Can Artificial Intelligence (ChatGPT 3.5) Improve Critical Thinking in Question Creation Elementary Education? Based on the Effectiveness of Citizenship Learning

Sultan Fakhrur Rassyi, Haryanto, Isro'ullaili, Akhi Ha Runi, Nurainy, Eko Marwanto, M. Yasid, M. Azhar

Abstract

The mid-semester exam evaluates the quality and feasibility of AI-generated question items, aiming to determine their validity, reliability, difficulty, distinguishability, and ability to enhance critical thinking. This quantitative descriptive study analyzes statistical data from 32 eighth-grade students using 40 multiple-choice questions on the Citizenship Education (PKn) exam. Findings reveal that 38 out of 40 questions are valid, with only questions on state institutions' functions and democratic processes being invalid. The Cronbach's alpha for the 38 valid questions is 0.65, indicating moderate reliability. Difficulty levels show 20 easy, 18 medium, and 2 difficult questions. Distinguishability results indicate 18 questions with low, 16 with sufficient, and 6 with good distinguishability. Student feedback shows 94% found the questions relevant, clear, accurate, and capable of stimulating critical thinking. Overall, the AI-generated questions are mostly valid, reliable, and effective in assessing students' knowledge and critical thinking skills.

Keywords: ChatGPT, Multiple Choice Questions, Midterm Exam, Citizenship Education, Critical Thinking.

The development of algorithms driven by machine-learning technology is now getting crazier and can no longer be stopped. This development is considered very worrying because it can replace human roles and tasks in every aspect of life, including education. One of the innovations born from the womb of machine-learning technology is ChatGPT. An interactive chatbot created by OpenAI, an artificial intelligence startup based in California [1]. OpenAI's ChatGPT is a comprehensive language model [2]. ChatGPT 3.5 is trained on large sets of text data using (deep learning algorithms) to create replies to every question humans ask (ChatGPT, 2023). The ChatGPT3.5 bot is now accessible at https://chat.openai.com.

Natural language processing technology in Artificial Intelligence, such as ChatGPT3.5, provides a means that allows computers to interact in two directions with human language [3]. An important stage in this technology is known as tokenization, which plays a role in converting unstructured information into organized text that is comprehensible and compatible with computing languages [4]. Its interactive nature is because ChatGPT 3.5 is able to understand what is requested and is able to convey it to humans as long as it complies with

Google policies and data bank availability [5]. For example, if we ask a search engine like Google to offer a list of questions related to a certain topic, Google will send links to websites that contain information relevant to the question we asked for. When asking the same command to ChatGPT 3.5, the application will provide answers as well as follow-up questions in that column [6].

The emergence of ChatGPT 3.5 is similar to the emergence of other new innovative technologies. If used properly, it has the potential to provide benefits to users, including in the world of education [7]. On the other hand, ChatGPT 3.5 also has the potential to be misused for things that are unacceptable in the academic field [8]. For example, students use ChatGPT 3.5 to complete assignments such as essays and answer multiple choice questions. Although, teachers may also be able to use AI to check which work was done by AI. Teachers can use ChatGPT 3.5 in a variety of ways, including asking questions regarding information, confirming the accuracy of data, reviewing a learning topic [9]. Teachers can also ask ChatGPT 3.5 to create multiple choice questions for exams. Of course, with the current version, ChatGPT 3.5 is not yet able to create an assessment instrument that is capable of measuring a learning objective accurately if explicit instructions are not given by an expert or teacher [10]. However, it is not impossible that in the future ChatGPT 3.5 will be able to generate complex questions if it has access to large data banks and has received extensive training [11].

