The Levels Of Difficulty And Discrimination Indices In Type A Multiple Choice Questions Of Pre-clinical Semester 1 Multidisciplinary Summative Tests

Mitra N K, Nagaraja H S, Ponnudurai G, Judson J P

Item analysis is the process of collecting, summarizing and using information from students' responses to assess the quality of test items. Difficulty index (P) and Discrimination index (D) are two parameters which help evaluate the standard of MCQ questions used in an examination, with abnormal values indicating poor quality. In this study, 120 test items of 12 Type A MCQ tests of Foundation 1 multi-disciplinary summative assessment from M2 / 2003 to M2 / 2006 cohorts of International Medical University were selected and their P-scores in percent and D-scores were estimated using Microsoft Office Excel. The relationship between the item difficulty index and discrimination index for each test item was determined by Pearson correlation analysis using SPSS 11.5. Mean difficulty index scores of the individual summative tests were in the range of 64% to 89%. One-third of total test items crossed the difficulty index of 80% indicating that those items were easy for the students. Sixty seven percent of the test items showed acceptable (> 0.2) discrimination index. Forty five out of 120 test items showed excellent discrimination index. Discrimination index correlated poorly with difficulty index (r = -0.325). In conclusion, a consistent level of test difficulty and discrimination indices was maintained from 2003 to 2006 in all the twelve summative type A MCQ tests.

IeJSME 2009: 3 (1): 2-7

Keywords: Item analysis, difficulty index, discrimination index, type A MCQ, summative tests

Introduction

Designing multiple choice questions (MCQ) is a complex and time consuming process in a multidisciplinary integrated curriculum. MCQs are used mostly for comprehensive assessment at the end of a semester or academic sessions[1,2] and provide feedback to the teachers on their educational actions[3]. Having constructed and assessed a test, a teacher needs to know, how good the test questions are and whether the test items were able to reflect students' performance in the course related to learning[4].

Because of their versatile character, multiple choice questions are the most commonly used tool type for assessing the knowledge capabilities of medical students[5,6]. There are different types of MCQs which have been used in the medical field. The most frequently used type of MCQ is the five choice completion type (type A MCQ)[1,7,8]. The assessment methods used for the Foundation 1 summative assessment course at the end of six months in the medical program of the International Medical University included short answer questions, objective structured practical examinations (OSPEs) and type A multiple choice questions. These test questions were taken from various disciplines like anatomy, physiology, biochemistry, genetics, statistics and behavioural science subjects. Therefore, the summative examination papers for the Foundation 1 course were multidisciplinary. The examination questions had been developed by the content experts who taught the respective disciplines and the questions had been vetted within the individual departments before being sent to the higher level vetting committee, which included senior academics of various disciplines.

One of the major concerns in the construction of test items for an examination is ensuring the reliability of the test items[8]. Item analysis is the process of collecting, summarizing and using information from students' responses to assess the quality of test items[9,10]. The item statistics can help to determine those items that are good and those that need improvement or deletion from the question bank. It allows any aberrant items to be given attention and reviewed. One of the most widely used method in investigating the reliability of a test item has been classical test theory (CT) item analysis[10,11]. This type of item analysis essentially determines test homogeneity. The more similar are the items given in the test; it is more likely that they measure the same kind of intended ability and therefore attaining higher reliability. Item difficulty index is the first item characteristic in classical test theory to be determined[12,13]. This is a common practice as tests are often not regarded as reliable measures of students'

Department of Human Biology, International Medical University, Bukit Jalil, 57000 Kuala Lumpur, Malaysia

Address for Correspondence:
Dr Nilesh Kumar Mitra, Senior Lecturer, Department of Human Biology, International Medical University, No. 126, Jalan 19/155B, Bukit Jalil, 57000 Kuala Lumpur, Malaysia
Email: nileshkumar_mitra@imu.edu.my

performance due to misfit of item difficulty with the ability of the students. In addition to item difficulty, item discrimination is an important index[10]. This provides information on how effectively the items in a given test discriminate between students who are higher in the ability measured and those who are low. Presence or absence of faults logically affects the values of discrimination. Items that discriminate poorly should be inspected for possible deficiencies[11,13,14].

Some basic forms of item analysis has been carried out routinely by the academic affairs department in International Medical University, but the data generated has not been used regularly to test the quality of the questions or for the development of multiple choice questions for the subsequent tests. Hence the present research study was taken up with an objective to analyze the quality of the multiple choice questions (type A) used in the summative assessments of seven cohorts of semester 1 students in the preclinical phase in the International medical University. We have also aimed to find out whether there was any relationship between the item difficulty and item discrimination indices of these MCQ items in the seven cohorts.

