Item-analysis of the multiple-choice questions used in the formative assessment of introductory posting examination in Medicine and Surgery at a medical university in Southern Nigeria.

Harry T.C[1], Bozimo G.E.L[2], Dimoko A.A[3], Ikuabe P[4], Kombo B.B[5]

[1]Provost, College of Medicine,
Bayelsa Medical University, Yenagoa, Bayelsa State, Nigeria

[2]Consultant Endocrinologist, Department of Internal Medicine,
Federal Medical Centre, Yenagoa, Bayelsa State, Nigeria

[3]Senior Lecturer/Consultant, General Department of Surgery,
Faculty of Clinical Sciences, Bayelsa Medical University,
Yenagoa, Bayelsa State, Nigeria

[4]Professor of Chest Medicine, Niger Delta University,
Wilberforce Island, Amassoma, Bayelsa State, Nigeria

[5]Head of Department, Department of Surgery
Faculty of Clinical Sciences, College of Medicine,
Bayelsa Medical University, Yenagoa, Bayelsa State, Nigeria

Corresponding author:
Professor Tubonye C Harry, FRCOG, FRCP(Lon)
Provost, College of Medicine,
Bayelsa Medical University, Yenagoa, Bayelsa State, Nigeria
Email: tubonye.harry@bmu.edu.ng
GSM: +2348066503973
Orcid ID: https://orcid.org/0000-0003-3773-5230

How to cite this article:
Harry T.C, Bozimo G.E.L, Dimoko A.A, Ikuabe P, Kombo B.B.: **Item-analysis of the multiple-choice questions used in the formative assessment of introductory posting examination in Medicine and Surgery at a medical university in Southern Nigeria.** NDJMS 2024; 4(2): 32-43

Received 23rd November, 2024 | **Accepted** 21st December, 2024 | **Published** 28th January, 2025

Abstract

Introduction: *The aim of this study was to undertake quality assurance as post-examination analysis of multiple-choice questions (MCQs) used in formative assessment.*

Materials & Methods: *Classical Test theory (CTT) and Item Response Theory (IRT) of 500 items (100 keys & 400 distractors) in single-best answer MCQs (A-type) in introductory medicine (IM)*

and introductory surgery (IS) from 62 medical students was done post-examination. Anonymised answer-scripts had item responses made binary as 0 and 1 and analysed using Microsoft Excel spreadsheet&JMetrik psychometric software to determine difficulty index, discrimination index, distractor efficiency and Cronbach's alpha coefficient.

Results: *The mean score in IM was 60.83 ± 9.48 (95% C.I. 58.42 – 63.24)] Fifty-four students (87.10%) [95% C.I. 76.15 – 94.26] passed and 8 (12.9%0 [95% C.I 5.74 - 23.85] failed. Thirteen (20.7%) attained a score of 70 and above.*

The mean score in IS was 63.5 ± 7.1 (95% C.I 61.70 – 65.32). Sixty (96.77%) [95% C.I 88.83 – 99.61] passed, and 2 (3.23%) [95% C.I 0.39 – 11.17] failed. Twelve (19.3%) scored 70 and above.
Difficulty index (DIF-I) of keys set at <0.3 (too hard) was 10% in IM & 14% in IS. DIF-I > 0.8 (too easy) was 22% in IM and 40% in IS respectively.

Discrimination index (DI) of keys set <0.1 (poor) was 44% in IM & 48% in IS and >0.3(good) was 10% in IM & 2% in IS respectively.
The Cronbach's alpha coefficient was 0.62 in IM & 0.45 in IS respectively.
Nonfunctioning and ineffective distractors (NFD) with a score of zero (0) was 28.8% in IM & 45.2% in IS respectively.

Conclusion: *Item analysis in this study showed many easy questions with poor discrimination, low reliability index and poor distractor efficiency. We recommend post-examination item-analysis as part of quality assurance matrix after formative assessment.*

KEYWORDS*:* item-analysis, discrimination index, difficulty index, distractor efficiency, key assignment, formative assessment.

