Novi Maulina, Rima Novirianthy. JPKI, 2020; 9(2): 131-137
doi: 10.22146/jpki.49006

ORIGINAL RESEARCH

JPKI

ITEM ANALYSIS AND PEER-REVIEW EVALUATION OF SPECIFIC HEALTH PROBLEMS AND APPLIED RESEARCH BLOCK EXAMINATION

Novi Maulina[1]*, Rima Novirianthy[1]
[1]Assesment Unit, Faculty of Medicine, Syiah Kuala University, Banda Aceh - INDONESIA

Submitted: 26 August 2019; Final Revision from Author: 22 June 2020; Accepted: 23 June 2020

ABSTRACT

Background: Assessment and evaluation for students is an essential component of teaching and learning process. Item analysis is the technique of collecting, summarizing, and using students' response data to assess the quality of the Multiple Choice Question (MCQ) test by measuring indices of difficulty and discrimination, also distracter efficiency. Peer review practices improve quality of assessment validity in evaluating student performance.

Method: We analyzed 150 student's responses for 100 MCQs in Block Examination for its difficulty index (p), discrimination index (D) and distractor efficiency (DE) using Microsoft excel formula. The Correlation of p and D was analyzed using Spearman correlation test by SPSS 23.0. The result was analyzed to evaluate the peer-review strategy.

Results: The median of difficulty index (p) was 54% or within the range of excellent level (p 40-60%) and the mean of discrimination index (D) was 0.24 which is reasonably good. There were 7 items with excellent p (40–60%) and excellent D (≥0.4). Nineteen of items had excellent discrimination index (D≥0.4). However, there were 9 items with negative discrimination index and 30 items with poor discrimination index, which should be fully revised. Forty-two of items had 4 functioning distracters (DE 0%) which suggested the teacher to be more precise and carefully creating the distracters.

Conclusion: Based on item analysis, there were items to be fully revised. For better test quality, feedback and suggestions for the item writer should also be performed as a part of peer-review process on the basis of item analysis.

Keywords: Item Analysis, Multiple Choice Questions (MCQs), Difficulty Index, Discrimination Index, Peer Review

PRACTICE POINTS

- Item analysis of MCQs not only offers useful data for item modification and future testing, but also a way to train the lecturers to develop better items.
- Peer-review process include training for staffs in developing items and become a peer reviewer, also enhancing the staff's skills through feedback they receive from the peer reviews and item analysis.
- Feedback and suggestions for the item writer and faculty should also be performed as a part of peer-review process on the basis of item analysis.

*corresponding author, contact: nofeemaulina@gmail.com

INTRODUCTION

Testing is the essential tool in the educational practice. It is crucial to be well versed in testing methods so that they can assess student progress reliably and validly.[1] A teacher should do item analysis to know how good the test items are and whether they represent student's knowledge related to the specific learning objectives given in a period, after administered and scored a test.

Multiple Choice Questions (MCQs) are commonly used for student evaluation. A typical MCQ includes a question called stem and a set of two or more alternative possible answers. The best answer is the key of the question and the remaining alternative options are referred as distracters. The key should be unambiguously correct and the distracters should be unambiguously incorrect.[4] Writing MCQs is relatively tough, particularly to generate excellent plausible distracters to enable items discriminating the student's capacity in learning materials. In fact, distracters availability will affect the quality of MCQs.[4]

The accuracy of the test results is a critical issue in generating MCQs for evaluations. Good MCQs are usually subject to a strictly rigorous item analysis process. Analysis of items is the technique of collecting, summarizing and using data from student's answers to evaluate the test quality. This process enables us to observe the features of a specific item and to guarantee that items are appropriate to be tested, or those items need improvement.[3] Revision is needed when the items are too difficult or too easy, or items could not able to differ student's ability, or items had plausible distractors. Removing non-discrimination items or altering the items or modifying the lead-in questions are several ways to correct any misunderstanding of teaching material, or teacher could also make adjustment in teaching method.[3,5]

