COMPARATIVE ANALYSIS OF AI AND TEACHER-GENERATED MATHEMATICS MULTIPLE-CHOICE ITEMS USING 2-PARAMETER MODEL

[1]BOKOLO, F. & [2]UWAH, V. I.
[1]Department of Educational Foundations, Isaac Jasper Boro College of Education, Sagbama, Bayelsa State. Nigeria.
[2]Department of Educational Psychology G/C University of Port Harcourt, Rivers State, Nigeria.
faithbokolo2003@gmail.com, idongesituwah6@gmail.com,

Abstract
The study was on comparative analysis of AI and teacher-generated mathematics multiple-choice items using 2-parameter model. Instrumentation design was used in the study. A sample of 178 SS2 students were drawn from a population 3,450 across public schools in Obio-Akpor L.G.A of Rivers State. Two versions of Mathematics Performance Test containing 50 items each of AI-generated (MPT-AI) and maually Generated (MPT-T) were used. Findings of this study showed that for difficulty index (Jaipurkar et'al, 2021 reasonable range of 0.30-0.70), MPT-AI format had 27 items with moderate difficulty indices meaning that 23 items were either two difficult or simple. For MPT-T, 38 items had moderate difficulty index indicating that 12 items were either too tough or simple in the test. For item discrimination (cut-off score of >0.40 proposed by Aljehani et'al, 2020) MAT-AI showed that 36 items discriminated adequately with 14 items discarded. MPT-T test also reveal that 26 items had better discrimination index with 24 being discarded. Comparatively, MPT-T test had greater number of items with good difficulty indices (38>27) than MPT-AI test. On the contrary, MPT-AI test had more items with good discrimination index (36>26) compared to MPT-T items. Additionally, KR_{20} was used in determining the reliability indices of 0.75 for MPT-AI and 0.96 for MPT-T format respectively. Based on the findings, it was recommended among others that test developers should compulsorily check the difficulty index of AI-generated items than discrimination while they should also compulsorily check discrimination index for teacher-generated test than the difficulty index.

Keywords: Test, Teacher-made test, AI-generated test, 2 parameter model

Introduction

A test whether formal or informal is an assessment process intended to measure student's knowledge, skill, aptitude as well as physical fitness. A test may be administered verbally, on paper, on a computer, or in a pre-determined area that requires a test taker to demonstrate or perform a set of skills. Opara (2021) defined a test as an instrument or procedure designed to measure the knowledge, intelligence,

ability, traits, skills, aptitude, interest, and attitude of an individual or group. Similarly, Uwah (2022) state that a test whether. Tests vary in style, rigor and requirements. A test may be developed and administered by an instructor, a clinician, a governing body, or a test provider. In some instances, the developer of the test may not be directly responsible for its administration. A teacher-made test is one constructed by the classroom teacher to measure the extent of performance of a specific objective within the class (Ukwuije, 2012). Kpolovie (2010) opine that not until the test reflects the true attributes, characteristics or ability of the students, then such test is invalid. But quite often, test scores as obtained by students on a subject with the same content items do vary significantly.

Mathematics as a core subject in the curriculum has been a problem for many students except few as reported by Obinna (2009). Student's performance in mathematics is the extent to which student master the subject matter. Students' performance in mathematics refers to the extent to which students demonstrate understanding, knowledge, and skills in mathematics, as measured by various assessments, evaluation and observations. Students at different point or the other may record high achievement in mathematics. At some point, they may as well record very poor grades. Tei-Firstman, (2011) argued that majority of students find it difficult to pass mathematics at first sitting. This trend is even worsened by the hate the students developed towards the subject. In a similar development.

Test items are constructed in various forms (teacher-made and standardized) measured by various assessments, evaluation and observations. However, because of the advent of technology, Artificial Intelligence can be used to generate test items. Kumar & Lal (2024) stated that a teacher-made test is an assessment created by an educator to evaluate students' understanding of specific learning objectives or course material. The characteristics of teacher made test according to Kumar & Lal (2024) include but not limited to ability to aligned with specific learning objectives or standards. The importance of teachers made test are enormous. According to Lee & William cited in Kumar & Lal (2024), the key to teacher-made tests is to make them a part of instruction and not separate from it. Robert in Copeland (2024) states that the other problem faced by classroom tests are that the teachers lack the skills of appraising the effectiveness of the test. As Chakanyuka (2000) advises, teachers should build a file of items for future use called an item bank and one of such ways is the application of artificial intelligence.