Introduction

The pioneer clinical class of the College of Medicine, Bayelsa Medical University, commenced their postings in introductory medicine and surgery on 23rd July 2024 in conformance to the regulatory standards of the Medical & Dental Council of Nigeria[1] after completing requirements of the first professional MBBS examination needed for progression as determined by the College of Medicine and ratified by Senate.

As part of the end of postings formative assessment we adopted the use of multiple-choice question(MCQs)[2]. Properly constructed MCQs can assess a wider aspects of the curriculum. Designing MCQs is complicated and time consuming and can be associated with flaws[3]. Multiple-choice questions (MCQs) can assess higher cognitive processing like interpretation, analysis and problem-solving of Bloom's taxonomy of learning [3,4,] when done properly with blueprinting of the curriculum and lesson planning. The choice of MCQs 5 could be either true/false (X-type), single best answer, also referred to as (A-type) with either 4 items (25% of getting right answer by guessing) or 5 items (20% chance of getting it correct) or extended matching questions (R-type) - have its advantages and disadvantages.

In this assessment we opted for 100 single-best answer MCQs with 5 items. Pure luck from guessing the correct answer diminishes when the questions6 are more than 20. The other controversial area is the scoring system[7]. The first is the formula scoring method, the correct answers are awarded a positive point and incorrect or omitted answers are given are given a negative score and the other is the number right scoring method, where incorrect and omitted answers are given no point–8 as a deterrent to guessing. We opted for the number right method in this formative assessment as this was not a high-stake exit summative assessment. We ensured that content expert in blueprinting of MCQs previewed the questions for flaws.

Formative assessment relates to how the assessments inform the students about their performance and enhances learning, underpinning Knowles Theory 9 of andragogy (adult-learning). The Knowles theory of self-directed learning helps learners develop the capacity for self-direction, self-evaluate and self-actualize.

This is a study of the analysis of the raw scores and item analysis of the responses to the multiple-choice questions in introductory medicine and surgery using psychometrics[10]. The aim of the study was to identify any gaps in the assessment and guide the conduct of subsequent assessments with MCQs. It will also provide objective feedback to the students and inform lecturers of unmet needs and deficiencies. One of the key goals of medical education, espoused[11] is to ensure that the true scores approaches the learners true score as reliably and validly as possible.

Materials & Methods:

This was a post-examination quality assurance study evaluating the outcome of formative assessment after two weeks posting in introductory medicine and surgery respectively using Classical Test Theory (CTT) and Item Response Theory (IRT) [11,12]. There were 62 students who participated in the assessment. There were 50 MCQs (A-Type) in Introductory Medicine and Introductory Surgery respectively with a lead-in question and correct option(key answer) to select with four incorrect answers (distractors)[13], to be completed in two hours. Conference marking was undertaken immediately after the examination. Pass mark was set at 50 and there was no negative marking. A score of 70 and above was considered as distinction.

The Departmental Board approved and published results after conduct of the examination. Unique identifiers of matriculation number were extracted from the published results with raw scores. Individual answer scripts containing the candidates chosen answers to the multiple-choice questions were extracted. The correct key items were also collected from the Examination Officer and entered. These data were collated into *Microsoft Excel spreadsheet as* 1 for the correct answer and 0 for a wrong answerand saved as comma separated values (csv) format file.

The file was then, exported to and analysed using *JMetrikpsychometric statistical software*[14] as binary items of (0 & 1). The binary items entered *Excel spreadsheet* was used to compute distractor efficiency as described by other workers[15,16.]

Descriptive statistics was performed including the use of box and whisker charts and histogram to display the scores. Student-t test was used in comparing means. For all statistical analysis $p<0.05$ was considered significant. Where appropriate 95% confidence interval (CI) was used, including 25% and 75% interquartile range (IQR). Discrimination and difficulty index was computed from the binary-items in the csv file using the *JMetrik psychometric statistical software* using Cronbach's alpha reliability coefficient.