Peer-review is an approach of a faculty development programs designed to support faculty members in MCQ writing and item quality evaluation. Studies showed the benefit of implementing review process to improve MCQ quality, especially for new and emerging medical schools.[6] A study described an education center performed this process by conducting sessions to become peer-reviewers, to construct good MCQs and to give feedback of review process to the lecturers. The MCQ peer review practices based on the National Board of Medical Examination (NBME) guidelines can result in significant improvements in the quality of items that can eventually improve the validity of assessment scores.[6] This method also promotes the faculty assessment to take benefit of item analysis to evaluate student performance aligns with the intended utility of items hypothesized by test developers.[7]

The present study aims to evaluate effectiveness of peer-review strategy by assessing the Specific Health problems and Applied Research Examination Block items quality by measuring index of difficulty and discrimination, and distracter efficiency.

METHODS

Responses of 150 third-grade students pursuing Bachelor in Medical program in block exam were collected. They participated in Specific Health Problem and Applied Research Block Examination in 18th block of medicine curricula which consist of 100 MCQs collected from 9 lecturers of block team in 100 minutes of time. The examination was conducted at Medical Faculty of Syiah Kuala University, Aceh, Indonesia, during the 2018/2019 academic year. Students' responses from the MCQs were analyzed using Microsoft Excel. The MCQ had five options, one of them is the key answer and the other four are distracters. A mark of 1 was awarded for a correct answer and no negative mark for the incorrect answer. Thus, the maximum possible score was 100 and the minimum was 0.

In order to evaluate our institution's peer-review strategy, the MCQs were analyzed for their difficulty index (p), power of discrimination or discrimination index (D) and distracter analysis for all incorrect options or Distractor Efficiency (DE). The results presented the test items' reliability and quality. Internal reliability value of test-scores was derived from The Kuder-Richardson formula (KR-21). Correlation of discrimination index and difficulty index was analyzed using Spearman correlation test by SPSS 23.0.

Difficulty index (p)

The index shows the percentage of the total correct answers to the test items, which were calculated using the formula, where p is the item difficulty index, R is the number of correct answers, and T is the total number of answers (includes both correct and incorrect answers).

The difficulty index or p (proportion) value ranges from 0 to 1. It converts to a percentage when multiplied by 100, which is the percentage of participants who got the item correct. The higher the p means the items is easier to be understood. It was considered as good and acceptable item if only p between 20 and 90%. And Items with p between 40 and 60% are considered excellent, because discrimination index (D) is maximum at this range.[8]

Items with difficulty index (p) <20% are considered too difficult, and those p >90% are considered too easy. They are included as not acceptable items and need to be revised or modified.[9]

Discrimination index (D)

The analysis of discrimination index (D) needs to rank the student's test score before separating them into 2 groups, which are the 25% of the uppers students and the 25% at the lower students. Then, D is number of upper group students correctly answered items minus number of the lower group students who answered the item correctly, divided by the total number of students. The formula for D is, where UG is the number of upper group students with correct answer and LG is the number of lower group students with correct answer and n is the total number of students.

The index used to discriminate student performance in a test.[10] Based on discrimination index (D), the items were classified as the item is satisfactorily functioned (if D ≥ 0.40); the item needs little or no revision (if 0.30 ≤ D ≤ 0.39); the item is marginal and needs revision (if 0.20 ≤ D ≤ 0.29); and the item should be eliminated or completely revised (if D ≤ 0.19).[11]

Distracter efficiency (DE)

The distracters are a non-dependent indicator of item functioning. Functioning distracters (FDs) are distracters which are chosen by one or more students (or ≥5% of participants), and those not chosen by anyone or chosen by <5% of participants are called non-functioning distracters (NFDs). The Distracter Efficiency (DE) is determined for each item based on the number of Non-Functioning Distracters (NFDs) in it and ranges from 0 to 100%. The item with no Non-functional distracter (NFD) will give DE-value 100%. However, the item with four, three, two or one non-functional distracters (NFDs) will result DE 0.75, 0.50 and 0.25, respectively.[12]