In the increasingly developing digital era, artificial intelligence (AI) technology has penetrated various aspects of life, including education. One AI application that is currently in the spotlight is ChatGPT, a language model developed by OpenAI. ChatGPT has the ability to produce text that is similar to human language, including creating exam questions. This potential opens up new opportunities in preparing evaluation questions, especially in Citizenship Education (PKn) subjects. Citizenship education plays an important role in shaping students' character and understanding of their rights and obligations as citizens. Apart from that, this subject also aims to improve students' critical thinking skills, which are very necessary in democratic life. Therefore, the quality of the exam questions used in evaluation must be able to accurately measure students' cognitive abilities, including critical thinking abilities. Mid-term exams are a form of evaluation that is commonly used to measure student learning achievements in the middle of the semester [12]. The feasibility and quality of the items in the Midterm exam must be tested to ensure that the instrument functions in accordance with the evaluation objectives. This includes testing validity, reliability, level of difficulty, and differentiability of questions. By implementing AI technology such as ChatGPT, it is hoped that the question preparation process can be more efficient and produce quality questions [13].

The application of AI technology, such as ChatGPT, provides additional value in the process of preparing exam questions [14]. AI can help in preparing questions that are more varied and in accordance with the desired academic standards [15]. In addition, AI can be used to automatically test the quality of questions, including identifying potential bias or errors in question construction. By utilizing AI technology, the process of developing and testing Midterm exam questions can become more efficient and effective [16]. This has the potential to increase the suitability of questions to the competencies expected of students, as well as reducing the time and effort required of teachers to develop high-quality evaluation instruments [17].

This research aims to analyze the application of ChatGPT 3.5 in producing Midterm exam questions for Civics subjects for grade 8 students at Elementary school 1 Praya. The focus of this research is to measure the validity, reliability, level of difficulty, distinguishability, and ability

of the questions produced by ChatGPT to improve students' critical thinking abilities. Thus, it is hoped that this research can make a real contribution to the use of AI technology in the field of education and provide an overview of the effectiveness of ChatGPT 3.5 in preparing quality evaluation questions.

Critical Thinking

Critical thinking is the ability to analyze, evaluate, and synthesize information in a logical and objective manner. In the context of elementary education, critical thinking is a crucial skill that helps students not only to receive information passively but also to understand, question, and apply that information in various situations [18]. Key components of critical thinking relevant to elementary education include analysis, evaluation, inference, logical reasoning, problem-solving, and creativity [19]. Analysis involves breaking down information into smaller parts and understanding the relationships between them. Evaluation teaches students to assess the credibility of sources, the reliability of data, and the strength of arguments. Inference enables students to draw conclusions based on existing evidence. Logical reasoning involves teaching students to use logic in their thinking and arguments while avoiding common logical fallacies. Problem-solving equips students with strategies to solve problems effectively and efficiently. Creativity, though sometimes seen as separate from critical thinking, is essential in finding new solutions or viewing problems from different perspectives. The concept of critical thinking refers to how AI can be utilized to develop these skills in elementary students. AI, such as Chat GPT 3.5, can be employed to create questions designed to stimulate critical thinking, providing challenges that require deep analysis, logical reasoning, and problem-solving. These questions can help students develop their ability to think critically from an early age, which will be highly beneficial for their future education.

Learning Effectiveness

In basic education, the effectiveness of learning is largely determined by the use of diverse learning methods that suit students' learning styles. Proper technology integration also plays an important role in increasing student engagement with learning material. The active role of parents in supporting children's learning processes, as well as promoting inclusive and collaboration-based classes, helps form an effective learning environment. In addition, an individualized approach that takes into account the needs as well as constructive feedback to students, helps in strengthening their understanding and skills. By paying attention to all of this, educators can create holistic learning experiences and support student development not only academically but also socially and emotionally [20].

Research Methods

This research is a quantitative descriptive study which aims to explore the validity and reliability of Midterm exam questions designed by ChatGPT 3.5. Before the research was conducted, a number of exam questions created by ChatGPT 3.5 were collected and given to students as samples. This process involves in-depth evaluation of each question to ensure the accuracy of the material presented and the level of difficulty is appropriate to the curriculum. Validity is measured by testing whether the questions accurately measure the desired subject matter, while reliability is measured to assess the consistency of the results of the questions [21]. In addition, this research also explores how these questions can improve students' critical skills, by requiring them to think analytically and conclude. It is hoped that the results of this research will not only validate the use of AI in education, but also provide valuable guidance for developing more effective curricula and supporting improving the quality of learning in the future. Next, researchers accessed the ChatGPT 3.5 website in 2024, created an account, and logged in to the application[22].