The difficulty index[17] is the proportion of learners who answered an item correctly and ranges from 0.0 to 1.0. It compares the performance of 27% high-scorers with that of 27% low-scorers. We adopted the University of Washington, Washington, USA, which classify difficulty index[18] as too hard when the value is less than 0.3, moderate when the range is between 0.3 – 0.8 and very easy when the value is greater than 0.8. Most authorities[19] recommend drop the item if is too difficult or easy.

Item discrimination refers to the ability of an item to differentiate among students because of how well they know the material being tested. It provides an estimate of the degree to which an individual item is measuring the same thing as the rest of the items[20]. Items with low discrimination index[21] are often ambiguously worded and should be examined. Items with negative indices should be examined to determine why a negative value was obtained. We adopted the University of Washington, Washington, USA, which classify item discrimination[18] as "good" if the index is above 0.30; "fair" if it is between 0.10 and 0.30; and "poor" if it is below 0.1 or negative.

Distractor efficiency analysis is the process of evaluating the performance of incorrect answers (distractors) in multiple-choice question items. It is used to access the credibility of distractors. A distractor[11][22] can be defined as functional when it is intended to be plausible for those students with low achievement, that is negative discrimination and to be selected by at least 5% of participants.

The Cronbach's alpha coefficient a measure of internal consistency and an estimator of test reliability[23] and accepted values are equal to or greater than 0.7 as benchmark.

Ethical committee approval was waived as this was a quality assurance study of post-examination answers analysed anonymously and in confidence without contact with any subjects.

Results:

There were 62 students who participated in the formative assessment. The raw scores for introductory medicine showed a mean score of 60.83 ± 9.48 (95% C.I. 58.42 – 63.24). The median score was 60 and IQR [54 -66]. Fifty-four students (87.10%) [95% C.I. 76.15 – 94.26] passed and 8 (12.9%0 [95% C.I 5.74 - 23.85] failed. Thirteen (20.7%) attained a score of 70 and above.

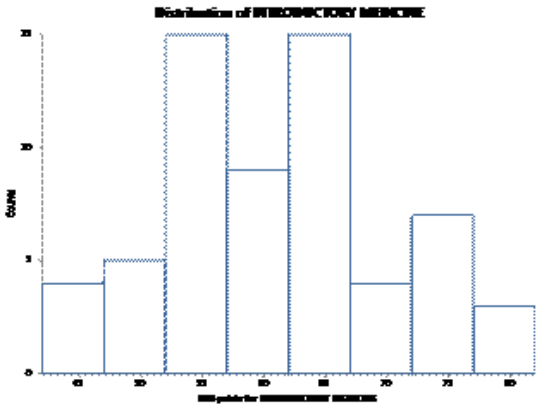

Figure 1

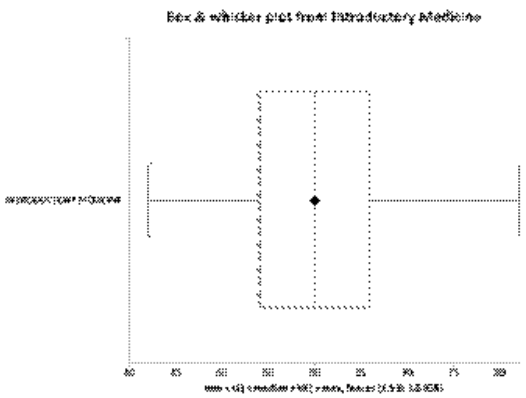

Figure 2

The raw scores of introductory surgery (see figure 3 & 4) showed a mean score of 63.5 ± 7.1 (95% *C.I* 61.70 – 65.32) and the median score was 63, IQR [58 – 68] and range of 48 – 82. Sixty (96.77%) [95% C.I 88.83 – 99.61] passed, and 2 (3.23%) [95% C.I 0.39 – 11.17] failed. Twelve (19.3%) scored 70 and above.