RESULTS AND DISCUSSION

Learning goals must be in-line with the test items to attain educational validity. This requires methods of developing and analyzing appropriate test items. To enhance teaching methods and test construction, teachers need to upgrade their understanding on how to use statistical analysis of test results.[13]

In this study, a total of 150 students participated in 100 one-best MCQs of Specific Health Problem and Applied Research topics of Block Examination. The scores ranged from 17.0 to 71.0 (of 100). The mean test score was 51.73 ± 10.12. To evaluate internal test scores reliability, we used the Kuder-Richardson formula (KR-21). It was 0.76 or in the range of teacher-made assessments. Most high-stakes tests have 0.90 or greater for their internal reliability, but usually teacher-made evaluations have values of 0.80 or lower. A previous study stated that a teacher-made evaluation requires roughly 0.50 or 0.60 reliability coefficients.[14]

The tools that teachers can use to confirm whether the MCQ items are well crafted or not is by using difficulty and discrimination index. Distracter efficiency (DE) is another tool for further analysis of quality distracters of an item.

The distribution of difficulty indices of the items (p) was showed in Table 1. The median was 0.54 ± IQR 0.44 or within the range of excellent difficulty level (p 40-60%). Eighty-one percent (81%) of items were acceptable with p 20-90%, while 46% of items among them had excellent difficulty index with p 30 - 70%. Thirteen (13%) items were too difficult (p<20%) and 6% was too easy (p>90%). Minimum p was 3% and maximum p was 96%.

Table 1. Distribution of Difficulty Index (p) of Items

Difficulty Index (p)	n (%)
<0.1	6
0.1 – 0.19	7
0.2 – 0.29	10
0.3 – 0.39	12
0.4 – 0.49	10
0.5 – 0.59	16
0.6 – 0.69	8
0.7 – 0.79	13
0.8 – 0.89	12
0.9 – 0.99	6
Total	100

The frequency distribution of discrimination indices of items was shown in Table 2. The mean D was 0.24±0.19 or acceptable. The highest D was 0.73 and the lowest D was -0.24. There were 19% of items with excellent discrimination index (D≥0.4). Nevertheless, there were also 9 items (9%) with negative discrimination index and 30 items (30%) with poor discrimination index.

Table 2. Distribution of Discrimination Index (D) of Items

Discrimination Index (D)	n (%)	Criteria
Negative	9	Defective Items / Wrong Key
0– 0.1	12	Poor
- 0.19	18	Poor
0.2 – 0.29	23	Acceptable
0.3 – 0.39	19	Good
≥0.4	19	Excellent
Total	100	

As presented in table 2, 39% of items should be fully revised due to very low discriminating power. The higher D correlates to the better item, because these kinds of items could differ the students with higher or lower scores better. However, those with low discrimination index items are often occur due to ambiguous terms worded and they should be re-checked. The current study demonstrates 9 items had negative discrimination index (D). This negative value could be due to incorrect key, or the questions framed implausible. The negative D items are not only useless; they also decrease the validity of the tests.[3] To increase the reliability of the test, negative discrimination index should be deleted or replaced. Zero or low positive discrimination index items should be replaced or rewrite. The lower discriminating power of an item, the harder or easier the items, although we often need such items to have appropriate and representative sampling of the learning content and its goal. Another reason is because the purpose of the item will affect the magnitude of its discriminating power in relation to the total test.[3,6]

The correlation of two indices was shown in Figure 1. Rank spearman correlation (r_s) was 0.37 with p-value 0.001. The scatter-plot chart shows that 74% of items have p 20-80% and D≥0.3. However, if only the items with excellent p (40–60%) and excellent D (≥0.4) are considered, there are 7 (7 %) items which could be Labeled as "excellent."

