Kurdish Studies
July 2024
Volume: 12, No: 4, pp 1808 - 1820
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.53555/ks.v12i4.3653

A Comparative Analysis of Psychometric Properties in AI-Generated and Teacher-Made MCQs Test

Muhammad Zeeshan[1], Misbah Iqbal[2], Sahibzada Shamim-ur-Rasul[3]*, Fareeha Sami[4], Ghulam Muhammad Malik[5], Deeba Imdad[6], Ghulam Zainab Sherazi[7]

[1]MPhil Scholar, Institute of Education, University of Sargodha, Pakistan, Email: Engr2mz@gmail.com
[2]Lecturer, Institute of Education, University of Sargodha, Pakistan, Email: misbah.iqbal@uos.edu.pk
[3]*Assistant Professor, Institute of Education, University of Sargodha, Pakistan, Email: shamim.rasool@uos.edu.pk
[4]Assistant Professor, Department of Education, University of Lahore, Sargodha Campus, Pakistan,
Email: fareeha.sami@ed.uol.edu.pk
[5]Lecturer, Institute of Education, University of Sargodha, Pakistan, Email: ghulam.muhammad@uos.edu.pk
[6]Visiting Lecturer, Institute of Education, University of Sargodha, Pakistan, Email: deebaimdad@gmail.com
[7]Lecturer, Institute of Education, University of Sargodha, Pakistan, Email: ghulam.zainab@uos.edu.pk

***Corresponding author:** Sahibzada Shamim-ur-Rasul
*Assistant Professor, Institute of Education, University of Sargodha, Pakistan, Email: shamim.rasool@uos.edu.pk

Abstract

This study aims to compare psychometric properties of AI-generated test and teacher made test. With the increasing use of AI in education, it has become essential to evaluate AI-generated assessments. Achievement test i.e. AI generated and teacher made test of 50 questions each, were used for this study. Validity was examined though three subject experts and three experienced teachers. Based on these experts judgement, relevance scores for items were consistently high across all experts for both tests. The reliability value of items using KR-20 for AI generated test and teacher made test were 0.9492 and 0.9271 respectively. Data was analyzed by using descriptive and inferential statistics. Descriptive statistics such as means scores, difficulty indices, discrimination indices and inferential statistics such as t-test was used for comparison of difficulty and discrimination indices of both tests. The major findings of the study were; on average, validity, reliability, item difficulty level and discrimination index of AI generated test is nearly equal to that of teacher made test. Recommendation of the study is by complimenting AI based test with teacher test, we can reduce work load of teachers while providing consistent assessment across different classes.

Keywords: Multiple choice questions, AI generated test, Teacher made test, Achievement test, AI in Education.

Introduction

Education is fundamental to personal and societal growth and it shapes individuals' knowledge, skills, and values. Education influences people's knowledge, abilities, and values, and is essential to both individual and community evolution. Testing and assessment are essential components of education because they help determine learning gaps and assess student progress. It also help in making informed instructional decisions. Artificial Intelligence (AI) and technology have become indispensable in the present era of education. They provide personalized learning, efficient evaluation, and creative tools that improve overall efficacy of teaching.

Artificial Intelligence (AI) is changing every aspect of our lives, including health care, employment, entertainment, and even education, sparking debates about personalized learning and the future of assessments (Farazouli et al., 2023). Specially, Integration of AI into education has potential in enhancing personalized learning, automating grading, time saving and increasing students satisfaction (Dempere et al., 2023; Sallam et al., 2023). Primary concern in this study is whether or not traditional, teacher-made multiple choice questions (MCQs) tests should be replaced by AI-generated tests.

In higher secondary education, assessment is essential for assessing students' understanding of the objectives of the curriculum. In addition to assessment being extremely expensive, particularly in terms of the time teachers must devote to it, it is also highly challenging to apply the same evaluating criteria to every student's response (Rodrigues & Oliveira, 2014).

There are various reasons why a comparison of tests created by teachers and those generated by AI is valuable. This approach makes it possible to evaluate the effectiveness of AI-generated tests to those created by traditional teachers in a systematic way. AI generated tests will be assessed for their accuracy in measuring what they are intended to measure i.e. validity and their consistency of results i.e. reliability. This comparison can conclude that whether AI generated tests can meet the standards of teacher made tests. By comparing these tests, educators can identify best practices in test creation. This can lead to the adoption of the most effective elements from both approaches. This may improve assessment strategies which is beneficial for both teachers and students. The results of comparative analysis can be used by policymaker and educational leaders for making well-informed judgments. Moreover, educational institutions can decide to what extent they want to include AI generated assessments into their current practices. Furthermore, this study will help in deciding to what extent we should allocate our resources for development and implementation of AI generated assessments.

Objectives of study

The major objectives were

1. What is the validity of AI generated test as compared to teacher made test in assessing student knowledge?
2. What is the reliability of AI generated test as compared to teacher made test in assessing student knowledge?
3. What are differences in term of difficulty indices between teacher made test and AI generated test?
4. What are differences in term of discrimination indices between teacher made test and AI generated test?
5. Is there a significant difference in students' performance between Artificial Intelligence-based tests and teacher-made tests?

Hypotheses of study

Ho1: There is no significant difference in the validity of AI-generated and teacher-made test.
Ho2: There is no significant difference in reliability of AI-generated and teacher-made test.
Ho3: There is no significant difference in the difficulty levels of AI-generated and teacher-made test.
Ho4: There is no significant difference in the discrimination indices of AI-generated and teacher-made test.

This study is significant to teachers, test developers and policy makers. First of all, it fills in a critical gap in the literature by thoroughly examining the use of Artificial Intelligence assessment tool in higher secondary education which adds to the expanding body of research on technology-enhanced learning (Cope et al., 2021). Second, the findings of this research could potentially assist instructors, curriculum designers, and educational officials in making well-informed decisions regarding the feasibility and effectiveness of integrating artificial intelligence into assessment protocols in higher secondary education (Johnson et al., 2016). By addressing the challenges of traditional assessments and using AI, educators may enhance the accuracy, efficiency, and fairness of their assessment.

Review of related literature

Educational assessment is the systematic process of collecting, analyzing, and interpreting data regarding students' performance, knowledge, skills, and capacities (Kubiszyn & Borich, 2024). It makes use of a wide range of techniques and resources to evaluate student learning. It also offer direction to learners in educational settings. Measurement of learning outcomes, identifying areas for improvement, accountability, guidance in curriculum development, support for educational policy, student feedback collection, ensuring fairness and equity, evaluating college and career readiness, and promoting continuous improvement are the goals of educational assessment (Newton, 2007).

These Chatbots powered by artificial intelligence, like ChatGPT and Google Gemini, have the power to revolutionize higher secondary education by making the current assessment procedures simpler and more effective and more objective (Keller et al., 2019).

ChatGPT is a general purpose AI based language model developed by openAI. It offers wide range of applications in education. AI chatbots are example of software applications (Deveci Topal et al., 2021; Salas-Pilco & Yang, 2022) that are able to produce output in no time as per input through natural language dialogue (Agarwal et al., 2022; Angelov et al., 2021). Chatbots merging together artificial intelligence (AI) and Natural Language Processing (NLP) to engage with human users via text or voice at any level of conversation (Pérez et al., 2020; Smutny & Schreiberova, 2020).

