education institutions is a widespread trend (Alomran & Chai, 2018), significantly saving time spent on feedback (Anakwe, 2008) and reducing the workload for teachers and lecturers. Based on various studies, no differences have been observed in the results of paper-based and computer-based tests (Hosseini et al., 2014; Hüseyin Öz, 2018; Logan, 2015; OECD, 2010; Piaw, 2012; Retnawati, 2015). Thanks to the internet, large groups of learners can now be assessed simultaneously. MOOCs (Massive Open Online Courses) have been an excellent example of how the number of learners per teacher can be significantly increased compared to the past, and this approach has been successfully exploited (Cubric & Tomic, 2020; Dong & Zhang, 2016).

Computer-based intelligent tutoring systems are also useful in supporting independent learning, as computer-supported homework improves learning efficiency (Kehrer et al., 2013). Immediate feedback is one of the major advantages of MCQs (Anakwe, 2008; Dong & Zhang, 2016; Hüseyin Öz, 2018). Immediate information on whether an answer was right or wrong is also considered feedback. It is believed that learning without feedback may not be effective for learners (Laurillard, 1993). In the absence of prompt feedback, learners may not even realise that they are making mistakes when solving tasks (Kehrer et al., 2013). However, the presence of feedback makes it more likely that students will not repeat the same mistakes (Kehrer et al., 2013; Scheeler et al., 2018).

In an MCQ, at least one of the answer options, known as the “key“, must be correct. One or more response options, referred to as “lure responses” and known as distractors, must be false (McNichols et al., 2023; Gierl et al., 2017). Generally, MCQs have 3-5 answer options. For instance, Roediger & Marsh (2005) demonstrated that questions with fewer answer options resulted in a higher percentage of correct responses, while Rodriguez (2005) argues in his meta-analysis that three answer options are sufficient and that an excessive number of answer options do not serve the purpose, as they may no longer meet the conditions required of distractors. In the case of MCQs, a question should be constructed in such a way that its content is as short, clear, precise, and unambiguous as possible (Kelly, 1916).

Creating a large number of distractors for questions becomes time-consuming for the test designer (Gierl et al., 2017; Hahn et al., 2021). There are several reasons for this. Firstly, incorrect answer options must be plausible (Gierl et al., 2017) and target students' most common misconceptions and knowledge gaps (McNichols et al., 2023). Options should neither be partially true nor false (Kelly, 1916), nor obviously false, to avoid the possibility of elimination (Gierl et al., 2017; McNichols et al., 2023). Additionally, it should be taken into account that technical and complex wording should be avoided when creating response options (Gierl et al., 2017). Generally, it is believed that creating distractors is easier for practising teachers, as their work experience has made them

more aware of the typical mistakes made by students (Gierl et al., 2017).

The use of computer-based MCQs serves two main objectives: 1) to support the learning process and increase motivation, and 2) to reduce teachers' workload through automated testing. The use of automated tests with multiple-choice answers is not very widespread in Estonia, as preparing tasks and answer options is time-consuming. The discussion around artificial intelligence and its application in education is also increasing in Estonia. If AI could generate questions and answer options in Estonian, it would significantly assist teachers in their work. This would not only reduce the workload of teachers in terms of task design and assessment but also allow them to devote more time to supporting students who need additional help or focusing on their own professional development. Additionally, it would make setting tasks to consolidate subject knowledge easier.

This paper aims to determine the suitability of GPT-4 for generating questions and answer options for MCQs using prompts in Estonian on vector topics, and to compare these with questions and answers created by a human expert.

The research question that guided the conception of the paper is as follows:

How appropriate are ChatGPT4-generated questions for multiple-choice questions in a vector task compared to human-expert-generated questions? Specifically, do the created questions:

- a) base on all sample types given in the prompt;
- b) have a single correct answer;
- c) have high quality distractors;
- d) provide sufficient information in clear Estonian language;
- e) contain correct syntax used in the Estonian education system.

## **Method**

This research may help shed more light on how AI can be used in education. The primary aim is to alleviate teachers' burdens by employing computerised platforms and MCQs to automate testing. In Estonia's educational landscape, the adoption of automated tests with MCQs remains limited due to the laborious task of crafting questions and answer options. With many educators in Estonia being older and hesitant to embrace digital tools – especially in mathematics, where English proficiency is limited—there is a reluctance to utilise computer-based programmes. However, if ChatGPT could generate MCQ answer options in Estonian, teachers would only need to add the questions and answer options to the test; this would significantly simplify the process. It would lead to a substantial reduction in teachers' workload, affording them more time to provide additional support to students who require it.

The National Curriculum for Upper Secondary Schools in Estonia, which was accepted in 2011 but the updated version of which was approved on 23 February 2023 (*Haridus- ja teadusministeerium, s.a*), states that students can choose between narrow or broad mathematics. The students who choose narrow mathematics study eight courses of mathematics during the three years of upper secondary school (one course consists of 35 lessons, 45 minutes each), and the students who choose broad mathematics study 14 courses of mathematics. The fifth course

in broad mathematics is “Vector on Plane. Equation of a line”. According to the national curriculum, a student who has completed the course will be able to do the following in relation to vectors:

- 1) explain the concepts of vector, unit vector, zero vector, reciprocal vector, vector coordinates, angle between two vectors;
- 2) add and subtract vectors and multiply a vector by a number, both geometrically and in coordinate form;
- 3) find the length of a vector, the coordinates of the midpoint of a segment, the scalar correlation of two vectors and apply them to geometry problems;
- 4) use the notions of perpendicularity and collinearity of vectors to solve geometry problems.

*(Gümnaasiumi riiklik õppekava–Riigi Teataja, s.a.)*

Analyses of the national secondary school mathematics examinations in Estonia show that scores for vector problems are among the lowest. For example, according to a national review of the spring 2022 mathematics state examination, the vector problem in the broad mathematics exam was the worst of the twelve problems, with an average score of 43.1%, while the average for the entire exam was 55.5% (Arismaa, 2022). However, the topic of vectors is crucial for science-related subjects in higher education. Students who have taken a broad mathematics course and passed the broad mathematics exam are more likely to advance to the next level of education in disciplines related to physics, engineering, etc., where knowledge of vectors will largely determine a student's academic success. Therefore, the authors of this article believe that vector-related learning in the Estonian educational landscape needs support in the form of teaching materials. One of the initial encounters with vectors in physics involves velocity and force, which are fundamental concepts of Newtonian mechanics. A significant portion of the undergraduate physics curriculum revolves around vector quantities, necessitating students to possess a strong understanding of vector concepts from a mathematical perspective. By the end of their first year in physics, students are expected to understand the integration of vector fields. Several studies have indicated that many students commence university studies with an inadequate grasp of vector concepts (Knight, 1995; Liu & Kottegoda, 2019; Nguyen & Meltzer, 2003; Tairab et al., 2020).

