


Using artificial intelligence to create biology multiple choice questions for higher education

Nanda Eska Anugrah Nasution ^{1*} 

¹ UIN Kiai Haji Achmad Siddiq Jember, Jawa Timur, INDONESIA

*Corresponding Author: nsteska@gmail.com

Citation: Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education*, 2(1), em002. <https://doi.org/10.29333/agrenvedu/13071>

ARTICLE INFO

Received: 11 Mar. 2023

Accepted: 11 Mar. 2023

ABSTRACT

This study aims to determine the validity, reliability, level of difficulty, and discrimination power of an artificial intelligence (AI)-generated collection of biology questions for higher education. Students' responses to AI-generated questions are also presented in this study. A sample of 272 students was selected using a random sampling technique to answer a series of multiple-choice questions and complete a questionnaire. Based on the research findings, 20 of the 21 questions generated by ChatGPT AI are valid. Cronbach's alpha coefficient was determined to be 0.65 (fairly reliable) for the twenty valid questions. Based on student responses to questions generated by ChatGPT's AI, it was determined that 79% of students indicated that the AI-generated questions were relevant to the class subject. 72% of students reported that the clarity of AI-generated questions was acceptable. 73% of students reported that the accuracy of AI-generated questions was good.

Keywords: ChatGPT, multiple choice questions, artificial intelligence, validity, reliability

INTRODUCTION

Algorithms driven by machine-learning technologies are now gaining maturity. ChatGPT is one such innovation. ChatGPT is an interactive chatbot created by OpenAI, a California-based artificial intelligence (AI) startup (Susnjak, 2022). OpenAI's ChatGPT is a comprehensive language model. ChatGPT AI was trained on a massive corpus of text data using a deep learning algorithm to create replies like those of a human for natural language questions (ChatGPT, 2023). ChatGPT AI bot is now accessible at <https://chat.openai.com/chat>.

AI natural language processing (NLP) technologies, such as ChatGPT AI, provide a means through which computers may engage with human language. A crucial stage in NLP, known as tokenization, is transforming unstructured information into organized text appropriate for computing (Hosseini et al., 2023). ChatGPT AI is interactive, able to comprehend what is being requested, and able to deliver it if it meets with application policies and data availability. For example, if you ask a search engine like Google to offer a list of questions connected to a particular topic, Google will send a link to a website that includes information relevant to the query you requested. When asking the same command to ChatGPT AI, the application will provide the question in that column.

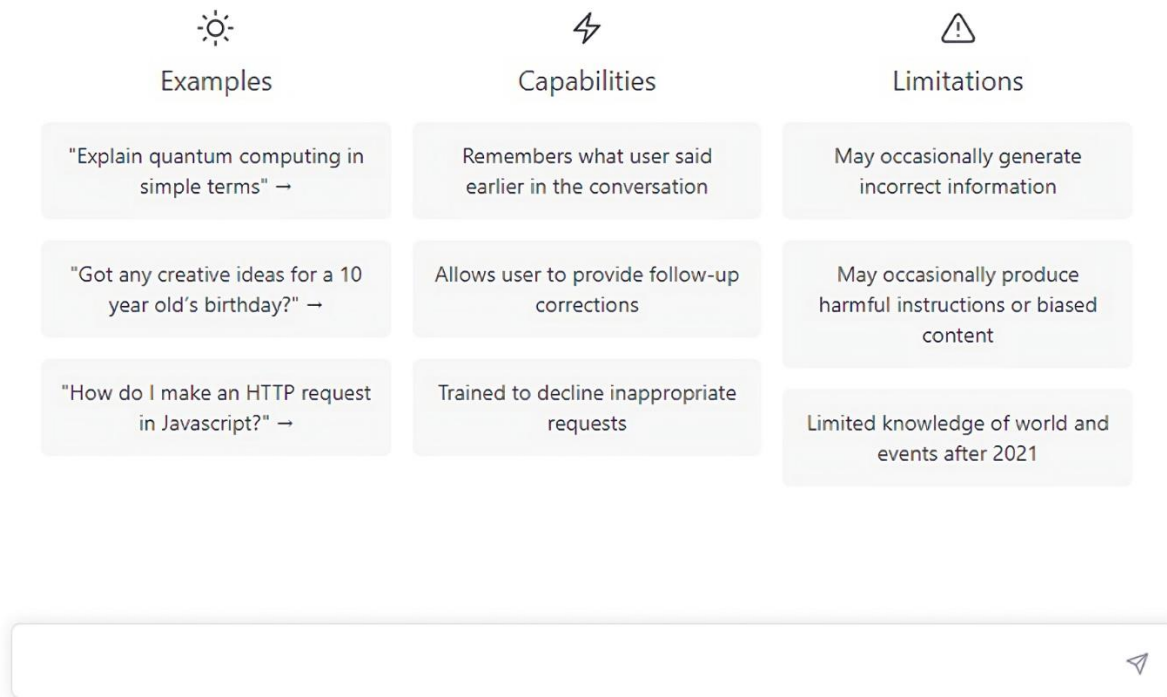
The emergence of ChatGPT AI is similar to the emergence of other new innovative technologies that, if used appropriately, have the potential to benefit education. Despite the fact that ChatGPT AI has the potential to be utilized for activities that are not acceptable in the academic sector. Students, for example, utilize ChatGPT AI to generate assignments such as essays. However, teachers may be able to use AI to spot AI-created works.

