AJSTME, Volume. 11, Issue 4. 48-58, November, 2025. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Comparison of Item Parameters of Traditionally Developed and Artificial Intelligence Generated Tests in Basic Science in Kafanchan Education Zone

Agah, J. John & Charles, Seth Bature
Department of Science Education, University of Nigeria, Nsukka
Corresponding Author: sethbature@gmail.com

Abstract

The study compared the Item Parameters of Traditionally Developed and Artificial Intelligence Generated Tests in Basic Science in Kafanchan Education Zone. An instrumentation research design was utilized. JS3 Basic Science Students in public secondary schools constituted the study's population. Multistage sampling technique was used and sample of 1200 respondents employed. The instruments utilized to collect the data were the AIGTIBS and TDTIBS, whose reliability coefficients were 0.92 and 0.86, respectively determined using Kuder-Richardson 20 (K-R20). jMetrik was used to determine the item parameters and ability estimates. The results showed that Basic Science test items for both instruments AI and TD item parameters were within a normal acceptable range. In conclusion, the study suggested that Item Response Theory (IRT) should be use for assessing the psychometric qualities of test items. The recommendations; includes workshops and seminars should be carried out by the government and stakeholders to train Basic Science teachers on the usage of IRT software such as JMetriks and R-package.

Keywords: Item Parameters, Artificial Intelligence Generated Test, Traditionally Developed Test, Item Response Theory.

Introduction

The universality of science and its content has made it possible to assess students all over the world using the same test items at whatsoever level has given rise to the ability of assessing Basic Science content for specific basic levels from any quarter. Thus, it behoves students exposed to test in form of formative and summative assessment to foster learning and acquisition of basic scientific skills. With AI technologies tremendously revolutionising educational sector it's becoming an attraction also in testing with its antecedent potential to enhance and minimise and or eliminates challenges bedevilling teaching and learning of Science and Basic Science achievement in particular (Campbell, Plangger, Sands, & Kietzmann, 2022).

As a result of the inconsistency and decline in students' academic achievement in Basic Science over the years evident by the Basic Science results in the study area on one hand and the dwindling interest towards science and technology as well as the application of scientific and technological knowledge and skills acquisition to meet societal needs (Enemarie, Ogbeba & Ajayi, 2019); Alebiosu & Ifamuyiwa, 2008). This has been attributed to vary variables among which, none in the area of study focused on the item development processes which ascertain the quality of test items, the psychometric quality is reliability and validity under IRT, whose assumptions must be taken into cognisance. The inclusion of AI in education testing brought about a new twist (Alam, 2022) that the researcher desire to explore and answer the following questions: do items generated and those developed possess the psychometric properties/qualities to be assumed valid and reliable? To what extent are the test items parameters being determine? Will the use of AI effectively assess and replaces the traditional test item development in basic science? Are the psychometric qualities of AI generated or traditional developed test items efficient to address students' challenges in basic science? Which of the two tests item will be more effective, accurate, to achieved goals of Basic Science item parameters?

This study hinges on the Item Response Theory (IRT) for assessing item parameters which are the traits and qualities of test items used in educational

evaluations. These traits include the item discrimination (*a*), item difficulty (*b*), guessing (*c*), and carelessness (*d*) parameters of assessment instruments, whose basic tenet is centred on unobserved traits of the examinee. They establish a link between the properties of items on an instrument, individuals responding to these items and the underlying trait being measured through item analysis.

Item analysis is the application of statistical methods in determining the quality of test items. Adom, Mensah and Dake (2020) assert that item analysis is a statistical tool used notably to identify defective test item and to indicate the area where learners have or have not mastered. For this study Item analysis is seen as a statistical method for evaluating the psychometric properties of a test instrument. It is a crucial stage in the creation and selection of test items. Good things are chosen for the final draft using a statistical process, whereas bad items are eliminated (Singh & Yadav, 2018). Thus, it examining the average individual scores on test items and it is used to evaluate the overall quality of the test and individual items within it. In addition, it helps teachers hone their test-developing skills.

Traditionally developed tests, created by human experts, have been widely used in educational assessment and rely heavily on the knowledge and experience of test developers who create items according to established norms and psychometric principles. Basic Science assessment instrument have two major sections, objectives (Multiple Choice Questions (MCQS) and essay sections). This process has been manual and subjective, relying on human judgment, which had led to inconsistencies and biases because it often rely on subjective evaluations, which can be time-consuming, inconsistent, and prone to bias (Tyagi, Fernandez, Mishra, & Kumari, 2020). Due to scientific and technological development people have found way of doing things with ease such as using Artificial Intelligence (AI). According to Dohn, Kafai, Mørch, and Ragni (2022), AI is defined as the capacity of a computer system or program to reason and gain knowledge through experience. Chen, Xiang, and Ren (2021) state that AI is a combination of Machine Learning and Deep Learning with enormous potential in assessment because it is machine-displayed intelligence that mimics human thinking and can be trained to solve particular problems. Assessments in the educational system, enables educators to evaluate students' knowledge and skills effectively and efficiently. The application of AI in this context offers the chance to improve the assessment process's effectiveness, accuracy, and fairness.

