Journal of Advanced Guidance and Counseling
Vol. 3 No. 1 (2022), 54-69
DOI: https://doi.org/10.21580/jagc.2022.3.1.11276
Journal Homepage

Comparative study of the psychometric properties of multiple-choice tests using confidence and number right scoring techniques

Jimoh Kasali[1]*, Adediwura Alaba Adeyemi[2]

[1,2]Obafemi Awolowo University, Nigeria
Email: jimoh.bukola@yahoo.com

Article Information:
Received:
6 June 2022
Revised:
13 June 2022
Accepted:
24 June 2022

Keywords:
Confidence scoring, number right, multiple choice items, item discrimination, item difficulty, distracter performance.

Abstract

Purpose - This study aims to ascertain the impact of several options on distractor performance when confidence scoring was used and established the impact of the numbers of options for multiple-choice test items on the reliability of coefficients that were not significant.

Method: This research used a descriptive survey design. There are two sampling methods used in this research, simple random sampling technique and purposive sampling. The instrument used for this study was an adapted version of the 2015 West Africa School Certificate Examination (WASCE) Economics test items. Data collected were analyzed using ANOVA, Kuder -Richardson Formula (KR-20) and Fisher's Z-Test with aid of FZT compotator.

Result - The results of the study showed that several options had a significant impact on distractors' performance when scored using confidence scoring ($F = 6.679$, $p < 0.05$). The results also showed that for each pairwise comparison of 3/4-options (zobt = 0.640), 3/5-options (zobt = 0.837) and 4/5-options (zobt = 0.196) at $p < 0.05$ the difference in the reliability coefficients were not significant.

Implication - This study suggests to improve scoring, the procedure should be encouraged and used in schools because it is effective in reducing the contribution of random guessing to testees' total score and in rewarding testees' partial knowledge on multiple-choice tests. Furthermore, the confidence scoring procedure significantly reduces the 'craze' for a do-or-die affair to pass an examination at all costs and thus should be used in all schools.

Originality - This research is the study to improve scoring procedures that should be encouraged and used in schools.

For citation: Kasali, J,. & Adeyemi, A.A. (2022). Comparative study of the Psychometric properties of multiple-choice tests using confidence and number right scoring techniques. *Journal of Advanced Guidance and Counseling*. *3*(1). 45-69. https://doi.org/10.21580/jagc.2022.3.1.11276
***Corresponding author**: Jimoh Kasali, (jimoh.bukola@yahoo.com), Department of Educational Foundations and Counseling, Obafemi Awolowo University, Ile, 220101, Ife, Nigeria.

Comparative study of the Psychometric properties of multiple-choice tests using confidence and number right scoring procedures

Keywords:
Skor kepercayaan diri, nomor kanan, item pilihan ganda, diskriminasi item, kesulitan item, kinerja pengalih perhatian.

Abstrak

Tujuan - Tujuan dari penelitian ini adalah memastikan dampak jumlah opsi pada kinerja pengecoh ketika penilaian kepercayaan digunakan dan menetapkan dampak jumlah opsi untuk item tes pilihan ganda pada keandalan koefisien tidak signifikan.

Metode - Penelitian ini menggunakan desain survei deskriptif. Ada dua metode pengambilan sampel yang digunakan dalam penelitian ini, yaitu teknik simple random sampling dan purposive sampling. Instrumen yang digunakan dalam penelitian ini adalah versi adaptasi dari soal-soal tes Ekonomi West Africa School Certificate Examination (WASCE) 2015. Data yang terkumpul dianalisis menggunakan ANOVA, Rumus Kuder-Richardson (KR-20) dan Fisher's Z-Test dengan bantuan kompotator FZT.

Hasil - Hasil penelitian menunjukkan bahwa jumlah opsi memiliki pengaruh yang signifikan terhadap kinerja pengecoh ketika diberi skor menggunakan penilaian kepercayaan ($F = 6.679$, $p < 0,05$). Hasil juga menunjukkan bahwa untuk setiap perbandingan berpasangan opsi 3/4 (zobt = 0,640), opsi 3/5 (zobt = 0,837) dan opsi 4/5 (zobt = 0,196) pada $p < 0,05$ selisih koefisien reliabilitas tidak signifikan.

