



**PSYCHOMETRICS PROPERTIES OF ARTIFICIAL INTELLIGENCE  
(CHATGPT) BREED ECONOMICS MULTIPLE CHOICE ITEMS**

<sup>1</sup>Udemba, Esther Chinenye <sup>2</sup>Jacob Esu Odiong <sup>3</sup>Oluwayemisi Damilola  
Akamolafe

<sup>1</sup>Department of Business Education, Faculty of Education, Federal University Oye -  
Ekiti.

email: estherudemba@surefootintlsch.com

<sup>2</sup> Department of Economics, Faculty of humanities and social sciences, Arthur  
Jarvis University, Akpabuyo, Cross River State, Nigeria.

email: jacobesu@gmail.com

<sup>1</sup>Department of Guidance and Counselling, Faculty of Education, Federal University  
Oye-Ekiti

<sup>3</sup>oluwayemisi.akomolafe@fuoye.edu.ng

**Abstract**

This research uncovers the validity, reliability, level of difficulty, and discrimination power of an artificial intelligence (AI) generated of Economics items for secondary students. Items can be prepared by either the teacher (teacher-made), examination bodies (standardised) or now AI-bred tests. Students' responses to AI-generated items were used in this research. A sample of 1,036 students was selected using a random sampling technique. The instrument used for the study was the AI Economics item. The validity and reliability of the instrument were conducted using content and face validity methods, while the Kuder Richardson 20 (Kr-20) method obtained a coefficient that was determined to be 0.76. Data were analysed using R-programming (coding) for p-value and d-value analysis. The rules of thumb used for the judgment on items analysis were P-values Indies from 0.40 to 0.60 = appropriate (easy), 0.00 – 0.39 and < 0.60 = inappropriate (very difficult and too easy), while d-values indices from 0.50 to 0.1 = Good, Indies < 50 and all negative indices = Bad. Results revealed that 22 items (73%) out of the 30 items were appropriately difficulty level (p-value), while eight items (27%) out of the 30 items were inappropriate difficulty (p-value). Also, from the computed d-values, seven items (23%) (30,27,22,12,6,5,2) were bad discriminators, while 23 items (77%) were good discriminators. This implies that Economics items generated by ChatGPT AI are appropriate and good in terms of their difficulty level and discrimination power. Based on the findings, it was concluded that ChatGPT's AI-made Economics items are reasonably appropriate in accuracy and satisfactory regarding the p-value and d-value indices. Based on these recommendations, test constructors should use ChatGPT or other AI in test preparation, administration, and scoring. Also, Economics students should use ChatGPT's AI as a practice tool when preparing for their test.

**Keywords:** Artificial intelligence for e-commerce items, Psychometrics PROPERTIES of ChatGPT-bred items, Item difficulty and discrimination Power.

## **Introduction**

Economics is a social science subject that studies human behaviour as a relationship between ends and scarce resources with alternative uses. It is the study of how society manages scarce resources, such as food, clothing, and housing, among others, and how humans relate to these scarce resources. It is also concerned with choice because humans face the problem of scarcity of resources with which to satisfy their wants, and hence, they are forced to choose which want to satisfy first and which want to satisfy last (Anyawuchi, 2008). The study of economics enables an individual to understand how humans behave when the price of a commodity is high and when it is low. Economics also helps produce, distribute, and consume scarce resources. Studying this subject is critical if Nigeria is to be among the top 20 world economies in 2030 Federal Republic of Nigeria, 2008). It is one of the core subjects for commercial students. The economics test items at the secondary school level are constructed either by the teacher (teacher-made test) or examination bodies (standardised test) such as the West African Examination Council (WAEC) and the National Examination Council (NECO).

