Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2022 February; 5 (1): 331-336

Measuring And Improving The Effectiveness Of Question Choices In Medical Assessments

Dr Pallavi[1*], Dr Saginala Ravichandra[2] , Dr Abhiram Kumar C[3], Jothieswari D[4]

[1*]Associate Professor, Sri Lakshmi Narayana Institute of Medical Sciences, Puducherry-605502
[2]Assistant Professor , Department of Paediatrics , Arunai Medical College and Hospital, Velu Nagar, Mathur, Tiruvannamalai, Tamilnadu, India-606603.
[3]Assistant Professor, Department of Paediatrics, Melmaruvathur Adhiparasakthi Institute of Medical Sciences and Research, Melmaruvathur, Tamil Nadu, India.
[4] Sri Venkateswara College of Pharmacy, RVS Nagar, Chittoor-517127, Andhra Pradesh, India.

ABSTRACT:

The purpose of this study was to examine the relationships between item difficulty, ability of the items to distinguish certain performance levels and the strength of the distractors in medical MCQs. The aim was to decide if selected questions needed to be kept, updated or disposed of. In addition, the research focused on deciding how many options to include per multiple-choice question for better quality and reliability. An examination of the data was done using a cross-sectional approach in the pediatric section of a teaching hospital. Post hoc evaluation of the study period showed 800 MCQs and 4,000 connecting distractors were used. The analyzed factors for each exam item were difficulty index, discrimination index and distractor performance. Most questions had a difficulty level between 36.70% and 73.14%, with discrimination averaging between 0.20 and 0.34. Distractor efficiency was above 66.50% for most cases. About 48.4% of the examined items contained no unwanted extras, 35.3% had one NFD, 11.4% had two, 3.9% had three and 1.1% had four. When the number of answer options was three or four, NFDs were found less often than when the number was five. The stronger the discrimination and the more difficulty an item required, the better the efficiency of its distractors became ($P < 0.005$). Using the Kuder-Richardson 20 calculation, the mean reliability was 0.76. Indeed, many MCQs were acceptable, but a small number required changes or should be replaced. If you have three or four answer options, there are less irrelevant choices and the reliability of the test rises.

KEYWORDS: Education for medical professionals, quality of assessments, multiple-choice questions, item analysis and the effectiveness of distractors are all included.

INTRODUCTION:

Besides checking knowledge, assessment is essential for boosting student learning and performance. An assessment is effective only if it shows validity, reliability and is objective. Also, any assessment should capture the differences in student achievement. Though designing a great MCQ exam takes time and effort, grading with this type of question is reliable and explains student performance well, making it preferable over easier forms of testing. Many people say that MCQs concentrate too much on remembering facts. When well made, MCQs can challenge students to use critical thinking skills that are defined by Bloom's taxonomy of learning objectives. From the very simple ability to remember facts, this taxonomy covers skills such as comprehension, applying learning, analyzing, synthesizing and evaluating. When questions include all three domains, teachers can better tell how well a student understands the material and how well they can solve problems. "Type A" MCQs are one of the most common forms of MCQs used in today's academic testing. These distractors have a big effect on how correctly the question can be answered. It's good for distractors to look easy enough for students who already understand the information but challenging for those who don't. Item analysis techniques are commonly used by educators after exams to make sure MCQs are both fair and of good quality. Statistics is used to analyze every question in order to find out how well it works and how valuable it is. Three important indicators in item analysis are DIFI, DI and DE. Difficulty index describes how many students answered a particular question correctly. Questions in an ideal test tend to have a DIFI between 30% and 70% which helps prevent them from being either too easy or too hard. Also called point biserial correlation, the discrimination index shows how successful a question is at telling apart middle-performing students from those who are strong and those who struggle. Acceptable value for DI is usually considered to be over 0.2, meaning the question distinguishes students by how much they understand. The fourth assessment measures how skillfully the wrong options act as distractors. Organizers consider a question with 100% DE to have made each answer choice different from the other distractors. At one institution, a detailed system for evaluating medical students was put in place during their 10-week pediatrics clinical rotation. As part of their testing, students completed multiple kinds of exams, including MCQ tests, evaluated patients through clinical consultations, answered short questions and were constantly evaluated throughout the rotation. In the MCQs,

there was just one correct answer and four phony ones. Mistakes did not affect the final grade and every exam was judged by comparing students to an established standard, not to each other. It was required that at least 60% of voters vote in favor. The MCQs in all exams were split, with half new and half taken from a regularly updated bank. Before using them again, the process of updating them was guided by what was revealed through earlier item analyses. By optimization of each question, the exam improved, though nothing assessed how hard each exam was, potentially affecting the same measure of fairness given to students each year. The researcher set out to retrospectively check MCQs used in four years of pediatric rotation exams. The key objective was to use the standard indices to assess each item and then make choices about whether to maintain, modify or discard certain questions. In addition, the work considered linkages among scoring difficulty, discrimination, distractor effects and choice options and looked for the best number of answer choices that balances learning benefits with testing cost. The purpose of these efforts was to strengthen the quality of assessments, encourage equal treatment and help continually improve student evaluation within medical programs.