This research was based on the 5<sup>th</sup> course of the Estonian upper secondary school national curriculum, “Vector on Plane. Equation of a Line,” from the 10<sup>th</sup>-grade mathematics textbook (Lepmann, 2011), incorporating theory, formulas, and examples of the topic of vectors alongside MCQs from the collection “Testid koolimatemaatikast VII Vektorid” (“Tests on School Mathematics VII Vectors”) (Lepmann, 1991). The topic of vectors was subdivided into five subtopics: coordinates of vectors, length of vectors, addition of vectors, subtraction of vectors, and scalar product of vectors. A total of 29 sample problems were presented across these subthemes.

For each sub-topic, a file was created to provide the necessary background information required to follow the instructions. This included the theory of designing MCQs, learning outcomes aligned with the current national curriculum, and the theoretical material on vectors based on the 10<sup>th</sup>-grade textbook (Lepmann, et al., 2011), which explained the basic concepts of the vector topic. Each file contained theoretical information, formulas, textbook examples, and a selection of MCQs from the collection “Testid koolimatemaatikast VII Vektorid” (“Tests on School Mathematics VII Vectors”) by Lea Lepmann, a mathematics didactician at the University of Tartu (Lepmann, 1991). It is important to note that the author of the collection also contributed to the textbook

used. For each of the five vector subtopics, GPT-4 was used to construct 10 questions with answer options, yielding a total of 50 questions with 200 answer options.

In this article, the ChatGPT model GPT-4 (OpenAI GPT API) is utilised, which permits the use of system prompts. However, a separate chat was employed in this study to generate the tasks for each sub-theme. At the beginning of each chat, the necessary background information (in the form of theoretical materials and sample tasks) was uploaded to the system in a PDF file. GPT-4 is more reliable and capable of handling much more nuanced instructions than GPT-3.5. Given that the reliable capabilities of GPT-4 are currently limited to text creation, the input files only contained tasks presented in coordinate form and did not include questions requiring the geometric representation of vectors. Additionally, GPT-4 is capable of processing information from files used as input.

The prompt specified that 10 MCQs had to be created for the given topic, with one correct and three incorrect answer options. The incorrect answer options were based on standard errors made by students. For the sub-topic of vector coordinates, it had also been previously specified that the stem had to include three types: given vector start and end point coordinates, given vector coordinates with start point coordinates, and given vector coordinates with end point coordinates. For tasks involving finding vector length, it was specified that the answer options could also include the square root, reflecting the authors' previous experience where GPT-4 did not do this automatically. In the example problems in the input file, the answer options were followed by a key (k) or a distractor (d). Although the prompts did not specify how the tasks should be formatted, it was assumed that GPT-4 would also mark the correct and incorrect answer options in the output based on a similar system. The study was carried out, and the stems and response options generated were analysed by four expert teachers: two of whom are practising teachers with at least five years of experience and who are conducting research, a didactic lecturer from a university involved in teacher training, and a lecturer who deals with university undergraduate students – students whose mathematical knowledge acquired in high school is extended from a higher perspective.

To answer the research questions, expert teachers evaluated the suitability of the tasks created by GPT-4, compared them with the sample tasks in the provided file, and analysed the created multiple-choice answers, assessing their correctness and appropriateness. For each answer option, the teachers attempted to identify the reasoning behind the error that led to each given answer. It is important to note that these are the subjective opinions of experts, which were not validated using GPT-4. The answer options and the mistakes identified were compared with the answers created by the didactic.

The two expert teachers evaluated the generated items for usability using a 4-point Likert scale, where 1 indicated “useless”; 2 signified “requires significant revision”; 3 denoted “needs minor edits”; and 4 represented “acceptable as is”.

## **Results**

This chapter gives an overview of the results obtained. The results are presented by vector sub-topics, referring to the generated stem and the answer options in each instance.

The usability ratings from the two evaluators demonstrated a correlation of 0.870 across all items. The high inter-rater reliability ( $r_{12} = 0.870$ ) can be largely attributed to the fact that the response variants contained only short stems and mathematical items, where only syntax and orthography errors (including punctuation) and substantive errors could arise. All stems were content-relevant, leading to minimal variation in the experts' ratings, with a maximum difference of one point per task. It is important to note that the designed questions did not require longer mathematical text in the answer variants, which might have resulted in greater discrepancies among the coders.

### **Coordinates of Vectors**

Based on the given prompt, GPT-4 created a total of 10 tasks, each with one correct and three incorrect answer options. GPT-4 utilised two out of the four predefined question types to generate these tasks. Tasks requiring additional knowledge on the given topic were left unused. In 100% of the cases, the key was correctly marked, i.e., the correct answer option was always present in the output and was marked accurately. According to the experts, the tasks were linguistically correct but syntactically typical of the English-language education system. The main error across all tasks was the difference in the notation of point and vector coordinates.

In Estonian school mathematics, the coordinates of a point and a vector are separated by a semicolon; however, in GPT-4's output, they are separated by a comma (see Figure 1). This discrepancy is particularly striking because, in the example exercises and theoretical materials, the vector coordinates adhere to the rules of Estonian school mathematics. The answer options did not indicate whether they were correct (k) or distractors (d). The correct answer was presented on a separate line; it was not formatted in the same manner as the example problems in the input file. It is important to note that this was the only instance in which GPT-4 failed to meet the formatting requirements—the later problems generated by GPT-4 conformed to the formatting rules of Estonian school mathematics. Another syntactic error, which occurred in six of the ten generated tasks, was related to GPT-4's failure to distinguish between a point and its coordinates.

