ANALYZING STUDENTS RESPONSES IN MULTIPLE CHOICE QUESTIONS (MCQS) USING ITEM ANALYSIS

Madhavi Sharma and Prof. C.R.K Murthy
IGNOU, New Delhi, India

Abstract-Assessment is an integral part of any academic programme. It indicates how much learning is done by students from the curriculum. Multiple Choice Questions (MCQs) are commonly used as formative and summative assessment components. Item Analysis is a technique for determining the quality of each individual item in a test. It helps to analyze the learning process in students and improve upon the Item writing techniques.The aim of this study is to evaluate items i.e. Multiple Choice Questions (MCQs) using Item Analysis parameters to (i) analyse the quality of items in a test (ii) examine the relation between difficulty index (p) and discrimination index (DI) (iii) find out the distractor(s).

A test having 20 questions was given to students of research scholarsenrolled in various disciplines. A total of 32 responses received from students. Each item i.e. each MCQ wasanalysed for Difficulty Index DIF I or p-value and Discrimination Index (DI) and to find the relationship between them. The Distractor analysis was done to identify the distractors effecting students choices.

The difficulty index DIF I or p-value of 12(60%) items lying in the range of(30%-70%) indicated that items were in acceptable range while another6(30%)items were found easy where difficulty index (DIF I >=70%) and 2(10%) items were hard with DI<30%. The internal consistency Kuder-Richardson'sreliability $KR_{21}$of the test was 0.72and Standard deviation was 3.68. The results showed that 4(20%) items having Discrimination Index DI<0.20 (poor items), 4(20%) items with DI in the range of0.20 to 0.24 (acceptable i.e. marginal items), 8(40%) items were in the range of DI (.25-.35) and remaining 4(20%) items having DI>=0.35 indicated that the items were excellent. The mean difficulty index and discrimination index were found as 59.72% and 0.25 respectively.

As majority of items were in excellent and acceptable range, indicates that test was well designed and students scored good scores. This helps faculty to analyze the students knowledge and understanding about the curriculum.

Key words : **DIF I – Difficulty Index, DI – Discrimination Index, MCQs – Multiple Choice Questions, Distractor Analysis, NFD – Non-Functional Distractor, FD – Functional Distractor**

Introduction

Multiple Choice Questions (MCQs) have been widely used in field of education, marketing, surveys, polls, trainings, opinions etc. In the field of education, MCQs have been used to assess student learning and performance in the academic programme in the class tests and the final examinations i.e. in the form of formative and summative assessment. The computerized evaluation of MCQs made it much more fascinated as the users can get the instant results. Hence, attracts students and other users to go through the MCQs. The faculty designed an objective type test paper having a number of questions with multiple choice questions (MCQs) including one correct choice. Sometimes MCQs carry negative marking also. Few patterns of designing a Test having MCQs are as below:

1. The students have to attempt all the questions in the sequence and submit response one by one till the last question. The result of correct and incorrect responses are displayed at the end of the completion of the test i.e. after the attempting the last question.
2. Students can attempt any question from the test without actually following the sequence of questions and finish the test by clicking submit button once only. The number of correct answers, incorrect answers and marks obtained are displayed instantly after finishing the test.
3. In this pattern, the correct responses are displayed after attempting each question by the student. In this case, learner become cautions, and comes to know his/her performance MCQs also motivates learners to do well in their programme.
4. In some cases, all the MCQs are compulsory to attempt while in others there is no such compulsion. Time limit is also an important constraint in MCQs especially in competitive exams.

Item Analysis is used to study the pattern, quality and behavior of MCQs in a test. Both Item and Test analysis are used to improve teaching practices and to improve the classification of students. After analyzing the students, it becomes easy for the teacher to find strength and weakness of students and take further improvements in teaching. Item Analysis is useful for analyzing the quality of questions used in the test. It will be useful to know how many questions are easy, hard or medium for faculty and students and whether they needs revision or discard.