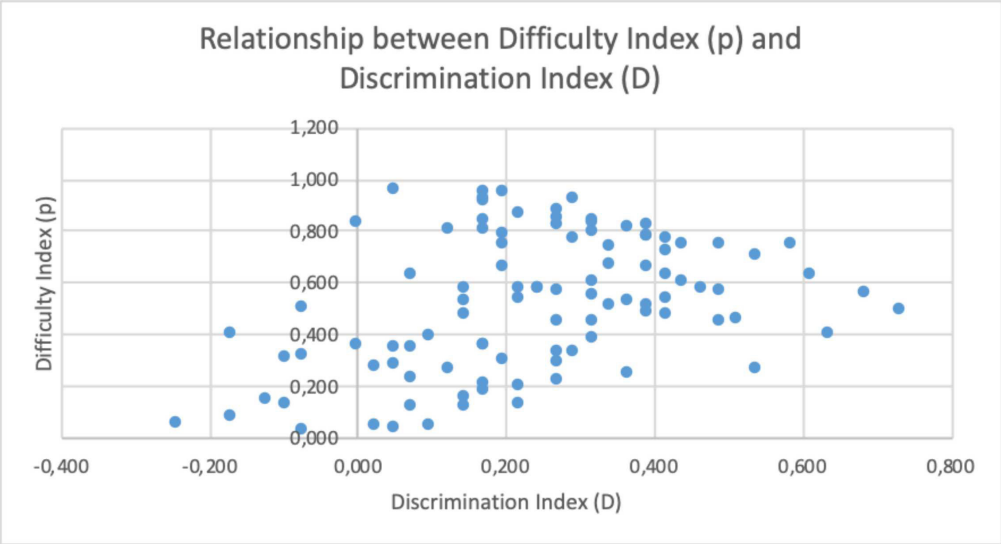

Figure 1. Relationship between Difficulty Index (p) and Discrimination Index (D)

Table 3 show the frequency of distracters of 100 items tested. Distracter efficiency (DE) is used to analyze the functioning and non-functioning distracters. A non-functioning distracter is as an alternative option answer with either a response frequency of <5% or a positive discriminating power and it has a positive correlation with the total score.

As presented in table 3, of all 400 distracters assessed, 111 distracters (27.75%) had a choice frequency of <5% (NFDs). However, 289 distractors (72.25%) had a frequency choice of ≥5 (FDs). The Distracter analysis enables us to identify non-functioning and functioning options, and to evaluate whether the distracters are well-constructed or not. It is tough to have options with equal plausibility to the key answer. Item quality, performance and test outcome are affected by the distractors' functionality, writing flaws and optimum number of choices, which are also interrelated each other. A previous study stated that two functioning distracters in an item were more difficult than three functioning distracters.[15] Current result reported that 42% of items had 4 functioning distracters with DE 0% which suggested the teacher should be more precise and carefully creating for the distracters.

Table 3. Frequency of Distracters

	Frequency	Percentage	DE
Number of Items	100	-	-
Number of Distracters	400	-	-
Distracters with frequency <5%	111	27.75	-
Distracters with frequency ≥5%	289	72.25	-
Functioning Distracters per Item	-	-	-
0	2	2	100%
1	14	14	75%
2	18	18	50%
3	24	24	25%
4	42	42	0%

There were a study suggested the peer-review process of MCQs to significantly improve the psychometric quality of items.[6] Peer reviews require faculty to be trained on standards of MCQ writing and interpreting post-exam item analysis, including the understanding of item characteristics (item difficulty, item discrimination), reliability indices, and performance of distractor options.[16] The process needs sessions of faculty meeting to train staffs to become peer-reviewers (in our institution it was done by the committee of assessment unit), to train the staffs to construct good quality of MCQs and to enhance the staff's skills through feedback they receive from the peer reviews. This process provided an essential quality assurance measure, proved with significant improvement in item discrimination, which in turn improved reliability of the test.

Assessment Unit of Medical Faculty of Syiah Kuala University has followed these strategies by conducting workshops on test development, item writing, and item analysis for all academic staff, to improve the quality of item writing and decrease the number of item flaws, following evidence-based international guidelines. There were also a committee assigned for each block exam to review the newly-developed MCQs from each lecturer, which was sent one week before administering the exam. Unfortunately, this committee do not give the review feedback or student's performance result (item analysis0 to the lecturer who develop items.