Ülesanne 9:  
On antud vektori koordinaadid (-3,5) ja alguspunkti koordinaadid A(2,-1), siis lõpp-punkti koordinaadid on:  
1. (-1, 4)  
2. (-5, 4)  
3. (-1, 6)  
4. (1, 6)  
Vastus: (-1, 4)

Figure 1. Example of Generated Questions in Estonian

In the didactic questions, the start and end points of vectors are not always marked with a letter, but GPT-4 always marked the points with a letter. In didactic tasks, vectors are always denoted by a letter or endpoints, yet GPT-4 never marked the vector in these tasks. The tasks generated by the didactician also included stems involving position vectors, which GPT-4 did not generate. On the other hand, GPT-4 generated tasks requiring the identification of origin coordinates, but there was no such stem in the tasks produced by the didactician. It is



important to note that this was required by GPT-4's prompt.

Table 1. Question Types on the Topic "Coordinates of Vectors"

Question type	No. of question	Example
The coordinates of the start and end points are given. The coordinates of the vector must be found.	1, 2, 7, 8	If the starting point of the vector is $A(-2,1)$ and the ending point is $B(3,5)$ , then the coordinates of the vector AB are: <ul style="list-style-type: none"> <li>• (5, 4)</li> <li>• (1, 4)</li> <li>• (5, 6)</li> <li>• (7, 9)</li> </ul> Answer: (5, 4)
The coordinates of the vector and the starting point of the vector are given. The coordinates of the end point of the vector must be found.	3, 4, 9	If the coordinates of the vector are (2,5) and the coordinates of the starting point are $A(1,1)$ , then the coordinates of the end point are: <ul style="list-style-type: none"> <li>• (3, 6)</li> <li>• (1, 6)</li> <li>• (3, 4)</li> <li>• (2, 7)</li> </ul> Answer: (3, 6)
The coordinates of the vector and the endpoint of the vector are given. The coordinates of the starting point of the vector must be found.	5, 6, 10	The coordinates of the vector (6,3) and the coordinates of the end point $B(10,5)$ are given, then the coordinates of the starting point are: <ul style="list-style-type: none"> <li>• (16, 8)</li> <li>• (4, 2)</li> <li>• (1, 1)</li> <li>• (5, 4)</li> </ul> Answer: (4, 2)

All the stems and answers generated were substantially correct; there were no overlapping stems, and there were no repetitive answer options. In addition, three different types of stems were generated (see Table 1), with at least three tasks of each type, ensuring an even distribution of tasks. The table also provides an example for each question type. For one type of task, the distractors were based broadly on the same misconceptions; for example, the coordinates of points were added when finding the coordinates of the vector, or a sign error was made in the

calculation. The location of the correct answer varied, as did the location of similar errors. The generated answer options also included bait answers where the type of error was not obvious.

### **Length of Vectors**

Based on the given prompt, GPT-4 created 10 tasks similar to those created by the didactician. GPT-4 was given six sample tasks, two of which had a vector as one of the lures instead of a number. Out of the 10 tasks created, this decoy option was never used, although all proposed standard stems were represented. When analyzing the distractors, the expert teachers noted that in the case of six tasks, the distractor answers were unclear and therefore considered of low quality.

Linguistically, according to experts, all tasks created by GPT-4 were recognized as correct. As mentioned previously, the notation of the points and vectors was consistent with the requirements of Estonian school mathematics. The vectors were correctly marked, and two correct formats were used: 1) vector marking with an arrow and a lowercase letter; 2) vector marking with endpoints and an arrow (see Figure 2).

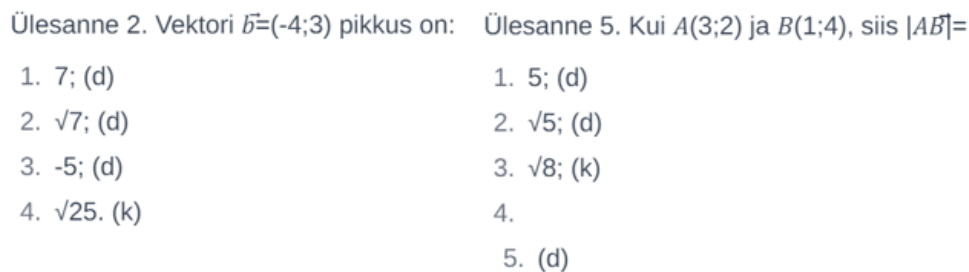


Figure 2. Vectors Marked on Two Correct Format

Furthermore, for this task, GPT-4 generated three different types of questions, all represented in the sample tasks. The distribution among the different types of stems was less even but similar to the distribution of the sample tasks in the input file (six tasks in total: three of the first type, two of the second type, and one of the third type, (see Table 2). There were no repetitive stems and no overlaps with the sample tasks provided by the didactician.

Out of the ten questions created, 30% had a wrong answer option marked as correct. Additionally, there were answer options using the square root, which made the answers more realistic. Compared to the vector coordinate generation tasks, GPT-4 had issues with the representation of output. In 50% of the cases, the fourth answer option was missing (blank), and the fifth answer option was missing a number, but indicated whether the given answer option was correct (k) or a distractor (d) (see Figure 3).

The prompts specified that one of the answers had to be correct and three had to be incorrect. If the output is copied to a text program, the fourth answer option appears blank, and underneath it, the fourth answer is presented without a sequence number (see Figure 3); however, this answer option is not displayed in the GPT-4 output itself. Each answer option was labelled as correct (k) or a distractor (d), as had been specified in the file of sample exercises.

Table 2. Question Types on the Topic “Length of Vectors”

Question type	No. of question	Example
The coordinates of the vector are given, the length of the vector must be found.	1, 2, 6, 7, 10	The length of the vector $\vec{a} = (-2; 4)$ is: <ul style="list-style-type: none"> <li>• 10; (d)</li> <li>• 2; (d)</li> <li>• <math>\sqrt{20}</math>; (k)</li> <li>• <math>\sqrt{15}</math>. (d)</li> </ul>
The coordinates of the starting and ending points of the vector are given. The length of the vector must be found.	3, 4, 5, 8	If $A(3; 2)$ and $B(1; 4)$ , then $ AB^{\rightarrow}  =$ <ul style="list-style-type: none"> <li>• 5; (d)</li> <li>• <math>\sqrt{5}</math>; (d)</li> <li>• <math>\sqrt{8}</math>; (k)</li> <li>• 8 (d)</li> </ul>
The coordinates of two points are given. The distance between the points must be found	9	* The distance between the points $A(-2; 4)$ and $B(6;-2)$ is: <ul style="list-style-type: none"> <li>• <math>\sqrt{36}</math>; (d)</li> <li>• 8; (d)</li> <li>• <math>\sqrt{80}</math>; (k)</li> <li>• 4 (d)</li> </ul>

There is no correct answer among the answer options.