Teachers can use ChatGPT AI in a variety of ways, including asking information-related questions, confirming the accuracy of data, reviewing topics, etc. Teachers can request ChatGPT AI to generate multiple-choice questions for tests. Obviously, with its current version, ChatGPT AI has not been able to create an assessment instrument that can accurately measure a learning objective if it is not given explicit instructions by an expert or teacher. However, it is not impossible that in the future ChatGPT AI may be able to generate complex questions if it has access to a huge amount of data and has received extensive training.

A question arises regarding the form of questions that the current version of ChatGPT AI is capable of compiling. How valid and reliable are the question sets generated by ChatGPT AI? What is the difficulty level of the questions created by the ChatGPT AI? What do students think about the questions created by ChatGPT AI? Is it easy to read or understand? Is it relevant to the material being studied? Is it comparable to questions posed by humans?

Reliability and validity are, at a minimum, the two most important and essential aspects to consider when evaluating any measurement instrument or tool used (Mohajan, 2017). A measurement instrument is valid when it measures what it is intended

ChatGPT



ChatGPT Jan 30 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Figure 1. View of the publicly available ChatGPT AI bot landing page after login (ChatGPT, 2023)

to measure (Muijs, 2011). In other words, if an instrument measures a required variable accurately, it is termed a valid instrument for that variable (Ghazali, 2016). In comparison, reliability is defined as “the degree to which test scores are free of measurement error” (Muijs, 2011). It is a measurement of the stability or internal consistency of an instrument used to measure particular variable (Jackson, 2003). Multiple-choice questions are regarded to have a high level of reliability since they are scored objectively (Considine et al., 2005; Haladyna, 1999). Validity and reliability are related. It is possible for an instrument to be reliable but not valid; however, it cannot be valid if it is not reliable (Jackson, 2003). In other words, a valid instrument must also be reliable (Ghazali, 2016).

The quality of a multiple-choice questions test instrument can be determined by its validity and reliability, as well as its level of difficulty and discrimination power (Considine et al., 2005; Friatma & Anhar, 2019; Setiawaty et al., 2017; Rao et al., 2016; Salwa, 2012). The item’s difficulty corresponds to the proportion of correct responses (McCowan & McCowan, 1999). It is the frequency with which test-takers select the appropriate response (Thorndike et al., 1991). Items with a higher difficulty index are less difficult. A question that was answered correctly by 75% of test-takers has a difficulty level of 0.75. A question that was answered correctly by 35% of test-takers has a difficulty level of 0.35 (McCowan & McCowan, 1999). Item discrimination contrasts the proportion of high scorers and low scorers who correctly answer a given item. It refers to the degree to which items discriminate between students in the high and low groups. The whole test and each individual item should assess the same concept. High performers should be more likely to answer a good question properly, but poor performers should be more likely to do so wrong (McCowan & McCowan, 1999).

This study aims to determine the validity, reliability, level of difficulty, and discrimination power of an AI-generated collection of biology questions for higher education. Students’ responses to AI-generated questions are also presented in this study.

METHODS

This research is a descriptive quantitative analysis to explain the validity and reliability of ChatGPT AI’s questions. Before conducting the research, questions obtained from ChatGPT AI were compiled and administered to students. The steps of research are described in more detail below.

Accessing ChatGPT Artificial Intelligence

The researcher accessed the ChatGPT AI website in 2023, created an account, and logged into the application. Version 30 January 2023 of ChatGPT AI is in use (Figure 1).

Table 1. Descriptive statistics and Cronbach's alpha coefficient value

Biology subject	Number of questions	Question number
Change and growth	3	1, 2, & 3
Cell	3	4, 5, & 6
Biodiversity	3	7, 8, & 9
Genetics	3	10, 11, & 12
Evolution	3	13, 14, & 15
Ecology	3	16, 17, & 18
Biotechnology	3	19, 20, & 21

Table 2. Questionnaire of student responses to AI-generated questions

Criteria	Question
Relevance	Is the question relevant to the biology subject you studied at university and high school?
Clarity	Is text of question easy to understand?
	Is it well-structured & logically organized?
	Does it use appropriate language & vocabulary for intended audience?
Accuracy	Are the questions correct (no incorrect questions or no answer keys)?
Precision	Is the question sufficiently explicit and detailed?
Depth	Is the question deep enough (not too simple to work on)?