With the rising popularity of artificial intelligence (AI), there has been increasing interest in exploring its potential applications in various domains, including test generation. The extent to which AI-generated tests match the quality of traditionally developed tests in the context of Basic Science education remains an area of exploration (Campbell, Plangger, Sands, & Kietzmann, 2022). Artificial Intelligence techniques, offer the possibility of automating test development process and improving item parameter estimation devoid of bias and other traditionally generated limitations. This study therefore, adopted IRT framework to determine the item qualities of Basic Education Certificate Examination in Basic Science in Kafanchan Education zone.

Literature Review

Artificial Intelligence (AI) Generated Tests

AI-generated tests involve using artificial intelligence and machine learning algorithms to automate various stages of the test development process. The integration of AI in test generation has various advantages, including enhanced efficiency, scalability, and the potential for adaptive testing tailored to individual test-takers (Ray, 2023). AI-generated tests are different from traditionally developed item parameters in that they use adaptive testing through real-time item selection based on test-takers' responses, item calibration and analysis to determine item difficulty, and discrimination based on test-takers' performance gathered from a large sample. Since it lessened prejudice in the processes of creating and reviewing items, all test takers' assessments were more equitable (Lee, Fyffe, Son, Jia & Yao, 2023). However, according to Dwivedi et al. (2021), there are drawbacks to using AI for test generation, including issues with item quality and diversity in the algorithms' training data. It also lacks the subjective knowledge and originality that human test developers provide to the process, which could result in a less comprehensive assessment.

Assessment plays a pivotal role in education by providing insights into students' comprehension, facilitating instructional improvements, and guiding educational decision-making. Before AI-generated assessments are used in real-world settings, they must be thoroughly vetted and matched with learning or evaluation objectives. The efficacy of assessments has a direct bearing on students' comprehension of basic

scientific ideas and their general interest in the subject matter when it comes to Basic Science education. For a long time, traditional assessment methods that depend on manually created item criteria have been utilized to measure the knowledge and abilities of students. On the other hand, Alam (2022) believes that as technology advances, the incorporation of artificial intelligence (AI) into teaching methods presents a chance for major improvements in methods of assessment.

Artificial intelligence, nonetheless, holds out the possibility of overcoming these restrictions by means of flexible and individualized assessment processes. AI-powered assessments can adjust the complexity of the questions in real-time in response to student answers, allowing the assessment process to be customized for each individual student (Abbas, Ali, Manzoor, Hussain, & Hussaini, 2023). These assessments use data analytics to find trends, patterns, and learning paths in pupils, making it possible to measure students' cognitive abilities and acquisition of knowledge more precisely and consistently.

Traditionally Developed Test Items

Traditionally developed item parameters refer to the process of designing and creating assessment items, such as test questions or items in a questionnaire, using conventional methods and human expertise. Irwing and Hughes (2018) state that the following are essential phases in the standard method of developing item parameters: item writing, item review, pilot testing, item analysis, and calibration. The lengthy process of developing a traditional item requires human competence at every turn. While it has been the standard for test development for decades, it can be labour-intensive and may have limits in generating huge volumes of items quickly. Traditionally, the process of developing item parameters involved the manual design of test items by subject matter experts. These subject-matter experts are well-versed in the cognitive demands of the test items.

While traditionally developed item parameters have proven to be valuable tools for educators to measure question difficulty, discrimination, and reliability, they are not without limitations. In order to assure validity and fairness, these characteristics frequently require professional judgment and in-depth statistical research. Because of this, the procedure might take a long time, and the subjectivity brought about by human interpretation gives rise to questions regarding possible biases in the assessment's content and results. Furthermore, standard tests are frequently static and do not take individual students' varied cognitive capacities, learning preferences, and competence levels into consideration.

The emergence of AI in assessment prompts an exploration of its potential impact on Basic Science education. This study's primary question is whether exams created by AI can outperform traditional assessments created with item specifications (Rudolph, Tan& Tan, 2023). The increasing prevalence of AI-driven assessments necessitates an evaluation of their validity, reliability, and general usefulness for assessing students' comprehension of Basic Science concepts.

Item Facility/Difficulty Indices (Item Threshold)

Item Difficulty (*b*) also known as item facility or threshold refers to the level of difficulty posed by a test item (question) within an assessment. According to Eleje and Esomonu (2017), the point on the S-shaped curve where the slope is steepest is the b parameter, which measures the difficulty of an item. It is a measure of the proportion of examinees who answered the item correctly (Rush, Rankin & White, 2016). In addition, they use the b-method for interpreting values for difficulty, which includes the Easy: -2.00 to -0.49; Average: 0.50 to 2.00 Simplified: -0.49 to 0.49 whereas more than 2.00 is considered very difficult. When designing tests, it plays a vital role in test development because it allows teachers to create assessments that are reflective of students' strengths and weaknesses.