OpenAI launched ChatGPT in November 2022. It uses the Generative Pre-trained Transformer (GPT) architecture. This is one of the largest models ever developed. The GPT architecture does natural language parsing based on the context of input text by using a neural network to generate responses. ChatGPT is able to produce human-like response based on input text. This has led to important advances in the language processing domain. It has also changed the way we interact with AI systems. AI simplified the evaluation process by automating the grading of homework, tests, and quizzes. (L. Chen et al., 2020). AI systems can read and assess written responses by using natural language processing. It also provide rapid and trustworthy feedback (Marrone et al., 2023). AI analyzes large volumes of data to provide insights on student performance. These are helpful for educators to find areas that need improvement and monitor student progress (Pedro et al., 2019).

Liu et al. (2022) mentioned that chatbots offer insights for students after they test the answers of students. This will help them improve their learning skills. They also mentioned that chatbots improve students thinking ability. AI can monitor students learning (Ait Baha et al., 2024). Chatbot helps teachers in evaluating students' assignments, scoring and providing timely feedback to students. (X. Chen et al., 2020; Cunningham-Nelson et al., 2019).

Traditional Teacher-Made Tests

Traditional multiple-choice questions (MCQs) are widely used in a range of educational contexts. These teacher made multiple choice questions must be of a high standard to be evaluated effectively. For instance, studies have been conducted in Ghana to evaluate the caliber of mathematics multiple-choice questions (MCQs) created by teachers (Abad et al., 2017). Catley (2014) found that whereas some tasks evaluated highly sophisticated cognitive thinking, others had reasonable degrees of difficulty and distinguished across various skill groups. In a variety of situations, the usage of student-generated MCQs modified by the instructor has been examined. These student-generated MCQs were used to create instructional games and set examinations, resulting in improved performance and student satisfaction (Gyamfi, 2022).

Challenges and Limitations of Teacher-Made Assessments

Designing and marking conventional Multiple-Choice Question (MCQ) assessments provides educators with a variety of challenges (Ryan et al., 2020) . While MCQs are extensively employed due to their efficiency and objectivity, these problems might have an impact on the overall success of the assessment process. Here is an analysis of the main issues faced by educators. Crafting well-constructed and effective MCQ items needs a profound understanding of the topic area, clarity in wording, and the ability to avoid ambiguity or prejudice. Poorly constructed questions can lead to misinterpretation, confusion, and an erroneous assessment of pupils' knowledge (Kaipa, 2021). Also educators should be aware of frequent item writing errors such as double negatives, confusing language, and misleading options. Flawed items can cause confusion and undermine the reliability and validity of assessment outcomes.

Traditional multiple-choice questions may have limits in measuring desired dimensions, particularly when assessing complex concepts or multidisciplinary knowledge (Haataja et al., 2023). The evaluation may lack validity in capturing the diverse nature of specific learning objectives.

AI-Based Educational Assessment

According to Jung et al. (2024), AI based automated grading and scoring display comparable performance to human scoring. This section explores how AI is changing the way we test students. It also explores strengths and weaknesses related to automated scoring.

Jukiewicz (2024) concluded that ChatGPT can grade programming related assignments with time efficiency, quality assessment, unbiased grading and generate feedback. Bsharat and Khlaif (2024) mentioned that AI based assessments offers efficiency and personalization. They also mentioned that AI is capable of providing tailored difficulty level, comprehensive data analysis, objective evaluations and reducing text anxiety. Owan et al. (2023) mentioned that AI can be programmed to recognize and account for common mistakes made by students. And they can also be programmed to highlight the areas that need improvements. AI Automated scoring can save a lot time of teachers by providing quick feedback (Ramesh & Sanampudi, 2022). AI assesses objectively. It can provide consistent feedback and it is unbiased (Kaur et al., 2022; Schwartz et al., 2022). AI plays a significant role in assessment. However, human supervision is required to authenticate complex evaluations and monitor the integrity of the grading process (Alzubaidi et al., 2023; Kaur et al., 2022). Some judgments may need a level of human judgment especially in subjective areas. AI systems struggle to accurately replicate human assessment (Kaur et al., 2022). Additionally, AI systems are prone to biases in training data. Roselli et al. (2019) and Schwartz et al. (2022) mentioned that there is need to discover and remove algorithmic biases which are affecting assessment outcomes.

According to Xu et al. (2023), AI can be used in construing personalized and accurate feedback systems for students. In contrast to traditional assessment, AI give tailored, real-time feedback. It reveals student's unique strengths and weaknesses in comprehension (Ahmad et al., 2020). These strategies are flexible. It can accommodate individual learning styles and preferences of students

Methodology

In our study comparing AI-generated and teacher made MCQs, we used a quantitative research approach. Both types of tests were administered to all students in a controlled setting. Administering both types of MCQs to all students allowed us to directly compare the findings and rule out any potential confounding variables. Furthermore, a quantitative approach enabled us to efficiently collect vast volumes of data and evaluate it using statistical approaches in order to reach meaningful conclusions.

The Population of the Study

The population size of the study was all secondary level students of district Sargodha, Pakistan.

Sampling and Sample of the study

The multi-stage random sampling technique was employed for the study. This sampling will ensure systematic and representative selection of sample from seven tehsils of district Sargodha.

At first stage, from seven tehsils of district Sargodha three tehsils were selected through random sampling method. Three selected tehsils are Sargodha, Bhalwal and Shahpur. At second sage, from these three selected tehsils, institutions offering Physics as elective subject at higher secondary level, were selected through random sampling technique. Lastly, all students enrolled in these institutions were included in the sample. Therefore, total of three hundred and eighty six students were selected from 14 higher secondary institutions altogether from 3 tehsils.

Research Tool

The instruments used in this study were achievement tests. One was generated by AI and other was designed by teacher.

Teacher-Made Test

50 items were developed by teacher. Three main criteria guided the creation of this instrument was the substance of the test, the learning objectives of the students, and the difficulty of the items. The national curriculum of Pakistan from 2006 specified the first two factors, while the researcher conceptually assessed each item's complexity. Subtopics, schemes, and topics comprised the test material. Student learning outcomes are statements that describe the specific knowledge students will gain, skill thy will develop, and abilities they will demonstrate by the conclusion of each topic or subtopic within a course or learning program. Items were developed to operationalize these learning outcomes in terms of item scores. Three level of difficulty i.e. easy moderate and difficulty were used to target the student learning outcomes.

At the beginning of the study, the researcher developed a test consisting of 50 multiple-choice items. Each question offered four answer choices labeled A, B, C, and D. Additionally, the researcher created a separate set of questions using a 4-point scale. Answers to these questions were graded on a descending scale, with 4 points awarded for the best answer and 1 point for the least accurate answer. The items were constructed from Physics curriculum as stated by the Ministry of federal education and professional training.