Creating questions

Researchers ask ChatGPT AI to create questions using the query "write me a multiple choice question with one correct answer option and four wrong answer options about <subject> for bachelor's degree, tag the correct answer". <subject> are seven basic biology studies discussed in high school and university biology subjects. The seven studies and the distribution of questions made by ChatGPT AI can be seen in **Table 1**.

In accordance with the request, ChatGPT AI successfully created 21 questions, each with five multiple-choice options, one of which was the correct answer and four of which were incorrect. ChatGPT AI also marks the correct answer for each question. The questions created by ChatGPT AI are written in English, which are then translated into Indonesian and evaluated by English and Indonesian lecturers with expertise in both languages. The researcher then compiled the 21 questions on the Google form and administered them to students in person. Students were presented with questions in both English and Indonesian. The test is administered under strict supervision and with closed books to ensure that students' responses are based solely on their own knowledge and not on the assistance of others or the internet/books. The exam is administered in 42 minutes, with two minutes allotted to each question.

Students' Responses to Artificial Intelligence-Generated Questions

We gathered student responses to AI-generated questions using the criteria developed by Susnjak (2022) in his research to assess AI responses. After completing the AI-generated questions, students work on this questionnaire. Students were told that the questions they had just worked on had been created by AI, and they were given 10 minutes to complete this questionnaire. Only students who are willing to complete this questionnaire will be eligible (not required for all students).

Table 2 displays the response questionnaire as well as the criteria.

Participants and Data collection

This study was carried out at the department of science education at a state university in East Java, Indonesia. A sample of 272 students were selected using random sampling technique from two study programs, namely biology education and natural science education. Not all pupils in class do the questions, only those who wish to. And only those students who choose to fill out a response questionnaire to students are asked to do so. 68% of the students that worked on the questions were those who worked on the response questionnaire, which only included 185 students.

Majority (38.97%, n=106) of the participants were aged 20 years. This was followed by 21 years (30.88%, n=84), 19 years (21.32%, n=58), 22 years (6.98%, n=19), 23 years (1.47%, n=4), and 24 years (.36%, n=1). Among the students, 231 (84.92%) were female and 41 (15%) were male. 133 participants (48.9%) are student of biology education program study, and 139 participants (51.1%) are student of natural science program study. The participants were picked from all levels in the undergraduate program. However, majority (50.37%, n=137) of them were third year students. The second year students were 47.42% (n=129), the fourth year students were 1.1% (n=3), the fifth year students were 0.7% (n=2), and the first year students were only 0.36% (n=1).

Statistical Analysis

All statistical analyses were calculated using IBM SPSS statistics 26 software. The validity of the questions was determined using Pearson product-moment correlation (Ahrens et al., 2020; Cho et al., 2006; Harahap et al., 2019; Mutmainah & Isdiati, 2022; Salwa, 2012). The reliability of the questions was determined using the Cronbach's alpha value (Ahrens et al., 2020; Cho et al., 2006; Harahap et al., 2019; Mutmainah & Isdiati, 2022; Salwa, 2012). The level of difficulty of the questions is determined by the following formula from McCowan and McCowan (1999):

Difficulty index (P)=# who answered an item correctly/total # tested.

Table 3. The results of the validity of all ChatGPT AI-generated questions

Question number	Pearson correlation	Sig. (2-tailed)	Description
1	.321**	.000	Valid
2	.211**	.000	Valid
3	.258**	.000	Valid
4	.313**	.000	Valid
5	.539**	.000	Valid
6	.401**	.000	Valid
7	.15**	.000	Valid
8	.365**	.000	Valid
9	.391**	.000	Valid
10	.502**	.000	Valid
11	.478**	.000	Valid
12	.421**	.000	Valid
13	.338**	.000	Valid
14	.318**	.000	Valid
15	.222**	.000	Valid
16	.467**	.000	Valid
17	.038	.533	Invalid
18	.429**	.000	Valid
19	.345**	.000	Valid
20	.39**	.000	Valid
21	.478**	.000	Valid

Table 4. The results of the reliability of all ChatGPT AI-generated questions (invalid question is not removed)

Cronbach's alpha	Number of items
.623	21

Table 5. The results of the reliability of all ChatGPT AI-generated questions (invalid question is removed)

Cronbach's alpha	Number of items
.655	20

The difficulty level of the questions is described, as follows: Difficult if below 0.3, medium if between 0.3 and 0.7, and easy if above 0.7. Using the following formula from Salwa (2012), the discrimination power level of the questions is determined:

Discrimination index (D)=# of top test takers who answered an item correctly/total # of top test takers tested (27% of all students)-# of bottom test takers who answered an item correctly/total # of bottom test takers tested (27% of all students).