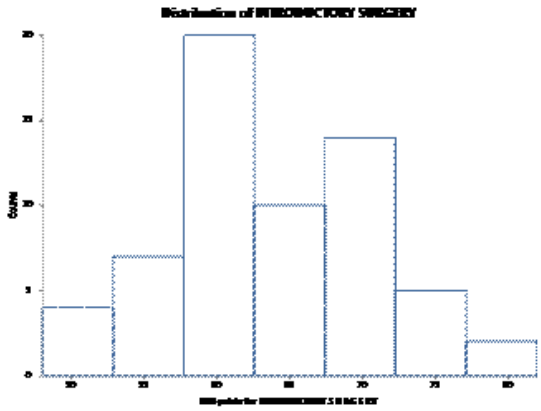

Figure 3

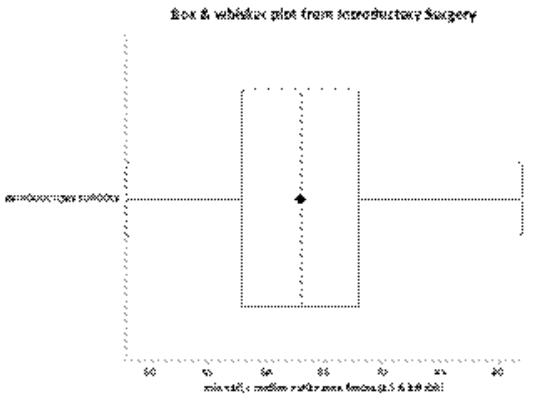

Figure 4

The mean scores in introductory medicine when compared to the mean scores in introductory surgery using the student-t test was statistically significantly different (one-sided p=0.02 and two-sided p=0.04).

Item-analysis for Introductory Medicine multiple-choice questions

In the 50 multiple-choice questions in introductory medicine with a single correct answer (key) the difficulty index of the keys is shown in table 1.

Difficulty Index (DIF-I)	**Difficulty level**	**Number of items**
<.3	Too hard	5 (10%)
0.3-0.8	Moderate	34 (68%)
>.8	Too easy	11 (22%)

Table 1

Items identified as difficult should be dropped from the question bank or re-examined for flaws, which could either be grammatical or poor distractor choice.

Discrimination Index (DI)	**Description**	**Number of items**
<0.1	Poor	22 (44%)
0.1 – 0.3	Fair	23 (46%)
>0.3	Good	5 (10%)

Table 2

Items with a key identified as poor should be discarded, while those identified as fair should be reviewed and modified. Those identified as good should be retained in the MCQ bank. All the items that scored zero should be expunged.

The Cronbach's alpha coefficient a measure of internal consistency and an estimator of test reliability[23] and accepted values are equal to or greater than 0.7 as benchmark. For the 50 keys analysed the Cronbach's alpha coefficient was calculated as 0.62. This is considered questionable in educational assessment where it is expected to be greater than 0.7.

Item analysis for Introductory Surgery multiple-choice questions:

Difficulty Index (DIF-I)

The analysis for the keys (correct answers) of 50 MCQs is as shown in table 3.
Difficulty index of the keys was set at <0.3 &> 0.8 respectively and showed 14% & 40% of the questions were either too hard or easy.

Difficulty Index (DIF-I)	**Difficulty level**	**Number of items**
<.3	Too hard	7 (14%)
0.3-0.8	Moderate	23(68%)
>.8	Too easy	20 (40%)

Table 3

Item keys identified as too hard will be expunged, while those identified as moderate will be retained. Items identified as too easy will be modified and if it is a high-stake summative examination that assesses true success, it should be critically reviewed.

Discrimination index (DI)

Of 50 keys in the MCQs, the correct answers showed a poor discrimination index of 48% in this examination (see table 4)

Discrimination Index (DI)	Description	Number of items
<0.1	Poor	24 (48%)
0.1 – 0.3	Fair	25 (50%)
>0.3	Good	1 (10%)

Table 4

Items identified as poor will be discarded and those shown to be fair will be reviewed and modified for retention. Those shown to be good will be retained.