Artificial intelligence (AI) is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings (Copeland, 2024). Russell & Norvig *(2021) defined* Artificial intelligence (AI) in its broadest sense, as intelligence exhibited by machines, particularly computer systems. According to *Kaplan & Haenlein (2019),* some high-profile applications of AI include advanced web search engines like Google Search, Waymo, generative and creative tools like ChatGPT, ChatBot etc. Milicević, Lazarova & Pavlović (2024) maintained that the uses of Artificial Intelligence has wider application with

limited questionable capabilities. In terms of testing, it is observed that most educational stakeholders like the teachers now have been involved in the use of AI in generating test items. Just like Cui, Chen, Shiri & Fan (2019) doubted, the questions remain if the AI generated items are capable of posing all the indices expected of a standard test to possess. According to Rossi (2023), GPT generates texts of any genre on any topic in seconds. Zong & Krishnamachari (2022) supported this fact by stating that it lacks clarity regarding text authorship. To Zong & Krishnamachari (2022), Using AI to write test items has many benefits that can significantly enhance the testing process. Attali, LaFlair and Runge (2023) opined that leveraging AI to write unit tests empowers developers to improve test coverage, enhance accuracy and efficiency, and optimize resource utilization

Item parameter represents the various continuum which a test item should possess. Ukwuije in Uwah (2022) maintained that items parameters symbolically represented as "a" parameter (discrimination), symbolically represented as "b" parameter (difficulty). Item Discrimination ("a") determines the rate at which the probability of endorsing a correct item changes given ability level. This parameter is imperative in differentiating between individuals possessing similar levels of the latent construct of interest. Items with low discrimination indices are often ambiguously worded and should be examined. According to ScorePak (2020), item discrimination is "good" if the index is above .30; "fair" if it is between .10 and .30; and "poor" if it is below .10.

Item difficulty ("b") as the parameter that determines the manner of which the item behaves along the ability scale. It is determined at the point of median probability i.e. the ability at which 50% of respondents endorse the correct answer. Opara in Uwah (2022) stated that item's difficulty is usually measured in terms of the percentage of examinees who answer the item correctly. It can range between 0.0 and 1.0, with a higher value indicating that a greater proportion of examinees responded to the item correctly, and it was thus an easier item. In general, tests developed and administered by individual instructors (teacher-made) likewise those generated by AI may not be properly standardized and this may have negative effect on the outcome of the test.

Teacher-generated test items as explained in this study is the manual and painstaking process of thinking out and writing down possible mathematics test items by teachers. This will involves the teacher sitting down with the course content and writing out questions from what he or she has taught in the class. In other words it could be seen as manually generated because the teacher in this process is not aided by any machine. The teacher in this process uses his or her mentality to coin the grammer of a question in line with the testees ability or level. On the other hand, AI-generated items are those which the teacher only "prompt" or command an artificial intelligence platform like Chatbot-AI or ChatGPT to generate such items for him. According to Binder (2024), AI generated test are modern teaching tools you can leverage to save time without sacrificing quality or personalization. An AI quiz

maker or an AI test maker will automatically create and prepare questions geared toward the subject you are teaching, with the added perk of suggesting answers, auto-grading, and more. In the current study, the resercahers put up a command or "prompted" the artificial intelligence app to just create or generate a particular number of multiple-choice items in mathematics. And that was done automatically. Hence, the aim of the study is to determine the:

1. difficulty index of multiple test items generated by teacher manually and by AI application.
2. discriminative index of multiple test items generated by teacher manually and by AI application
3. reliability indices of multiple test items generated by teacher manually and by AI application

Research Questions

The following research questions were asked;

1. What the difficulty indices of multiple test items generated by teacher and by AI application?
2. What are discriminative indices of multiple test items generated by teacher and by AI application?
3. What is are the reliability indices of multiple test items generated by teacher and by AI application?