The discrimination power of the questions is defined, as follows: poor if below 0.2, adequate if between 0.2 and 0.4, good if between 0.4 and 0.7, and excellent if over 0.7. If the result is negative, the discrimination power level of the item is inadequate, and the item must be eliminated. Students' responses to AI-generated questions were analyzed descriptively.

RESULTS

ChatGPT Artificial Intelligence-Generated Questions

ChatGPT AI successfully generated 21 questions. **Appendix A** shows a list of all questions.

Validity

The results of the validity test of all ChatGPT AI-generated questions can be seen in **Table 3**.

Reliability

The results of the reliability test for all ChatGPT AI-generated questions may be viewed in **Table 4** if the invalid question is not removed (question no. 17), and in **Table 5** if the invalid question is removed.

Level of Difficulty and Discrimination Power

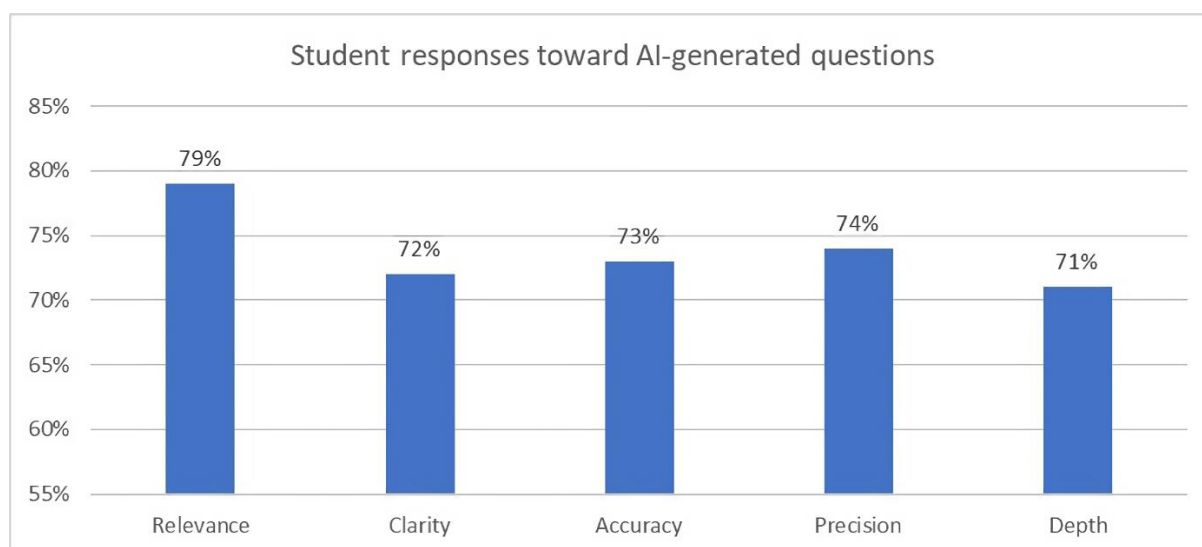
The results of the level of difficulty and discrimination power of all ChatGPT AI-generated questions can be seen in **Table 6**.

Student responses to Artificial Intelligence-Generated Questions

The percentage of student responses to questions generated by ChatGPT AI can be seen in **Figure 2**.

Table 6. The results of the level of difficulty and discrimination power of all ChatGPT AI-generated questions

Question number	Level of difficulty		Discrimination power	
	Scale	Description	Scale	Description
1	0.87	Easy	0.19	Poor
2	0.15	Difficult	0.21	Adequate
3	0.36	Medium	0.32	Adequate
4	0.74	Easy	0.4	Good
5	0.39	Medium	0.7	Good
6	0.88	Easy	0.29	Adequate
7	0.17	Difficult	0.08	Poor
8	0.71	Easy	0.44	Good
9	0.82	Easy	0.36	Adequate
10	0.85	Easy	0.45	Good
11	0.57	Medium	0.63	Good
12	0.56	Medium	0.55	Good
13	0.70	Medium	0.34	Adequate
14	0.82	Medium	0.29	Adequate
15	0.39	Medium	0.19	Poor
16	0.92	Easy	0.21	Adequate
17	0.33	Medium	0	Poor
18	0.38	Medium	0.53	Good
19	0.88	Easy	0.23	Adequate
20	0.41	Medium	0.51	Good
21	0.93	Easy	0.22	Adequate

**Figure 2.** The percentage of student responses to questions generated by ChatGPT AI (Source: Author's own elaboration)

DISCUSSION

According to Kimberlin and Winterstein (2008), validity is generally described as the degree to which an instrument measures what it claims to measure. It is necessary for an instrument to be valid so that it may be used to measure its intended subject. Using Pearson product moment correlation method to assess the validity of the questions, it was determined that 20 out of 21 items were valid, while one item was invalid. The invalid question is number 17, which is related to ecology. The results of the validity test indicate that 20 of the 21 questions generated by AI are valid and may be used.