Cronbach's alpha coefficient was reported as 0.45. This is rated as poor, as the accepted level for reliability is greater than or equal to 0.7. The seemingly good performance of a pass rate of 60 out of 62 students (97%) is from an unreliable formative assessment.

Frequency analysis of placement of keys in the answers in Introductory Medicine:
The correct answer keys A, B, C, D, E frequencies within the 50 MCQs are as shown:

Value	Frequency	Relative %
A	6	12
B	6	12
C	13	26
D	12	24
E	13	26
TOTAL	**50**	**100%**

Table 5

The key placement should be homogenous13, but here it is reported as heterogenous.

Item analysis of Distractor Efficiency in Introductory Medicine:
The distractor analysis of 250 items (50 keys and 200 distractors), the index of effectiveness of each item distractor is shown in table 6:

Distractor effectiveness index (DE)	Frequency	Percentage
Negative index	109	43.6%
Zero (0)	72	28.8%
Positive index between (0.09 – .36)	69	27.6%

Table 6

The distractors with a negative index, 43.6% is reassuring. It shows the distractors are plausible. More students in the lower group selected these than those in the higher group. Distractors with a zero index are non-functional and ineffective. The distractors with a positive index (0.09 – 0.36) are not plausible. More students in the higher group selected the distractor.

Frequency analysis of placement of keys in the answers in Introductory Surgery:
The correct answer keys A, B, C, D, E frequencies within the 50 MCQs are as shown:

Value	Frequency	Relative %
A	11	22
B	7	14
C	10	20
D	13	26
E	9	18
Total	**50**	**100%**

Table 7

The key placement should be homogenous13, but here it is reported as heterogenous.
Item analysis of Distractor Efficiency in Introductory Surgery
The distractor analysis was undertaken from 250 items (50keys and 200 distractors) in the introductory surgery MCQs. (see table 8)

The proportion of non-functional and ineffective distractors in the MCQs was 45.2%. This high proportion is of concern and calls for remedial action from the trainers in the construct of MCQs.

Distractor effectiveness index (DE)	Frequency	Percentage
Negative index	78	31.2%
Zero (0)	113	45.2%
Positive index between 0.09 – 0.36	59	23.6%

Table 8

The distractors with a negative index, 31.2% is reassuring. It shows the distractors are plausible. More students in the lower group selected these than those in the higher group. Distractors with a zero index are non-functional and ineffective and was reported as 45.2%. The distractors with a positive index (0.09 – 0.36) are not plausible. More students in the higher group selected the distractor.

Discussion

Item analysis of MCQs is desirable after assessments to identify any flaws and confirm the reliability and validity. Multiple-choice questions (MCQs) can assess higher cognitive processing like interpretation, analysis and problem-solving of Bloom's taxonomy of learning[3,4,] when done properly with blueprinting of the curriculum and lesson planning. Constructive alignment [24] an approach to curriculum design which is focused on closely aligning teaching and assessment to intended learning outcomes should be reflected in aligning the MCQs with learning objectives. Designing good MCQs are difficult but can be overcome when consciously done. These should be done early at the stage of lesson planning and submitted to the MCQ bank.

Nonetheless, MCQs is now a favoured assessment tool for both formative and summative assessment and in a recent study[25], students show more preference to it. One of the key goals of assessment in medical education is the minimisation of all errors influencing a test to produce an observed score which approaches a learner's 'true' score, as reliably and validly as possible[19]. To achieve this, assessors need to be aware of the potential biases that can influence all components of the assessment cycle from question creation to the interpretation of exam scores[19].

In low stake examinations as in formative assessment, it is important to undertake item analysis, which can be revealing. In the evaluation undertaken here, the high pass rate and seemingly good performance of the pioneer students is not vindicated by the item-analysis.