Designing Item Analysis is a time taking process. Item analysis involves a number of tests which provides useful information to analyze and improve the quality of each item and accuracy of multiple choice questions. It examines student responses to each question to assess the quality of the items and test as a whole. According to (Thompson & Levitov, 1985), Item Analysis

investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test.It uses this information to improve item and test quality. There are three elements of Item Analysis:

1) Item Difficulty (DIF I)
2) Item Discrimination (DI)
3) Item Distractors

Item analysis implies how students have perceived the items and responded. The analysis of distractors further explores to think why students would have chosen a particular option and not the right option. Also, to explore the reasons for choosing the mixed responses i.e. whether distractors were written logically similar or students couldn't understand them fully. Hence, the construction of distractors is equally important to be designed carefully so that the students understand them clearly.

Methodology

A test paper containing 20 Multiple choice questions was designed on Research Methodology. A total of 32research students were participated. The learners were given a set of 20 questions to respond on Research Methodology. The MCQs are coded in Google Form and sent to learners through email. The learners responses are analysed in Excel. Also, it was not a time bound. The learners were given 2 days time to respond back. Each item contains a stem and 4 choices including one correct choice and three other (incorrect) distractors. Each correct answer scored 1 mark and there is no negative marking for incorrect response. The evaluation was done out of 20 marks. Pre-validation of the questions in test paper is validated by the faculty and designers of question paper. The post-validation of the test paper was done by item analysis. The questions in the test paper were evaluated with Answer key and final scores were calculated. The scores of all the learners were sorted in descending order i.e. from highest to lowest score. Based on Kelly's (1939) derivation, the upper and lower 27% rule is usually used in Item analysis.

The upper 27% scores are taken as high group (H) and low 27% scores (L) are taken as low group i.e. H = 9 and L = 9 in the present study.

Each of the 20 items were analysed and based on data, following indices were calculated:

(i) Difficulty Index (DIF I)

Difficulty index of each item is calculated as a percentage of the total number of correct responses to all the test items.

DIF I = (H + L) / N x 100

Where H = No. of students opted right answer in High group

L = No. of students opted right answer in Low group

N = Total No. of students in both the groups

Difficulty Index is also calculated as:

p = R / T

Where p = Item Difficulty Index, R is the number of correct responses and T is the total responses (correct and incorrect). The value of p lies between 0 and 1 (Hotiu 2006). This value is multiplied by 100 to get the percentage of students who responded correct answers. Hence, p-value ranges between 0 and 100% (Table 1). Higher the value of p, easier the item is and vice-versa. An ideal item will be one which has average difficulty index between 30%-70%.

Difficulty Index (p)	Quality of Item
<30%	Difficult
30% -70%	Acceptable
>70%	Easy

Table 1 : Interpretation of Difficulty Index

(ii) Discrimination Index (DI)

Discrimination provides the idea of rough validity of an item in the test. It is a measure of an item's validity to discriminate between those who scored high on the test to the students who scored low. In other words, DI is ability to differentiate between students of higher and lower abilities.

The difference between the correct responses given by students in High group and Low group of the entire group indicates whether an item has discriminated between students in high group and low group in a test.TheEbel's (1972) Guidelines on Classical test theory item analysis described the items were categorized in their discrimination indices.Item discrimination (also called Point Biserial) is a measure to differentiate between the performance of students in the high score group and those in the low score group and ranges between 0 and 1.

DI = H – L / N

The items were categorized (Table 2)in their discriminating indices as below:

Discrimination Index (DI)	Quality of Item	Remarks
>0.35	Very Good items, Excellent Discrimination	Satisfactory item
0.25 to 0.35	Reasonably Good, Good Discrimination	No or little revision required
0.20 to 0.24	Marginal item, Acceptable Discrimination	Needs revision
<0.20	Poor item, Poor Discrimination	Discard item, needs complete revision

Table 2: Interpretation of Discrimination Index

(iii) Distractor Analysis

All other options except the correct answer in a MCQ are known as Distractors. The distractor analysis helps in finding out plausible distractors from implausible ones. A zero response indicates that the distractor was not responded by any of the students i.e. implausible distractor.

A Non-Functional distractor (NFD) in an item is defined as an option selected by less than 5% ofthe students. While Functional distractor (FD) is an option selected by >=5% of students among the non-correct options in a MCQ. DE ranges from 0 to 100% based on NFDs in an item. Also, DE would be 0,33.3%, 66.6%, 100% for 3,2,1 or no NFDs in an item.