METHODOLOGY:

Within the Department of Paediatrics at a learning institution, the study completed all MCQs from summative examinations that students answered between November 2013 and June 2016. Annually, students answered 50 MCQs on each paper, leading to 3,200 nieve questions and a total of 12,800 distractor answers in the year. Over the study period, the assessment studied involved 608 students and an average of 38 students in every session. No MCQ items were looked at with students after the exams and they were only used for summative assessment. Content and construct validity were guaranteed by a devoted Examination Committee consisting of five paediatrics experts and consultants. All examinations were built to follow a planned structure that supported the set learning objectives and included all important areas of information and skills. After examination, the structured database management system was used for item analysis and MCQs were either maintained as is, updated or removed depending on the findings, obvious issues or how often they had been used before. Students handed in their work on optical sheets which were then scanned to be scored automatically. The quality of each MCQ was measured by using the difficulty index (DIFI), the discrimination index (DI) and distractor efficiency (DE). The DIFI ranged from 0% when no students got it right to 100% when all students got it right. DIFI scores under 30% were judged as making the task difficult, between 30% and 70% were labeled acceptable and scores over 70% were considered easy. The difference between high- and low-achieving students was calculated using Kelley's method. When the DI is high, the test is more able to tell patients apart. Ratings on DI varied between −1 for those who were only correct when low-achieving and +1 for those who were only correct when high-achieving. DI scores of 0.35 or more were considered excellent, scores between 0. and 0.34 were acceptable and all scores under 0.2 were classed as poor for discriminating relationships. The efficiency of distractors was computed by taking into account the options selected by less than 5% of examinees. To score the DE for an item, it was given 0% if no NFDs were found and 25%, 50%, 75% or 100% if one, two, three or four NFDs were seen. The consistency of every examination was assessed by Kuder-Richardson Formula 20 (KR-20), a statistical method that tests the viability of multiple-choice questions. A KR-20 score can reach 1 and any value close to this means the test is highly reliable. Scores below 0.3 are regarded as poor and any score of 0.7 or above is fine. Tests usually lose accuracy when lots of the items have scores that stand out as either very high or very low or when the DI is very low. Researchers used Version 23.0 of a statistical software package for data analysis. The findings were reported using means and standard deviations. A Pearson correlation coefficient was applied to check how DIFI and DI were related on a linear level. In addition, the influence of both DIFI and DI on DE was analyzed by applying a two-way ANOVA. Results with a p-value of less than 0.050 were considered statistically significant. Appropriate authorities gave permission for the study and provided the access to the examination data. No information about student identities was revealed and individual data was anonymous at all times. Since the researchers worked only with existing records, no people participated directly.

Table 1. Item Analysis findings for Paediatric Clerkship Multiple-Choice Questions Examinations by year and examination session.

Year	Exam	DIFI % (Mean ± SD)	DI (Mean ± SD)	DE % (Mean ± SD)
2013	1	61.45 ± 21.67	0.31 ± 0.20	68.50 ± 27.90
	2	69.72 ± 18.30	0.33 ± 0.17	70.00 ± 28.00
	3	66.10 ± 20.50	0.29 ± 0.19	71.20 ± 26.75
	4	55.88 ± 22.75	0.21 ± 0.18	77.80 ± 20.80
	Total	**63.79 ± 21.56**	**0.28 ± 0.19**	**71.38 ± 25.86**
2014	5	50.60 ± 20.70	0.25 ± 0.15	80.50 ± 18.75
	6	48.90 ± 19.85	0.26 ± 0.24	74.00 ± 22.00
	7	47.40 ± 23.10	0.22 ± 0.23	81.00 ± 16.50
	8	49.85 ± 20.45	0.27 ± 0.19	83.80 ± 18.30