Item Discrimination Indices (Item Slope)

The discrimination (slope) index (*a*) is the ability to differentiate high and low scoring learners. The discrimination index (a) of an item is a measure of its ability to differentiate between students with high and poor scores; an item with a value closer to 1 is considered to have good discriminatory power (Karadag & Sahin, 2016). When an item on a test is discriminatory in a positive sense, it means that students with higher ability have a better possibility of getting the question right, whereas students with inferior ability have a poorer chance. Negative discrimination works in the same way. They further, make use of Norms Regarding Discrimination (a) Results: 0.01-0.34-based from 0.35 to 1.34 Very high: 2.01 and higher Minimum: 1.35 to 2.00 Very high. Discrimination index determine whether or not the

items are functioning properly will increase the validity and reliability of the tests.

Item Guessing Factor

The guessing parameter or index (*c*) represents the likelihood that a test-taker with very low ability will still answer a multiple-choice item correctly purely by chance, without actually understanding the item content. According to Royal and Stockdale (2017) stated that guessing is a strategy employed by examinees to earn more marks. The more the items in the examination are prone to guessing, the more difficult it becomes to determine the true abilities of the examinees. Thus, guessing is a strategy examinees employ to obtain higher marks unduly. The 2-parameter logistic model (2PLM), the 3PLM, and 4PLM are few examples of IRT models that commonly employ the guessing parameter (Bichi& Talib, 2018). It affects the psychometric properties of the test and increases the variance error of test scores and reduces its reliability.

The Carelessness or Upper Asymptote parameter

The dichotomous IRT model was developed with the proposal of the carelessness or upper asymptote parameter. This parameter forms a four-parameter logistic (4PL) model and accounts for the case where a high-ability examinee incorrectly answers a test item when they should have the ability to correctly answer it (Curran & Denison, 2019). The 4PL IRT model provides an explanation for carelessness, slippage, and guessing. According to Pardede, Santoso, Diki, Retnawati, Rafi, Apino, and Rosyada (2023), the 4PL IRT model of carelessness takes into consideration extra factors that could cause a high-ability examinee to give an incorrect answer on a test item. The 4PLM limits the range of possible positive reactions to an object to -3 to +3, making ensuring that not even those with extremely high levels of a latent trait have a probability close to zero or one. Careless behaviour is defined by Jasper (2024) as when people give answers that don't reflect their true preferences. The 4PLM places a cap on the range of possible positive responses to an item from -3 to +3, not even those with the most extreme latent traits have a chance that is near to zero or one.

Research Questions

1. What are the difficulty indices of AI-generated test items and traditionally developed items in Basic Science?
2. What are the discrimination indices of AI-generated test items, and traditionally developed items in Basic Science?
3. What are the guessing factors of AI-generated test items and traditionally developed test items in Basic Science?
4. What are the carelessness indices of AI-generated test items and traditionally developed test items in Basic Science?

Methods

The design adopted for this study was instrumentation design. Instrumentation design according to Emaikwu (2013) is the development of tools, improved lesson plans, or instructional strategies. The study was conducted in Kafanchan Education zone of Kaduna state. The sample size for the study was 1200 students. The researcher adopted the multi-stage sampling procedure

The instruments for data collection are Artificial Intelligence (AI) generated Basic Science test items and traditionally developed Basic Science test items developed by the researcher. The two instruments were subjected to face validity by five (5) lecturers with the requisite expertise, both the AI generated test items and traditionally developed (TD) test items were subjected to face and content validation, three from Measurement and Evaluation and two from Science Education all in the Department of Science Education, Faculty of Education, University of Nigeria, Nsukka. The students' responses were collected and analysed using Kuder-Richardson 20 (K–R 20), because it is a dichotomously scored. Reliability estimates of 0.92 and 0.86 were obtained for AI generated test items and TD test items respectively, indicating that the instruments were highly appropriate for the level of students (examinees) they are meant for.

AJSTME, Volume. 11, Issue 4. 48-58, November, 2025. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Results

Research Question One: What are the difficulty indices of AI-generated test items and traditionally developed items in Basic Science?

Table 1: Difficulty indices (b) of AI-generated test items and Traditionally Developed items in Basic Science