AI-Generated Test

50 items were generated by AI. At the beginning of the study, the researcher generated a test consisting of 50 multiple-choice items using ChatGPT 3.5 which was freely available by OpenAI. Each question offered four answer choices labeled A, B, C, and D. Additionally, the researcher created a separate set of questions using a 4-point scale. Answers to these questions were graded on a descending scale, with 4 points awarded for the best answer and 1 point for the least accurate answer. The items were constructed from Physics curriculum as stated by the Ministry of federal education and professional training.

Validity of the Instruments

Teacher-Made Test

This instrument consisted of 50 items which were developed by researcher with Test Blue Print. To ensure the test accurately reflects the curriculum and it is appropriate for a multiple-choice test format, test items were reviewed by three subject matter experts and three experienced teachers. The review focused on aligning the items (drawn from the higher secondary school Physics curriculum using a table of specification) with the curriculum, their suitability for the format, and the overall quality of the topics. This was necessary to ensure both content and face validity.

AI-Generated Test

This instrument consisted of 50 items which were generated by AI according to Test Blue Print. To ensure the test accurately reflects the curriculum and it is appropriate for a multiple-choice test format, AI generated test items were reviewed by three subject matter experts and three experienced teachers. The review focused on aligning the items (drawn from the higher secondary school Physics curriculum using a table of specification) with the curriculum, their suitability for the format, and the overall quality of the topics. This was necessary to ensure both content and face validity.

Data Description

The collected data is used as reference in analyzing the quality of both teacher made and AI generated test in Physics subject using IBM SPSS statistics 25 and Microsoft Excel program. Characteristics of the items produced include the level of item difficulty, discrimination power, distractor efficiency, for Physics subjects in academic year 2023-2024.

Item difficulty

Table 1 and Table 2 shows difficulty index for teacher made and AI generated test.

Table 1: Facility index of teacher made test

Item #	Facility Index F1	Facility Index F2	Item #	Facility Index F1	Facility Index F2	Item #	Facility Index F1	Facility Index F2
1	80.15	75.48	18	39.69	52.40	35	42.53	52.40
2	73.20	69.71	19	41.49	54.33	36	13.92	12.02
3	43.30	46.63	20	47.94	56.73	37	66.75	63.94
4	64.43	57.69	21	57.47	58.65	38	40.46	49.04
5	72.16	65.38	22	29.90	25.96	39	44.07	56.73
6	57.99	58.65	23	56.19	59.62	40	51.03	60.10
7	66.49	65.38	24	51.80	57.69	41	23.71	20.67
8	38.14	47.60	25	42.78	52.40	42	54.38	61.06
9	56.96	58.65	26	29.90	25.96	43	39.95	30.29
10	39.95	50.96	27	42.78	56.25	44	32.73	26.92
11	43.30	48.08	28	50.77	59.62	45	47.68	36.54
12	65.72	65.38	29	20.36	15.87	46	60.57	62.50
13	46.91	51.92	30	38.92	56.25	47	49.48	56.73
14	57.47	63.94	31	51.29	58.65	48	44.07	60.10
15	71.65	64.90	32	57.47	62.98	49	55.41	64.42
16	56.19	60.58	33	28.87	26.44	50	40.72	47.60
17	51.03	56.25	34	26.55	39.90			

Table 1 shows the facility index for 50 test items of teacher made test. It ranges from 0 to 100. Higher values indicates easier items and lower values indicates more difficult items. Items with facility index above 80% are considered easy. Because large percentage of students answered them correctly. Only item number *1*, that is italic, has facility index of 80.15. Items with facility index between 20% and 80% are considered as moderate. Forty seven items in Table 1 have facility index between 20% and 80% which are 2-28, 30-35, 37-50.

Items with facility index lower than 20% are considered as difficult. Because small number of students answered them correct. Only item number **29**, **36**, that are bold, have facility index of 15.87 and 12.02.

Table 2: Facility index of AI generated test

Item #	Facility Index F1	Facility Index F2	Item #	Facility Index F1	Facility Index F2	Item #	Facility Index F1	Facility Index F2
51	59.54	57.00	68	49.23	57.50	85	67.78	62.50
52	44.33	29.00	69	25.26	14.50	86	36.34	49.50
53	65.72	62.00	70	63.92	67.50	87	48.71	55.00
54	66.49	62.50	71	56.70	60.50	88	51.55	58.00
55	63.92	64.00	72	70.88	64.50	89	27.32	11.50
56	54.12	56.00	73	54.12	58.00	90	49.74	54.00
57	42.53	44.00	74	58.25	60.50	91	40.72	52.00
58	21.91	17.50	75	20.62	14.00	92	53.87	55.00
59	38.66	41.50	76	67.27	62.00	93	55.41	57.50
60	38.92	49.50	77	46.13	57.50	94	51.29	58.00
61	69.85	63.50	78	53.35	57.00	95	18.56	13.50
62	62.37	69.50	79	40.46	43.50	96	43.30	50.50
63	42.53	55.50	80	61.08	59.00	97	24.48	17.50
64	32.22	41.50	81	57.73	57.00	98	47.42	57.00
65	46.65	52.00	82	38.14	48.50	99	25.77	14.50
66	15.98	9.50	83	23.45	21.50	100	38.40	27.00
67	50.00	54.50	84	52.32	54.00			

Table 2 shows the facility index for 50 test items of AI generated test. It ranges from 0 to 100. Items with facility index above 80% are considered easy. Because large percentage of students answered them correctly. No item falls in this range. Items with facility index between 20% and 80% are considered as moderate. Forty seven items in Table 2 have facility index between 20% and 80% which are 51-57, 59-65, 67-68, 70-74, 76-88, 90-94, 96, 98, 100.

Items with facility index lower than 20% are considered as difficult. Because small number of students answered them correct. Item number **58**, **66**, **69**, **75**, **89**, **95**, **97**, **99**, that are bold, have facility index of 17.50, 15,98, 14,50, 14.00, 11.50, 13.50, 17.50 and 14.50 respectively.

Description for the result of item difficulty for both teacher made test and AI made test can be seen in the following charts.