Artificial intelligence (AI) is a rapidly evolving field of technology that involves the development of intelligent machines that can perform tasks that typically require human intelligence, such as understanding natural language, recognising patterns, and making decisions based on data. AI refers to the field of computer science that involves creating computer programs capable of imitating intelligent behavior and ideally enhancing human-like abilities (Naqvi, 2020). AI-powered tools and applications are now being used in many industries, including education, to enhance the quality of services provided to students and teachers. AI tools such as Bing and ChatGPT have been referred to as objects individuals can think with, especially in the teaching-learning situation for learners to enhance their ability to think critically and reflectively, foster creativity, acquire problem-solving skills, and grasp concepts effectively (Vasconcelos et al., 2023).

The integration of AI in teaching effectively realised learner-centered learning (Huang, 2018). AI-powered tools and applications improve educational measurement, including testing, assessment, and evaluation. These tools can provide educators valuable insights into student performance, learning outcomes, and instructional effectiveness. For example, AI-powered assessment tools can analyse student responses to assignments and provide personalised feedback to help students identify areas of strengths and weaknesses (Nazaretsky et al., 2022). AI can transform the assessment process in education by examining the use of AI in different stages of test development, such as test purpose determination, test item generation, test administration, scoring, interpretation of test results, and reporting. In addition, AI-powered tools can help automate many aspects of the assessment process, saving time and reducing the burden on teachers. For example, AI-powered grading tools can analyse students' essays and provide feedback on grammar, structure, and content, reducing teachers' time grading assignments.

Testing has remained a vital tool in the hands of teachers. It is an instrument in the hands of any teacher because it is a very relevant tool in educational measurement and has received endless interest in education. The tool enables teachers to place judgment, make decisions, check performance, get response style or picture and determine students' ability. To measure the learning and teaching of any school subject, the measurement instrument must be planned, evaluated and tested to ensure it meets reliability, validity and usability (Ubi & Udemba, 2021). The importance of

tests is high for all the stakeholders in education, such as assessment, placement, accreditation, monitoring, decision making, feedback or reporting, and certification, to mention but a few. Because of the usefulness of tests in education, it is very important that tests developed present quality test items to achieve educational testing purposes.

Reliability and validity are the two most essential aspects to consider when assessing the qualities of any test tool. The reliability of a test measures the degree of consistency of the items in measuring what it is meant to measure. The result should be consistent when a test is presented to any student repeatedly. Reliability refers to a test instrument's internal consistency or stability over time. Thus, it is the relative absence of measurement errors in a test tool. This can be done by applying the split-half method, test-retest method, equivalent form or parallel form method, and internal (inter-item) consistency method. Reliability can be achieved regardless of correctness, i.e., a measurement instrument can give incorrect results so consistently/reliably, to the point that it can be effectively used (Joshua, 2013)

Validity is the degree to which a measuring instrument accurately measures what it is designed to measure. This is why a measuring instrument designed to measure length (e.g., a meter rule) cannot be used to measure weight because the results would not be valid. Validity measures the agreement of test results with the test intended to measure. For instance, somebody can decide to use the height of individuals as a measurement for confidence. Validity and reliability are related. An instrument can be reliable but not valid; however, it cannot be valid if it is unreliable (Joshua, 2013). Educational testing has different kinds of tests, such as essay and objective tests. Multiple-choice questions are regarded to have a high level of reliability since they are scored objectively. The quality of a multiple-choice test instrument can be obtained by its validity and reliability, as well as its level of difficulty and discrimination power.

Psychometric properties measure test items' validity, reliability and item analysis (item difficulty, discrimination power and option distractor). This is done to produce quality items for educational testing. The item's difficulty corresponds to the proportion of correct responses. It is the frequency with which test-takers select the appropriate response. This is done to ascertain whether an item is too easy or hard for the testees. This is known as the p-value, or the facility index, the proportion of examinees that responded correctly to the item (s). The item difficulty index equals the number of students who scored that item right divided by the number of students who attempted the item (Omirin, 2022). The indices (p-value) vary from zero (0) for a complicated item (nobody got it right) to 1(1.00) for a very easy item (everybody got it right). The difficulty coefficient is based on the assumption that a higher difficulty coefficient indicates a more demanding test score. Also, an increase in a test difficulty level results in an increased variability of the test, which is why when a p-value is low, the items are primarily difficult and vice versa.