Ülesanne 8. Kui $A(3;-1)$ ja $B(5;3)$ , siis $ AB^{\rightarrow} =$	Ülesanne 8. Kui $A(3;-1)$ ja $B(5;3)$ , siis $ AB^{\rightarrow} =$
1. $\sqrt{20}$ ; (k)	1. $\sqrt{20}$ ; (k)
2. 20; (d)	2. 20; (d)
3. $\sqrt{15}$ ; (d)	3. $\sqrt{15}$ ; (d)
4.	4.
5. (d)	15. (d)

Figure 3. GPT-4 Output with Faulty Fourth Answer Option

In one instance, GPT-4 provided the correct answer with a square root ( $\sqrt{25}$ ), but it could have given the exact answer ( $\sqrt{25} = 5$ ). This answer option can be interpreted in several ways: either GPT-4 has given an incomplete answer (although it is mathematically correct), or it can be viewed as an effective distractor to test the learners’ attention. In one case, GPT-4 incorrectly marked a wrong answer as correct, while the actual correct answer was presented as a distractor. In two instances, no correct answer option was included among the answers, and a wrong answer was marked as correct.

### Addition of Vectors

Based on the given prompt, GPT-4 created a total of 10 tasks similar to those created by the didactician. Of the two sample tasks provided, GPT-4 used one as a model. There were no repetitive questions or answer options. In the stems, two vectors are given, and the task is to find the sum of these vectors (see Table 3). The vectors were correctly labelled – in lowercase with an arrow. It was specified for each answer option whether it was correct (k) or a distractor (d), as recorded in the file of sample problems. An important observation was that in the GPT-4 generated stems the given vectors were labelled with the letters  $\vec{a}$  and  $\vec{b}$ , and all the tasks were  $\vec{a} + \vec{b}$ , but in the tasks generated by the didactician the labels of the vectors varied (also, for example,  $\vec{u}$  and  $\vec{v}$ ). In the tasks created by the teacher, the addition of a vector was also multiplied by a number ( $0.5\vec{u} + \vec{v}$ ). Although this case was represented in the sample tasks provided for GPT-4, such operations were not required in the input prompt.

11. Ülesanne: On antud vektorid $\vec{a}=(0;4)$ ja $\vec{b}=(-2;-1)$ , Leia $\vec{a}+\vec{b}$ .	3. Ülesanne: On antud vektorid $\vec{a}=(0;4)$ ja $\vec{b}=(-2;-1)$ , Leia $\vec{a}+\vec{b}$ .
12. (-2;5) (d)	1) (-2;5) (d)
13. (2;3) (k)	2) (2;3) (k)
14. (-2;3) (d)	3) (-2;3) (d)
15. (2;5) (d)	4) (2;5) (d)

Figure 4. Question Numbering in GPT-4 and in Text File

GPT-4 displays the numbering of the questions and answer options incorrectly, but everything is corrected when copied to a text file (see Figure 4). At first glance, the types of stems appear similar, but upon closer examination, three types of tasks emerged. All of the tasks featured both positive and negative vector coordinates, but there were no tasks where vector coordinates were exclusively positive or exclusively negative (see Table 3).

Table 3. Question Types on the Topic “Addition of Vectors”

Question type	No. of question	Example
The coordinates of two vectors are given: the coordinates of one vector are both positive, the coordinates of the other vector are both negative. The coordinates of the sum vector of vectors must be found	2, 3, 4, 5, 7, 9	Given vectors $\vec{a} = (-2; -3)$ and $\vec{b} = (5; 1)$ , Find $\vec{a} + \vec{b}$ . <ul style="list-style-type: none"> <li>• (7; -2) (d)</li> <li>• (3; -2) (k)</li> <li>• (-7; 2) (d)</li> <li>• (3; 2) (d)</li> </ul>
The coordinates of two vectors are given: the coordinates of one vector are both positive, the coordinates of the other vector have different signs. The coordinates of the sum vector of vectors	1, 6, 10	If $\vec{a} = (3; 2)$ and $\vec{b} = (1; -1)$ , then $\vec{a} + \vec{b} =$ <ul style="list-style-type: none"> <li>• (4; 3) (d)</li> <li>• (2; 1) (d)</li> <li>• (4; 1) (k)</li> <li>• (2; 3) (d)</li> </ul>

Question type	No. of question	Example
must be found		
The coordinates of two vectors are given: the coordinates of both vectors have different signs. The coordinates of the sum vector of vectors must be found	8	<p>Given vectors <math>\vec{a} = (-2; 3)</math> and <math>\vec{b} = (4; -1)</math>, Find <math>\vec{a} + \vec{b}</math>.</p> <ul style="list-style-type: none"> <li>• (6; 4) (d)</li> <li>• (6; -4) (d)</li> <li>• (2; 2) (k)</li> <li>• (-6; -2) (d)</li> </ul>

Out of the ten tasks created, 90% were correct, i.e., the correct answer option was marked as (k). In one question, one answer option was suggested twice, first as (d) and then as (k) (see Figure 5, answer options 1 and 4). In the creation of dummy answers, the main type of error was a notation error, i.e., incorrectly adding/subtracting a negative coordinate or adding vector coordinates instead of subtracting them. In terms of linguistic correctness, the expert teachers would have presented six out of the ten tasks with more accurate wording; in the same six tasks, an orthography error was also detected.

5. Ülesanne: On antud vektorid  $a^{\vec{}}=(-3;-1)$  ja  $b^{\vec{}}=(2;4)$ , Leia  $a^{\vec{}}+b^{\vec{}}$ .