Implikasi - Penelitian ini menyarankan untuk meningkatkan prosedur penilaian harus didorong dan digunakan di sekolah-sekolah karena telah terbukti efektif dalam mengurangi kontribusi tebakan acak terhadap skor total testi dan dalam menghargai pengetahuan parsial testi pada tes pilihan ganda. Selanjutnya, prosedur penilaian kepercayaan secara signifikan mengurangi 'kegemaran' untuk urusan do-or-die untuk lulus ujian di semua biaya, dan dengan demikian harus digunakan di semua sekolah.

Orisinalitas - Penelitian ini adalah studi untuk meningkatkan prosedur penilaian harus didorong dan digunakan di sekolah.

Introduction

In accordance with the rest of the globe, the Nigerian educational institution has prioritized multiple-choice questions as a means of evaluating teaching and learning. Items are one of the most basic notions in classroom processes. Multiple-choice exams have several benefits which have helped them become highly common in psychometrics assessment. Testa, Toscano & Rosato (2018) referred to multiple-choice items as the most commonly used instruments for assessing students' knowledge and skills. A key aspect of this type of assessment is the presence of functioning distractors, i.e, wrong options intended to be plausible for students with lower achievement. These can be tailored to a variety of areas of learning and tasks, and they can be used to assess both rote memory and advanced abilities. These elements' adaptability and versatility are undoubtedly responsible

for their inclusion in many constitutes achievements and ability exams. Multiple-choice exams have a lot of benefits that have helped them become quite popular in psychometrics assessment. According to DiBattista & Kurzawa, (2011) it was reported that the use of standardized and computerized tests for learning evaluation is an interesting and relevant topic for those involved in the learning process, evaluation, and instruction. As far as student assessment is concerned, it is often possible to assemble a pool of multiple-choice elements (MCIs) to be administered during an exam. Given its advantage in reducing testing time, this form of evaluation has become popular and is frequently used in very large university classes.

Multiple-choice items allow testees and testers the most equal chance to demonstrate their knowledge and fairness. Multiple-choice tests are often considered to be the most relevant, adaptable, and helpful sort of objective test items. Multiple-choice examinations are typically regarded as the most dependable due to their uniformity in scoring and fairness to all candidates (Osunde, 2009). Multiple-choice questions assessments deter students from anticipating likely issues and instead urge them to cover all of the subject material in their studies.

Multiple-choice items are efficient to administer; they are easy to score objectively; they can be used to sample a wide range of content; they require a relatively short time to administer (Rodriguez, 2016). Another benefit of multiple-choice examinations was the simplicity of scoring, which means that teachers, scoring machines, and computers can evaluate MC items quickly, precisely, and objectively. Olutola (2015) also believes that multi-choice items are fair in terms of development and grading because they cover a larger range of school content and teaching outcomes.

Because of these important benefits, multiple-choice items continue to have broad appeal and, hence, application in education, despite some potential drawbacks, such as guessing effects and unintentionally exposing students' to wrong information. These benefits have led to the usage of MC testing in sizable applications. A stem that poses an issue, the correct or best response, and various distractors are the three sections of a typical multiple-choice item (ie the wrong or

less appropriate option). Foils or distractors are terms used to describe improper answers. The correct answer is referred to as the key. A straight question or an unfinished statement might be used as the stem. In classroom processes, an item is categorized by the figures based on how the stems are expected to be answered. Examinees must first comprehend the stem before recognizing and selecting the proper solution from among numerous options on recognition items.

True/false, multiple-choice (MC), and matching-type items are the most prevalent types of recognition items. There are five alternatives in some multiple-choice items, some others use four, and some use three. In general, the more options you have, the less likely you are to estimate correctly. In a true/false multiple choice test, for example, the likelihood of guessing right is 12 or 50%, but the probability of guessing properly is 1/5 or 20% if the possibilities are five. So, if there are 50 items, a candidate who is very poor and guesses in all of them is unlikely to get more than 20% of them correct, whilst if there are four options, the possibility of correctly guessing increases to 1/4 or 25%, and if the options are three, the chances increase to 1/3 or 33%. Two distractors, on the other hand, are easier to create or come up with than four. Furthermore, a meta-analysis carried out by Vyas and Supe (2008) showed that the 3-option test does not have any significant advantage/disadvantage in its psychometric properties over 4- and 5-option tests. Generally, researchers who supported the 3-option format argued that developing many response options increases the testing time and is energy- and time-consuming for the authors.