Items S/N	AI(b)	TD(b)	Items S/N	AI(b)	TD(b)	Items S/N	AI(b)	TD(b)
1	0.53	0.57	**35**	0.81	0.83	**69**	-0.30	0.65
2	0.72	0.60	**36**	-0.55	-0.56	**70**	-0.15	-0.93
3	-41.87	-1.04	**37**	19.34	27.49	**71**	0.54	0.41
4	-0.09	-0.04	**38**	-0.20	-0.15	**72**	27.27	0.33
5	-5.17	-1.99	**39**	0.69	0.65	**73**	0.44	0.64
6	-0.44	-0.11	**40**	-1.00	-0.93	**74**	0.71	0.23
7	-0.41	-0.27	**41**	-0.31	0.41	**75**	19.75	0.37
8	0.21	0.17	**42**	0.57	0.33	**76**	0.08	0.14
9	-0.30	-0.25	**43**	1.77	0.64	**77**	-0.84	-0.23
10	-0.15	0.24	**44**	0.23	0.23	**78**	-1.98	-0.75
11	0.54	0.43	**45**	0.21	0.37	**79**	0.20	-0.53
12	27.27	27.16	**46**	0.28	0.14	**80**	1.08	0.00
13	-0.44	-0.24	**47**	-0.48	-0.23	**81**	0.53	0.57
14	0.71	0.48	**48**	-0.79	-0.76	**82**	0.72	0.60
15	19.75	18.87	**49**	-0.84	-0.53	**83**	-41.87	-1.04
16	0.08	-0.11	**50**	0.12	0.00	**84**	-0.09	-0.04
17	-0.84	-0.67	**51**	-0.17	-0.20	**85**	-5.17	-1.99
18	-1.98	-0.03	**52**	-0.98	-0.75	**86**	-0.44	-0.11
19	0.20	0.07	**53**	5.93	5.87	**87**	-0.41	-0.27
20	1.08	0.83	**54**	-0.21	0.24	**88**	0.21	0.17
21	-0.41	-0.38	**55**	-1.03	-0.68	**89**	-0.30	-0.25
22	0.77	0.80	**56**	-0.774	-0.36	**90**	-0.15	0.24
23	0.45	0.57	**57**	0.17	0.32	**91**	0.54	0.43
24	0.30	0.43	**58**	0.47	0.53	**92**	27.27	27.16
25	0.11	0.45	**59**	1.83	1.79	**93**	-0.44	-0.24
26	-0.22	-0.23	**60**	-0.12	-0.08	**94**	0.71	0.48
27	-0.03	0.09	**61**	0.53	-0.22	**95**	19.75	18.87
28	-0.44	0.47	**62**	0.72	0.20	**96**	0.08	-0.11
29	0.79	-0.51	**63**	-41.87	0.01	**97**	-0.84	-0.67
30	-0.39	-0.18	**64**	-0.09	-0.11	**98**	-1.98	-0.03
31	-0.59	-0.22	**65**	-5.17	0.83	**99**	0.20	0.07
32	0.04	0.20	**66**	-0.44	-0.56	**100**	1.08	0.83
33	-0.09	0.01	**67**	-0.41	27.49			
34	-0.12	-0.11	**68**	0.21	-0.15			

AI = artificial intelligence test items; TD = traditionally develop test items; b= difficulty index

The result in Table 1 shows the 4PLM difficulty indices (b) of AI-generated test items and traditionally developed (TD) test items in Basic Science. The item difficulty values for AI test ranged from -41.87 (items 3, 63, and 83) to +27.27 (item 12, 72 and 92) while item difficulty values for TD test ranged from -1.99 (item 5, 85) to +27.49 (item 67).Eighty-seven (87) items representing 87.0% of AI-generated test and Ninety-three (93) items representing 93.0% of TD developed test fall within the acceptable range of item difficulty -3 to +3. On the other hand, thirteen AI items 3, 5, 12, 37, 53, 63, 65, 72, 75, 83, 85, 92, and 95 and seven TD items 12, 15, 37, 53, 67, 92, and 95 constituting 13.0% and 7.0% respectively were outside the acceptable range. Specifically, items 3, 5, 63, 83, and 85 were very easy while items 12, 15, 37, 53, 72, 75, 92, and 95 were very difficult for the students in AI-generated test. Furthermore, all the items that were

outside the acceptable range for TD test were very difficult for the students. The finding of the study therefore shows that the mean item difficulty indices for AI-generated test items is 0.24 with SD of 9.63 while the mean of TD test items is 1.53 with SD of 5.97 for Basic Science test.

Research Question Two: What are the discrimination indices of AI-generated test items, and traditionally developed items in Basic Science?

Table 2: Discrimination indices of AI-generated test items, and traditionally developed items in Basic Science