1) (-1;3) (d)

2) (-1;5) (d)

3) (-5;3) (d)

4) (-1;3) (k)

Figure 5. Two Repeated Answer Options

### Subtraction of Vectors

Based on the given prompt, GPT-4 created a total of 10 tasks, similar to one task created by the didactician. However, a total of five sample tasks were presented. There were no repetitive questions or answer options. In the stems, two vectors are given, and the task is to find the difference between them. The vectors were correctly labelled – in lowercase with an arrow. It was indicated for each answer option whether it was correct (k) or a distractor (d), as was the case in the sample exercises in the input file. An important observation was that in the GPT-4 generated stems the given vectors were labelled with the letters  $\vec{a}$  and  $\vec{b}$ , and all the tasks were  $\vec{a} - \vec{b}$ , but in the tasks generated by the didactician the labels of the vectors varied (including, for example,  $\vec{u}$  and  $\vec{v}$ ) and the order of their subtraction was different. The didactician's questions also contained a multiplication of the vector by a number when subtracting (for example  $2\vec{a} - \vec{b}$ ), but although this was present in the sample tasks provided for GPT-4, such operations were not required in the input prompt. There are language errors. In 30% of the cases, GPT-4 uses “и” instead of “and” in the generated stems, which has the same meaning, but the "и" is in Russian (see Figure 6).

Ülesanne 2. Kui  $\vec{a}=(3;-1)$  и  $\vec{b}=(1;2)$ , siis  $\vec{a}-\vec{b}=?$

1. (2;3) (d)
2. (2;-3) (k)
3. (-2;-3) (d)
4. (-4;1) (d)

Figure 6. Language Error “и” in Task

GPT-4 created 10 problems, all with a similar task: given two vectors  $\vec{a}$  and  $\vec{b}$ , find the coordinates of their intermediate vector  $\vec{a} - \vec{b}$  (see Table 4). Since the main error in subtracting vectors is sign errors, it was decided to divide the problems into three, according to the signs of the coordinates of the reducer.

Table 4. Question Numbering in “Subtraction of Vectors”

Question type	No. of question	Example
The coordinates of the two vectors are given, the coordinates of the reducer are positive. The coordinates of the intermediate vector $\vec{a} - \vec{b}$ must be found.	1, 2, 5, 10	If $\vec{a} = (1; 2)$ and $\vec{b} = (2; 1)$ , then $\vec{a} - \vec{b} =$ <ul style="list-style-type: none"> <li>• <math>(-1; 1)</math> (k)</li> <li>• <math>(-4; 2)</math> (d)</li> <li>• <math>(-1; 2)</math> (d)</li> <li>• <math>(-2; 3)</math> (d)</li> </ul>
The coordinates of the two vectors are given, the coordinates of the reducer are negative. The coordinates of the intermediate vector $\vec{a} - \vec{b}$ must be found.	7, 8	If $\vec{a} = (-1; 2)$ and $\vec{b} = (-2; -1)$ , then $\vec{a} - \vec{b} =$ <ul style="list-style-type: none"> <li>• <math>(1; 3)</math> (k)</li> <li>• <math>(-1; -3)</math> (d)</li> <li>• <math>(1; -3)</math> (d)</li> <li>• <math>(-1; 3)</math> (d)</li> </ul>
The coordinates of the two vectors are given, the coordinates of the reducer have different signs. The coordinates of the intermediate vector $\vec{a} - \vec{b}$ must be found.	3, 4, 6, 9	If $\vec{a} = (-2; -1)$ and $\vec{b} = (1; -3)$ , then $\vec{a} - \vec{b} =$ <ul style="list-style-type: none"> <li>• <math>(-3; 2)</math> (k)</li> <li>• <math>(-3; -2)</math> (d)</li> <li>• <math>(1; -4)</math> (d)</li> <li>• <math>(3; -2)</math> (d)</li> </ul>

Out of the ten tasks created, 90% were correct, i.e., the correct answer option was marked as (k). In one of the tasks where GPT-4 marked the wrong answer option as correct, the correct answer option was marked as (d). The creation of the distractors relied on typical errors made by the students, for example, the coordinates of the vectors were wrongly subtracted ( $\vec{a} - \vec{b}$  instead of  $\vec{b} - \vec{a}$ ) or the coordinates of the vectors were added ( $\vec{a} + \vec{b}$ ). There were also answer options where one of the coordinates was correct but the other was found using  $\vec{b} - \vec{a}$  or  $\vec{a} + \vec{b}$ , or a notation error was made in the calculation. In the case of the individual latching responses, it was not possible to understand how the answer option had been obtained.

### Scalar Product on Vectors

On the basis of the given prompt, GPT-4 created a total of 10 tasks similar to those created by the didactician. There were no repetitive questions and no repetitive answer options. In the stems, two vectors are given and their dot product has to be found. The vectors were correctly labelled - lowercase with an arrow. For each answer option, it was indicated whether it was (k) or (d), as was the case in the sample exercises in the input file. An important observation was that in both the GPT-4 generated tasks and in the didactician's tasks, the given vectors were marked with the letters a and b, and all the tasks were  $\vec{a} \cdot \vec{b}$ . Of the ten tasks generated, 80% were correct, i.e., the correct answer option was marked as (k). In one task where GPT-4 marked the wrong answer option as correct, the correct answer option was marked as (d) (see Figure 7a) and in one task the correct answer option was missing (see Figure 7b).

Ülesanne 5. Kui  $\vec{a}=(1;-2)$  ja  $\vec{b}=(3;1)$ , siis  $\vec{a} \cdot \vec{b} =$

1. -1; (k)
2. 1; (d)
3. (3;-2); (d)
4.  $\sqrt{1}$ . (d)

Figure 7a. Correct Answer Option marked as a Distractor

Ülesanne 6. Kui  $\vec{a}=(-2;1)$  ja  $\vec{b}=(4;-3)$ , siis  $\vec{a} \cdot \vec{b} =$

1. 10; (k)
2. -10; (d)
3. (-8;-3); (d)
4.  $\sqrt{100}$ . (d)

Figure 7b. No Correct Answer Option

In the majority of cases (80%), the first answer option was marked with (k). In some of those instances, the actual correct answer option was marked as false or there was no correct answer option at all. The creation of bait answers was based on the types of errors that students had made. The answer options were similar to those generated by the teacher. The suggested answer options were integers, square roots, and coordinates (see Table 5).