Firstly, the very high pass rate in introductory surgery is flawed by the very low reliability index from the Cronbach's alpha coefficient of 0.45. The recommended reliability index from most educational assessments is a value above 0.7.

Secondly, the high proportion of easy questions in both the introductory medicine (22%) and surgery (40%) MCQs could have gone unnoticed.

Thirdly the poor discrimination index of 44% in introductory medicine and 48% in introductory surgery would have gone unnoticed.

Lastly the poor distractor efficiency of 45.2%, items scoring zero (0) in introductory surgery, seemingly accounted for the very high pass rate of 96.7% and high median score of 63%. The answer key assignment was D (26%) and A (22%) as the correct option in the items. The distractor efficiency was better in introductory medicine at 28.8%, items scoring zero (0).

A preview of all MCQs in the bank must be undertaken periodically by content experts, critically evaluating flaws in distractors. Heterogeneity of answer keys must be avoided as was the case here. Homogenous answer keys must be done consciously by the examination officer by random allocation. For example, amongst the 50 MCQs the keys A to E should be represented evenly at 20% in a 5-stem single-best answer MCQ.

This post-examination study has shown the level of difficulty in writing single-best answer MCQs and identifying factors leading to poor discrimination index as part of the teaching and learning culture, also shown by others[26]. The distractor efficiency of zero (0), score of 28.8% in introductory medicine and 45.2% in introductory surgery

respectively is worrying. When absolute pass scores are used and set at a fixed percentage (i.e., 50%), as they were in our centre, such a high proportion of easy items will likely result in many borderline candidates passing.[27]

Limitations:

The limitation of this study includes not analysing in-depth poorly functioning distractors 28 with a relatively low choice frequency of <5%. Our post-examination analysis cannot be generalised, but the process of item analysis should be widely adopted. We used American standards to set for difficulty and discrimination indexes as there was a paucity of literature from Nigerian Medical Schools. The only study from Nigeria29, set difficulty index at (0.03 – 0.75) and discrimination index at >0.2, lower than the American standards used in this study.

There was no preview of the examination MCQs by content experts to exclude grammatical and logical cues.

Conclusion

We recommend item-analysis of MCQs be undertaken routinely after formative and summative assessment in medical schools in Nigeria to ensure content validity and reliability as part of quality assurance matrix. Content experts must preview MCQ bank items to ensure consistency and appropriateness.

Declarations of conflict of interest

There was no declaration of conflicts of interest from any of the authors.

AcknowledgementWe are grateful for the critique of the draft manuscript by Prof Ibi Erekosima, Consultant Renal Physician, Salford Royal NHS Foundation Trust, Salford, M6 8HD, England and Visiting Professor of Medicine & Medical Education, University of Port-Harcourt, Port Harcourt, Rivers State, Nigeria

Funding:

No funding was obtained for this study.

Contributions of the authors:

GB, PI, AD, BK, TCH all contributed to the conceptualisation of the study and participated in the draft of the manuscript. TCH did the data entry and analysis.

References

1.Guidelines on minimum standards for undergraduate medical & dental education in Nigeria. Medical & Dental Council of Nigeria. Accessed 23rd November 2024

2.Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos Univ Med J. Aug 2010;10(2):203-9.

3.Bhat SK, Prasad KHL. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. Indian Journal of Ophthalmology. 2021;69(2)

4.Wood TJ, Cunnington JP, Norman GR. Assessing the measurement properties of a clinical reasoning exercise. Teach Learn Med. Fall 2000;12(4):196-200. doi:10.1207/S15328015TLM1204_6

5.Coughlin PA, Featherstone CR. How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. European Journal of Vascular and Endovascular Surgery. 2017;54(5):654-658. doi:10.1016/j.ejvs.2017.07.012

6.Ventista O. Time to increase the quality of the multiple-choice questions you use! 2017; Accessed 23rd November 2024

7.Ndu IK, Ekwochi U, Di Osuorah C, et al. Negative Marking and the Student Physician--A Descriptive Study of Nigerian Medical Schools. J Med Educ Curric Dev. Jan-Dec 2016;3doi:10.4137/jmecd.S40705

8.Lesage E, Valcke M, Sabbe E. Scoring methods for multiple choice assessment in higher education Is it still a matter of number right scoring or negative marking? Studies in Educational Evaluation. 2013/09/01/ 2013;39(3):188-193. doi:

9.Knowles M. The modern practice of adult education: From pedagogy to andragogy. Englewood Cliffs, NJ. Cambridge Adult Education; 1980.