	Total	**49.69 ± 21.03**	**0.25 ± 0.20**	**79.83 ± 18.89**
2015	9	37.80 ± 19.95	0.24 ± 0.17	86.20 ± 14.25
	10	41.10 ± 21.40	0.22 ± 0.15	85.50 ± 18.40
	11	39.85 ± 21.00	0.26 ± 0.17	82.00 ± 16.00
	12	42.15 ± 22.80	0.20 ± 0.16	78.90 ± 19.50
	Total	**40.23 ± 21.29**	**0.23 ± 0.16**	**83.15 ± 17.53**
2016	13	39.55 ± 20.65	0.25 ± 0.14	82.30 ± 18.10
	14	35.20 ± 21.30	0.21 ± 0.17	84.00 ± 15.40
	15	50.90 ± 17.45	0.30 ± 0.13	80.50 ± 19.30
	16	45.05 ± 20.70	0.22 ± 0.14	85.10 ± 14.70
	Total	**42.68 ± 20.53**	**0.25 ± 0.15**	**82.73 ± 16.88**
Overall Average		**51.60 ± 22.85**	**0.25 ± 0.18**	**79.77 ± 20.71**

Table 2. Number of Non-Functional Distractors per Item and Related Mean DIFI and DI shown by Year

Year	Parameter	Number of NFDs per Item	Total
		0	1
2013	n (%)	60 (30.0)	76 (38.0)
	Mean DIFI %	50.45	64.75
	Mean DI	0.31	0.30
2014	n (%)	85 (42.5)	87 (43.5)
	Mean DIFI %	46.80	55.10
	Mean DI	0.24	0.27
2015	n (%)	112 (56.0)	68 (34.0)
	Mean DIFI %	38.70	44.85
	Mean DI	0.22	0.25
2016	n (%)	117 (58.5)	58 (29.0)
	Mean DIFI %	39.55	47.90
	Mean DI	0.26	0.23
Total	n (%)	374 (46.8)	289 (36.1)
	Mean DIFI %	41.88	53.65
	Mean DI	0.26	0.26

Table 3. Item Classification uses the DIFI, Discrimination Index (DI) and Distractor Efficiency (DE) while listing actions to take.

Index	n (%)	DE %	P value	Proposed Action
DIFI				
Difficult	172 (21.5)	89.40	<0.005*	Review
Acceptable	420 (52.5)	87.10		Store and review
Easy	208 (26.0)	62.75		Discard
DI				
Poor	260 (32.5)	79.20	<0.005†	Discard
Acceptable	260 (32.5)	81.50		Store and review
Excellent	260 (35.0)	84.00		Store

Figure 1: Pediatric Clerkship MCQ Examination Analysis (2013-2016), Difficulty Index, Discrimination Index, and Effectiveness Trends

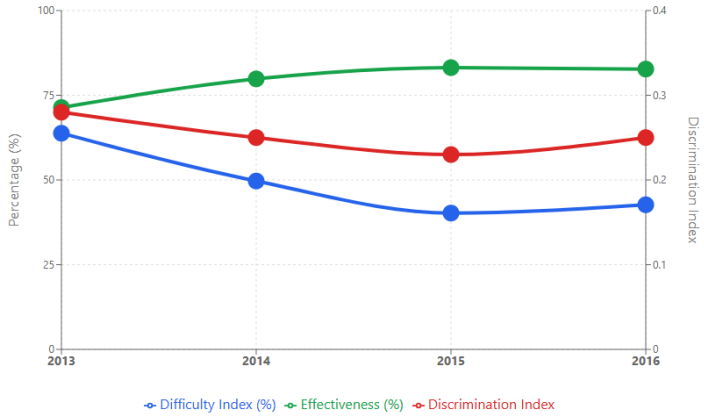
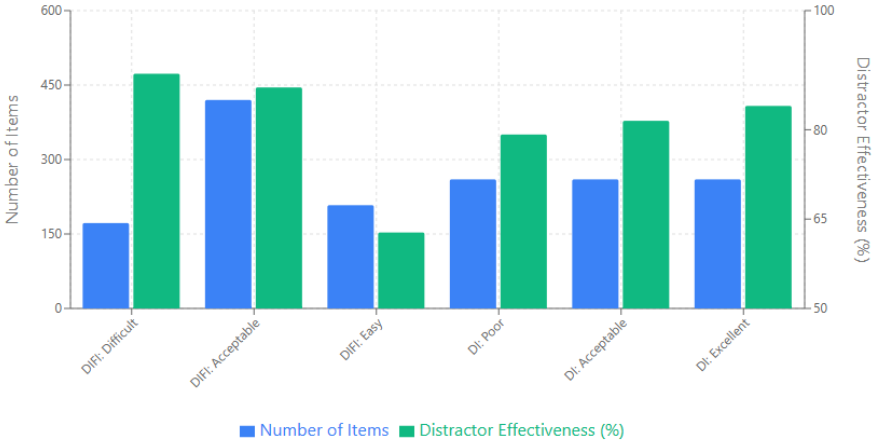