Figure 1: Item difficulty for teacher made test

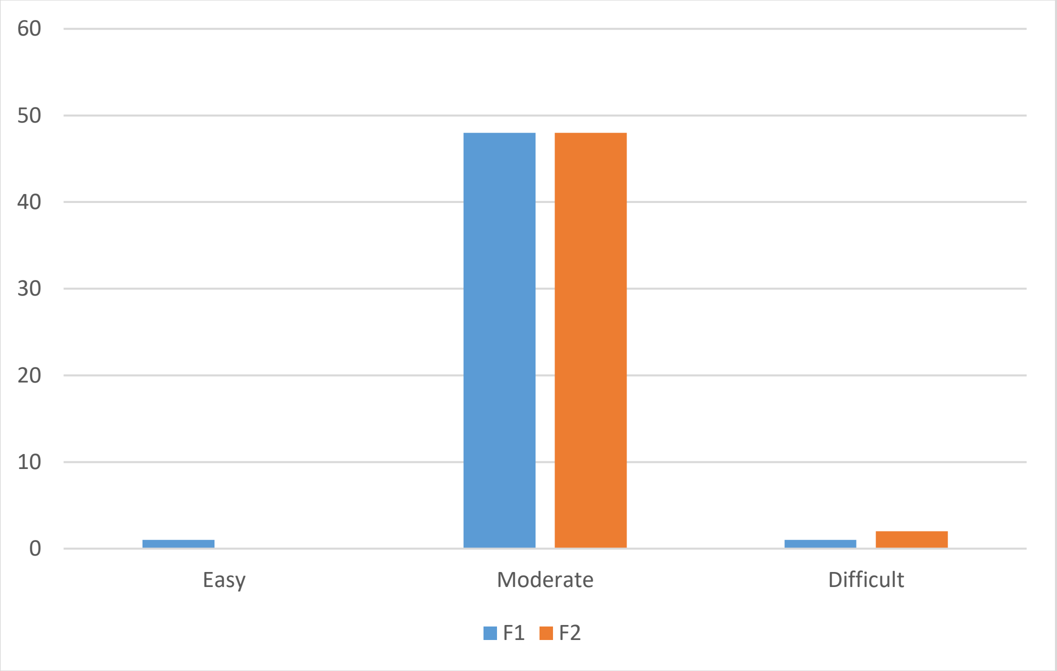
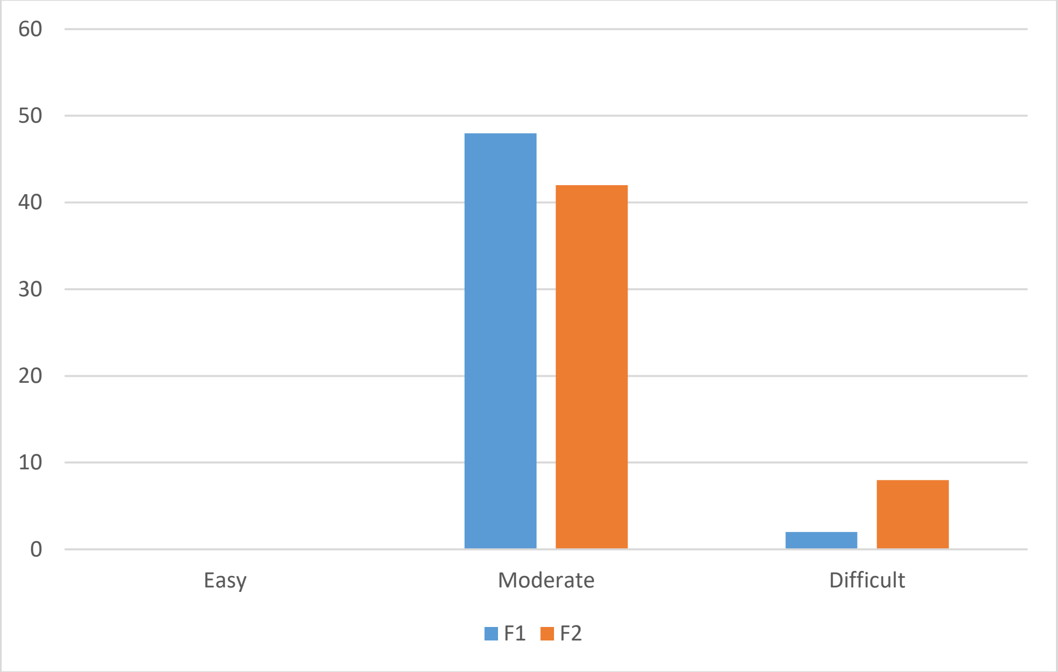


Figure 2: Item difficulty for AI generated test



Based on the data above, it can be seen in teacher made test that highest number of items are moderate 47 (94%). Whereas only one item (2%) can be categorized as easy and two items (4%) can be categorized as difficult. In AI generated test, highest number of items are moderate i.e. 42 (84%). Whereas 8 (16%) item can be categorized as difficult and no item can be categorized as easy.

Table 3: Item difficulty results for teacher made test

Sr. No.	Category	Items	Frequency	Percentage
1	Difficult	29,36	2	4%
2	Moderate	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16, 17,18,19,20,21,22,23,24,25,26,27,28, 30,31,32,33,34,35,37,38,39,40,41,42, 43,44,45,46,47,48,49,50	47	94%
3	Easy	1	1	2%

Table 3 shows that highest number of items are moderate 47 (94%). Whereas only one item (2%) can be categorized as easy and two items (4%) can be categorized as difficult in teacher made test.

Table 4: Item difficulty results for AI generated test

Sr. No.	Category	Items	Frequency	Percentage
1	Difficult	58,66,69,75,89,95,97,99	8	16%
2	Moderate	51,52,53,54,55,56,57,59,60,61,62,63,64, 65,67,68,70,71,72,73,74,76,77,78,79,80, 81,82,83,84,85,86,87,88,90,91,92,93,94, 96,98,100	42	84%
3	Easy	Nil	0	0%

Table 4 shows that highest number of items are moderate i.e. 42 (84%). Whereas 8 (16%) item can be categorized as difficult and no item can be categorized as easy.

From the Table 3 and Table 4, it can be seen that 94% items are moderate in teacher made test and 84% are moderate in AI generated test. In teacher made test only 4% items are difficult whereas 16% items are difficult in AI generated test. Therefore, it can be concluded that teacher made test has performed well in term of difficulty indices of the items.

Item discrimination index

Here are results for item discrimination index for both teacher made test and AI generated test. Now let's analyze each item's discrimination according to the criteria mention in Table 5. This will help us see that which item has difficult, moderate or high discrimination.

Table 5: Item discrimination results for teacher made test

Item #	Discrimination D1	Discrimination D2	Item #	Discrimination D1	Discrimination D2
1	0.47	0.55	26	0.06	0.07
2	0.61	0.66	27	0.80	0.80
3	0.70	0.70	28	0.69	0.71
4	0.79	0.80	29	0.03	-0.36
5	0.69	0.73	30	0.74	0.75
6	0.58	0.59	31	0.75	0.76
7	0.69	0.73	32	0.72	0.75
8	0.49	0.49	33	-0.01	-0.01
9	0.75	0.76	34	0.51	0.52
10	0.56	0.56	35	0.72	0.72
11	0.58	0.58	36	0.02	-0.31
12	0.60	0.63	37	0.57	0.59
13	0.75	0.75	38	0.94	0.94
14	0.47	0.49	39	0.35	0.35
15	0.64	0.67	40	0.59	0.60
16	0.62	0.63	41	-0.28	-0.34
17	0.61	0.61	42	0.74	0.76
18	0.63	0.63	43	0.13	0.14
19	0.61	0.61	44	0.04	0.04
20	0.56	0.56	45	0.06	0.06
21	0.63	0.64	46	0.60	0.62
22	-0.10	-0.11	47	0.56	0.56
23	0.81	0.82	48	0.80	0.81
24	0.38	0.39	49	0.63	0.66
25	0.80	0.80	50	0.51	0.51

Table 5 shows the discrimination indices for 50 test items of teacher made test. It ranges from -1 to 1. Higher values indicates better discrimination between high scorer and low scorer students and lower values indicates poor discrimination. Items with discrimination index above 0.4 have high discrimination. Because these items differentiate between high scorer and low scorer. Thirty nine items fall in this range. These items are 1-21, 23, 25, 27-28, 30-32, 34, 37-38, 40, 42-43, 46-50.