The psychometric traits and qualities of multiple-choice tests determine their utility in achieving testing objectives. The significance of psychometric features of multiple-choice items cannot be overstated, since they are critical in determining item difficulty and discrimination indices. Traditional items and sample-dependent statistics, such as item difficulty and item discrimination, are used in classical test theory. The two statistics that form the cornerstone of classical test theory are item difficulty and item discrimination. The difficulty index of an item is defined by Adewuyi and Olutoun (2001) as the percentage of testees who properly answer the item.) It has a value that ranges from 0 to 1.00. Items with higher difficulty indexes are easier. An item with a difficulty level of 0.75 has been answered

correctly by 75% of the examinees. An item with a difficulty level of 0.35 is properly answered by 35% of the examinees. The Difficulty Index indicates how simple the item was for the students in that group. The question becomes easier as the difficulty index rises, and vice versa. According to Abiri (2006), multiple-choice tests with fewer options have higher difficulty indices than those with a larger number of options. On the other hand, the discrimination index measures the capacity to distinguish between bright and poor students (Alonge 2013). The performance of achievement items is typically assessed in terms of difficulty and discrimination power depending on the theoretical approach, difficulty is assessed differently and is defined as the percentage of correct answers (P-value) in the Classical Test Theory (CTT) approach and as the skill level required to have a 0.5 chance of giving the correct answer in the Rasch modeling approach (De Ayala, 2013). Discrimination power refers to the ability to distinguish between high and low achievers. The right answer must have positive discrimination (Tarrant et al., 2009). When the test consists of MCIs, the performance of distractors must also be considered: implausible options lengthening the duration of the test without improving the accuracy of the assessments (DiBattista and Kurzawa, 2011). The quality of the distractor can be evaluated by the frequency of selection and discrimination. A distractor can be defined as functional when it is intended to be plausible for those students with low achievement. For this reason, a distractor is expected to have negative discrimination and to be selected by at least 5% of the participants (Hingorjo & Jaleel, 2012).

The examiner will be unable to distinguish a test taker who was undecided between only two answers from the one who had no idea what the proper answer was. There have been attempts to address this problem by suggesting novel structures to replace typical MC items or Number Right scoring. Conversely, other scoring systems are used for multiple-choice items, like the traditional number right method, in which learners are directed to provide an option as the answer and are given one point for each right answer. This strategy promotes guesswork, doesn't take partial knowledge into account, and can't identify misunderstandings. Full knowledge and lucky guesses are considered correct, but partial knowledge, ignorance, and preconceptions are considered incorrect (Lau, 2010). A few

learners, according to Richard and Joseph (2013), are simply better at multiple-choice tests than others, and this aptitude can transfer to higher marks even in subject areas where they have little understanding. This could bias assessments and obscure relevant information about a student's level of understanding.

The confidence scoring process is one of the attempts to improve the classic number correct MC format. In reliability, validity, and measurement accuracy, Omirin (2021) recommended that the confidence scoring procedure should be encouraged and used at schools as it is more effective in eliminating random guessing. The discriminating indices of multiple-choice tests are improved by using the confidence scoring approach. The confidence scoring approach revealed that the discriminating values of 3-alternative, 4-alternative, and 5-alternative multiple-choice tests were statistically insignificant (Omirin, 2021).

Confidence scoring procedure is a pedagogical practice involving a modification to the usual ways of conducting low-stakes. Students are asked to state a confidence rating alongside each of their answers to express how certain they are that each answer is correct (Foster, 2016). Each student's score is then calculated as the sum of the confidence ratings for the items that they answered correctly, minus the sum of the confidence ratings for the items that they answered incorrectly. There are many similar and overlapping constructs in the literature relating to confidence in the fine grain size of individual items on an assessment (Marsh et al., 2019). For confidence assessment, a pupil's "confidence of response" may be defined as "how certain they are that the answer that they have just given is correct"(Foster,2016), and this may be represented on a scale from 0 (completely uncertain; ie, just guessing) to 10 (absolutely certain). Since students' scores are calculated by summing these ratings (positively for correct answers and negatively for incorrect answers), it may be reasonable to treat this as a linear scale

Some argue that confidence testing discourages guessing since the scoring methods for some confidence testing systems require an examinee to show his genuine level of assurance in replying to maximize his projected score. To evaluate confidence testing, it must be demonstrated that the technique provides more ability variation to the system than error variation and that any increase in the quantity of knowledge obtained is worth the effort. One of the advantages of CA is

that it is easy to implement, as it does not require redesigning assessment instruments (Barton, 2019; Foster, 2016, Foster et al., 2021).