Methodology

The researcher in this study used instrumentation design in the study. Instrumentation design is used here because the study involves the development of a test instrument. Uwah (2022) noted that this involves the *design* of the *assessment instrument* is extremely important in the ability to collect valid and reliable information the skills and knowledge of students. Hence, this use of the design was justified as it will be used in collecting of data on mathematics performance. The targeted population of the study consisted of 8,450 SS2 students limited to Obio-Akpor L.G.A of Rivers State, Nigeria. The sample size of the study consisted of 178 Senior Secondary School 2 (SSS2) students drawn across public schools in Rivers State using the multi-stage sampling procedure. In stage 1, the researcher used stratified sampling technique to stratify the schools in state based on the three senatorial districts. At stage two, simple random sampling technique by ballot to select two Local Government Areas (LGAs) from each of the senatorial districts. This gave a total of 6 LGAs. At stage three, simple random sampling technique was also to select two schools from each of the LGA. This gave a total of 12 schools in all. At stage 4, in any of the school visited, 15 SSS2 students were sampled using simple random sampling. This gave a total of 180 students in all.

Two versions of the Mathematics Performance Test containing 50 items each of AI-generated (MPT-AI) and Teacher Generated (MPT-T) will be used as instruments in

the study. The MPT-T is a 50-item test generated by the classroom teachers. It is a cognitive test which is patterned using the multiple-choice items format in selected topics in mathematics using the course content from SS2. These topics include algebra, geometry, statistics, number and operations as well as probability, Similarly, the MPT-AI is also a multiple-choice items test which the researcher generated using ChatGPT. The test also contains 50-items in all. Table of specification in line with Bloom taxonomy of educational objective domain was used in determining the quality of test items. Data was analyzed using Excel and application of item difficulty and discrimination formulae respective. The use of Excel aided in the automatic arrangement of the items based on specific command. Like the extraction of items that are above certain index. The parameter for acceptability of an item based on difficulty was 0.30 to 0.70 as stated by Jaipurkar et' al, (2021). In terms of item discrimination, a cut-off score of >0.40 as proposed by Aljehani et'al, (2020) was used. KR_{20} will be used in determining the reliability estimates of MAT-AI and MAT-T format respectively

Results

Research Question One: What are the difficulty indices of multiple test items generated by teacher and by AI application?

Table 1: Items of MPT-T and MPT-AI and their respective difficulty Indices

Questions	MPT-T	Remarks	MPT-AI	Remarks
1	0.96*	Too easy item	0.96*	Too easy item
2	..	**Moderate Item**	0.94*	Too easy item
3	0.27**	Too Difficult	**0.41**√	**Moderate Item**
4	**0.60**√	**Moderate Item**	0.97*	Too easy item
5	**0.41**√	**Moderate item**	**0.51**√	**Moderate Item**
6	**0.37**√	**Moderate item**	0.79*	Too easy item
7	**0.42**√	**Moderate Item**	**0.32**√	**Moderate Item**
8	**0.46**√	**Moderate Item**	0.96*	Too easy item
9	**0.49**√	**Moderate Item**	0.79*	Too easy item
10	**0.61**√	**Moderate Item**	**0.38**√	**Moderate Item**
11	0.97*	Too easy item	0.97*	Too easy item
12	**.51**√	**Moderate Item**	**0.41**√	**Moderate Item**
13	0.89*	Too easy item	**0.49**√	**Moderate Item**
14	**0.46**√	**Moderate Item**	0.96*	Too easy item
15	**0.48**√	**Moderate Item**	**0.44**√	**Moderate Item**
16	**0.51**√	**Moderate Item**	0.90*	Too easy item
17	**0.63**√	**Moderate Item**	0.98*	Too easy item
18	0.94*	Too easy item	**0.39**√	**Moderate Item**
19	**0.33**√	**Moderate Item**	0.93*	Too easy item
20	**0.45**√	**Moderate Item**	0.89*	Too easy item
21	**0.68**√	**Moderate Item**	**0.68**√	**Moderate Item**
22	**0.50**√	**Moderate Item**	**0.40**√	**Moderate Item**
23	**0.52**√	**Moderate Item**	0.82*	Too easy item
24	0.75*	Too easy item	**0.55**√	**Moderate Item**
25	**0.36**√	**Moderate Item**	0.06**	Too difficult item
26	**0.51**√	**Moderate Item**	**0.51**√	**Moderate Item**
27	**0.57**√	**Moderate Item**	**0.57**√	**Moderate Item**
28	**0.54**√	**Moderate Item**	**0.54**√	**Moderate Item**
29	**0.43**√	**Moderate Item**	0.73*	Too easy item
30	**0.49**√	**Moderate Item**	0.79*	Too easy item
31	**0.69**√	**Moderate Item**	**0.69**√	**Moderate Item**
32	**0.33**√	**Moderate Item**	**0.33**√	**Moderate Item**
33	**0.35**√	**Moderate Item**	0.25**	Too difficult item
34	**0.57**√	**Moderate Item**	**0.57**√	**Moderate Item**
35	**0.60**√	**Moderate Item**	**0.60**√	**Moderate Item**
36	**0.61**√	**Moderate Item**	**0.61**√	**Moderate Item**
37	**0.54**√	**Moderate Item**	**0.54**√	**Moderate Item**
38	**0.40**√	**Moderate Item**	0.10**	Too difficult item
39	0.01**	Too difficult item	**0.41**√	**Moderate Item**
40	**0.70**√	**Moderate Item**	**0.70**√	**Moderate Item**
41	**0.44**√	**Moderate Item**	0.14**	Too difficult item
42	**0.43**√	**Moderate Item**	**0.43**√	**Moderate Item**
43	0.79*	Too Easy item	0.19**	Too difficult item