Items with discrimination index between 0.3 and 0.4 are considered moderate discriminating. These items are moderately discriminating between higher scorer and low scorer. Item number **24, 39**, that are bold, have discrimination index of **0.38** and **0.35**.

Items with discrimination index between 0.20 and 0 have low discrimination. These items cannot effectively distinguish between high scorer and low scorer. Item number *26,29,36,43,44,45,* that is bold, have low discrimination index of *0.06, 0.03, 0.02, 0.13, 0.04, 0.06.*

Items with discrimination index below 0 have no discrimination. Negative values of discrimination index indicates that low scorer students are more likely to answer this item correctly than high scorer students. Therefor these items are problematic and should be deleted. Item number ***22, 33, 41,*** that are bold and italic have discrimination index of ***-0.10, -0.01,*** -0.28.

Table 6: Item discrimination results for AI generated test

Item #	Discrimination D1	Discrimination D2	Item #	Discrimination D1	Discrimination D1
51	0.56	0.85	76	0.65	0.72
52	0.22	0.29	77	0.63	0.66
53	0.74	0.78	78	0.77	0.81
54	0.70	0.75	79	0.36	0.39
55	0.67	0.75	80	0.75	0.81
56	0.85	0.87	81	0.75	0.79
57	0.66	0.66	82	0.60	0.63
58	-0.30	-0.28	83	-0.28	-0.28
59	0.16	0.74	84	0.84	0.88
60	0.32	0.37	85	0.67	0.73
61	0.70	0.76	86	0.84	0.85
62	0.56	0.64	87	0.90	0.90
63	0.80	0.84	88	0.78	0.81
64	0.46	0.50	89	-0.07	-0.08
65	0.72	0.74	90	0.84	0.86
66	-0.15	-0.22	91	0.80	0.82
67	0.86	0.89	92	0.84	0.84
68	0.68	0.72	93	0.80	0.84
69	-0.33	-0.38	94	0.75	0.77
70	0.62	0.69	95	-0.28	-0.37

71	0.75	0.79	96	0.56	0.59
72	0.68	0.74	97	0.14	0.20
73	0.79	0.81	98	0.72	0.73
74	0.76	0.79	99	0.04	0.10
75	-0.14	-0.14	100	0.11	0.14

Table 6 shows the discrimination indices for 50 test items of AI generated test. It ranges from -1 to 1. Higher values indicates better discrimination between high scorer and low scorer students and lower values indicates poor discrimination. Items with discrimination index above 0.4 have high discrimination. Because these items differentiate between high scorer and low scorer. Thirty seven items fall in this range. These items are 51, 53-57, 59, 61-65, 67-68, 70-74, 76-78, 80-82, 84-88, 90-94, 96, 98.

Items with discrimination index between 0.3 and 0.4 are considered moderate discriminating. These items are moderately discriminating between higher scorer and low scorer. Item number **52, 60, 79,** that are bold, have discrimination index of **0.22, 0.32,** and **0.36** respectively.

Items with discrimination index between 0.20 and 0 have low discrimination. These items cannot effectively distinguish between high scorer and low scorer. Item number 97, *99, 100* that are italic, have low discrimination index of *0.14, 0.04. 0.11*.

Items with discrimination index below 0 have no discrimination. Negative values of discrimination index indicates that low scorer students are more likely to answer this item correctly than high scorer students. Therefor these items are problematic and should be deleted. Item number ***58, 66, 69, 75, 83, 89, 95*** that are bold and italic have discrimination index of -0.30, -0.15, -0.33, -0.14, -0.28, -0.07, -0.28.

Description for the result of item discrimination for both teacher made test and AI made test can be seen in the following charts.