Table 5. Question Types on the Topic “Scalar Product of Vectors”

Question type	No. of question	Example
The coordinates of two vectors are given: the coordinates of one vector are positive, the coordinates of the other vector have negative values. We need to find the scalar product of vectors.	1, 2, 5, 8, 10	If $\vec{a} = (3; -2)$ and $\vec{b} = (1; 5)$ , then $\vec{a} \cdot \vec{b} =$ <ul style="list-style-type: none"> <li>• 1; (d)</li> <li>• -7; (k)</li> <li>• (3; -10); (d)</li> <li>• <math>\sqrt{13}</math>. (d)</li> </ul>
The coordinates of two vectors are given: the coordinates of both vectors have different signs. We need to find the scalar product of vectors.	3, 4, 6, 7, 9	If $\vec{a} = (-4; 2)$ and $\vec{b} = (3; -5)$ , then $\vec{a} \cdot \vec{b} =$ <ul style="list-style-type: none"> <li>• -22; (k)</li> <li>• 22; (d)</li> <li>• (-12; -10); (d)</li> <li>• <math>\sqrt{164}</math>. (d)</li> </ul>

**Summary of Results**

A summary of the tasks generated by GPT-4 during the study is presented in Table 6. In total, GPT-4 generated 50 tasks, all of which are essentially correct and could be used either to consolidate the topic of vectors or to test knowledge. Of the 29 tasks provided as samples, GPT-4 used 15 (54%) to create tasks similar in content. Out of the 50 tasks, 42 (84%) included correct multiple-choice answer options, but only 30 (60%) of them had correct syntax and no linguistic errors, allowing them to be used in tests in an unedited form and without further modification.

Table 6. Summary of Generated Questions

Topic	Suitable tasks	Tasks with suitable answer options				
		Correct	There is no correct answer	The correct answer option is, but it is marked as a distractor	Correct language	Correct syntax
Coordinates of vectors	10	10	0	0	10	4
Length of the vectors	10	7	2	1	10	10
Addition of vectors	10	8	0	1+1*	4	10
Subtraction of vectors	10	9	0	1	7	10
Scalar product on vectors	10	8	1	1	10	10

\* In the task marked with an asterisk, the correct answer was given twice, one of which was counted as correct, the other as incorrect.

**Discussion**

The aim of the work was to use GPT-4 to create MCQs that could be used in secondary school mathematics, in 10<sup>th</sup> grade, to teach the topic of vectors. The use of MCQs can be a useful way to consolidate and test knowledge. Testing has been the main way to check the level of knowledge of learners. However, the use of AI tools could make this activity more convenient and less time-consuming for the teacher (Kersting et al., 2014). In this study,



a total of 50 MCQs were generated, each question with 4 possible answers where the correct answer had already been found. According to the expert teachers, it can be argued that generating the tasks was indeed faster than doing it manually. At the same time, it should be noted that other types of tasks, related to AI tools and GPT-4 in particular, can be time-consuming for the teacher: creating prompts and arranging theory material and sample exercises in such a way that the resulting materials meet the (formatting) requirements of a particular country's school mathematics. The same was experienced in the context of this work, where, in the case of the vector coordinate finding tasks, the GPT-generated tasks consistently failed to meet the formal requirements of Estonian school mathematics. Adherence to the formatting is also of crucial importance when creating such tasks. Mathematics is by its very nature a traditional subject with a long history, and rules and conventions vary from country to country. In the course of this investigation, GPT-4 erred in formatting, specifically in the notation of vector coordinates, separating the coordinates with a comma rather than a semicolon. It can be assumed that the AI was following the English formatting requirements here and not the input file provided. Although comma-separation may be common in some countries, it is confusing for learners in Estonia because the decimal place is also comma-separated. Thus, a simple difference in notation can lead to a situation that is unnecessarily confusing for the learner and/or requires additional effort from the teacher to change the notation.

In order to lighten the teacher's workload, it would be important that the AI tools used are easy to use and reliable. In the course of the work, it became clear that the creation of a prompt was critical. Although comprehensive input files were created for the study, the generation of the tasks revealed a number of shortcomings. Some of these were obviously related to the prompt that was generated (e.g., it was not specified that each distractor should have an explanation of the type of error on which the answer option was based), while there were also errors that the authors did not anticipate (e.g., various formatting issues that could have been clear from the input file). Thus, it seems that, in order to make creating MCQs with AI more efficient, teachers would need more precise guidance on what input information GPT-4 would need in the first place. In the context of this work, reliability was expressed primarily in terms of whether the tasks generated were age- and topic-appropriate, and whether the corresponding answer options were appropriate, i.e., whether there was a correctly marked correct answer and realistic decoy responses. Out of the 50 questions created in the study, GPT erred in marking a total of 8 tasks - either the correct answer was not present in the answer options (3 tasks); the correct answer option was present in the answers but was marked as a decoy answer (4 tasks); or the correct answer option was presented twice, being marked as correct once and the other time as a decoy answer (1 task). Thus, it can be said that GPT-4 was wrong 16% of the time. Although this is a small percentage, it is still too large an error to be used in class, as it is these small errors that are particularly labour-intensive to look for because they require a lot of attention. Thus, the authors of the paper believe that it is more important that the AI tools are able to correctly determine the correct answer, rather than be more convincing in creating baiting errors. While 84% of cases were satisfactory in terms of mathematical content and the provided answer options, experts estimated that due to linguistic and syntactical errors, the number of immediately usable MCQs decreased to 60%.

The examples of potential tasks provided to the AI were deliberately selected to avoid requiring assessment at higher levels of Bloom's taxonomy. Consequently, we did not need to classify the generated tasks into different cognitive levels. However, it is crucial for AI to be capable of considering cognitive complexity, which would

necessitate additional input refinement from the teacher. This study serves as a preliminary exploration into how ChatGPT-4 can generate educational items. It is anticipated that future iterations of ChatGPT-4 will be more adept at generating higher-level tasks with fewer errors, thereby better aligning with the upper levels of Bloom's taxonomy. Nevertheless, this study demonstrates that ChatGPT-4 still lacks the stringent rules necessary for generating accurate mathematical tasks. With appropriate example tasks, it could potentially produce mathematically and linguistically correct outputs. Additionally, it is important to note that, despite the provided examples, ChatGPT-4 struggled to generate sufficiently varied tasks. This lack of variety is often a challenge for educators, as diverse task phrasing is essential for fostering a deeper understanding of the subject matter.