Any classroom formative assessment method in which students write their answers on paper, or even on mini-whiteboards (McCrea,2019) can easily be modified by asking the students to write a confidence rating from 0 (low) to 10 (high) alongside each answer to indicate how sure they are that they are correct.

Absolute confidence, imperfect knowledge, and random guessing are the three levels. Absolute confidence is a response given with certainty that the evidence released is correct, ie, the testee answers the item based on his confidence in the answer. Partial knowledge is an answer offered with some reservations based on the information provided, whereas blind guessing is a response chosen randomly without any prior knowledge. This scoring technique is excessively time-consuming and inconvenient. One of the most important aspects of an answer is the confidence that comes with it. Misrepresenting one's degree of confidence in a response might have disastrous consequences. Confidence evaluation, on the other hand, is rarely stressed in Nigerian secondary school teaching. Students are encouraged to think about their answers in new ways and to assess their confidence in the answers using the confidence-based scoring approach presented. Each answer is scored on whether it is correct or incorrect, as well as if the student is confident in that answer.

The number of alternatives that should be written for each item is one of the most commonly discussed standards for option development in multiple-choice examinations. To lessen the effect of guessing, it is customarily recommended to use four or five selections per item. Multiple-choice tests are an alternative for evaluating, however, there are some differing viewpoints.

Since the beginning of the objective format's use, one of the most common criticisms and concerns has been that students' results do not accurately reflect genuine achievement unless the scores are modified in some way to lessen the negative impacts of guessing. Guessing and cheating are two serious risks to the validity and reliability of testing. The evaluation is not legitimate, reliable, or fair if examinees answer questions without knowing the content of the items and get

them right only by guessing and cheating. Despite the widespread usage of four or five alternatives per item advocated by many writers and test makers, the majority of research conducted to determine the ideal number of options has concluded that the use of four or five options is the best option. Multiple-choice assessments can be effective tools for assessing pupil comprehension. They are simple and inexpensive to administer and score, provide objective scoring, and may be statistically analyzed to compare student demographics or educational techniques. One significant disadvantage is that the responses alone do not reveal all of the cognitive processes. However, when used in conjunction with student interviews, well-designed examinations can be effective educational assessment instruments. Multiple-choice items exams have long been blamed for having several flaws, including a reduction in reliability and validity due to blind guessing and a failure to acknowledge partial knowledge, particularly when the number right scoring technique is utilized. As a result, it was impossible to distinguish between clever and low kids. Because confidence scoring is not widely used among Nigerian classroom teachers, its capacity to improve the psychometric quality of three-, four-, and five-option multiple-choice examinations has not been thoroughly established empirically, necessitating this study. The objectives of the study are to determine the difference in item difficulty of three, four, and five options multiple-choice test items using confidences and number right scoring procedures, and to determine the difference in discriminating indices of three, four, and five options multiple-choice test items using a confidence scoring and number right scoring procedures. To achieve the stated objectives, two research hypotheses were posed.

Research Method

Research Design

The descriptive survey design was adopted for this study. This entails the process of gathering information from a representative sample of a population. The population for the study comprised Senior Secondary School Students in Osun State. There are 410 Senior Secondary Schools in Osun State. The student population consisted of a total number of 137,083, that consisted of 115.881 from

public schools and 21,402 from private schools with a total number of 69,372 males and 67, 708 females.

The study sample consisted of 360 students selected using a multistage sampling technique from the three senatorial districts of the state. The three senatorial districts in the State include the Osun Central Senatorial District, the Osun East Senatorial District, and the Osun West Senatorial District. From the three senatorial districts in the State, two Local Government Areas (LGAs) were selected using simple random sampling technique. From each of the two LGAs selected, three schools were also selected randomly to make a total of 18 schools. From each school 20 Senior Secondary two (SSII) were selected using the purposive sampling technique, being the best 20 students in a pre-test in each school for the study.