Items S/N	AI(a)	TD(a)	Items S/N	AI(a)	TD(a)	Items S/N	AI(a)	TD(a)
1	1.82	2.47	**35**	1.61	1.78	**69**	2.60	1.12
2	2.71	2.50	**36**	2.40	2.40	**70**	2.47	1.37
3	0.03	0.31	**37**	0.82	0.82	**71**	2.21	1.40
4	2.19	2.00	**38**	1.95	1.95	**72**	0.82	2.54
5	0.25	0.32	**39**	2.48	1.12	**73**	2.17	0.91
6	2.65	2.44	**40**	2.17	1.37	**74**	2.67	1.31
7	2.60	2.64	**41**	2.06	1.40	**75**	0.82	1.38
8	2.26	1.92	**42**	2.28	2.54	**76**	1.74	1.19
9	2.60	2.75	**43**	1.33	0.91	**77**	2.79	2.59
10	2.47	1.44	**44**	1.53	1.31	**78**	0.54	1.68
11	2.21	1.98	**45**	1.03	1.38	**79**	1.76	1.55
12	0.82	0.82	**46**	1.08	1.19	**80**	2.64	2.41
13	2.17	1.72	**47**	2.57	2.59	**81**	1.82	2.47
14	2.67	2.22	**48**	1.64	1.68	**82**	2.71	2.50
15	0.82	0.82	**49**	1.82	1.55	**83**	0.03	0.31
16	1.74	2.22	**50**	1.75	2.41	**84**	2.19	2.00
17	2.79	2.70	**51**	1.25	1.65	**85**	0.25	0.32
18	0.54	2.74	**52**	2.30	2.33	**86**	2.65	2.44
19	1.76	2.17	**53**	0.68	1.19	**87**	2.60	2.64
20	2.64	2.60	**54**	1.66	1.42	**88**	2.26	1.92
21	1.87	1.68	**55**	2.78	2.13	**89**	2.60	2.75
22	2.48	2.55	**56**	2.52	2.19	**90**	2.47	1.44
23	1.18	1.18	**57**	0.44	0.44	**91**	2.21	1.98
24	1.71	1.85	**58**	0.64	0.65	**92**	0.82	0.82
25	1.96	1.85	**59**	1.29	1.11	**93**	2.17	1.72
26	2.34	2.03	**60**	2.13	2.45	**94**	2.67	2.22
27	2.17	2.16	**61**	1.82	1.71	**95**	0.82	0.82
28	2.53	1.36	**62**	2.71	1.92	**96**	1.74	2.22
29	0.63	2.42	**63**	0.03	1.70	**97**	2.79	2.70
30	1.58	2.37	**64**	2.19	2.37	**98**	0.54	2.74
31	2.45	1.71	**65**	0.25	1.78	**99**	1.76	2.17
32	1.71	1.92	**66**	2.65	2.40	**100**	2.64	2.60
33	1.52	1.70	**67**	2.60	0.82			
34	1.52	2.37	**68**	2.26	1.95			

a = discrimination indices

The result in Table 2 shows the 4PLM discrimination indices (a) of AI-generated test items and traditionally developed (TD) test items in Basic Science. The item discrimination indices for AI test ranged from 0.03 (items 3 and 83) to 2.79 (item 97) while item discrimination indices for TD test ranged from 0.31 (item 3, 83) to 2.75 (item 89).Ninety-four (94) items representing 95.0% of AI-generated test and Ninety-six (96) items representing 96.0% of TD developed test fall within the acceptable range of item discrimination 0 to +3. On the other hand, six AI items 3, 5, 63, 65, 83 and 85 and four TD items 3, 5, 83 and

85 constituting 6.0% and 4.0% respectively were outside the acceptable range. Particularly, items 3, 5, 63, 65, 83 and 85 of AI and items3, 5, 83 and 85 of TD test were not effective in discriminating the students, Furthermore, all the items that were outside the acceptable range for AI-generated test and TD test were not able to adequately discriminates among the students. The finding of the study therefore indicates that the mean item discrimination indices for AI-generated test items is 1.83 with SD of 0.79 whereas the mean of TD test items is 1.81 with SD of 0.66 for Basic Science test.

Research Question Three: What are the guessing factors of AI-generated test items and traditionally developed test items in Basic Science?

Table 3: the guessing factors of AI-generated test items and traditionally developed test items in Basic Science

Items S/N	AI(c)	TD(c)	Items /N	AI(c)	TD(c)	Items S/N	AI(c)	TD(c)
1	0.20	0.26	**35**	0.50	0.50	**69**	0.02	0.50
2	0.62	0.60	**36**	0.03	0.02	**70**	0.02	0.08
3	0.23	0.50	**37**	0.11	0.19	**71**	0.03	0.29
4	0.01	0.03	**38**	0.03	0.03	**72**	0.10	0.33
5	0.25	0.24	**39**	0.53	0.50	**73**	0.02	0.11
6	0.02	0.02	**40**	0.24	0.08	**74**	0.44	0.11
7	0.02	0.02	**41**	0.32	0.29	**75**	0.19	0.23
8	0.04	0.02	**42**	0.35	0.33	**76**	0.12	0.11
9	0.02	0.02	**43**	0.40	0.11	**77**	0.02	0.08
10	0.02	0.03	**44**	0.19	0.11	**78**	0.37	0.26
11	0.03	0.04	**45**	0.30	0.23	**79**	0.02	0.09
12	0.10	0.10	**46**	0.18	0.11	**80**	0.53	0.24
13	0.02	0.03	**47**	0.09	0.08	**81**	0.20	0.26
14	0.44	0.35	**48**	0.33	0.26	**82**	0.62	0.60
15	0.19	0.17	**49**	0.20	0.09	**83**	0.23	0.50
16	0.12	0.11	**50**	0.33	0.24	**84**	0.01	0.03
17	0.02	0.05	**51**	0.07	0.04	**85**	0.25	0.24
18	0.37	0.58	**52**	0.07	0.04	**86**	0.02	0.02
19	0.02	0.05	**53**	0.09	0.12	**87**	0.02	0.02
20	0.53	0.51	**54**	0.02	0.03	**88**	0.04	0.02
21	0.03	0.05	**55**	0.06	0.00	**89**	0.02	0.02
22.	0.63	0.59	**56**	0.03	0.04	**90**	0.02	0.03
23	0.02	0.03	**57**	0.33	0.33	**91**	0.03	0.04
24	0.04	0.08	**58**	0.50	0.50	**92**	0.10	0.10
25	0.04	0.05	**59**	0.06	0.13	**93**	0.02	0.03
26	0.05	0.03	**60**	0.03	0.05	**94**	0.44	0.35
27	0.00	0.02	**61**	0.20	0.03	**95**	0.19	0.17
28	0.05	0.04	**62**	0.62	0.03	**96**	0.12	0.11
29	0.05	0.03	**63**	0.23	0.03	**97**	0.02	0.05
30	0.08	0.04	**64**	0.01	0.36	**98**	0.37	0.58
31	0.02	0.03	**65**	0.25	0.50	**99**	0.02	0.05
32	0.03	0.03	**66**	0.02	0.02	**100**	0.53	0.51
33	0.07	0.03	**67**	0.02	0.19			
34	0.50	0.36	**68**	0.04	0.03			