Criterion 0.30-0.70 (Jaipurkar et' al, 2021)
=Too Easy Item
****=Too Difficult Item***

From the Table 1, out of the fifty mathematics items generated by the teacher (MPT-T), five items (3, 39, 45, 49 and 50) marked (**) were considered too difficult for the students and were thus eliminated from the test. Items 1, 11, 13, 18, 24, 43 and 48 marked (*) were considered too easy and thus were also removed from the test. Hence, 38 items which involves items 2, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 17, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 42, 44, 46 and 47 respectively marked "√" fell in the moderate index range and were considered the good items test.

For the MPT-AI generated test, fifteen items (1, 2, 4, 6, 8, 9, 11, 14, 16, 17, 19, 20, 23, 29 and 30) marked (*) were too easy and were removed from the test. Eight items (25, 33, 38, 41, 43, 45, 47 and 49) were marked (**) and were considered too difficult for the testees and were removed from the test. Hence, 27 items including items 3, 5, 7, 10, 12, 13, 15, 18, 21, 22, 24, 26, 27, 28, 31, 32, 34, 35, 36, 37, 39, 40, 42, 44, 46, 48 and 50 marked (√) were considered moderately and good for the testees. In all, it indicates that MPT-T had 38 items with acceptable difficulty while the MPT-AI had 27 items. It could be seen that MPT-T was better in producing items with moderate difficulty Than MPT-AI.

Research Question Two: What are discrimination indices of multiple test items generated by teacher and by AI application?

Table 2: Items of MPT-T and MPT-AI and their respective discrimination indices

Questions	MPT-T	Remarks	MPT-AI	Remarks
1	0.05	Poor	...	Good
2	0.51√	Good	0.47√	Good
3	0.02**	Poor	0.12*	Poor
4	-0.04	Poor	0.45√	Good
5	0.42√	Good	0.49√	Good
6	0.09**	Poor	0.55√	Good
7	0.44√	Good	0.62√	Good
8	0.10	Poor	0.49√	Good
9	0.18	Poor	0.62√	Good
10	0.41√	Good	0.49√	Good
11	0.53√	Good	0.47√	Good
12	0.10	Poor	0.51√	Good
13	-0.01	Poor	0.08**	Poor
14	0.51√	Good	0.09**	Poor
15	0.40√	Good	0.52√	Good
16	0.42√	Good	0.55√	Good
17	0.13	Poor	0.43√	Good
18	0.44√	Good	0.20	Poor
19	0.02**	Poor	0.47√	Good
20	0.12*	Poor	0.57√	Good
21	0.45√	Good	0.31√	Good
22	0.49√	Good	0.40√	Good
23	0.55√	Good	0.02**	Poor
24	-0.12**	Poor	0.57√	Good
25	-0.19**	Poor	0.48√	Good
26	0.62√	Good	0.14*	Poor
27	0.49√	Good	0.52√	Good
28	0.47√	Good	0.43√	Good
29	0.51√	Good	0.40√	Good
30	0.08**	Poor	-0.09**	Poor
31	0.09**	Poor	0.50√	Good
32	0.52√	Good	0.41√	Good
33	0.15	Poor	-0.09**	Poor
34	0.43√	Good	0.05	Poor
35	0.20	Poor	0.51√	Good
36	0.47√	Good	0.54√	Good
37	0.57√	Good	-0.04	Poor
38	0.31*	Poor	0.42√	Good
39	-0.05**	Poor	0.09**	Poor
40	0.02**	Poor	0.44√	Good
41	0.57√	Good	0.10	Poor
42	0.48√	Good	0.42√	Good
43	0.14*	Poor	0.41√	Good