Materials and Methods

Data Collection

Multiple choice question items were taken from the twelve Foundation 1 summative assessment test papers from the years 2003 – 2006 (with each year having two cohorts except 2003). Items from two tests held in each cohort from 2003 to 2005 and 1 test held in two cohorts of 2006 were used for analysis. Each of these summative examinations was held during the first six months of the preclinical phase and the test paper was multidisciplinary. A total of 120 test items were selected for the item analysis. Each type A MCQ consisted of a stem and five choices and the students were to select one best answer from these five choices. A correct answer was awarded 2 marks and there were no negative marks for the incorrect answer.

Item Analysis

The results of the examinees' performance in the summative tests were used to analyze the difficulty index and discrimination index of each multiple choice question item. The item difficulty index is calculated as percentage of the total number of correct responses to the test item[15,16]. It is calculated using the formula P=R/T, where P is the item difficulty index, R is the number of correct responses and T is the total number of responses (which includes both correct and incorrect responses). An item was considered difficult when the difficulty index value was less than 30% and the item was considered easy when the index value was greater than 80%. The item discrimination index measures the differences between the percentages of students in the upper group with that of the lower group who obtained the correct responses. At first 27% of the total number (n) of students (varied between 195 to 217 students in different cohorts) were counted[16,18]. Then the total number of students in the upper 27% (UG) who obtained the correct response and the lower 27% (LG) who obtained the correct response was counted. The discrimination index was calculated using the formula D=(UG-LG) / n. The higher the discrimination index, the test item can discriminate better between students with higher test scores and those with lower test scores. Based on Ebel's (1972) guidelines on classical test theory item analysis, items were categorized in their discrimination indices[10,19]. The item with negative discrimination index (D) was considered to be discarded; D: 0.0 – 0.19 – poor item – to be revised; D: 0.2 – 0.29 – acceptable; D: 0.3 – 0.39 – good; D: >0.4 – excellent.

Statistical Analysis

All data were expressed as mean ± standard deviation of the total number of items. The relationship between the item difficulty index and discrimination index for each test item was determined by Pearson correlation analysis using SPSS 11.5.

Results

Figure 1 shows the difficulty index values from the twelve summative tests held from 2003 to 2006. The mean difficulty index of the nine tests were found in the range between 64% and 79% and only two tests (held in the year 2004) had mean difficulty index value more than 80%, which could be considered as easy test paper. M2 / 2006 cohort summative examination MCQs had the mean difficulty index of 64.05, which has been the least among all the tests. Only 40% of the test items in this study crossed the difficulty index of 80% indicating that one-third of the items were easy for the students.

Figure1. Bar chart showing the mean difficulty indices (± S. D) of the individual summative MCQ tests

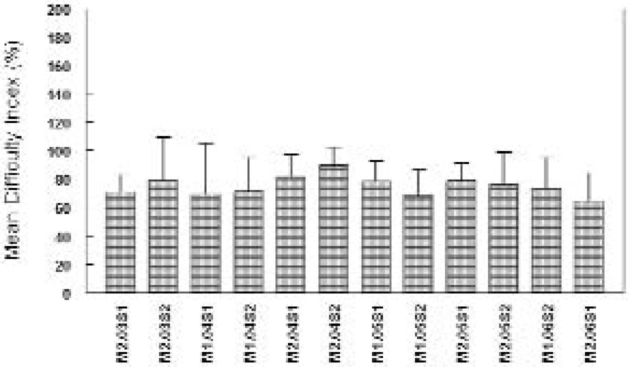


Figure 2 represents the discrimination index values of the twelve summative tests. Except for one test for the cohort M2/2003 held in 2003, all the other tests showed the discrimination index values of 0.2 or higher indicating that the test items were of acceptable discrimination quality.

Eighty (66.7%) out of total 120 test items were found with discrimination index level of 0.2 of higher and were able to discriminate good and weak students. Out of the twelve tests, only one test showed poor mean discrimination index. Seven out of twelve tests showed excellent mean discrimination index, equal to or more than 0.4.

Figure 2. Bar chart showing the mean discrimination indices (± S. D) of the individual summative MCQ tests

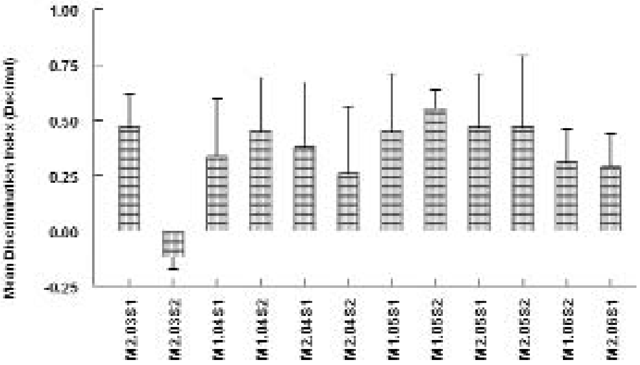
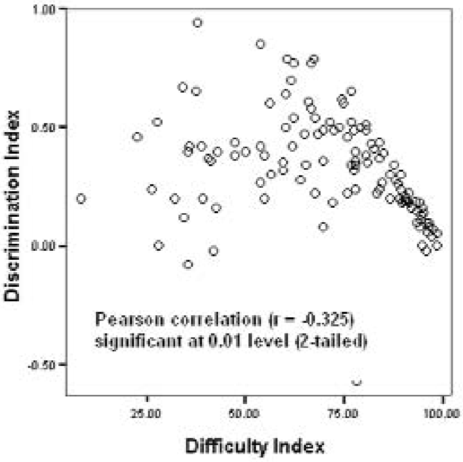