It is clear that artificial intelligence today cannot take over the role of the teacher in creating MCQs. The results show that GPT-4, one of the most powerful tools of its kind, sticks very firmly to the boundaries given to it by the input file. Thus, the tasks are rather monotonic (in particular, for example, in the generation of vector sum and vector difference tasks), with no noticeable variability, for example, in the form of simple modifications (labeling vectors with different letters, presenting sum and difference in a different order for variation). Care is required with regard to the suggested answer options - although AI can easily generate a large number of tasks with multiple answer options, the teacher needs to be careful about the quality of the generated answers (it is not guaranteed that the correct answer is always included or marked as such) and about the variability (there is a pattern in the answer options, e.g., in scalar product tasks most correct answers are given as the first option). The authors do not intend to suggest that the use of AI tools for creating multiple-choice questions and enhancing learning is unjustified. However, it is crucial that students are not left to learn independently or to assess their knowledge using AI-generated tasks that have not been reviewed by a subject matter expert.

## **Conclusion**

In this work, GPT-4 was used to see if it would make creating MCQs easier and faster. It is obvious that the evolution of technology has an impact on the educational landscape, and its potential to offer a range of options to better organise learning for both the learner and the teacher needs to be skilfully exploited. The most important findings are that GPT-4 is able to generate high-quality questions based on the background information and examples given in the input file. The questions are on the same level as the tasks created by the didactician, Lea Lepmann. The second important finding is that GPT-4 was able to generate answer options, but in 16% of the cases the key was missing or was marked incorrectly. It can be seen that GPT-4 is an aid to the teacher, but the percentage of errors in generating MCQ answer options is still too high for it to be considered as a substitute for the teacher. In the prompt it was stated that the distractors should be based on the types of errors made by students, but for some of the distractors it was not clear how they were obtained, making it difficult to accurately assess the appropriateness of the responses. However, according to the authors, it is more important for the AI to be able to correctly determine the correct answer, as inadequate distractors would do less harm than presenting the wrong answer option as incorrect. A third important result is that GPT-4 is able to receive as input and process Estonian-language learning material, including mathematical formulas and examples. In doing so, it is important to make sure that sufficient information about all the requirements is provided in advance in the prompt.

In future work, artificial intelligence experiments should be carried out on the tasks and answer options generated,

given the typical errors made by the students, specifying in more detail how the generated tasks should be structured and formatted. The appropriateness of the generated tasks and the distractors should then be assessed. In addition, input files on typographical errors should be generated for all the mathematics topics related to the national curriculum. The limitation of the present work is that GPT-4 requires a paid subscription, but GPT-3.5 does not support the use of input files. In the research carried out, GPT-4 did provide an output that produced a brief summary of the input file, but it is not clear whether this was relevant and necessary for the tasks created.

## **Recommendations**

In the light of this study, the researchers wanted to shed light on the learning and teaching opportunities offered by artificial intelligence and test the ability of the GPT-4 language model to generate MCQs. The results showed that GPT-4 is capable of generating questions based on Estonian input. Attention needs to be paid to the correctness of the answer options generated – whether the answer option marked as the key is actually correct. Therefore, it can be said that the AI is not able to replace the teacher and, in the GPT-4 example, MCQ generation can become even more burdensome and cumbersome due to the need to create an input file and review the answer options. In order to prevent the student from memorising incorrect information after taking MCQ tests, a key element here could be the feedback from the answers, which is not limited to correct-or-false feedback, but also explains the place or reason for the error for each incorrect answer option. We believe that feedback is an important aspect to focus on in future studies.

## **References**

- Alissa, R. A. S., & Hamadneh, M. A. (2023). The Level of Science and Mathematics Teachers' Employment of Artificial Intelligence Applications in the Educational Process. *International Journal of Education in Mathematics, Science and Technology*, 11(6), 1597–1608. <https://doi.org/10.46328/ijemst.3806>
- Alomran, M., & Chai, D. (2018). Automated Scoring System for Multiple Choice Test with Quick Feedback. *International Journal of Information and Education Technology*, 8(8), 538–545. <https://doi.org/10.18178/ijiet.2018.8.8.1096>
- Anakwe, B. (2008). Comparison of Student Performance in Paper-Based Versus Computer-Based Testing. *Journal of Education for Business*, 84(1), 13–17. <https://doi.org/10.3200/JOEB.84.1.13-17>
- Arismaa, H. (2022). Matemaatika riigieksami analüüs 2022. Haridus- ja Noorteamet. <https://projektid.edu.ee/pages/viewpage.action?pageId=142575079>
- 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, Article 903077. <https://doi.org/10.3389/frai.2022.903077>
- 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](https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf)
- Bartram, D., & Hambleton, R. K. (2006). Computer-based testing and the internet: Issues and advances. John Wiley & Sons Ltd.
- Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2024). Principled Instructions Are All You Need for Questioning