c = guessing indices

The result in Table 3 shows the 4PLM guessing indices (c) of AI-generated test items and traditionally developed (TD) test items in Basic Science. The AI-generated test item, guessing indices ranges from 0.00

(item 27) to 0.63 (item 22), while for the TD test, they range from 0.00 (item 55) to 0.60 (items 2, 82). The permissible range of item guessing indices is 0 to +3, and one hundred (100) items, representing 100% of the AI-generated test and TD test, fall within this range. In addition, none of the items were outside the acceptable range for AI-generated and TD test were easy for the students to guess. The finding of the study hence shows that the mean item guessing indices for AI-generated test items is 0.168 with SD of 0.182 whereas the mean of TD test items is 0.169 with SD of 0.177 for Basic Science test.

Research Question Four: What are the carelessness parameters of AI-generated test items and traditionally developed test items in Basic Science?

Table 4: the carelessness parameters of AI-generated test items and traditionally developed test items in Basic Science

Items S/N	AI(u)	TD(u)	Items S/N	AI(u)	TD(u)	Items S/N	AI(u)	TD(u)
1	0.97	0.90	**35**	0.92	0.91	**69**	0.97	0.97
2	0.98	0.95	**36**	0.60	0.60	**70**	0.87	0.98
3	0.91	0.88	**37**	0.60	0.91	**71**	0.97	0.97
4	0.94	0.94	**38**	0.87	0.96	**72**	0.91	1.00
5	0.95	1.00	**39**	0.92	0.97	**73**	0.92	0.89
6	0.94	0.95	**40**	0.90	0.98	**74**	1.00	0.92
7	0.95	0.95	**41**	0.70	0.97	**75**	0.91	0.95
8	0.98	0.90	**42**	0.96	1.00	**76**	0.96	0.94
9	0.97	0.96	**43**	0.70	0.89	**77**	0.97	0.63
10	0.87	0.93	**44**	0.90	0.92	**78**	0.98	0.92
11	0.97	0.92	**45**	0.88	0.95	**79**	0.96	0.98
12	0.91	0.91	**46**	0.92	0.94	**80**	0.98	0.96
13	0.92	0.98	**47**	0.60	0.63	**81**	0.97	0.90
14	1.00	0.97	**48**	0.91	0.92	**82**	0.98	0.95
15	0.91	0.60	**49**	0.92	0.98	**83**	0.91	0.88
16	0.96	0.91	**50**	0.93	0.96	**84**	0.94	0.94
17	0.97	0.95	**51**	0.91	0.70	**85**	0.95	1.00
18	0.98	0.99	**52**	0.94	0.92	**86**	0.94	0.95
19	0.96	0.97	**53**	0.91	0.91	**87**	0.95	0.95
20	0.98	0.91	**54**	0.81	0.91	**88**	0.98	0.90
21	0.96	0.96	**55**	0.70	0.70	**89**	0.97	0.96
22	0.97	0.94	**56**	0.90	0.94	**90**	0.87	0.93
23	0.88	0.94	**57**	0.70	0.70	**91**	0.97	0.92
24	0.88	0.96	**58**	0.60	0.60	**92**	0.91	0.91
25	0.92	0.97	**59**	0.93	0.70	**93**	0.92	0.98
26	0.89	0.93	**60**	0.92	0.90	**94**	1.00	0.97
27	0.92	0.94	**61**	0.97	0.98	**95**	0.91	0.60
28	0.66	0.96	**62**	0.98	0.91	**96**	0.96	0.91
29	0.70	0.60	**63**	0.91	0.96	**97**	0.97	0.95
30	0.91	0.93	**64**	0.94	0.92	**98**	0.98	0.99
31	0.91	0.98	**65**	0.95	0.91	**99**	0.96	0.97
32	0.92	0.91	**66**	0.94	0.60	**100**	0.98	0.91
33	0.95	0.96	**67**	0.95	0.91			
34	0.95	0.92	**68**	0.98	0.96			