ChatGPT 3.5 version May 25, 2024 was used to understand the research process in implementing GPT AI chat.

This research was conducted at Elementary school 1 Praya. A total of 32 grade 8 students were selected using saturated sampling techniques as samples. All students in class 8 work on questions produced by ChatGPT as Midterm exam. And students are also asked to fill out a questionnaire to provide their responses regarding the questions generated by ChatGPT students at the end of the exam [23]. The majority of students' age range was 13 years (87.5%, n=28) followed by 14 years (12.5%, n=4). The comparative number of students is 17 (58%) are women and 15 (42%) are men. And statistical analysis was calculated using IBM SPSS Stats 25 software. Question validity was determined using Pearson product-moment correlation. Question reliability was determined using Cronbach's alpha value[24]. The level of difficulty of the questions is determined by the following formula from [25].

Result and Discussion

5.1. Result

The researcher asked ChatGPT 3.5 to create questions using the command "Write 40 examples of multiple choice questions with answers for the Mid-Semester Exam for Civic Education class 8 odd semester middle school with the Merdeka curriculum. In the material chapter Pancasila as the Foundation of the State; Indonesian Government System; Rights and Obligations of Citizens; Democracy and Involvement in Elections". The four distributions of research questions related to question design produced by ChatGPT 3.5 can be seen in Table 1.

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

Civics Learning Materials	Number of Questions	Number of Questions
Pancasila as the Foundation of the State	10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Indonesian Government System	10	11, 12, 13, 14, 15, 16, 17, 18, 19, 20
Rights and Obligations of Citizens	10	21, 22, 23, 24, 25, 26, 27, 28, 29, 30
Democracy and Involvement in Elections	10	31, 32, 33, 34, 35, 36, 37, 38, 39, 40

Researchers also collected student responses to questions generated by ChatGPT 3.5 using criteria developed by [26] in his research to assess AI responses. After finishing answering the AI-generated questions, students complete this questionnaire . Students were told that the questions they had just worked on had been generated by AI, and they were given 10 minutes to complete the questionnaire. Table 2 displays the response questionnaire and its criteria.

Table 2. Questionnaire student responses to AI-generated questions

Question	Criteria
Relevance	Is this question relevant to the Civics subject you study in class/school?
Clarity	Is the question text easy to understand? Is it well structured and logically organized? Does it use appropriate language & vocabulary for its intended audience?
Accuracy	Are the questions correct (no wrong questions or no answer key)?
Precision	Are the questions explicit and detailed enough?
Depth	Are the questions deep enough (not too simple to do)?

After creating questions according to the structure, researchers carry out a test to validate the questions. The validity test results of all questions from ChatGPT 3.5 can be seen in Table 3.

Table 3. Validity results of all questions generated by ChatGPT3.5

R table	R Count	Information	No	R table	R Count	Information
0.482	0.739	Valid	21	0.482	0.560	Valid
0.482	0.763	Valid	22	0.482	0.598	Valid
0.482	00.757	Valid	23	0.482	0.621	Valid
0.482	0.677	Valid	24	0.482	0.647	Valid
0.482	0.567	Valid	25	0.482	0.754	Valid
0.482	0.627	Valid	26	0.482	0.756	Valid
0.482	0.654	Valid	27	0.482	0.662	Valid
0.482	0.590	Valid	28	0.482	0.654	Valid
0.482	0.598	Valid	29	0.482	0.603	Valid
0.482	0.394	No Valid	30	0.482	0.690	Valid
0.482	0.684	Valid	31	0.482	0.598	Valid
0.482	0.599	Valid	32	0.482	0.590	Valid
0.482	0.578	Valid	33	0.482	0.598	Valid
0.482	0.543	Valid	34	0.482	0.434	No Valid
0.482	0.603	Valid	35	0.482	0.739	Valid
0.482	0.699	Valid	36	0.482	0.763	Valid
0.482	0.598	Valid	37	0.482	0.757	Valid
0.482	0.621	Valid	38	0.482	0.677	Valid
0.482	0.647	Valid	39	0.482	0.7567	Valid
0.482	0.754	Valid	40	0.482	0.739	Valid

The results of the reliability test for all questions produced by ChatGPT 3.5 can be seen in Table 4, both by retaining all questions including those that were invalid (questions no. 10 and 34) or by eliminating these questions.