Figure 3: Item discrimination for teacher made test

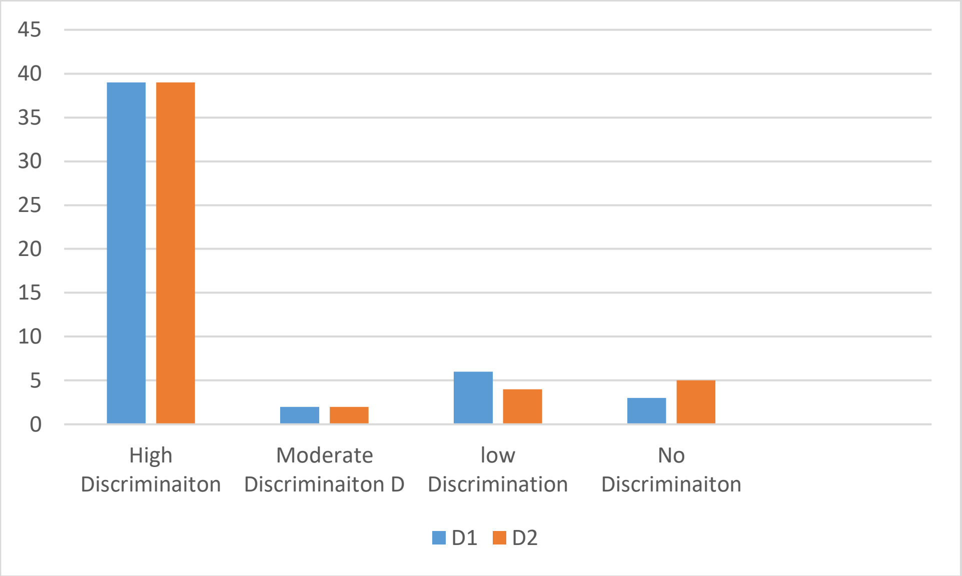


Figure 4: Item discrimination for AI generated test

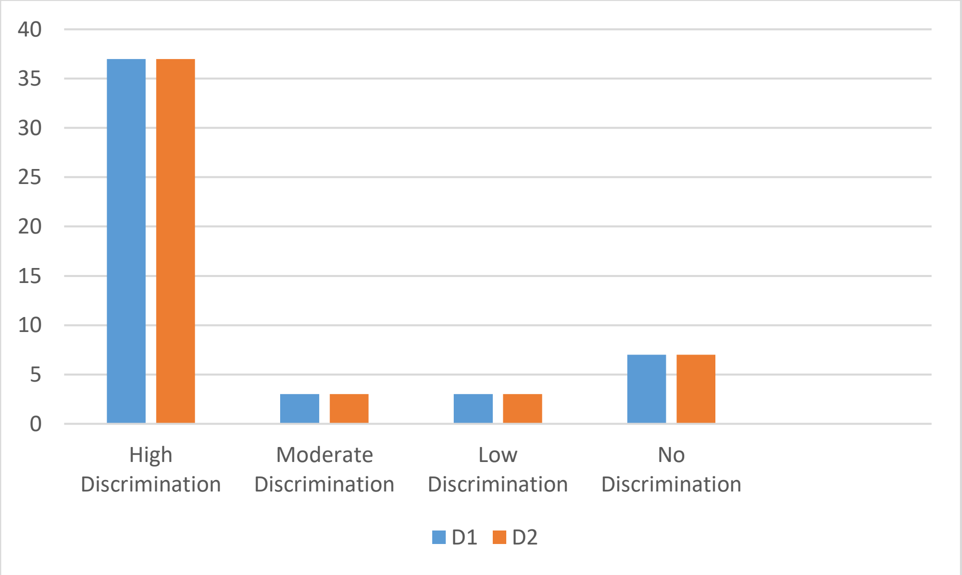


Table 7: Item discrimination results for teacher made test

Sr. No.	Category	Items	Frequency	Percentage
1	High Discrimination	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,18,19,20,21,23,25,27,28,30, 31,32,34,37,38,40,42,43,46,47, 48,49,50	39	78%
2	Moderate Discrimination	24,39	2	4%
3	Low Discrimination	26,29,36,43,44,45	6	12%
4	No Discrimination	22, 33, 41	3	6%

Table 7 shows that majority of items i.e. 39 (78%) have high discrimination. 2 items (4%) have moderate discrimination and 6 items (12%) have low discrimination. Only 3 items (6%) cannot discriminate between high scorer and low scorer.

Table 8: Item discrimination results for AI generated test

Sr. No.	Category	Items	Frequency	Percentage
1	High Discrimination	51, 53, 54, 55, 56, 57, 59, 61, 62, 63, 64, 65, 67, 68, 70, 71, 72, 73, 74, 76, 77, 78, 80, 81, 82, 84, 85, 86, 87, 88, 90, 91, 92, 93, 94, 96, 98	37	74%
2	Moderate Discrimination	52, 60, 79,	3	8%
3	Low Discrimination	97, 99, 100	3	4%
4	No Discrimination	58, 66, 69, 75, 83, 89, 95	7	14%

Table 8 shows that majority of items i.e. 37 (74%) have high discrimination. 3 items (6%) have moderate discrimination and 3 items (6%) have low discrimination. 7 items (14%) cannot discriminate between high scorer and low scorer.

Reliability

Reliability is a characteristic of a good test in which a test consistently measures what it is said to be measure. Reliability of a test can be measured through reliability coefficient which ranges from 0.00 to +1.00. A higher value indicates greater reliability.

Reliability of Teacher made test

Table 9: Values of K, ∑pq and σ^2

K	∑qp	σ^2
50	9.15	100.077

Reliability $= \frac{k}{k-1}\left(1-\frac{\Sigma pq}{\sigma^2}\right)$

Where k = number of items & σ^2 = variance

$Kr_{20} = \frac{50}{50-1}\left(1-\frac{9.15}{100.077}\right)$

$Kr_{20} = \frac{50}{49}(1-0.091429)$

$Kr_{20} = 1.0204\ (0.90857)$

$Kr_{20} = 0.9271$

Reliability of AI base test

Table 10: Values of K, ∑pq and σ2

K	∑qp	σ^2
50	8.00884	114.852

Reliability $= \frac{k}{k-1}\left(1-\frac{\Sigma pq}{\sigma^2}\right)$

Where k = number of items & σ^2 = variance

$Kr_{20} = \frac{50}{50-1}\left(1-\frac{8.00884}{114.852}\right)$

$Kr_{20} = \frac{50}{49}(1-0.06973)$

$Kr_{20} = 1.0204\ (0.93027)$

$Kr_{20} = 0.9492$

Both teacher made and AI generated test values are above 0.90, which indicate that both tests are highly reliable. Since AI generated test also scored more than 0.90 which means that it can produce consistent and repeatable results.

Quantitative Analysis

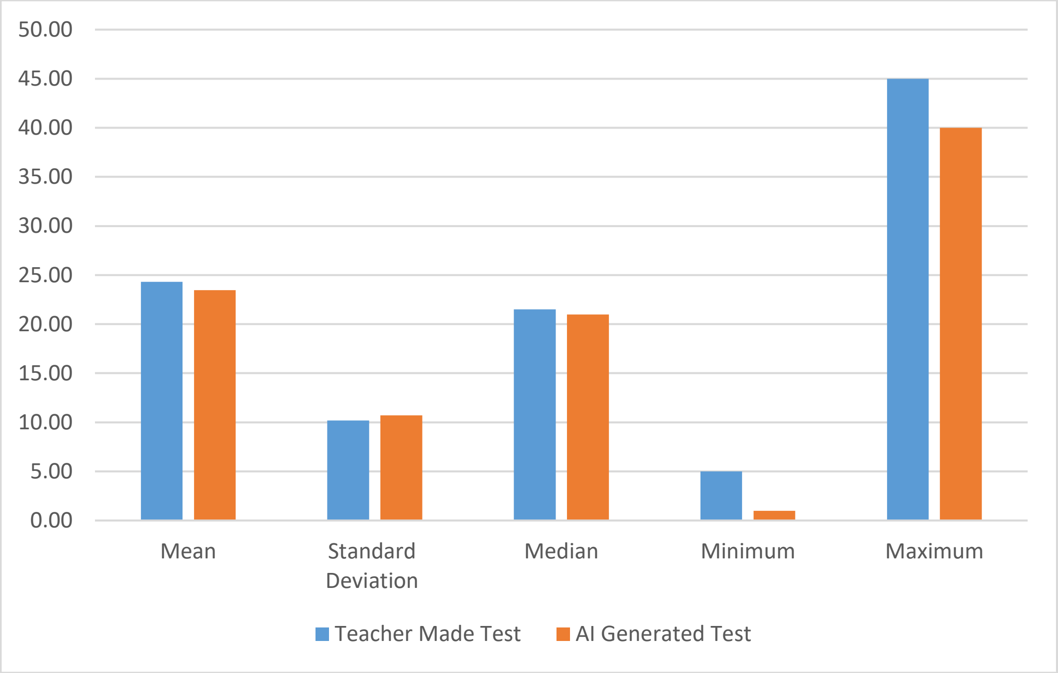
Comparison of Test Scores

Table 13 reports the descriptive statistics for teacher made test and AI generated test scores of higher secondary students. Considering score means, on average the scores for teacher made test are slightly higher than the scores for the AI generated test. Considering standard deviation, both teacher made and AI generated test have similar levels of variability, but the AI generated test has a slightly higher standard deviation which indicates a bit more spread in the scores. The median scores are very close, suggesting that the central point of the distribution of scores is similar for both tests. The lowest score and the highest score for teacher made test is higher than AI generated test.

Table 11: Descriptive statistics for teacher made test and AI generated test

Statistic	Teacher Made Test	AI Generated Test
Mean	24.30	23.47
Standard Deviation	10.19	10.72
Median	21.50	21.00
Minimum	5.00	1.00
Maximum	45.00	40.00

Figure 5: Descriptive statistics for teacher made and AI generated test



Statistical Tests for difficulty and discrimination

Since the scores from the teacher made test and AI generated test are from the same group of students, therefore, samples are not independent. The paired sample t-test is used for this dependency by considering the paired nature of the data.