Research Instruments

The main instrument for this study was the 2015 West Africa School Certificate Examination (WASCE) Economics items. The instrument was entitled the "Economics Achievement Test" (EAT). The instrument used for the study was designed into 4 versions which include; a selection test, three-choice test formats of two scoring methods namely, 3-options, 4-options, and 5-options called Type A, Type B, and Type C. Number right and Confidence scoring methods were used to score the 3-options, 4-options and 5-options EAT. The items in the instruments were both adopted and adapted from 2015 WASCE Economics past questions and it covered the entire Economics syllabus from SS1 to SS2. The process of option reduction (ie removal of non-functional or least functioning distractors based on item analysis data) was used to modify the 4-options to the 3-options while the fifth option added to the original 4-options was added by the researcher based on the plausibility of the response to the stem as distractors.

Validation of Instrument

The EAT, which was the 4-options format (Type B) was adapted from the 2015 WASCE Economics items that had been validated for use by the West African Examinations Council. The validity of the other types (3-options and 5-options formats) were determined and scrutinized by experts in Tests and Measurement and Economics teachers in the secondary school to judge their face and content

validity as well as item arrangement. The corrections were incorporated into the final version of the instruments. The EAT instrument of 3-options and 5-options was validated in a pilot study conducted using 40 senior secondary school II Economics students who were not part of the final sample size but in a different study area with similar characteristics. Given the responses of the respondents used in the pilot study, the 3-options and 5-options which consisted of 50 items were subjected to a measure of internal consistency using Kurder-Richardson 21 (K21) to ascertain the reliability of the instrument. The result of the K-21 for both 3-options and 5-options yielded coefficients of 0.79 and 0.83 respectively. This indicated that the 3-options and 5-options EAT still remains internally reliable despite the adaptations made by the researcher in the course of the study.

Methods for Data Collection

The researcher visited the selected secondary schools in Osun State. Letter of introduction was presented to the school principals and if the principals were not around, the Vice principals' academics were given such a letter. The researcher with the permission of school principals and the assistance of the Economics teachers in the selected schools administered the EAT to students offering Economics. The research assistants were teachers from the selected schools, with a minimum qualification of a Bachelor's Degree (B.Sc.Ed and B.Sc). The test administration was conducted under strict examination conditions. The students were thoroughly briefed about the essence of the study and they were encouraged to answer all items in the questions. Initially, an EAT selection test was administered to the Economics students in each school selected for the study with the aid of their Economics teachers. The result obtained from the EAT selection test was used as a pre-test purposely to select the best twenty Economics students. Furthermore, to ensure proper preparation by the selected students, adequate notification was given to the selected students on the specific date of another test administration for subsequent tests. These students were properly monitored and instructed by the researcher on how the questions should be answered and questions asked by the students were properly attended to by the researcher after which, the researcher make sure all questions were collected back from the

students and marked by the researcher. Data collection for the study is expected for 3 weeks.

Techniques for Data Analysis

The data collected from the administered 2015 Economics Achievement Test (EAT) were analyzed using ANOVA. To test hypotheses one, two, and three, students' responses to the 50 items of three, four, and five options were scored using confidence and number right scoring methods. Item analysis was carried out using the Microsoft Excel Package. The results of the item indices were then analyzed using two-way analysis of variance (ANOVA) while hypothesis four was tested using Kuder-Richardson Formula 20 reliability index. The differences in the estimated reliability were determined using Fisher's Z-Test with the aid of FZT compotator. All hypotheses were tested at a 5% level of significance.

Results

Hypothesis 1: There is no significant difference in the item difficulty index of three, four, and five options multiple-choice test items using confidence and number right scoring procedures.

Table 1: Difference difficulty Index of three, four, and five option multiples choice test items using confidence and number right scoring methods

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6753.170 [a]	5	1350.634	10,927	.000
Intercept	351711223	1	351711223	2845,347	.000
Options	5381,616	2	2690,808	21.769	.000
Scoring	511,801	1	511,801	4.140	.043
Option*Scoring	859,753	2	429,876	3.478	.032
Error	36341.115	294	123,609		
Total	394805.508	300			
Corrected Total	43094.285	299			

R Squared = .157 (Adjusted R Squared = .142)

The results as presented in table 1 showed that there is a significant main effect of some options on the difficulty index of multiple-choice test items ($F_{(2, 294)} = 21.77$, $p < 0.05$). This is an indication that the difficulty index of multiple-choice test items significantly depends on the number of options. The result also showed a

significant main effect of the scoring method on the multiple-choice test item difficulty index ($F_{(1, 294)} = 4.14$, $p<0.05$). The scoring method significantly has effects on the difficulty index of multiple-choice test items. Furthermore, the result also showed a significant interaction effect between some options and the scoring method of multiple-choice test items ($F_{(2, 294)} = 3.47$, $p < 0.05$). There was a significant difference in item difficulty of three, four, and five options multiple-choice test items using confidence and number right scoring procedures. Thus, the hypothesis that "there is no significant difference in the item difficulty index of three, four and five options multiple-choice test item using confidence and number right scoring procedures' is rejected.