The question number 17, which is invalid, asks students to choose the term used to describe the way by which organisms obtain energy from their environment. Option D (photosynthesis) is the correct answer, selected by 90 students (or 33%). Option A (metabolism), which 113 students (41.54%) selected, option B (ecosystem), which 40 students (14.7%) selected, option C (biodiversity), which 21 students (7.7%) selected, and option E (biogeography), which eight students (2.9%) selected, are all incorrect answer choices. Based on student choices, it was determined that a higher number of students selected option A (incorrect answer option) than the right answer. It is possible to derive that, first, answer choice A is an excellent diversion, or, second, there is a problem with question number 17. According to follow-up interviews with three random students who claimed to have chosen answer option A, they were confused by the question sentences. If the question is 'how do plants get energy from their environment' or 'how do organisms obtain energy from nature', it is likely that the student would choose option D (the correct one). Yet, given the wording in question 17 is how organisms get energy from their environment, students have misinterpreted the organisms at issue as animals and plants, and the environment referred to are other living species (by way of prey).

In addition, language and sentence issues that may be present in multiple-choice questions created by ChatGPT AI, such as question number 17 in this research, may be corrected by experts using content and face validity, as suggested Considine et al. (2005) and Harahap and Nasution (2022). Nevertheless, we did not do so in our research because we wanted to ensure that the questions generated by ChatGPT AI were free from any human adjustment.

Cronbach's alpha was used to assess the scale's internal consistency. Cronbach's alpha coefficient was determined to be 0.623% if item 17 (invalid item) was not removed, and 0.655% if item 17 (invalid item) was removed. The acceptable values for Cronbach's alpha vary according to the source. According to van Griethuijsen et al. (2014), the acceptable values of Cronbach's alpha are 0.7 or 0.6. Arulogun et al. (2020), George and Mallery (2003), Morgan et al. (2004), Rii et al. (2020), Taber (2018), and Wongpakaran and Wongpakaran (2012) emphasized the same point, that a Cronbach's alpha above 0.6 can be recognized as a reliable instrument. If this value is adhered to, then the multiple-choice questions generated by ChatGPT AI in this study may be deemed reliable. Several other resources, however, say that the allowable values for Cronbach's alpha are 0.8 or even 0.9; if this figure is used, the multiple-choice questions created by ChatGPT AI in this study may be regarded unreliable.

By evaluating the level of difficulty of the questions, it was determined that, of the 21 questions created by ChatGPT AI, nine were classified as easy, 10 were categorized as medium, and two were classified as difficult. It is preferable to use a proportionate distribution of easy, medium, and difficult multiple-choice questions. In this context, proportionate means that there should be at least twice as many questions at the medium level as at the easy and difficult levels, with an equal number of questions at the easy and difficult levels. ChatGPT AI developed multiple choice questions having nearly identical easy and medium levels, and only two items (9.5%) are classified as difficult. It is preferred if questions with easy and challenging difficulty levels be revised to be more proportional, or to become questions with a medium difficulty level. Rao et al. (2016) stated that ideally multiple choice questions have a medium level of difficulty. Of course, this should be revised depending on the aim of the assessment.

By assessing the discriminating power of the questions, it was determined that, of the 21 questions created by ChatGPT AI, 4 had low discrimination power, nine had adequate discrimination power, and the remaining eight had good discrimination power. Questions with low discrimination power should be modified to have discrimination power that is adequate or greater. There are no items with negative discrimination power, suggesting that there are no questions that should be deleted based on the discriminating power analysis. One of the items, however, has a discrimination value of zero, indicating that this item has very poor discriminatory power, since the number of students who answered this item correctly in the upper group and the lower group are identical. This question turned out to be number 17, which was classified as invalid based on the validity test, thus it is not unexpected that this question has very poor discriminating power. Moreover the difficulty index and discrimination index are reciprocally related (Chauhan et al., 2013; Mehta & Mokhasi, 2014; Rao et al., 2016; Suruchi & Rana, 2014). For instance, if a question is determined to have a low level of difficulty and poor discriminating power, the question should be revised (Rao et al., 2016).