Item discrimination contrasts the proportion of high and low scorers who are correctly answering a given item. It refers to the degree to which items discriminate between students in the high and low groups. The whole test and each item should assess the same concept. High performers should be more likely to answer a good question correctly, but poor performers should be more likely to do so incorrectly (McCowan & McCowan, 1999). This paper is focused on the validity, reliability, level of difficulty, and discrimination power of AI-generated Economics items.

Day in and day out, many criticisms on educational tests unfold irrespective of who constructed the test. There is criticism that teacher-made tests are poorly constructed and that results obtained from such poorly constructed tests are not valid

and reliable (Wakjissa, 2010). Again, there is criticism that standardised tests are biased and complicated and do not discriminate between dull and bright students (Ubi & Udemba, 2012). This could be one of the reasons for students' poor performance in external examinations. There is, therefore, the need to develop a quality teacher-made achievement test to continuously assess students in the subject to improve students' performance. Some reasons that could be held responsible for students' poor performance in the subject include teachers' poor knowledge of test construction skills which results in poor quality teacher-made economics achievement test EAT that are used in assessing students' achievement in the classroom, teachers' attitude, students' attitude, commitment and teacher's qualification among others.

Based on these criticisms, the need to develop and present quality test items has been pressing in the 21st-century educational assessment. This is why Algorithms driven by machine-learning technologies are now gaining maturity. ChatGPT is one such innovation. ChatGPT is an interactive chatbot created by OpenAI, a California-based artificial intelligence (AI) startup (Anyawuchi, 2018). OpenAI's ChatGPT is a comprehensive language model. ChatGPT AI was trained on a massive corpus of text data using a deep learning algorithm to create replies like a human's for natural language questions (ChatGPT, 2023).

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

The emergence of ChatGPT AI is similar to the emergence of other new innovative technologies that, if used appropriately, have the potential to benefit education. However, ChatGPT AI has the potential to be utilised for activities that are not acceptable in the academic sector. Students, for example, utilise ChatGPT AI to generate assignments such as essays. However, teachers may be able to use AI to spot AI-created works. Test developers can use ChatGPT AI in various ways, including asking information-related questions, confirming data accuracy, reviewing topics, etc. Teachers can request ChatGPT AI to generate multiple-choice questions for tests. Obviously, with its current version, ChatGPT AI has not been able to create an assessment instrument that can accurately measure a learning objective if an expert or teacher does not give it explicit instructions. However, ChatGPT AI can generate complex questions in the future if it has access to a huge amount of data and has received extensive training.

### **Research Question**

What are the difficulty level indices and discrimination power of the Economics items generated by ChatGPT AI?

### **Methodology**

This study's area was Oye-Ekiti, Ekiti State, Nigeria. It adopted a survey research design. The study population comprised all three senior secondary school (SS3) students in public schools in the study area. As of the time of this research, the target

population was 4,674 students in 3 public secondary schools in Oye-Ekiti (Source: Planning Research and Statistics, State Secondary Education Board - February 2024 Academic Session). The study sample comprised 1,036 students selected using a random sampling technique.

The instrument used for collecting data was 30 multiple-choice economics items constructed by ChatGPT AI. Test validity and reliability were not carried out since they are part of the study's analysis. The responses collected from the sample were analysed using content and face validity methods. At the same time, reliability was done using the Kuder Richardson 20 (Kr-20) reliability method, and item analysis was done using R-programming (coding). The rules of thumb used for the judgment on items analysis were P-values indices from 0.40 to 0.60 =appropriate (easy), 0.00 – 0.39 and < 0.60 = inappropriate (very difficult and too easy), while d- values indices from 0.50 to 0.1 = Good, Indies < 50 and all negative indices = Bad.