Figure 2:Impact of Non-Functional Distractors on Difficulty Index, Mean Difficulty Index by Number of NFDs (2013-2016)

Figure 3:Item Classification Analysis, Distribution and Effectiveness by DIFI and Discrimination Index Classifications



RESULT:

The item analysis of multiple-choice questions (MCQs) from paediatric clerkship examinations over four years revealed important trends in item performance and quality indicators. The difficulty index (DIFI) showed variability across years and examination sessions, with mean values ranging from approximately 35% to 70%. Earlier years tended to have higher mean DIFI values, indicating relatively easier items, while later years demonstrated lower DIFI values, reflecting more challenging questions. The overall average DIFI was 51.60%, suggesting that, on average, items had moderate difficulty. The discrimination index (DI), measuring how well items differentiate between high- and low-performing students, showed consistent moderate values around 0.25 across the study period. This indicates that most questions had acceptable discriminative power. The distractor efficiency (DE), representing how effectively incorrect options attracted students away from the correct answer, averaged nearly 80%, highlighting generally well-constructed distractors. DE tended to be higher in more recent years, suggesting improved question design. Analysis of non-functional distractors (NFDs) revealed that nearly half of the items had zero or one NFD per question, indicating good quality distractors overall. Items with no NFDs had higher mean DIFI and DI values compared to those with one NFD, underscoring the importance of effective distractors in item performance. Classification based on DIFI showed that 21.5% of items were difficult and recommended for review, while 52.5% were of acceptable difficulty and stored for future use with review, and 26% were deemed easy and discarded. Regarding discrimination, 32.5% of items were poor and discarded, while 32.5% were acceptable and 35% excellent, with both groups stored for future use. These findings support ongoing refinement of the MCQ bank, emphasizing the need to review difficult and poor discriminating items to maintain examination quality and fairness.

DISCUSSION:

The study reported that the difficulty level of the 800 items tested in 16 summative examinations was acceptable. DIFIs in 2013 were mostly lower than those seen in 2015 and 2016. Probably, the examination committee has recently improved the difficulty level of MCQs. Even though incentives and how they are tested may differ, the outcomes from the studies are broadly similar. Previously, research found mean DIFIs from 39% to 89% in many kinds of assessments. According to a study, most items or around 61%, were acceptable, 24% were easy and a further 15% were rated as difficult. Meanwhile, 53.4% of items fit in the first category, 25.9% fit into the second and 20.8% were part of the third. Choose straightforward MCQs for major concepts and put harder versions towards the end of the examination to highlight which students performed best. A similar pattern to other reports was seen in the DI, where the counts of poor, acceptable and excellent were nearly the same. Flawed keys, uncertain questions or challenging content are often the reason for poor DI. Consequently, these items should be excluded because they cannot differentiate between students' abilities. Reducing the number of unnecessary distractors and designing good distractors play a big role in improving MCQ quality. Very few questions had more than two NFDs in this study and distractor efficiency (DE) got better each year from 2013 to 2016. Growing numbers of zero NFDs and declining numbers of items with three or more NFDs demonstrate that efforts to improve quality are ongoing. Assessments with lots of NFDs often have a high DIFI and a low DI, so they do not measure student learning as well. Big-DE materials deserve additional efforts to delete them if their difficulty level is SSS and they must be reviewed again if they are LLL, yet simple small-DE items should not be in the course. Acceptable DIFI and DE items should be retained to make them even better. Take time to look over questions where people often answer wrong more than right. Results indicate that fewer NFDs in an item were related to better discrimination, making it necessary to discard questions with low indices and keep those with high ones. Results showed that objects with just a few NFDs had good marks for both DIFI and DI, whereas items with three or more NFDs had poorer scores. As long as you hold onto important topics, you can check how well your students are learning. Almost without exception, studies agree that three-option MCQs are superior because they improve reliability, reduce the time needed to write questions and allow for more questions in the exam. A dome-shaped connection was detected between DIFI and DI and higher DI was measured at mid-level DI scores. According to the results, this test achieved an acceptable level of reliability with a coefficient of 0.76. Simply put, accurate assessment of students' progress depends on good MCQ questions. While NFDs don't challenge K students much, less guessing and more use of functional features makes the exams better for everyone. If an examination committee and special training are included, the development of questions can be even more effective. This research should be complemented by efforts to keep improving the model and applying it to other disciplines.