u= carelessness parameter

The result in Table 4 shows the 4PLM carelessness parameters (u) of AI-generated test items and traditionally developed (TD) test items in Basic Science. The item carelessness parameters for AI test ranged from 0.60 (items 36, 37, 47 and 58) to 1.00 (items 14, 74 and 94) while item carelessness parameters for TD test ranged from 0.60 (items 15, 36, 58, 66 and 95) to 1.00 (items 5, 42, 72 and 85).One hundred (100) items representing 100.0% of AI-generated test and TD developed test fall within the acceptable range of item carelessness -3 to +3. On the other hand, none of AI-generated and TD test items were outside the acceptable range. The finding of the study thus indicates that the mean item carelessness indices for AI-generated test items is 0.910 with SD of 0.931 while the mean of TD test items is 0.905 with SD of 0.105 for the test.

Discussion of the Findings

That the Basic Science test items for AI and TD fall within the acceptable range of item difficulty 87% and 93% respectively. Test items for both AI-generated and traditionally developed spread across wide range of scale (easy, moderate and difficult) due to the wide coverage of the instruments, the items are drawn from Basic Science curriculum which factor instructional procedures, and the varying level of students' abilities also contributing the outcome of the result. The difficulty level of the items indicates whether items are too easy or too hard for the intended test population. The finding agrees with Rudolph, et al (2019) that item difficulty is a crucial parameter in various models used to estimate examinees' ability and item characteristics. It also aligns with Okoye and Ndubuzor (2016) that item difficulty enhances high quality of test items when they fall within the acceptable range. The finding also corroborates Brown and Abdulnabi (2017) that the more difficult it is for a student to have a 50% chance of correctly answering an item, the higher the ability level needed to achieve this goal.

The items of AI and TD can adequately discriminate between the high ability students from the low ability students.AI-generated and TD test items consists of good items that can effectively discriminates and assess Basic Science content among JSS 3 students as represented by the item discrimination indices for AI-generated and TD test items average mean for Basic Science test. This indicates that the discrimination indices of AI-generated test items and traditionally developed items in Basic Science as both AI and TD can effectively discriminates between the high and low ability students. The findings align with the conclusions drawn by Reynolds, Altman and Allen (2021) regarding the necessity of distinguishing between examinees who exhibit comparable levels of the latent construct in question. In a similar vein, it concurs with the assertions made by Ashraf and Jaseem (2020) affirming that the item discrimination index serves as an indicator of an item's efficacy in differentiating between examinees who possess knowledge and those who do not. Thus, the parameter is crucial for distinguishing between individuals who exhibit comparable levels of the latent construct in question.

Guessing parameters of all the items in Basic Science for the AI and TD were good and the students cannot easily use guessing to provide correct answer to the items. Consequently, it means that examinees are not randomly guessing or providing responses that does not align with their actual knowledge or skill level. The findings agree with Soland, Jensen, Keys, Bi and Wolk (2019) that Guessing raises measurement error by increasing the variance error of test scores and reduce the reliability of the test. Corroborating, Al-zboon and Alharayzeh (2023) that applying the correction equation for the guessing effect increases the value of explained variance and decreases the values of the standard error in estimating the parameters of the equation when guessing has an impact. It also aligns withRoyal and Stockdale (2017) that guessing is a strategy employed by examinees to earn more marks, as the more the items in the examination are prone to guessing, the more difficult it becomes to determine the true abilities of the examinees. Thus, the guessing parameter is essential for accurately interpreting of test scores and making valid inferences about examinee abilities. Therefore, both format of test item development checkmate guessing chances by low ability students.

Carelessness or upper asymptote parameter of all the items for the AI and TD were good and the high ability students cannot easily fail to correctly answer to the items. The study's findings indicate that AI and TD achievement test items are effective tools for validating the assessment of Basic Science content among students enhancing performance. Additionally, they mitigate situations in which a high-ability examinee provides an incorrect response to a test item that they should be able to answer correctly. The finding aligns with Jasper (2024) that careless behaviour manifests when individuals answer questions in a manner inconsistent with their true

preferences, and that the upper asymptote performs better in datasets offering good choice for parameter estimation. Similarly agrees with Nathaniel (2023) that when respondents engage in careless responding, however, responses contain little to no validity. Eventually both format of test item development for Basic Science should be encourage and embraced among teachers for testing scientific content.

Conclusions

The AI-generated test items and traditionally developed (TD) test items if used for the Basic Education Certificate Examination align with the content and objectives of Basic Science in Kafanchan Education Zone would produce a consistent result when used in different test occasions.The items parameters for both AI-generated and TD test items have been established to be of good item difficulties and are suitable to the level of the students they are designed for. The test items equally can sufficiently distinguish between the high ability students from the low ability students, as such has good psychometric qualities. The AI-generated test items and TD test items for both lower (guessing) and upper (carelessness) asymptote fall within the acceptable range of ability estimates. Consequently, the emergence of AI has unified science curriculum to enable students think globally as well as get acquainted to innovations and advancement in science. It behoves the adoption of AI in transforming the educational landscape the world over.