After the paired sample t-test was performed, the results of the descriptive analysis of the processed data were obtained in Table 12.

Table 12: Pair samples statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Teacher_Difficulty	0.5205	50	0.14745	0.02085
	AI_Difficulty	0.4758	50	0.17502	0.02475
Pair 2	Teacher_Discriminaiton	0.5276	50	0.32313	0.04570
	AI_Discrimination	0.5618	50	0.38278	0.5413

The mean value shows the average value of the data. Average value for teacher made test is slightly higher than the AI generated test for both pairs i.e. item difficulty and discrimination index. The value N indicates amount of data for each variable. In this study N indicates number of items for which the difficulty and discrimination have been calculated.

The results show that the standard deviation of item difficulty for AI generated test is 0.17502 and item difficulty for teacher made test is 0.14745. And standard deviation of item discrimination for AI generated test is 0.38278 and item discrimination for teacher made test is 0.32313. The values of standard deviation is smaller than the average value which means average value represent the overall data well. Table 15 shows paired sample t test results.

Paired samples test output

Pair 1: Teacher made test difficulty and AI generated test difficulty

Table 13 shows that the average difficulty difference between teacher made test and AI generated test is 0.04470. A positive average difference of 0.04470 indicates that, on average the difficulty level for the teacher made test are slightly higher than those from AI generated test. The standard deviation of the differences between the paired scores is 0.22942, indicating variability. Lower value is 0.02050 and upper value is 0.10990 of 95% confidence interval of the difference. It is the range within which we can be 95% confident that eh true mean difference lies. Sine this interval includes zero, it indicates that the true mean difference could be zero. The t-value for the paired samples t-test is 1.378 which indicates that the number of standard errors the mean difference is away from zero. The p-value for the test is 0.175. This value indicates the probability that the observed mean difference is due to random chance. A p-value greater than 0.05 means the difference is not statistically significant.

The mean difference (0.04470) suggests that difficulty scores of teacher made test are slightly higher than AI generated test, but the p-value (0.175) indicates that this difference is not statistically significant. Therefore, it can be concluded that difficulty of AI generated test is nearly equal to the difficulty of teacher made test.

Pair 2: Teacher made test discrimination and AI generated test discrimination

Table 13 shows that the average discrimination difference between teacher made test and AI generated test is -0.0342 which means that on average teacher made test discrimination are 0.0342 lower than AI generated test discrimination scores. A negative average difference of 0.0342 indicates that, on average the discrimination level for the teacher made test are slightly lower than those from AI generated test. The standard deviation of the differences between the paired scores is 0.46734, indicating variability. Lower value is -0.16702 and upper value is 0.09862 of 95% confidence interval of the difference. It is the range within which we can be 95% confident that eh true mean difference lies. Sine this interval includes zero, it indicates that the true mean difference could be zero. The t-value for the paired samples t-test is -0.517 which indicates that the number of standard errors the mean difference is away from zero. The p-value for the test is 0.607. This value indicates the probability that the observed mean difference is due to random chance. A p-value greater than 0.05 means the difference is not statistically significant.

The mean difference (-0.03420) suggests that discrimination scores of teacher made test are slightly higher than AI generated test, but the p-value (0.607) indicates that this difference is not statistically significant. Therefore, it can be concluded that discrimination of AI generated test is nearly equal to the difficulty of teacher made test.

Table 13: Paired Samples Test for difficulty and discrimination

		Paired Differences							
					95% Confidence Interval				
			Std.	Std. Error	of the Difference				
		Mean	Deviation	Mean	Lower	Upper	T	df	Sig. (2-tailed)
Pair 1	Teacher_Difficulty AI_Difficulty	-.04470	.22942	.03244	-.02050	.10990	1.378	49	.175
Pair 2	Teacher_Discrimination - AI_Discrimination	-.03420	.46734	.06609	-.16702	.09862	-.517	49	.607

Conclusion

1. In teacher made test, majority of items (47 out of 50) have moderate difficulty level. This suggest that test is appropriately challenging for most of students. Two items were found to be difficult, and one item was identified as easy. These items require revision. Overall teacher made test have a reasonable balance of difficulty level.
2. In AI generated test, majority of items (42 out of 50) have moderate difficulty level. This suggest that test is appropriately challenging for most of students. Eight items were identified as difficult. These items require revision. Overall AI based test have a reasonable balance of difficulty level.
3. There was no significant difference in difficulty scores of teacher made test and AI based test. Therefore, it can be concluded that difficulty of AI based test is nearly equal to the difficulty of teacher made test.
4. In teacher made test, majority of items (39 out of 50) have high discrimination index. And two items have moderate discrimination index. These items can effectively differentiate between high scorer and low scorer. Whereas six items have low discrimination index and three items have negative discrimination index. These items are problematic. Overall teacher made test has significant number of well discriminating items.
5. In AI based test, majority of items (37 out of 50) have high discrimination index. And three items have moderate discrimination index. These items can effectively differentiate between high scorer and low scorer. Whereas three items have low discrimination index and seven items have negative discrimination index. These items are problematic. Overall AI based test has significant number of well discriminating items.
6. There was no significant difference in discrimination indices of teacher made test and AI based test. Therefore, it can be concluded that discrimination indices of AI based test is nearly equal to the discrimination indices of teacher made test.
7. Both the AI generated test and the teacher made test show excellent reliability in assessing student knowledge. The AI-generated test, demonstrates a marginally higher level of reliability compared to the teacher made test. This suggests that AI generated tests can be highly reliable and potentially even more consistent than traditional teacher-made tests in evaluating student knowledge.

Discussions

There was no significant difference in difficulty scores of teacher made test and AI based test. Therefore, it can be concluded that difficulty of AI based test is nearly equal to the difficulty of teacher made test. This has also been verified by Rezigalla (2024) in his study " AI in medical education: uses of AI in construction type A MCQs" reported that most items fall in moderately difficult range and many items had excellent discrimination indices. These results are contradictory to the findings of Agarwal et al. (2023) in their study " Analysing the applicability of ChatGPT, Bard, and Bing to generate resonating based Multiple-Choice Questions in Medical Physiology" that ChatGPT generated most valid MCQs but least difficult ones. The possible reason for this contradiction in the findings might be that ChatGPT yet lacks the capability of generating MCQs in the field of Physiology at higher education.

Both tests show good content validity, but the teacher test demonstrates higher expert agreement on item relevance. This suggests that AI generated test performed well in term of validity and it can be used for generating assessments at high secondary level for the subject of physics. This has also been verified by Cheung et al. (2023) in their study " ChatGPT versus human in generating medical graduate exam multiple choice questions – A multinational perspective study (Hong Kong SAR, Singapore, Ireland and the United Kingdom)". They reported that questions generated by AI yielded a wider range of scores, while human made questions were consistent and within narrower range.

Both the AI generated test and the teacher made test show excellent reliability in assessing student knowledge. The AI-generated test, with a KR-20 value of 0.9492. It demonstrates a marginally higher level of reliability compared to the teacher made test, which has a KR-20 value of 0.9271. This suggests that AI generated tests can be highly reliable and potentially even more consistent than traditional teacher-made tests in evaluating student knowledge.