CONCLUSION:

Data from the study proves that the right kind of MCQs are very important for assessing students accurately. The results indicate that questionnaire items which meet certain conditions play a key role in the reliability and validity of the examinations. After passing time, the quality of questions grew as NFDs were reduced while DE was increased, demonstrating that assessment quality is being constantly pursued. Items that contained fewer bad distractors were able to tell the difference between top and bottom students, unlike those that had many of these distractors which went down in ease and usefulness for telling ability apart. The results show that it's important to regularly review and modify MCQs to maintain a strong library of questions. Giving easier questions at the beginning of the test can motivate students and putting tougher questions at the close helps sort out how well they did. According to the findings, offering fewer but similar options can help assessments remain accurate and also lower the time needed to create them. Improvement can be achieved by frequently analyzing items, with the help of dedicated committees and by teaching or retraining item writers. Working on applying these practices to other subjects will secure standardized yet effective ways to measure what students have learned.

REFERENCE:

1. Case SM, Swanson DB. Extended-matching items: A practical alternative to free-response questions. Tech Learn Med. 1993;5:107–15.
2. Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psycol Meas. 2004;64:391–418.
3. Bloom BS, Hastings JT, Madaus GF. Handbook on Formative and Summative Evaluation of Student Learning. New York, USA: McGraw-Hill; 1971. p. 103.
4. Skakun EN, Nanson EM, Kling S, Taylor WC. A preliminary investigation of three types of multiple choice questions. Med Educ. 1979;13:91–6.
5. Skakun EN, Nanson EM, Taylor WC, Kling S. An investigation of three types of multiple choice questions. Annu Conf Res Med Educ. 1977;16:111–16.
6. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. J Pak Med Assoc. 2012;62:142–7.

7. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. BMC Med Educ. 2009;9:40.
8. Kelley TL. The selection of upper and lower groups for validation of test items. J Educ Psychol. 1939;30:17–24.
9. Mehta G, Mokhasi V. Item analysis of multiple choice questions: An assessment of the assessment tool. Int J Health Sci Res. 2014;4:197–202.
10. Bland JM, Altman DG. Cronbach's alpha. BMJ. 1997;314:572. doi: 10.1136/bmj.314.7080.572.
11. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York, USA: McGraw-Hill; 1994.
12. Mitra NK, Nagaraja HS, Ponnudurai G, Judson JP. The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. Int EJ Sci Med Educ. 2009;3:2–7.
13. Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahemdabad, Gujarat. Indian J Community Med. 2014;39:17–20.
14. Karelia BN, Pillai A, Vegada BN. The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of year II M.B.B.S students. Int EJ Sci Med Educ. 2013;7:41–6.
15. Sharif M, Rahimi SM, Rajabi M, Sayyah M. Computer software application in item analysis of exams in a college of medicine. ARPN J Sci Tech. 2014;4:565–9.
16. Lin LC, Tseng HM, Wu SC. Item analysis of the registered nurse license exam by nursing candidates from vocational nursing high schools in Taiwan. Proc Natl Sci CouncRepub China D. 1999;9:24–31.
17. Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. Int J Appl Basic Med Res. 2016;6:170–3.
18. Mackenzie J. Vague and ambiguous questions on multiple-choice exercises: The case for. Educ Philos Theory. 1994;26:23–33. doi: 10.1111/j.1469-5812.1994.tb00198.x. [DOI] [Google Scholar]
19. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract. 2005;24:3–13. doi: 10.1111/j.1745-3992.2005.00006.x. [DOI] [Google Scholar]
20. Tomak L, Bek Y. Item analysis and evaluation in the examinations in the Faculty of Medicine at OndokuzMayis University. Niger J Clin Pract. 2015;18:387–94. doi: 10.4103/1119-3077.151720. [DOI] [PubMed] [Google Scholar]
21. Mukherjee P, Lahiri SK. Analysis of multiple choice questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. IOSR J Dent Med Sci. 2015;14:47–52. doi: 10.9790/0853-141264752. [DOI] [Google Scholar]
22. Nwadinigwe PI, Naibi L. The number of options in a multiple-choice test item and the psychometric characteristics. J Educ Pract. 2013;4:189–96. [Google Scholar]
23. Vegada B, Shukla A, Khilnani A, Charan J, Desai C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. Indian J Pharmacol. 2016;48:571–5. doi: 10.4103/0253-7613.190757. [DOI] [PMC free article] [PubMed] [Google Scholar]
24. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Educ Today. 2010;30:539–43. doi: 10.1016/j.nedt.2009.11.002. [DOI] [PubMed] [Google Scholar]
25. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore. 2006;35:67–71. [PubMed] [Google Scholar]