Hypothesis 2: There is no significant difference in the item discrimination index of three, four, and five options multiple-choice test items using confidence and number right scoring procedures.

Table 2: Difference discrimination Index of three, four, and five option multiples choice test items using confidence and number right scoring methods

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1,391 [a]	5	0.278	6.19	.000
Intercept	23,415	1	23.42	520.552	.000
Options	.299	2	.15	3.32	.037
Scoring	1.088	1	1.09	24.18	.000
option * Scoring	.005	2	.002	.053	.949
Error	13.225	294	.045		
Total	38,031	300			
Corrected Total	14,615	299			

a. R Squared = .095 (Adjusted R Squared = .080)

The result as presented in table 2 showed that there is a significant main effect of some options on the discriminating power of multiple-choice test items ($F_{(2, 294)} = 3.32$, $p < 0.05$). This is an indication that the number of options has a significant effect on the discriminating power of multiple-choice test items. The results as presented in table 2 also showed a significant main effect of the scoring method on multiple-choice test item discriminating power ($F_{(1, 294)} = 24.18$, $p<0.05$). The scoring method significantly has effects on the discriminating power of multiple-choice test items. However, there is no significant interaction effect between some options and the scoring method of multiple-choice test items ($F_{(2, 294)} = 0.053$, $p > 0.05$). The

effect of the scoring method on the item discriminating power of a multiple-choice test item does not depend on some options. Thus, the hypothesis that "there is no significant difference in the item discriminating power of three, four and five options multiple-choice test item using confidence and number right scoring procedures' is accepted.

Discussion of the Findings

The findings of the current study analysis of hypothesis one revealed that the difficulty index of multiple-choice test items is greatly influenced by the number of alternatives when employing the two scoring techniques (confidence and number right scoring procedure). This observation is consistent with the findings of research conducted by (Haladyna, Downing, & Rodriguez, 2002). In terms of the number of alternatives, they discovered 22 relevant papers in 1989 and another 7 in 2002. They discovered that adjusting the number of alternatives has a considerable impact on item difficulty. This result is consistent with the findings of Owolabi and Olatunji (2009), who discovered that the number of options had a substantial impact on the difficulty and discrimination indices of NECO multiple-choice exam problems in Economics. However, in comparable research made by Atalmis and Kingston (2017), he concluded that item difficulty does not differ significantly between MCIs with four options, three options, and none of the above options. Students' test scores on a test with four alternatives are nearly identical to those on a test with three options and none of the above options. This result also revealed that scoring systems (such as confidence and number right) have a considerable impact on the difficulty index of multiple-choice test items. This result corroborated the work of Caldwell and Pate (2013) they reported that items containing none of the above as the correct alternative increased item difficulty but not discrimination power

Conclusion

According to the findings, the difficulty index of multiple-choice test items is significantly influenced by the number of options available when utilizing the two scoring techniques (number right and confidence scoring). The discriminating power of multiple-choice test items is significantly affected by scoring procedures.

Confidence Scoring Method, on the other hand, is better at capturing students' cognitive status in multiple-choice tests and increasing the skills and knowledge of multiple-choice items exams to bring maximum evaluation fairness, effective examination, authentic testing, precise estimation, and higher construct validity and reliability than the number of correct answers.

The study recommended that the confidence scoring procedure should be encouraged and used in schools because it is effective in reducing the contribution of random guessing to testees' total scores and in rewarding testees' partial knowledge on multiple-choice tests. Furthermore, the confidence scoring procedure significantly reduces the 'craze' for a do-or-die affair to pass an examination at all costs and thus should be used in all schools.

References

Abiri, JOO (2006). *Elements of Evaluation Measurement and Statistical Techniques in Education.* Ilorin: Library and Publication committee, University of Ilorin, Nigeria.

Alonge, MF (2013) Assessment and Examination: The Pathways to Educational Development. *Inaugural Lecture.* University of Ado Ekiti.