Figure 3. Scatter plot showing relationship between difficulty index and discrimination index of items. Also showed is the Pearson Correlation value. Correlation was tested between individual item's difficulty index and discrimination index score.



When difficulty index was analyzed along with discrimination index, 74% of the test items with poor discrimination index had the difficulty index ranging between 89 – 98%. Forty five out of 120 test items showed excellent discrimination index (0.4 or >0.4). Out of these items 82.5% had the difficulty index between 50 – 79%. 50% of the items with difficulty index greater than 80% were with poor or negative discrimination index.

Pearson correlation between difficulty and discrimination indices showed that discrimination index correlate poorly with difficulty index (r = -0.325). The correlation is significant at 0.01 level (2-tailed). Negative correlation signifies that with increasing difficulty index values, there is decrease in discrimination index. As the items get easy (above 75%), the level of discrimination index decreases consistently (Figure 3).

Discussion

Out of the 12 summative tests conducted from 2003 to 2006 for Foundation 1 course, the mean difficulty index scores of the individual tests were ranging from 64% to 89%. Only two tests had the mean difficulty index value more than 80%. These two tests had many easy multiple choice questions where most of the students got full score in the tests. Thus majority of these tests (with mean difficulty index scores between 64% and 79%) had good MCQ test items. Only 40% of the total test items had difficulty index scores crossing 80%. This observation was similar to a study of year two examinations of a medical school reported by Si Mui Sim et al (2006), who found that about 40% of the MCQ items crossed difficulty index 70% showing that the test items were easy for the examinees[16]. Li Chan Lin *et al* (1999) while doing item analysis of Basic Medical Science items of Registered Nurse Licensure Examination in Taiwan found item difficulty in the range of 10% – 93% with a mean of 48%[17]. Brown (1983) and Algina (1986) have reported that any discrimination index of 0.2 or higher is acceptable and the test item would be able to differentiate between the weak and good students[20,21]. In the present study, we have seen that 80% of the MCQ from the twelve tests had the discrimination index of more than 0.2. Thus it showed that most of the multiple choice questions used in all these summative tests were good or satisfactory questions which would not need any modifications or editing as they were able to differentiate good and weak students. Seven of the 12 tests showed mean discrimination index equal to or more than 0.4, indicating that these MCQ items were excellent test items for differentiating between poor and good performers. Li Chan Lin *et al* (1999) reported that 29% of the multiple choice test items in the Basic Medical Sciences Nursing Licensure examination in Taiwan had the discrimination less than 0.2[17].

Negative correlation between difficulty and discrimination index indicated that with increase in difficulty index, there is decrease in discrimination index. As the test items get easier, the discrimination index decreases, thus it fails to differentiate weak and good students. Si Mui Sim *et al* (2006) found that maximum discrimination occurred with difficulty index between 40 – 74%[16]. In the present study, 82.5% of the test items with difficulty index between 50% and 79% had excellent discrimination index.

For calculation of the discrimination index our study used the method adopted by Kelley (1939) where upper and lower 27% performers were selected[18]. The only limitation of this test is that it cannot be used for a smaller sample size. But in our study, the sample size was from 160 to 205 in various groups and hence the observed results truly reflect the discriminative power of the test items used. One inadequacy of only analyzing a question in terms of its difficulty index is the inability to differentiate between students of widely differing abilities. Subjective judgment of item difficulty by item writer and the vetting committee may allow faulty items to be selected in the item bank. Items with poor discrimination index and too low or too high difficulty index should be reviewed by the respective content experts[22]. This serves as an effective feedback to the respective departments in a Medical school about the quality control of various tests. When the difficulty index is very small, indicating difficult question, it may be that the test item is not taught well or is difficult for the students to grasp. It also may indicate that the topic tested is inappropriate at that level for the students[23]. In the scatter plot, there is a wide variation in the discrimination indices with similar levels of difficulty index below 75%. The negative marking for incorrect responses has not been started in the summative MCQ tests in our university. Hence guessing practices by the students might have caused the wide variation in discrimination indices.