## Results

### Research Question One:

#### What is the difficulty level and discrimination power of the Economics items generated by ChatGPT AI?

The responses from students were keyed into R-programming software to calculate the p-value and d-values, respectively.

**Rules of thumb** used for the judgment on items analysis were

**P-values indices** 0.40 to 0.60 = appropriate

0.00 – 0.39 and < 0.60 = Inappropriate (very difficult and too easy).

**d- value indices** from 0.50 to 0.1 = Good,

0.00 to 0.49 and all negative indices = Bad.

**Table 1: P –value and d-values indices of Economics items generated by ChatGPT AI?**

Item	P-values	Description	Item	d-values	Description
1	0.45	Appropriate	1	0.65	GOOD

2	0.53	Appropriate	2	-0.23	BAD
3	0.54	Appropriate	3	0.89	GOOD
4	1.00	Not appropriate	4	0.90	GOOD
5	0.01	Not appropriate	5	0.01	BAD
6	0.54	Appropriate	6	0.11	BAD
7	0.47	Appropriate	7	0.98	GOOD
8	0.46	Appropriate	8	0.88	GOOD
9	0.56	Appropriate	9	0.77	GOOD
10	0.59	Appropriate	10	0.69	GOOD
11	0.60	Appropriate	11	0.76	GOOD
12	0.57	Appropriate	12	-0.56	BAD
13	0.89	Not appropriate	13	0.85	GOOD
14	0.90	Not appropriate	14	0.98	GOOD
15	0.52	Appropriate	15	0.75	GOOD
16	0.49	Appropriate	16	0.67	GOOD
17	0.58	Appropriate	17	0.66	GOOD
18	0.59	Appropriate	18	0.51	GOOD
19	0.49	Appropriate	19	0.71	GOOD
20	0.53	Appropriate	20	0.87	GOOD
21	0.91	Not appropriate	21	0.78	GOOD
22	0.92	Not appropriate	22	0.01	BAD
23	0.89	Not appropriate	23	0.89	GOOD
24	0.55	Appropriate	24	0.59	GOOD
25	0.49	Appropriate	25	0.98	GOOD
26	0.10	Not appropriate	26	0.56	GOOD
27	0.59	Appropriate	27	0.22	BAD
28	0.49	Appropriate	28	0.89	GOOD
29	0.46	Appropriate	29	0.69	GOOD
30	0.52	Appropriate	30	0.03	BAD

Results from Table 1 indicated that out of the 30 items, eight items (27%)- (26,23,22,21,14, 13,5,4) had very high or very low p-values, which considered them inappropriate. The remaining 22 items (73%) were appropriate with average p-values. Also, from the computed d-values, seven items (30,27,22,12,6,5,2) were bad discriminators, while the remaining items were good discriminators. This implies that Economics items generated by ChatGPT AI are appropriate and good in terms of their difficulty level and discrimination power.

### Discussion

The study revealed that the 30 Economics items generated by ChatGPT are appropriate and good in terms of difficulty and discrimination. The findings support Ubi and Udemba (2021) who studied the influence of age differentials on item difficulty and discrimination indices of West African School Certificate (WAEC) English Language Objective test for May/June 2014 taken by students in Nigeria and results concluded that Results showed that when items are difficulty and are bad discriminators indices the response of students are affected the quality of a test influence the responses (performance) of students. This result also supports Frank

(2023), who studied the psychometric properties of 2022 NECO mathematics items. Results showed that out of the 40 items, 34 items discriminated the dull from the bright students appropriately, while 16 items did not discriminate correctly among the dull and bright students. It was concluded that the 2022 NECO mathematics items discriminated appropriately among the testees.

### **Conclusion and Recommendations**

Based on students' responses to Economic items generated by ChatGPT AI, 21 of the 30 items were appropriate with average p-values. In comparison, nine items were inappropriate. Only seven items were bad discriminators, and 23 were good discriminators. Based on the findings, it was recommended that;

- (i) Test developers, examination bodies, and teachers should employ AI or AI software in test construction to produce high-quality items.
- (ii) Students should use AIs as a guide in preparing for their various examinations.
- (iii) Other researchers should research how AI can improve educational assessment.