10.Roine I, Molina Y, Caneo M. A psychometric appraisal of the dundee ready education environment measure in a medical school in Chile. Educ Health (Abingdon). Sep-Dec 2018; 31(3):148-154. doi:10.4103/efh.EfH_17_18

11.Tavakol M, Dennick R. Post-examination analysis of objective tests. Medical Teacher. 2011/06/01 2011;33(6):447-458. doi:10.3109/0142159X.2011.564682

12.Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Medical Teacher. 2012/03/01 2012;34(3):e161-e175. doi:10.3109/0142159X.2012.651178

13.Gasmalla HEE, Mohamed Tahir MEM. A-Type MCQs. In: Gasmalla HEE, Ibrahim AAM, Wadi MM, Taha MH, eds. Written Assessment in Medical Education. Springer International Publishing; 2023:73-89.

14.Meyer JP. JMetrik (Version 4.1) Computer Software. 2014; Retrieved from .

15.Brown JD. Testing In Language Programs: A Comprehensive Guide To English Language Assessment. Upper Saddle River, NJ. McGraw-Hill College; 2005.

16.Etobro BA, Taiwo O, Alawaye M. Pragmatic Perspective of Item Analysis Using Microsoft Office Excel Data Analysis Tools. Ilorin Journal of Education. 04/08 2024;44(2):279 - 292.

17.Reynolds CR, Altmann RA, Allen DN. Item Analysis: Methods for Fitting the Right Items to the Right Test. In: Reynolds CR, Altmann RA, Allen DN, eds. Mastering Modern Psychological Testing: Theory and Methods. Springer International Publishing; 2021:263-289.

18.Understanding Item Analyses. University Of Washington; Accessed 23rd November 2024

19.Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33(6):447-58. doi:10.3109/0142159X.2011.564682

20.Lord FM. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika. 1952;17(2):181-194. doi:10.1007/BF02288781

21.Ebel RL, Frisbie DA. Essentials of Educational Measurement. Prentice-Hall; 1986.

22.Testa S, Toscano A, Rosato R. Distractor Efficiency in an Item Pool for a Statistics Classroom Exam: Assessing Its Relation With Item Cognitive Level Classified According to Blooms Taxonomy. Original Research. Frontiers in Psychology. 2018-August-28 2018;9doi:10.3389/fpsyg.2018.01585

23.Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16, 297-334

24.Biggs JB, Tang CKC. Teaching for quality learning at university: what the student does. Maidenhead. McGraw-Hill; 2011.

25.Vegi VAK, Sudhakar PV, Bhimarasetty DM, et al. Multiple-choice questions in assessment: Perceptions of medical students from low-resource setting. J Educ Health Promot. 2022;11:103. doi:10.4103/jehp.jehp_621_21

26.Loh KY, Elsayed I, Nurjahan MI, Roland GS. Item Difficulty and Discrimina-tion Index in Single Best Answer MCQ: End of Semester Examinations at Taylors Clinical School. Springer Singapore; 2018:167-171.

27.Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Medical Education. 2009/07/07 2009;9 (1):40. doi:10.1186/1472-6920-9-40

28.Pan Q, Jiang Z. Examining distractor qualities of pediatrics subject tests from a national assessment. Front Med (Lausanne). 2022;9:921719. doi:10.3389/fmed.2022.921719

29.Nnodim JO. Multiple-choice testing in anatomy. Med Educ. Jul 1992;26 (4):301-9. doi:10.1111/j.1365-2923.1992.tb00173.x