Recommendations

1. Difficulty level, discrimination index and reliability of both teacher made test and AI based test is nearly equal, therefore, it is recommended that AI based test can be used along with teacher made test at higher secondary level.
2. By complimenting AI based assessment with tradition assessments, can reduce work load of teachers and can provide consistent assessment across different classes and schools.
3. Policy makers may encourage teachers to incorporate AI generated tests into traditional teacher made assessments. By combining pros of both test and leaving cons, can be used to offer a through assessments of students' performance. Moreover, concrete steps may be ensured by policy makers and item developers for teachers to improve their skills in test design and analysis by using AI tools. This can assist educators in developing evaluations of the highest caliber that are in line with both student learning needs and academic standards.
4. Detailed standards and guidelines for educational assessments may be created and put into effect by policy makers the incorporation of AI based assessment. This will help to improve reliability, fairness, and conformity when using AI-generated assessments in various educational contexts.
5. AI developers may work closely with teachers and educational experts to get technical details and improve assessments. Furthermore, AI developers can get insights from subject experts and psychometricians to make assessment technically sound at different standards.

REFERNCES

1. Abad, E., Gil, J., & Suárez, P. (2017). A game-based educational method relying on student-generated questions. *The International journal of engineering education*, *33*(6), 1786-1797.
2. Agarwal, M., Sharma, P., & Goswami, A. (2023). Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus*, *15*(6).
3. Agarwal, S., Agarwal, B., & Gupta, R. (2022). Chatbots and virtual assistants: a bibliometric analysis. *Library Hi Tech*, *40*(4), 1013-1030.
4. Ahmad, K., Qadir, J., Al-Fuqaha, A., Iqbal, W., El-Hassan, A., Benhaddou, D., & Ayyash, M. (2020). Data-driven artificial intelligence in education: A comprehensive review.
5. Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2024). The impact of educational chatbot on student learning experience. *Education and Information Technologies*, *29*(8), 10153-10176.
6. Alzubaidi, L., Al-Sabaawi, A., Bai, J., Dukhan, A., Alkenani, A. H., Al-Asadi, A., Alwzwazy, H. A., Manoufali, M., Fadhel, M. A., & Albahri, A. (2023). Towards risk-free trustworthy artificial intelligence: Significance and requirements. *International Journal of Intelligent Systems*, *2023*.
7. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.
8. Bsharat, T., & Khlaif, Z. (2024). Generative AI-Powered Adaptive Assessment. In (pp. Pages: 430). https://doi.org/10.4018/979-8-3693-6397-3
9. Catley, P. (2014). Online formative MCQs to supplement traditional teaching: a very significant positive impact on student performance in the short and long run. *Brookes E-Journal of Learning and Teaching*, *6*(1).
10. Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, *8*, 75264-75278.
11. Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100002. https://doi.org/https://doi.org/10.1016/j.caeai.2020.100002
12. Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T.-H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). *PloS one*, *18*(8), e0290691.
13. Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, *53*(12), 1229-1245.
14. Cunningham-Nelson, S., Boles, W., Trouton, L., & Margerison, E. (2019). A review of chatbots in education: practical steps forward. 30th annual conference for the australasian association for engineering education (AAEE 2019): educators becoming agents of change: innovate, integrate, motivate,
15. Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. Frontiers in Education,

16. Deveci Topal, A., Dilek Eren, C., & Kolburan Geçer, A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies*, *26*(5), 6241-6265. https://doi.org/https://doi.org/10.1007/s10639-021-10627-8
17. Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2023). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, 1-13.
18. Gyamfi, A. (2022). Application of Classical Test Theory to the validation of teacher made Mathematics Multiple Choice Test (MMCT) items. *Asian Journal of Advanced Research and Reports*, *16*, 1-12. https://doi.org/10.9734/AJARR/2022/v16i11434
19. Haataja, E. S., Tolvanen, A., Vilppu, H., Kallio, M., Peltonen, J., & Metsäpelto, R.-L. (2023). Measuring higher-order cognitive skills with multiple choice questions–potentials and pitfalls of Finnish teacher education entrance. *Teaching and Teacher Education*, *122*, 103943.
20. Johnson, L., Becker, S. A., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). *NMC horizon report: 2016 higher education edition*. The New Media Consortium.
21. Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity*, *52*, 101522.
22. Jung, J. Y., Tyack, L., & von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *Large-scale Assessments in Education*, *12*(1), 10.
23. Kaipa, R. M. (2021). Multiple choice questions and essay questions in curriculum. *Journal of Applied Research in Higher Education*, *13*(1), 16-32.
24. Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, *55*(2), 1-38.
25. Keller, B., Baleis, J., Starke, C., & Marcinkowski, F. (2019). Machine learning and artificial intelligence in higher education: a state-of-the-art report on the German University landscape. *Heinrich-Heine-Universität Düsseldorf*, 1-31.
26. Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons.
27. Liu, L., Subbareddy, R., & Raghavendra, C. (2022). AI intelligence Chatbot to improve students learning in the higher education platform. *Journal of Interconnection Networks*, *22*(Supp02), 2143032.
28. Marrone, R., Cropley, D. H., & Wang, Z. (2023). Automatic assessment of mathematical creativity using natural language processing. *Creativity Research Journal*, *35*(4), 661-676.
29. Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in education*, *14*(2), 149-170.
30. Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(8), em2307.
31. Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). Artificial intelligence in education: Challenges and opportunities for sustainable development.
32. Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549-1565. https://doi.org/ https://doi.org/10.1002/cae.22326
33. Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495-2527.
34. Rezigalla, A. A. (2024). AI in medical education: uses of AI in construction type A MCQs. *BMC Medical Education*, *24*(1), 247.
35. Rodrigues, F., & Oliveira, P. (2014). A system for formative assessment and monitoring of students' progress. *Computers & Education*, *76*, 30-41.
36. Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in AI. Companion Proceedings of The 2019 World Wide Web Conference,
37. Ryan, A., Judd, T., Swanson, D., Larsen, D. P., Elliott, S., Tzanetos, K., & Kulasegaram, K. (2020). Beyond right or wrong: More effective feedback for formative multiple-choice tests. *Perspectives on Medical Education*, *9*, 307-313.
38. Salas-Pilco, S. Z., & Yang, Y. (2022). Artificial intelligence applications in Latin American higher education: a systematic review. *International Journal of Educational Technology in Higher Education*, *19*(1), 1-20.
39. Sallam, M., Salim, N. A., Barakat, M., & Ala'a, B. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*, *3*(1).
40. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *NIST special publication*, *1270*(10.6028).
41. Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, *151*, 103862. https://doi.org/https://doi.org/10.1016/j.compedu.2020.103862
42. Xu, W., Meng, J., Raja, S. K. S., Priya, M. P., & Kiruthiga Devi, M. (2023). Artificial intelligence in constructing personalized and accurate feedback systems for students. *International Journal of Modeling, Simulation, and Scientific Computing*, *14*(01), 2341001.