Based on student responses to questions produced by ChatGPT's AI, it was determined that 79% of students indicated the AI-generated questions were relevant to the departmental subject they study. This finding suggests that ChatGPT AI is capable of generating questions pertaining to the specified subject, in this case biology in natural science, including change and growth, cell, biodiversity, genetics, evolution, ecology, and biotechnology. 72% of students reported that the questions generated by AI were clear. This suggests that the majority of students are capable of comprehending the questions posed by ChatGPT AI. The clarity of questions is determined by three survey items. The first item on the questionnaire asks whether the questions generated by ChatGPT AI are simple to comprehend. 66% of students indicated that the questions were straightforward. The second question asks if the questions generated by ChatGPT AI are logically structured and ordered. According to 76% of students, the questions were well-structured and logically ordered. The last question asks if questions generated by ChatGPT AI employ the proper language. 73% of students feel the question language is suitable. The questions in an assessment must be clear and concise. Difficult-to-understand questions will surely make it harder for students to answer, and there is a chance that students will respond erroneously not because of their incompetence but because of an error in the question.

73% of students stated that the AI-generated questions were accurate. This means that the majority of students consider the questions created by AI to be accurate; they see no grammatical or conceptual errors in the questions. However, you can't depend just on students' opinions to confirm the accuracy of a question. Several experts should be consulted to validate the question. Nevertheless, as stated previously, the questions in this study were not evaluated by professionals in order to determine how the questions were generated by AI.

74% of students indicated that the questions generated by AI were precise. This suggests that the majority of students consider AI-generated questions to be explicit and detailed. Students comprehend the intent of the questions and the required responses. If questions are not made clear and explicit, it is possible that students may have difficulties answering.

71% of students indicated that the questions posed by AI were of sufficient depth. The majority of students found that the questions generated by ChatGPT AI were challenging, not overly simple, and appropriate for their college or university level. As was done in this study, measuring the difficulty level of the questions is another method for determining if the questions are too easy or too difficult. Just two of the twenty-one questions generated by AI are difficult, while nine are quite easy.

The majority of students responded positively to the questions generated by ChatGPT's AI, according to the results of the student response questionnaire. Therefore, the teacher can use AI to assist him construct an assessment tool, but this must be complemented by the teacher's capacity to provide AI with clear instructions and to verify and optimize the resulting assessment tool as needed. Further study is required to determine if students can differentiate between questions developed by AI and those created by humans, as well as their perspectives on the conditions for AI-created questions.

Given that constructing multiple-choice questions is a complex and time-consuming process (Rao et al., 2016), it would be highly beneficial if AI could aid teachers or the education sector in the future in developing standardized and high-quality multiple-

choice questions. Nevertheless, the present version of ChatGPT AI has several limitations, as mentioned by OpenAI on its website (ChatGPT, 2023), such as the possibility of producing wrong information, harmful instructions, or biased material, and limited awareness of the world and events after 2021. Quite likely, ChatGPT AI will acquire more data and better training over time, allowing it to assist its users more effectively.

CONCLUSION

Based on the research findings, twenty of the twenty-one questions generated by ChatGPT AI are valid. Ecology-related questions are the only question that is invalid. Cronbach's alpha coefficient was determined to be 0.65 for the twenty valid questions. By assessing the level of difficulty of the questions, it was determined that, of the 21 questions created by ChatGPT AI, nine were rated as easy, 10 were classified as medium, and two were classified as difficult. By assessing the discriminating power of the questions, it was determined that, of the 21 questions created by ChatGPT AI, four had low discrimination power, nine had adequate discrimination power, and the remaining eight had good discrimination power. Based on student responses to questions generated by ChatGPT's AI, it was determined that 79% of students indicated that the AI-generated questions were relevant to the class subject. 72% of students reported that the clarity of AI-generated questions was acceptable. 73% of students reported that the accuracy of AI-generated questions was good. According to 74% of pupils, the accuracy of AI-generated questions was good. 71% of students reported that the depth of the questions generated by AI was acceptable.

Funding: No funding source is reported for this study.

Ethical statement: Author stated that all participants were over the age of 18 and that their participation was entirely voluntary. The author also stated that since no personal data was analyzed and pseudonyms were used in this article, no ethics committee approval was required.

Declaration of interest: No conflict of interest is declared by the author.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the author.