Table 4. Reliability test results for all questions

Cronbach's alpha	Number of question items
0.639	38
0.546	40

The reliability test results show two different scenarios for the collection of questions produced by ChatGPT 3.5 for the mid-semester exam in the Citizenship Education subject. First, by retaining all questions including invalid ones, Cronbach's alpha reached 0.639 from a total of 38 questions. Even though it shows a moderate level of consistency, there is diversity in the quality of the questions that needs further attention. The second scenario, after eliminating invalid questions (questions no. 10 and 34), shows an increase in the number of questions to 40, but reliability decreases to 0.546. This highlights the need for further review of the quality and consistency of the questions produced, as well as improvements in validity in order to more accurately measure students' critical thinking abilities in the context of Citizenship Education learning.

5.2. Discussion

According [7], validity is generally described as the extent to which an instrument measures what it wants to measure. An instrument must be valid so that it can be used to measure the intended subject. Testing using the Pearson product moment correlation method to assess the validity of the questions, it was determined that 38 of the 40 questions were valid, while 2 items were invalid. The invalid questions are numbers 10 and 34. The validity test results show that 38 of the 40 questions produced by AI are valid and can be used and can improve students' critical thinking. Question

number 10, which is invalid, asks students to choose answers related to the function of state institutions and the democratic process. It also confirms the existence of language and sentence problems that may arise in multiple choice questions created by ChatGPT3.5. Question number 10 As in this research, it should be corrected by experts using content validity as suggested [28]. However, we did not do this in our study because we wanted to ensure that the questions generated by ChatGPT3.5 were free from human adjustments.

Cronbach's alpha was used to assess the internal consistency of the scale Kilic, S. (2016). The reliability of the question instrument is known to have a Cronbach's alpha coefficient of 0.65% if 2 invalid question items (questions 10, 34) are removed. This is in line with the acceptable Cronbach's alpha value according to [29]. [30] confirmed the same thing, that Cronbach's alpha above 0.6 can be recognized as a reliable instrument. If these values are adhered to, then the multiple choice questions generated by ChatGPT3.5 in this study may be considered reliable. However, several other sources say that the allowable value for Cronbach's alpha is 0.79 [31]. If these numbers were used, the multiple choice questions generated by ChatGPT3.5 in this study might be considered unreliable.

Based on the assessment of the difficulty level of the questions, it is known that of the 38 questions (invalid items excluded) created by ChatGPT3.5, 20 questions are classified as easy, 16 questions are classified as medium, and 2 questions are classified as difficult. A good question is one that uses a proportional division of easy, medium and difficult multiple choice questions [32]. In this context, proportional means that the number of questions at the medium level is at least twice as many as at the easy and difficult levels, with the same number of questions at the easy and difficult levels. ChatGPT3.5 develops multiple choice questions that have almost the same level of easy and medium, and only two questions (9.5%) are classified as difficult so that each question can improve students' critical thinking.

Rao stated that ideally multiple choice questions have a medium level of difficulty. Of course, this must be revised depending on the objectives of the assessment. By assessing the discriminating power of questions, it was determined that, of the 40 questions created by ChatGPT3.5, 3 questions had low discriminating power, 5 questions had adequate discriminating power, and the other 33 questions had good discriminating power. Questions with low discriminatory power should be modified to have adequate or greater discriminatory power. There were no items that had negative discriminant power, thus indicating that no questions should be deleted based on the discriminant power analysis. However, one of the two items has a discrimination value of zero, which indicates that the item has very poor discriminative power. This is because the number of students who answered this item correctly in the upper group and lower group was the same. This question turned out to be numbers 10 and 34 which were classified as invalid based on the validity test, so it was unexpected that this question had very poor discriminating power. Moreover, the adversity index and the discrimination index are interconnected [33]. For example, if a question is considered to have a low level of difficulty and poor discriminating power, then the question should be revised [34].