A consistent level of test difficulty and discrimination indices appears to be maintained from 2003 to 2006 in all the twelve summative tests. This observation could be due to the fact that the test items went through a series of vetting before being selected for the examinations. The quality of test items may be further improved based on action taken in reviewing the distractors by the item writer based on the calculated discrimination and difficulty index values. Few common causes for the poor discrimination are ambiguous wording, grey areas of opinion, wrong keys and areas of controversy[24,25]. Items showing poor discrimination should be referred back to the content experts for revision to improve the standard of these test items. It is important to evaluate the test items to see how effective they are in assessing the knowledge of the students based on the difficulty and discrimination indices of the test items.

Conclusion

There is a consistent spread of difficulty in type A MCQ items used for three years in 12 summative tests. The test items that demonstrated excellent discrimination tend to be in the moderately difficult range. Factors other than the difficulty, like the faulty test item constructions, are not significant at the summative tests in the preclinical Foundation I summative examinations at International Medical University. The results of this study should initiate a change in the way MCQ test items are selected for any examination and there should be a proper assessment strategy as part of the curriculum development. Much more of these kinds of analysis should be carried out after each examination to identify the areas of potential weakness in the one best answer type of MCQ tests to improve the standard of assessment.

Acknowledgement

Academic affairs department of International Medical University.

REFERENCES

1. Skakun EN, Nanson EM, Kling S, Taylor WC. A preliminary investigation of three types of multiple choice questions. Med Edu 1979; 13: 91-96.
2. Weech AA. Multiple choice examinations in medicine: A guide for examiners and examinee. Pediatrics. 1961; 28:106.
3. Cunnigham S. Can pediatric medical students devise a standard for their colleagues? Arch Dis Child 1999; 80: 573-575.
4. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of 'fact recall' with 'higher order questions in multiple choice examinations as predictors of clinical performance of medical students. Acad Med 1990; 65: S59-S60.
5. Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian 2005; 12: 19-24.
6. Irish B. Preparing for multiple choice paper. Practitioner 2005; 249: 124-127.
7. Skakun EN, Nanson EM, Taylor WC, Kling S. An investigation of three types of multiple choice questions. Ann Conf Res Med Educ 1977; 16: 111-116.
8. Senanayake MP, Mettananda DSG. Standards medical students set for themselves when preparing for the final MBBS examination. Annals Acad Med 2005; 34: 483-485.
9. Kehoe J. Basic item analysis for multiple choice tests. Practical Assessment, Research, Evaluation. 1995; 4: 10.
10. Zubairi AM, Kassim NLA. Classical and Rasch analysis of dichotomously scored reading comprehension test items. Malaysian J of ELT Res 2006; 2: 1-20.
11. Davies A. Principles of language testing. Cambridge. Oxford: Basil Blackwell Ltd. 1990.
12. Warburton WI, Conole GC. Key Findings from recent literature on Computer-aided Assessment. In, *ALT-C 2003, Sheffield*, 2002.
13. McAlpine M, Hesketh I. Multiple Response Questions- Allowing for Chance in Authentic Assessments. In 7th International CAA Conference (Ed, Christie, J.) Loughborough University, Loughborough. 2003.
14. Bachman LF. Fundamental considerations in language testing. Oxford: Oxford University Press. 1990.
15. Johnston A. LTSN physical sciences practice guide effective practice in objective assessment the skills of fixed response testing. http://www.heacademy.ac.uk/physsci/home/pedagogicthemes/assessment 2003
16. Si-Mui Sim, 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.
17. Lin LC, Tseng HM, Shiao-Chi WU. Item analysis of the registered nurse licensure exam taken by nursing candidates from vocational nursing high school in Taiwan. Proc Natl Sci Counc 1999; 9: 24-31.

18. Kelley TL. The selection of upper and lower groups for validation of test items. J Educ Psychol 1939; 30: 17-24.
19. Ebel RL. Essentials of educational measurement. 1st Ed. New Jersey; Prentice Hall. 1972.
20. Brown FG. Principles of educational and psychological testing. 3rd Ed. New York: Holt, Rinehart and Winston. 1983.
21. Crocker L, Algina J. Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston. 1986.
22. Meshkani Z, Abadie H. Multivariate analysis of factors influencing reliability of teacher made tests. Journal of Med Ed Winter. 2005; 6: 149-159.
23. Marso RN, Pigge FL. An analysis of teacher made tests and item construction errors. J Contemp Edu Psych 1991; 16: 279-286.
24. Baner D, Kopp V, Fischer MR. Answer changing in multiple choice assessment change that answer when in doubt and spread the word. BMC Med Education. 2007; 7: 28.
25. Meckenzie J. Vague and ambiguous questions on multiple choice exercises: the case for educational philosophy and theory. 1994; 26: 23-33.