REFERENCES

- Arulogun, O. T., Akande, O. N. Akindele, A. T., & Badmus, T. A. (2020). Survey dataset on open and distance learning students' intention to use social media and emerging technologies for online facilitation. *Data in Brief*, 31, 105929. <https://doi.org/10.1016/j.dib.2020.105929>
- ChatGPT. (2023). *ChatGPT*. <https://chat.openai.com/chat>
- Chauhan, P. R., Ratrhod, S. P., Chauhan, B. R., Chauhan, G. R., Adhvaryu, A., & Chauhan, A. P. (2013). Study of difficulty level and discriminating index of stem type multiple choice questions of anatomy in Rajkot. *Biomirror*, 4(6), 1-4.
- Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4), 891-901. <https://doi.org/10.1037/0022-0663.98.4.891>
- Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12(1), 19-24. [https://doi.org/10.1016/S1322-7696\(08\)60478-3](https://doi.org/10.1016/S1322-7696(08)60478-3)
- de Barros Ahrens, R., da Silva Lirani, L., & de Francisco, A. C. (2020). Construct validity and reliability of the work environment assessment instrument WE-10. *International Journal of Environmental Research and Public Health*, 17(20), 7364. <https://doi.org/10.3390/ijerph17207364>
- Friatma, A., & Anhar, A. (2019). Analysis of validity, reliability, discrimination, difficulty and distraction effectiveness in learning assessment. *Journal of Physics: Conference Series*, 1387, 012063. <https://doi.org/10.1088/1742-6596/1387/1/012063>
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference 11.0 update*. Allyn & Bacon.
- Ghazali, N. H. M. (2016). A reliability and validity of an instrument to evaluate the school-based assessment system: A pilot study. *International Journal of Evaluation and Research in Education*, 5(2), 148-157. <http://doi.org/10.11591/ijere.v5i2.4533>
- Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Lawrence Erlbaum.
- Harahap, F., Nasution, N. E. A., & Manurung, B. (2019). The effect of blended learning on student's learning achievement and science process skills in plant tissue culture course. *International Journal of Instruction*, 12(1), 521-538. <https://doi.org/10.29333/iji.2019.12134a>
- Harahap, M. P., & Nasution, N. E. A. (2022). Validity of computer based learning media to improve junior high school students' learning outcomes on ecosystem topics. *META: Journal of Science and Technological Education*, 1(1), 31-45.
- Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications. *Accountability in Research*, 1-9. <https://doi.org/10.1080/08989621.2023.2168535>
- Jackson, S. L. (2003). *Research methods and statistics: A critical thinking approach*. Thomson Wadsworth.
- Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 65(23), 2276-2284. <https://doi.org/10.2146/ajhp070364>
- McCowan, R. J., & McCowan, S. C. (1999). Item analysis for criterion-Referenced tests. *Center for Development of Human Services*. <https://files.eric.ed.gov/fulltext/ED501716.pdf>

- Mehta, G., & Mokhasi, V. (2014). Item analysis of multiple choice questions—An assessment of the assessment tool. *International Journal of Health Sciences and Research*, 4(7), 197-202. <https://doi.org/10.1016/j.mjafi.2020.11.007>
- Mohajan, A. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University*, 17(3), 58-82. <https://doi.org/10.26458/1746>
- Morgan, P. J., Cleave-Hogg, D., DeSousa, S., & Tarshis, J. (2004). High-fidelity patient simulation: Validation of performance checklists. *BJA: British Journal of Anaesthesia*, 92(3), 388-392. <https://doi.org/10.1093/bja/ae081>
- Muijs, D. (2011). *Doing quantitative research in education with SPSS*. SAGE. <https://doi.org/10.4135/9781849203241>
- Mutmainah, I., & Isdiati, A. (2022). Validity and reliability test of a written English test online-based using Google Form. *INTERACTION: Jurnal Pendidikan Bahasa [INTERACTION: Journal of Language Education]*, 9(1), 89-100.
- Rao, C., Kishan, P. H. L., Sajitha, K., Permi, H., & Shetty, J. (2016). Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *International Journal of Education and Psychological Research*, 2, 201-214. <https://doi.org/10.4103/2395-2296.189670>
- Rii, K. B., Choi, L. K., Shino, Y., Kenta, H., & Adianita, I. R. (2020). Application of iLearning Education in Learning Methods for Entrepreneurship and Elementary School Student Innovation. *Aptisi Transactions on Technopreneurship*, 2(2), 131-142. <https://doi.org/10.34306/att.v2i2.90>
- Salwa, A. (2012). *The validity, reliability, level of difficulty and appropriateness of curriculum of the English test* [PhD thesis, Diponegoro University].
- Setiawaty, R., Sulistyorini, T. B., Margono, & Rahmawati, L. E. (2017). Validity test and reliability of Indonesian language multiple choice in final term examination. In *Proceedings of the 1st International Seminar on Language, Literature and Education* (pp. 43-50). <https://doi.org/10.18502/kss.v3i9.2609>
- Suruchi, S., & Rana, S. S. (2014). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in biology. *Paripex-Indian Journal of Research*, 3(6), 56-58. <https://doi.org/10.15373/22501991/JUNE2014/18>
- Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv*. <https://doi.org/10.48550/arXiv.2212.09292>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48, 1273-1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. MacMillan.
- van Griethuisen, R. A. L. F., van Eijck, M. W., Haste, H., den Brok, P. J., Skinner, N. C., Mansour, N., Gencer, A. S., & BouJaoude, S. (2014). Global patterns in students' views of science and interest in science. *Research in Science Education*, 45(4), 581-603. <http://doi.org/10.1007/s11165-014-9438-6>
- Wongpakaran, T., & Wongpakaran, N. (2012). A comparison of reliability and construct validity between the original and revised versions of the Rosenberg self-esteem scale. *Psychiatry Investigation*, 9(1), 54-58. <https://doi.org/10.4306/pi.2012.9.1.54>