Based on student responses to ChatGPT's AI-generated questions, it was determined that 94% of students thought the AI-generated questions were relevant to the subjects they studied in class. These findings show that ChatGPT3.5 is able to generate questions related to certain subjects [35]. Most students (94%) reported that the AI-generated questions were clear. This shows that most students are able to understand the questions asked by ChatGPT3.5. The clarity of the questions was determined by three survey items. The first item on the questionnaire asks whether the questions generated by ChatGPT3.5 are easy to

understand. Most students (94%) also stated that the questions were easy. The second question asks whether the questions generated by ChatGPT3.5 are structured and logically ordered. According to 94% of students, the questions are well structured and logically arranged. The final question asks whether the questions generated by ChatGPT3.5 use the right language. Most students (94%) felt that the language of the questions was appropriate. The questions in the assessment should be clear and concise. Questions that are difficult to understand will certainly make it difficult for students to answer them, and it is very likely that students will answer incorrectly not because of their incompetence but because of errors in the question. Not a few students (94%) stated that the questions generated by AI were accurate. This means that most students found the AI-generated questions to be accurate. They don't see any grammatical or conceptual errors in the questions. However, you can't just rely on students' opinions to ensure the accuracy of a question. Several experts should be consulted to validate the question. However, the questions in this study were not evaluated by professionals to determine how they were generated by AI.

Additionally, Most students (94%) also agreed that the questions generated by AI were appropriate. This shows that most students found the AI-generated questions to be explicit and detailed. Students understand the meaning of the question and the required response. If questions are not made clearly and unambiguously, it is likely that students will have difficulty answering them. Lastly, Most students (94%) thought that the questions asked by AI were quite insightful. The majority of students found the questions generated by ChatGPT3.5 to be challenging, not too simple, and appropriate for their grade/school level. As was done in this study, measuring the difficulty level of questions is another method for determining whether questions are too easy or too difficult. Only two of the two AI-generated questions were difficult, while the remaining 38 questions were fairly easy and moderate. The majority of students responded positively to questions generated by AI ChatGPT, as revealed by the results of the student response questionnaire. This shows that the use of AI in developing assessment tools can be an effective alternative for teachers. However, it is important for teachers to provide clear instructions to AI so that the questions generated are in line with the desired learning objectives. Further studies are needed to evaluate whether students can differentiate between AI-developed and human-generated questions. It is important to understand how the use of AI technology can interact with students' perceptions of their educational evaluation process.

Previous research by [36] showed that ChatGPT3.5 was able to generate valid and reliable questions. These findings support the potential of AI technology in assisting teachers in compiling Midterm Exam questions that meet the required evaluation standards. Overall, the use of AI in education, especially in developing evaluation instruments such as midterm exam questions, offers the possibility of increasing the efficiency and quality of the evaluation process. However, further research needs to be carried out to ensure that the use of AI not only meets technical criteria, but also pays attention to its impact on students' perceptions and overall learning experience. However, other studies such as [37] also show very different things. That AI technologies such as ChatGPT and Bard are unable to achieve appropriate and minimum scores, especially in the fields of endocrinology and diabetes/diabetes technology. This study shows that AI technology has the potential to facilitate students but still requires more up-to-date information and fresh data to support the validity and reliability of the questions (

So it is true what was stated by Suppadungsuk, that using ChatGPT as the only source for identifying literature review references is not recommended. Future research could look for ways to improve the performance of AI language models in identifying relevant literature. But on the other hand, considering that

composing multiple choice questions is a complex and time consuming process. It would be very beneficial if AI could help teachers or the education sector in the future in developing standardized and high-quality multiple choice questions. This efficiency allows individuals to focus more on other aspects of teaching, research, or content development. Lastly, the current version of ChatGPT3.5 has several limitations such as the possibility of generating misinformation, harmful instructions, or biased material. There may be times when the questions are ambiguous, misleading, or do not adequately test understanding of the material. This can have an impact on the effectiveness of assessments or learning outcomes. We can see this from the author's findings on 2 invalid question items (questions 10, 34) which need to be removed. Over time, ChatGPT3.5 will gain more data and better training that will allow it to help its users more effectively.