However, based on National Board of Medical Examiners (NBME) guidelines, the assessment committee should identify item flaws, recommend removal or medication of items using the guideline, and also provide feedback and recommendations to the item writer and faculty for future improvement of item writing.[16]

Although the result of this study indicated that indices of difficulty and discrimination were acceptable, there were still 39% items with very low discriminating power (means those items couldn't differ student's capacity) and they need to be fully revised. The result also reported that 42% of items had 4 functioning distracters with DE 0% which suggested the teacher should be more precise and carefully searching for the distracters. With the benefit of the MCQ analysis result, we assume that completing 3 aspects of peer-review process will help a lot in improving the MCQ quality. Under this rule, assessment unit of Syiah Kuala University should present outcome of item analysis, including difficulty and discrimination indices, distracter efficiency (DE), to all faculty staff, especially lecturers who develop items, as feedback and suggestions to the writer and faculty. The outcome of item analysis will enable the faculty to evaluate the testing method and also reward those who crafted the excellent items, which hopefully improve the quality of learning and testing method.

CONCLUSION

It takes time and careful selection of items and distracters to construct multi-choice test items for an end of block examination. Item analysis not only offers useful data for item modification and future testing, but also a way to train the lecturers to develop better items.

Quality control is essential for test development in medical schools. The study presented item analysis of 100 specific health problem and applied research block items for 150 third-grade medical students. The analysis showed that majority items were acceptable with small percentage items to be reviewed. For better test quality, feedback and suggestions for the item writer and faculty should also be performed as a part of peer-review process on the basis of item analysis

RECOMMENDATION

The result of item analysis should be presented to item writer in peer-review process, as it could be a consideration to improve the future testing quality.

COMPETING INTERESTS

The authors declare that there are no competing interests related to the study.

AUTHORS' CONTRIBUTION

Novi Maulina – contributed in analyzing data, writing and revising manuscript.

Rima Novirianthy – had full access to data used in the study, also contributed in analyzing data.

REFERENCE

1. Anamuah-Mensah J, Quagrain KA. Teacher competence in the use of essay tests. The Oguaa Educator University of Cape Coast. 1998;12:31-42.
2. Popham WJ. Classroom assessment: What teachers need to know. Allyn & Bacon, A Viacom Company, 160 Gould St., Needham Heights, MA 02194; World Wide Web: http://www. abacon. com; 1999.
3. Quaigrain K, Arhin AK. Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. Cogent Education. 2017 Jan 1;4(1):1301013.
4. Bacon DR. Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. Journal of Marketing Education. 2003 Apr;25(1):31-6.
5. Gronlund NE. Assessment of student achievement. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele; 1998.
6. Abozaid H, Park YS, Tekian A. Peer review improves psychometric characteristics of multiple choice questions. Medical teacher. 2017 Mar 16;39(sup1):S50-4.
7. Shakil M. Assessing Student Performance Using Test Item Analysis and its Relevance to the State Exit Final Exams of MAT0024 Classes-An Action Research Project. In A Paper presented on MDC Conference Day 2008 Mar 6.
8. Hotiu A. The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course (Doctoral dissertation, Florida Atlantic University).
9. Thorndike RM, Cunningham GK, Thorndike RL, Hagen EP. Measurement and evaluation in psychology and education. Macmillan Publishing Co, Inc; 1991.
10. 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. IeJSME. 2009;3(1):2-7.
11. Pande SS, Pande SR, Parate VR, Nikam AP, Agrekar SH. Correlation between difficulty and discrimination indices of MCQs in formative exam in Physiology. South East Asian J Med Educ. 2013;7:45-50.
12. 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 Ahmedabad, Gujarat. Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine. 2014 Jan;39(1):17.
13. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. JPMA-Journal of the Pakistan Medical Association. 2012 Feb 1;62(2):142.
14. Rudner LM, Schafer WD. What teachers need to know about assessment. 2002.
15. 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 Dec;9(1):40.
16. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners; 1998 Oct 20.