Recommendations

1. Item response theory (IRT) is recommended as the framework to be used in determining the psychometric qualities of tests.
2. Educational Resource Department (ERD) and other relevant bodies should obtain the psychometric qualities of test items in order to sufficiently make useful comparison and reliable decisions on the quality and effectiveness of the examinees that took those tests.
3. Workshops and seminars should be organized by the government and stakeholders to train Basic Science teachers, researchers on the use of IRT software such as jMetriks and R-package for estimate item parameters and ability estimates.
4. That both format of test item development for Basic Science should be embraced among teachers to encourage the universality of testing scientific test content among Basic Science students.

References

Abbas, N., Ali, I., Manzoor, R., Hussain, T., & Hussaini, M. H. A. (2023). Role of Artificial Intelligence Tools in Enhancing Students' Educational Performance at Higher Levels. *Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172*, *3*(05), 36-49.

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education*, *9*(1), 109-119.

Alam, A. (2022). Employing adaptive learning and intelligent tutoring robots for virtual classrooms and smart campuses: reforming education in the age of artificial intelligence. In *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022* (pp. 395-406). Singapore: Springer Nature Singapore.

Al-Zboon, H. S., &Alharayzeh, M. M. I. (2023). The impact of guessing on the accuracy of estimating simple linear regression equation parameters and the ability to predict. *International journal of instruction*, *16*(2).

Ashraf, Z. A. & Jaseem, K. (2020). Classical and modern methods in item analysis of test tools. *International Journal of Research and Review*, *7*(5), 397-403.

Bichi, A. A. & Talib, R. (2018). Item response theory: an introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education*, *7*(2), 142-151.

Brown, G. T. & Abdulnabi, H. H. (2017) Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education* 2, p. 24

Campbell, C., Plangger, K., Sands, S., & Kietzmann, J. (2022) Preparing for an era of deep fakes and AI-generated ads: *A Framework for Understanding Responses to Manipulated Advertising. Journal of Advertising*, 51(1), 22-38.

Chen, S., Xiang, J., & Ren, Z. (2021) To Leverage the Innovation Capability by Co-creation of Human-Machine. *Forest Chemicals Review*, 345-360.

Dohn, N. B., Kafai, Y., Mørch, A. & Ragni, M. (2022) Survey: artificial intelligence, computational thinking and learning. *KI-KünstlicheIntelligenz*, *36*(1), 5-16.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T. & Williams, M. D. (2021).

Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Eleje, L. I. &Esomonu N. P.M. (2017) Diagnostic Quantitative Economics Skill Test for Secondary Schools: Development and Validation Using Item Response Theory Journal of Education and Practice www.iiste.orgISSN 2222-1735 (Paper) ISSN 2222-288X (Online) Vol.8, No.22, 2017

Hongwen, G., Ru, L., Matthew S. J., Dan, F., McCaffrey E. & Princeton, N.J (2022) Alternative methods for item parameter estimation: *From CTT to IRT ETS Research Report Series* doi:10.1002/ets2.12355 ETS RR–22-12

Irwing, P. & Hughes, D. J. (2018) *Test development*. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, 1-47.

Karadag, N. & Sahin, M. D. (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context. *Turkish Online Journal of Educational Technology-TOJET*, *15* (4), 16-24.

Lee, P., Fyffe, S., Son, M., Jia, Z. & Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, *38* (1), 163-190.

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.

Okoye, R. O., &Ndubuzor, J. N. (2016) Empirical Analysis of item difficulty and discrimination indices of national business and technical examination board (NABTEB) economics essay test items from 2013-2015.

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., &Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *REID (Research and Evaluation in Education)*, *9*(1), 86-117.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Reynolds, C. R., Altmann, R. A. & Allen, D. N. (2021) The problem of bias in psychological assessment. In *Mastering Modern Psychological Testing: Theory and Methods*, 573-613 Cham: Springer International Publishing.

Royal, K. D. & Stockdale, M. R. (2017) The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *Journal of Advances in Medical Education & Professionalism*, *5*(2), 84

Rudolph, J., Tan, S. & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6* (1)

Rush, B. R., Rankin, D. C. & White, B. J. (2016) The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC medical education*, *16* (1), 1-10.

Singh, A. & Yadav, D. (2018) Construction and Standardization of achievement test in

Soland, J., Jensen, N., Keys, T. D., Bi, S. Z. & Wolk, E. (2019) Are test and academic disengagement related? Implications for measurement and practice. *Educational Assessment*, *24*(2), 119-134.

Souza, A. C. D., Alexandre, N. M. C. &Guirardello, E. D. B. (2017) Psychometric properties in instruments evaluation of reliability and validity. *Epidemiologia e servicos de saude*, *26*, 649-659.

Taherdoost, H. (2016) Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *How to test the validation of a questionnaire/survey in a research (August 10, 2016)*

Tyagi, A. K., Fernandez, T. F., Mishra, S. & Kumari, S. (2020) Intelligent automation systems at the core of industry 4.0. In *International conference on intelligent systems design and applications* 1-18. Cham: Springer International Publishing.