APPENDIX A: ChatGPT AI-GENERATED QUESTIONS

1. Which of the following is NOT a factor that affects aging and biological changes in organisms?
 - a. Environmental factors
 - b. Genetic factors
 - c. Lifestyle choices
 - d. Planetary alignments**
 - e. Hormonal imbalances
2. Which of the following hormones is primarily responsible for stimulating cell division and growth in animals?
 - a. Adrenaline
 - b. Insulin
 - c. Thyroxine
 - d. Testosterone
 - e. Human growth hormone**
3. What is the process by which a single cell develops into a multicellular organism called?
 - a. Differentiation**
 - b. Fertilization
 - c. Metamorphosis
 - d. Germination
 - e. Replication
4. What is the primary function of the endoplasmic reticulum in a eukaryotic cell?
 - a. Energy production
 - b. Protein synthesis**
 - c. Waste removal
 - d. Storage of lipids
 - e. DNA replication
5. What is the primary function of lysosomes in a eukaryotic cell?
 - a. Energy production
 - b. Protein synthesis
 - c. Waste removal**
 - d. Storage of lipids
 - e. DNA replication
6. What is the basic unit of life?
 - a. Atom
 - b. Molecule
 - c. Cell**
 - d. Tissue
 - e. Organ
7. What is the main cause of loss of biodiversity?
 - a. Climate change
 - b. Deforestation
 - c. Overfishing
 - d. Land use change**
 - e. Increased CO₂ emissions
8. What is the definition of biodiversity?
 - a. The number of different species in a particular habitat
 - b. The total amount of living matter in a given area
 - c. The variety of life forms in a particular ecosystem**
 - d. The total mass of all living organisms in a given area
 - e. The study of different types of plants and animals in a given area

9. What is the term used to describe the variety of life on Earth, including the variety of species, ecosystems, and genes?
- Biodiversity**
 - Geodiversity
 - Climate diversity
 - Technology diversity
 - Cultural diversity
10. What is the study of heredity and variation in organisms called?
- Botany
 - Zoology
 - Genetics**
 - Paleontology
 - Geology
11. What type of molecule carries genetic information in cells?
- Carbohydrates
 - Lipids
 - Proteins
 - Nucleic acids**
 - Vitamins
12. What is the scientific method used to determine the sequence of nucleotides in a DNA molecule?
- Electrophoresis
 - PCR
 - Sequencing**
 - Microscopy
 - Chromatography
13. What is the process by which species of organisms change over generations through natural selection?
- Regression
 - Mutation
 - Adaptation
 - Evolution**
 - Stagnation
14. What theory was developed by Charles Darwin and Alfred Russel Wallace to explain the mechanism of evolution?
- Plate Tectonics Theory
 - General Relativity
 - Theory of Evolution by Natural Selection**
 - Quantum Mechanics
 - The Big Bang Theory
15. What is the main driving force behind evolution?
- Divine intervention
 - Survival of the fittest**
 - A fixed process
 - Mutation of species at random
 - The environment alone
16. What is the study of relationships between organisms and their environment known as?
- Sociology
 - Geology
 - Ecology**
 - Meteorology
 - Archaeology

-
17. What is the term used to describe the way organisms obtain energy from their environment?
- Metabolism
 - Ecosystem
 - Biodiversity
 - Photosynthesis**
 - Biogeography
18. What is the term used to describe the variety of life found in a specific location?
- Biodiversity**
 - Ecosystem
 - Biogeography
 - Metabolism
 - Photosynthesis
19. What is the application of scientific and engineering principles to the processing of materials by biological agents to provide goods and services?
- Nanotechnology
 - Biotechnology**
 - Robotics
 - Geotechnology
 - Information technology
20. What is the process of genetically modifying an organism to produce a desired trait or characteristic called?
- Cloning**
 - Fermentation
 - Hybridization
 - Transfection
 - Gene editing
21. What is the production of a large amount of a specific protein using microorganisms such as bacteria or yeast in bioreactors called?
- Cloning
 - Fermentation**
 - Hybridization
 - Transfection
 - Gene editing