Criterion= (0.40, Aljehani et'al, 2020)

From Table 2, 26 items in MPT-T had good discrimination. They included items 2, 5, 7, 10, 11, 14, 15, 16, 18, 21, 22, 23, 26, 27, 28, 29, 32, 34, 36, 37, 41, 42, 44, 46, 48 and 49. On the other hand, item 1, 3, 4, 6, 8, 9, 12, 13, 17, 19, 20, 24, 25, 30, 31, 33, 35, 38, 39, 40, 43, 44, 45 and 47 respectively discriminated poorly. For MPT-AI, 36 items had good discrimination indices. They included items 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 19, 20, 21, 22, 24, 25, 27, 28, 29, 31, 32, 35, 36, 38, 40, 42, 43, 44, 45, 47, 48 and 49 respectively. On the contrary, items 3, 13, 14, 18, 23, 26, 30, 33, 34, 37, 39, 41, 46 and 50 had poor discrimination indices. Comparatively, this implies that MPT-AI had better discrimination (36 >24) than MPT-T (24 < 36)

Research Question Three: What are the reliability indices of multiple test items generated by teacher and by AI application?

Table 4; Internal consistency of MPT-T and MPT-AI using Kr_{20}

Test Type	N		$ð^2$	∑Pq	KR_{20}
MPT-T	50	12.50	156.25	6.58	0.96
MPT-AI	50	5.31	28.19	7.31	0.75

From Table 3, reliability estimates using KR_{20} showed that MPT-T has reliability index of 0.96 while MPT-AI had reliability index of 0.75. This shows that MPT-T has better reliability coefficient compared to MPT-AI.

Discussion of Findings

From finding one, it is revealed that MPT-T had 38 items that are moderately difficult compared to MPT-AI which had 27 items. It could be that the students are very familiar with these set of questions. It could also be that they were simply focused within the time limit of response to the item. It could simply be that the students did not understand these questions or they may have been too confused to know the exact answer and miss out attempting them due to time. In another point of view, it could be that the majority avoided or skipped these items or may have had no time to attempt them thereby seeming unanswered and difficult. It may also be that such easiness index or difficulty level may come as a result of other factors which may have accounted for the variation in performance of students. For the MPT-AI that had lower level of moderate items, it could be that the artificial intelligence used may not have been able to distinguished some possible items based on the target testees. The current findings is in agreement with the findings of Zong & Krishnamachiari (2022) who reported that AI generated test lacks significant clarity which serve as a disadvantage to students when compared with manually generated ones. However, it should also be noted that the present standard for adjudging the levels of difficulty may vary which the present one was based on Jaipurkar et'al, (2021) range of 0.30 to 0.70. Hence, if other researcher's follow other standards, this may give slight outcomes. The analysis also reveals that some of the items are

negative. A negative index indicates a mis-match for the items. This means that some of the items may not have measured what it purports to measure and this could only be explained by the fact that some of them use foreign symbols and analogies which could come mostly from the AI generate items which may be totally unfamiliar with the respondents.

From the second findings, based on Aljehani et'al, (2020) scale, item discrimination is "good" if the index is above .40. From this premise, it is clear that for MPT-T, 26 items had good discrimination while 24 items showed poor discrimination. This means that majority of the items of MPT-T were not able to distinguish between the high performing students and the lower performing one. This implies that there is a difference between students who are intelligent and those who are less intelligent. From MPT-AI, 36 items had good discrimination while 14 items showed poor discrimination. Comparatively, it indicates that MPT-AI had better discriminating power than MPT-T. The outcome of the finding here could only be explained that by the fact that the teachers had no knowledge of developing items that can identify individuals' students with abilities in specific construct. It could also be that the AI have better algorithms in delineating specific constructs more compared to individual teachers' abilities. This finding supports the assertion of Attali, LaFlair and Runge (2023) who noted that leveraging AI to write unit tests empowers developers to improve test coverage, enhance accuracy and efficiency, and optimize resource utilization.