- LLaMA-1/2, GPT-3.5/4 (arXiv:2312.16171). arXiv. <http://arxiv.org/abs/2312.16171>
- Cubric, M., & Tomic, M. (2020). Design and evaluation of an ontology-based tool for generating multiple-choice questions. *Interactive Technology and Smart Education*, 17(2), 109–131. <https://doi.org/10.1108/ITSE-05-2019-0023>
- Dong, F., & Zhang, Y. (2016). Automatic Features for Essay Scoring – An Empirical Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1072–1077. <https://doi.org/10.18653/v1/D16-1115>
- Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1–15. <https://doi.org/10.1080/14703297.2023.2195846>
- Fenwick, T. (2018). Pondering purposes, propelling forwards. *Studies in Continuing Education*, 40(3), 367–380.
- 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>
- Gümnaasiumi riiklik õppekava–Riigi Teataja. (s.a.). Viimati vaadatud 25. aprill 2024, <https://www.riigiteataja.ee/akt/123042021011>
- Hahn, M. G., Navarro, S. M. B., De La Fuente Valentin, L., & Burgos, D. (2021). A Systematic Review of the Effects of Automatic Scoring and Automatic Feedback in Educational Settings. *IEEE Access*, 9, 108190–108198. <https://doi.org/10.1109/ACCESS.2021.3100890>
- Haridus- ja teadusministeerium. (2023). Valitsus kiitis heaks ajakohastatud riiklikud õppekavad. <https://www.hm.ee/uudised/valitsus-kiitis-heaks-ajakohastatud-riiklikud-oppekavad>
- Hosseini, M., Abidin, M. J. Z., & Baghdarnia, M. (2014). Comparability of Test Results of Computer based Tests (CBT) and Paper and Pencil Tests (PPT) among English Language Learners in Iran. *Procedia - Social and Behavioral Sciences*, 98, 659–667. <https://doi.org/10.1016/j.sbspro.2014.03.465>
- Hüseyin Öz. (2018). Computer-based and Paper-based Testing: Does the Test Administration Mode Influence the Reliability and Validity of Achievement Tests?
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- 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>
- Kehrer, P., Kelly, K., & Heffernan, N. (2013). Does Immediate Feedback While Doing Homework Improve Learning? *Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference*, 542–545.
- 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>
- Knight, R. D. (1995). The vector knowledge of beginning physics students. *The Physics Teacher*, 33(2), 74–77. <https://doi.org/10.1119/1.2344143>
- Lacity, M., & Willcocks, L. P. (2017). *Robotic process automation and risk mitigation: The definitive guide*. SB Publishing.
- Laurillard, D. (1993). *Rethinking university teaching: A framework for the effective use of educational technology*. London: Routledge.
- Lepmann, L. (1991). *Testid koolimatemaatikast VII Vektorid*. Eesti õppekirjanduse keskus. Tallinn
- Lepmann, L., Lepmann, T., Velsker, K. (2011). *Matemaatika 10. Klassile*. Koolibri. Tallinn.
- Liu, D., & Kottegoda, Y. (2019). Disconnect between undergraduates' understanding of the algebraic and geometric aspects of vectors. *European Journal of Physics*, 40(3), 035702. <https://doi.org/10.1088/1361-6404/ab0509>
- 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>
- Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal*, 27(4), 423–441. <https://doi.org/10.1007/s13394-015-0143-1>
- 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>
- Mead, A. D., & Zhou, C. (2024). Evaluating the Quality of AI-Generated Items for a Certification Exam. *Journal of Applied Testing Technology*. Retrieved from <https://jattjournal.net/index.php/atp/article/view/173204>
- Moreno, J., & Pineda, A. F. (2020). A Framework for Automated Formative Assessment in Mathematics Courses. *IEEE Access*, 8, 30152–30159. <https://doi.org/10.1109/ACCESS.2020.2973026>
- Nguyen, N.-L., & Meltzer, D. E. (2003). Initial understanding of vector concepts among students in introductory physics courses. *American Journal of Physics*, 71(6), 630–638. <https://doi.org/10.1119/1.1571831>
- OECD. (2010). *PISA Computer-Based Assessment of Student Skills in Science*. OECD. <https://doi.org/10.1787/9789264082038-en>
- OpenAI. (2023). GPT-4. [https://openai.com/research/gpt-4?fbclid=IwZXh0bgNhZW0CMTEAAAR0XtZmm3eNcwprxPNSm4Sy0DWwxIu7XYQ848z90ILyGoxWV0HzMevWUo\\_aem\\_AfXWbE0vEUNpAzM-6J47eTNTlrM2qP5Wth\\_5U1Cnb6ugu8LcBlfSaOJCpM8oqazbDJxI8aFsN3OrEUvisLD62JWC](https://openai.com/research/gpt-4?fbclid=IwZXh0bgNhZW0CMTEAAAR0XtZmm3eNcwprxPNSm4Sy0DWwxIu7XYQ848z90ILyGoxWV0HzMevWUo_aem_AfXWbE0vEUNpAzM-6J47eTNTlrM2qP5Wth_5U1Cnb6ugu8LcBlfSaOJCpM8oqazbDJxI8aFsN3OrEUvisLD62JWC)
- Piaw, C. Y. (2012). Replacing Paper-based Testing with Computer-based Testing in Assessment: Are we Doing Wrong? *Procedia - Social and Behavioral Sciences*, 64, 655–664. <https://doi.org/10.1016/j.sbspro.2012.11.077>
- Retnawati, H. (2015). The Comparison of Accuracy Scores on the Paper and Pencil Testing vs. Computer- Based Testing. *The Turkish Online Journal of Educational Technology*, 14(4).
- Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years


- of Research. Educational Measurement: Issues and Practice, 24(2), 3–13. <https://doi.org/10.1111/j.1745-3992.2005.00006.x>
- Roediger, H. L., & Marsh, E. J. (2005). The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155–1159. <https://doi.org/10.1037/0278-7393.31.5.1155>
- Scheeler, M. C., Morano, S., & Lee, D. L. (2018). Effects of Immediate Feedback Using Bug-in-Ear With Paraeducators Working With Students With Autism. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, 41(1), 24–38. <https://doi.org/10.1177/0888406416666645>
- Tairab, H., Al Arabi, K., Rabbani, L., & Hamad, S. (2020). Examining Grade 11 science students' difficulties in learning about vector operations. *Physics Education*, 55(5), 055029. <https://doi.org/10.1088/1361-6552/aba107>
- von Davier, M. (2019). Training Optimus Prime, M.D.: Generating Medical Certification Items by Fine-Tuning OpenAI's gpt2 Transformer Model (arXiv:1908.08594). arXiv. <http://arxiv.org/abs/1908.08594>
- Wiliam, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist*, 45(2), 107–122. <https://doi.org/10.1080/00461521003703060>
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021, 1-18.

---

### Author Information

---

#### Laura Kuusemets

 <https://orcid.org/0009-0006-4714-4265>


University of Tartu

Ülikooli 18, 50090 Tartu

Estonia

Contact e-mail: [laura.kuusemets1@ut.ee](mailto:laura.kuusemets1@ut.ee)

#### Kristin Parve


 <https://orcid.org/0000-0002-8683-7017>

Tallinn University

Narva mnt 25, 10120 Tallinn

Estonia

#### Kati Ain


 <https://orcid.org/0009-0004-0680-2319>

University of Tartu

Ülikooli 18, 50090 Tartu

Estonia

#### Tiina Kraav

 <https://orcid.org/0000-0002-4893-8415>

University of Tartu

Ülikooli 18, 50090 Tartu

Estonia

---