Conclusion

Based on the research findings presented, it can be concluded that the application of ChatGPT 3.5 in producing mid-semester exam questions for Citizenship Education subjects has several relevant characteristics. Of the 40 questions generated, 38 questions can be considered valid, with the only invalid question relating to the function of state institutions and the democratic process. Cronbach's alpha coefficient of 0.65 indicates an adequate level of reliability for 38 valid questions, indicating consistency between one question and another. In terms of the level of difficulty of the questions, research shows that 20 questions are classified as easy, 18 questions are classified as medium, and 2 questions are classified as difficult. Meanwhile, in terms of discrimination power, 18 questions had low discrimination power, 16 questions had adequate discrimination power, and 6 other questions had good discrimination power. So the application of ChatGPT 3.5 in producing mid-semester exam questions for Citizenship Education subjects has the potential to improve students' critical thinking abilities, although it is necessary to make adjustments to invalid questions and increase the discriminatory power of some questions to ensure a better test pedagogically.

Acknowledgments

We would like to express our thanks to the supervisor who has directed us in writing this scientific article and we would like to express our deepest gratitude to BPPT (Center For Higher Education Funding) and LPDP (Indonesia Endowment Fund For Education) of the Republic of Indonesia for providing the Indonesian Education Scholarship (BPI) so that we can complete the publication of this scientific article.

WORKS CITED

[1] Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. In Research in Social and Administrative Pharmacy (Vol. 15, Issue 2, pp. 214–221). Elsevier Inc. https://doi.org/10.1016/j.sapharm.2018.03.066

[2] Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. In Journal of AI (Vol. 52, Issue 7).

[3] Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in Obstetrics & Gynecology, 31(4), 466–475. https://doi.org/10.1002/uog.5256

[4] Beckstead, J. W. (2009). Content validity is naught. International Journal of Nursing Studies, 46(9), 1274–1283. https://doi.org/10.1016/j.ijnurstu.2009.04.014

[5] Bialocerkowski, A. (2008). Measurement error and reliability testing: Application to rehabilitation Journal Title International Journal of Therapy and Rehabilitation Copyright Statement Link to published