### **References**

- Anyawuchi, R. A. J. (2008). *Fundamentals of economics for senior secondary schools*. Onisha, African First Publishers Limited
- Adebule, S. O & Adaramola, F. M (2020). Analysis of Distractor Indices of Mathematics Items in Ekiti State Unified Examination. *Journal of Psychometry and Assessment Techniques (JOPAT)*. Institute of Education Ekiti State University, Ado-Ekiti, Nigeria.1(1)34- 40.
- ChatGPT. (2023). *ChatGPT*. <https://chat.openai.com/chat>
- Frank U. (2023) Psychometric Properties of 2022 NECO Mathematics items. Unpublished M.Ed Thesis, Faculty of Education University of Calabar.
- Fatima and Anny (2019) Psychometric Properties of Biology teacher made test items. *International Journal of Psychometry vol 1(1)4- 10*.
- Federal Republic of Nigeria (2008). Policy on Education
- Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications. *Accountability in Research*, 1-9. <https://doi.org/10.1080/08989621.2023.2168535>
- Gill, et .al (2021) Validity at of 2018 WASSCE English Languages items. *Journal Psychometrics Assessment*. 2(2), 32-39
- Isaac, I. O. (2011). *Development and validation of psycho-productive skills multiple choice test items in agricultural science for students in secondary schools*. Retrieved 16 August 2024 from: [www.unn.edu.ng](http://www.unn.edu.ng).
- Joshua, M. T.(2013). *Fundamentals of Test and Measurement in Education*. ANITA Press 8 Eyo- Ita Street, Calabar, Nigeria.
- Lewis, et.al (2021) Validity of English language items constructed by teachers Port Harcourt Government Secondary School. *Journal of Educational Assessment*, vol3(2) 12-21
- Naqvi, A. (2020). *Artificial intelligence for audit, forensic accounting, and valuation: A strategic perspective*. John Wiley & Sons. <https://doi.org/10.1002/9781119601906>
- Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program

- to improve it. *British Journal of Educational Technology*, 53(4), 914-931.  
<https://doi.org/10.1111/bjet.13232>
- Suh, W., & Ahn, S. (2022). Development and validation of a scale measuring student attitudes toward artificial intelligence. *Sage Open*, 12(2), 21582440221100463.  
<https://doi.org/10.1177/21582440221100463>
- Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv*.  
<https://doi.org/10.48550/arXiv.2212.09292>
- Tovani-Palone, M. R. (2023). Some challenges and limitations of using ChatGPT in medicine. *Electronic Journal of General Medicine*, 20(5), em503.  
<https://doi.org/10.29333/ejgm/13263>
- Ubi I . O. and Udemba E .C. (2021) Age Differentials In Calibrated Items Of WAEC English Language Objective Test Taken By Students in Nigeria. *Global Journal of Education Research Vol 20, Issue 1 Pages 45-54. 6<sup>th</sup> August.* Publisher AJOL [www.globaljournalseries.com](http://www.globaljournalseries.com); [globaljournalseries@gmail.com](mailto:globaljournalseries@gmail.com).
- Udo O. (2022) Item parameter of 2021 Promotion test items in Akwa Ibom State. Unpublished PhD thesis Faculty of Education University of Jos.
- Vasconcelos, M. A. R., & dos Santos, R. P. (2023). Enhancing STEM learning with ChatGPT and Bing Chat as objects to think with: A case study. *EURASIA Journal of Mathematics, Science and Technology Education*, 19(7), em2296.  
<https://doi.org/10.29333/ejmste/13313>
- Wakjissa, S. G. (2010). Appraisal of senior secondary 2 geography teachers' competency in assessing students Blooms levels of cognitive objectives in plateau state Nigeria. *Journal of Assessment in African*, 5(1), 177 – 188.