The report on reliability indicates that mathematics performance test generated by the teacher manually has better reliability coefficient compared to those generated by artificial intelligence. The findings here means that may be attributed to the fact that AI algorithms can analyze vast amounts of data and generate test items that are more consistent and accurate. Additionally, AI-generated test items can be designed to meet specific learning objectives and can be tailored to individual students' needs. It also suggests that teachers can focus more on teaching and less on test preparation, as AI can take care of generating reliable test items. Secondly, it highlights the potential for AI to support personalized learning, as AI-generated test items can be tailored to individual students' needs and abilities. The findings of the study however is in disagreement with the finding reported earlier by Dogan, Goru Dogan and Bozkurt (2023) who reported significant teacher made test has higher reliability that the AI generated test.

Conclusion

Mathematics performance test organized by teachers are fairly difficult compared to those generated by artificial intelligence. However, the result of the difficulty level differs depending on the standard of standard of judgement applied. Comparatively, mathematics test developed using AI have better discrimination power that those generated by the teacher manually. Finally, the use of either approach or methods has their unique merit and demerit.

Recommendations

Based on the findings of the study, the following recommendations are made;

1. Because teacher generated mathematics test have more moderate difficulty index that the AI generated, it is recommended that test developers and teachers should always ensure they generate test items manually.
2. Based on the finding that mathematics test developed using AI have better discrimination power that those generated by the teacher manually, it is recommended that test constructors and teachers should also ensure that they generate test items manually in order to have good discrimination index.
3. Based on the findings that AI-generated items have better reliability index, it is recommended that teachers should employ AI occasionally especially when trying to develop test that will be reuse in future dates.

References

Attali, Y., LaFlair, G., & Runge, A. (2023). A new paradigm for test development [Duolingo webinar series].

https://www.youtube.com/watch?v=rRc96oe9bzk&t=2s

Chakanyuka, S. (2000). Measurement and evaluation. Harare: Zimbabwe Open University press.

Copeland, B. J. (2024). Artificial intelligence. Retrieved from https://www.britannica. *com/technology/artificial-intelligence on 12/5/2024.*

Kaplan, J. D. & Haenlein, G. (2019). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS2020).*https://proceedings.neurips.cc/paper/2020/hash/1457c0d6b fcb4967418bf b 8a c142f64a-Abstract.html.

Kpolovie P., J. (2010) Test, Measurement and Evaluation in Research (end Ed.)owerri. Springfeild Publishers.

Kumar, R. & Lal, K. (2024).Teacher Made Test.*International Journal of Creative Research Thoughts. 12*, (3) 227-238.

Lee, C. and William, D. (2005). Studying changes in the practice of two developing assessment learning. Teacher Development, 20(5): 265-283.

Migs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational research and Evaluation, 12*(1): 53-74.

Opara, I. M. (2014). Psychological Testing: Principles and Techniques. Owerri: Career publishers.

Rossi, O. (nd). Using AI For Test Item Generation: opportunities and challenges.

Russell, S. J. & Norvig, P. (2021). Artificial Intelligence: A Modern Approach. (4th ed.). Hoboken: Pearson.

Tei-Firstman R. I (2011) Test item Arrangements on Students Test Scores. Unpublished Thesis University of Port Harcourt).

Ukwuije, R.P.I. (2012). Educational Assessment: A Sine Qua Non for Quality Education. *83rd Inaugural Lecture*, University of Port Harcourt, Choba.

Uwah, I. V. (2022). Revalidation of Renault Adaptable Intelligence Test for Use in Nigeria. *Journal of Education in Developing Areas, 14* (2) 31-39.

Weingartz, R. & Sulevmano, N. (nd).How to write better tests with AI. Retrieved from https://aqua-cloud.io/ai-to-write-tests/

Zong, M., & Krishnamachari, B. (2022). A survey on GPT-3. Preprint. https://doi.org/10.48550 /arXiv.2212.00857.

Milicević, V., Lazarova, L. K. & Pavlović, M. J. (2024). The Application of Artificial Intelligence in Education – The Current State and Trends, *International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE), 12*(2), 259-272

Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences, 120*(3/4), 208-227.

Dogan, M. E., Goru Dogan, T., & Bozkurt, A. (2023). The use of artificial intelligence (AI) in online learning and distance education processes: A systematic review of empirical studies. *Applied Sciences, 13*(5), 3056