version. http://hdl.handle.net/10072/47277http://www.ijtr.co.uk/cgi-bin/go.pl/library/article.cgi?uid=31210;article=IJTR_15_10_422_427
[6] Christmann, A., & Van Aelst, S. (2006). Robust estimation of Cronbach's alpha. Journal of Multivariate Analysis, 97(7), 1660–1674. https://doi.org/10.1016/j.jmva.2005.05.012
[7] De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. In Medical Education (Vol. 44, Issue 1, pp. 109–117). https://doi.org/10.1111/j.1365-2923.2009.03425.x
[8] Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. Journal of Organizational Behavior, 36(1), 3–15. https://doi.org/10.1002/job.1960
[9] Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? Journal of Responsible Technology, 13, 100060. https://doi.org/10.1016/j.jrt.2023.100060
[10] Faiz, A., & Kurniawaty, I. (2023). Tantangan Penggunaan ChatGPT dalam Pendidikan Ditinjau dari Sudut Pandang Moral. EDUKATIF : JURNAL ILMU PENDIDIKAN, 5(1), 456–463. https://doi.org/10.31004/edukatif.v5i1.4779
[11] 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. https://doi.org/10.1080/14703297.2023.2195846
[12] Gogia, P. P., Braatz, J. H., Rose, S. J., & Norton, B. J. (1987). Reliability and Validity of Goniometric Measurements at the Knee PHYSICAL THERAPY.
[13] Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. Computers & Education, 206, 104898. https://doi.org/10.1016/j.compedu.2023.104898
[14] Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. Contemporary Educational Technology, 15(2). https://doi.org/10.30935/cedtech/13036
[15] Huang, S. (2014). Relevance of IT Integration into Teaching to Learning Satisfaction and Learning Effectiveness. World Journal of Education, 4, 1-11.
[16] 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
[17] Kalla, D. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. In International Journal of Innovative Science and Research Technology (Vol. 8, Issue 3). www.ijisrt.com
[18] Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions.
[19] 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. In Learning and Individual Differences (Vol. 103). Elsevier Ltd. https://doi.org/10.1016/j.lindif.2023.102274
[20] Kehoe, J. (1994). Basic Item Analysis for Multiple-Choice Tests. Practical Assessment, Research, and Evaluation, 4, 10. https://doi.org/10.7275/07zg-h235
[21] Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. Medical Journal Armed Forces India, 77, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007
[22] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198
[23] Pahrudin, A., Zahra, Y.F., Supriadi, N., Sugiharta, I., Farida, F., & Suherman, S. (2021). Assessing moodle-assisted e-learning for students' concept understanding and critical thinking skills in algebra. Al-Jabar : Jurnal Pendidikan Matematika.
[24] Sullivan, G. M. (2011). A Primer on the Validity of Assessment Instruments. Journal of Graduate Medical Education, 3(2), 119–120. https://doi.org/10.4300/jgme-d-11-00075.1
[25] Lambrinou, E., Sourtzi, P., Kalokerinou, A., & Lemonidou, C. (2005). CARE OF OLDER PEOPLE Reliability and validity of the Greek version of Kogan's Old People Scale.

[26] Limna, P., Kraiwanit, T., Jangjarat, K., Klayklung, P., & Chocksathaporn, P. (2023). The use of ChatGPT in the digital era: Perspectives on chatbot implementation. Journal of Applied Learning and Teaching, 6(1), 64–74. https://doi.org/10.37074/jalt.2023.6.1.32

[27] Marwati, I., Usman, M.U., & Prastiti, T.D. (2023). The Effect of Concept Achievement on Critical Thinking and Creative Mathematical In Elementary School Students. Journal Research of Social Science, Economics, and Management.

[28] Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? https://ssrn.com/abstract=4333415

[29] Marshall, G. (2005). The purpose, design and administration of a questionnaire for data collection. In Radiography (Vol. 11, Issue 2, pp. 131–136). W.B. Saunders Ltd. https://doi.org/10.1016/j.radi.2004.09.002

[30] McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. In Medical Teacher (Vol. 26, Issue 8, pp. 709–712). https://doi.org/10.1080/01421590400013495

[31] 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

[32] Ouyang, T., Nguyen-Son, H.-Q., Nguyen, H. H., Echizen, I., & Seo, Y. (2023). Quality Assurance of A GPT-based Sentiment Analysis System: Adversarial Review Data Generation and Detection. http://arxiv.org/abs/2310.05312

[33] Mccowan, R. J., & Mccowan, S. C. (1999). Item Analysis for Criterion-Referenced Tests. http://www.bsc-cdhs.org

[34] Mohajan, H. K. (2017). TWO CRITERIA FOR GOOD MEASUREMENTS IN RESEARCH: VALIDITY AND RELIABILITY. Annals of Spiru Haret University. Economic Series, 17(4), 59–82. https://doi.org/10.26458/1746

[35] Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice, 20(2). https://doi.org/10.53761/1.20.02.07

[36] Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). IS CHATGPT A GENERAL-PURPOSE NATURAL LANGUAGE PROCESSING TASK SOLVER?

[37] Rao, C., Kishan Prasad, H., Sajitha, K., Permi, H., & Shetty, J. (2016). Item analysis of multiple choice questions: Assessing an assessment tool in medical students. International Journal of Educational and Psychological Researches, 2(4), 201. https://doi.org/10.4103/2395-2296.189670.