Atalmis, E. H, & Kingston, N. M (2017). Three, four and none of the above options in multiple choice items. *Turkish Journal of Education, 6(* 4 *), 143-157.*

Barton, C. (2019). Conference takeaways: Research Ed Blackpool 2019. [Audio podcast episode]. Accessed August 4, 2021 fromhttp://www.mrbartonmaths.com/blog/conference-takeaways-researched-blackpool-2019

Caldwell, DJ, and Pate, AN (2013). Effects of question formats on student and item performance. Am. J. Pharm. Educ. 77:71. doi: 10.5688/ajpe77471

De Ayala, RJ (2013). *Theory and Practice of Item Response Theory*. New York, NY: Guilford Press

DiBattista, D., and Kurzawa, L. (2011). Examination of the quality of multiple choice items on classroom tests. can. J. *Scholarsh. Teach. Learn*. 2:4. doi:10.5206/cjsotl-rcacea.2011.2.4

Downing, SM (2006). "SelectFoster, C. (2016). Confidence and competence with mathematical procedures. *Educational Studies in Mathematics, 91(* 2 *), 271–288* .https://doi.org/10.1007/s10649-015-9660-9

Foster, C., Woodhead, S., Barton, C., & Clark-Wilson, A. (2021). School students' confidence when answering diagnostic questions online. *Educational Studies in Mathematics*. https://doi.org/10.1007/ s10649-021-10084-7

Haladyna, TM, Downing, SM, & Rodriguez MC (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education* , *15* (3), 309-334.

Hingorjo, MR, and Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *J. sir. Med. Assoc*. *62*, 142–147.

Lau, SH (2010). *Practicality and robustness of Number Right Elimination Testing (NRET) for multiple-choice items in paper-and-pencil testing (PPT) and computer-based testing (CBT).* Unpublished doctoral dissertation, Universiti Malaysia Sarawak, Malaysia.

Marsh, HW, Pekrun, R., Parker, PD, Murayama, K., Guo, J., Dicke, T., & Arens, AK (2019). The murkydistinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111* (2), *331–353.* https://doi.org/10.1037/edu0000281 .

McCrea, E. (2019). *Making every maths lesson count: Six principles to support great maths teaching*. Crown House Publishing Limited.

Olutola, AT (2015). *Empirical Analysis of Item difficulty and discrimination indices of senior school certificate multiple choice Biology test in Nigeria.* A paper presented at the 41th Annual conference of International Association of Educational Assessment (IAEA) held on 11 th – 15 th October, 2015 at University of Kausa, Lawrence, Kausa USA

Omirin, MS (2021). Discrimination indices of three multiple choice tests using the confidence scoring procedure. *International Journal of Education Research and Reviews 9(* 1 *),*1-4

Osunde, A. (2009). *Essay and multiple choice tests: Bridging the Gap* . Workshop papers on multiple choice test items. Writing procedures for academic staff, University of Ilorin, Ilorin On Monday 4 th Monday, 2009. pp. 14-24

Owolabi, H. O & Olatunji, M. (2009). Characteristics of Multiple Choice Items. Workshop Papers On Multiple Choice Tests Item Writing Procedures For Academic Staff, University Of Ilorin. Ilorin, On Monday, 4th , 2009.

Richard, BG, & Joseph, ML (2013). Inherent limitation of multiple-choice testing. *Academic Radiology* , *20* (10), *1319–1321.*

Rodriguez, MC (2016). "Selected-response item development," in Handbook of Test Development, 2nd Edn, eds S. Lane, M. Raymond, and T. Haladyna (New York, NY: Routledge), 259–273.

Tarrant, M., Ware, J., & Mohammed, AM (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med. Educ. 9:1–8.* doi:10.1186/1472-6920-9-40

Testa S, Toscano A & Rosato R (2018) Distractor Efficiency in an Item Pool for a Statistics Classroom Exam: Assessing Its Relationship With Item Cognitive Level Classified According to Bloom's Taxonomy. Front. Psychol. 9:1585. doi: 10.3389/fpsyg.2018.01585

Thorndike, RM, & Thorndike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education* (8th Ed). Upper Saddle River, NJ: Pearson/Merril Prentice Hall.

Vyas, R., and Supe, A. (2008). Multiple choice questions: a literature review on the optimal number of options. *christmas. Med. J. India 21, 130–133*