Statistical Analysis

The internal consistency Reliability test was conducted using Kuder-Richardson formula

$KR_{21} = (n/n-1) * [1 – ((M-(M * M/N)/(SD * SD)]$

And found the value of KR_{21} was 0.61

Results

A test containing 20 MCQ items was given to students. The responses were recorded and analysed. The marks of 32 students are ranged from 6 to 18. The difficulty index DIF I ranged from 0 to 100%. The mean Difficulty Index p was found as62.5%.The Discrimination Index ranged from 0 to 0.35. The Mean Discrimination Index DI was found as 0.23.Whereas,the Mean and Standard Deviation of correct responses was 10.75 and 3.68 respectively. The 60% of the items were found in acceptable range having difficulty index between 30%-70% and 30% items having Difficulty index more than 70% were found as easy items while 10% items were hard. Such items may be rejected or revised.

Based on Difficulty Index DIF I, the quality of the item as seen as below:

P value	No. of items (%)	Interpretation	Action
<30%	2 (10%)	Difficult	Review for confusing language
30% - 70%	12 (60%)	Average/ Acceptable	Keep
>70%	6 (30%)	Easy	Discard/ Revise

Table 3 : Difficulty Index based on students responses

The items between 50-60% are considered as ideal, good quality.As 2(10%) items were hard , hence items need revision for language change or construction of MCQ. Graphically, it can be seen as below:

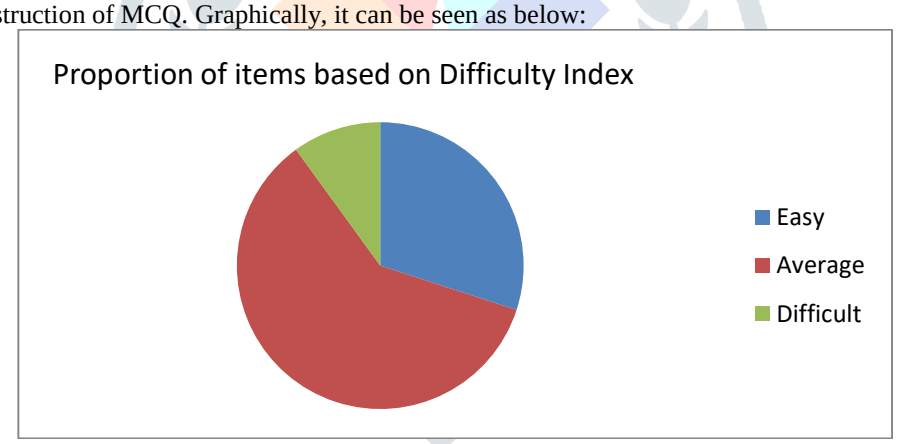

Figure 1 : Proportion of Difficulty Index in the test

The Discrimination Index can be seen as:

DI Range	No. of Items	% Items	Interpretation	Action
>0.35	4	20%	Excellent	Keep/ review
>=0.25 and <0.35	8	20%	Good	Keep
>=0.20 and <0.24	4	40%	Marginal	Revise
<=0.20	4	20%	Poor	Discard/revise

Table 4 : Number of Items in Difficulty index Range

Some other indices can be seen as below:

Parameter	Mean value
Difficulty Index (DIF I)	59.72%
Discrimination Index (DI)	0.25
Standard Deviation (SD)	3.68
KD_{21}	0.72

Table 5: Analysis of various parameters

The relation between difficulty index (DIF I) and discrimination index (DI) is shown as graphics as below:

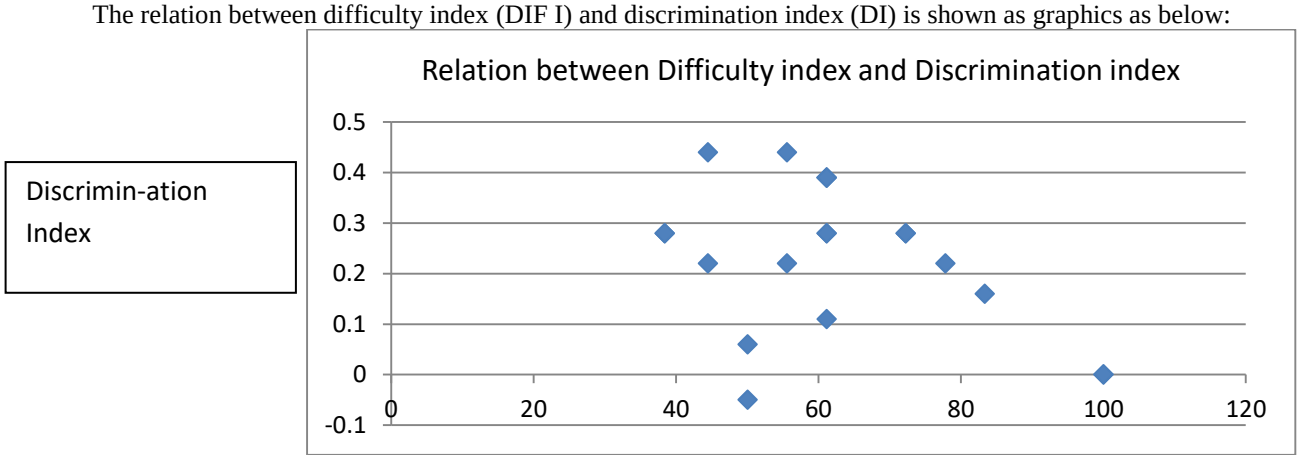

Figure2: Relation between Difficulty index and Discrimination index of items

A total of 20 items had 60 distractors. Out of these , about 15% items had Non-functional distractors (NFDs) and about 75% items had Functional distractors (FDs). The distribution range was from 0% to 100%. There were 2 (10%) items with 0-NFD, 3(15%) items with 1-NFDs and FDs were present in 15(75%) of items.

Discussion

It is very important that a method of assessment to be evaluated regularly. In curriculum development, the development of assessment strategy is very crucial as this helps in evaluating learners knowledge. During the course sessions, test with MCQs motivates students to move ahead in their courses. Also, MCQs reducing the burden of writing the descriptive responses on students but at the same time providing information to teacher about students learning. Post exam analysis of MCQs help to assess the quality of each test item and test as a whole.

The Mean and Standard Deviation were found to be 10.75 and 3.68 respectively. The items with Difficulty index DIF I between 30-70% were considered as acceptable. The Mean DIF I in this study was 59.72% which lies in acceptable range of (31-70%). Hard items where DIF I < 30% resulting low scores and low motivation among as students gained less marks. The easy items where DIF I >70% inspired students to take test more interestingly. The difficult items may be reviewed for possible difficult language used, wrong answer key or tricky construction of the item.

Higher the Discrimination Index DI, better the item is. The item is more able to discriminate between students of higher and lower groups. Items with low discrimination value are ambiguously written and needs to be examined. However, removing items with low DI values and negative values affects seriously on the validity of the test. There might be various reasons for low value of DI (Meherson and Lehman, 1991) as sometimes such items are required to have adequate sampling of the course contents and course objectives. The purpose of the item in relation to the entire test has effect on the magnitude of its discrimination power. The value of DI as 1 is considered as ideal as it exactly discriminates between students in high and low group. The negative values of DI i.e. DI<0 indicates that students in low group answer more correctly than students in high group. In this case, there may be many reasons. The students in low group might have guessed the correct answer whereas students in high group couldn't reach the conclusion to make the right choice.

The too easy items were responded correct by most of the students (DI>35) while items with poor (DI <.20) i.e. too hard items were answered wrongly by majority of students. Its value ranges between 0 and 1. In the present study, mean DI was 0.25. Also, 12(60%) Items in the range of 0.20 to 0.34 are more reliable and are good for test reliability. The items with poor discriminatory index may be reviewed in order to improve students performance. There was negative DI in one item i.e. students in low group answered correctly compared to the students on high group, perhaps by chance. Also, all the students responded correctly for 2(10%) items i.e. DI was 0 in this case and hence no discrimination found between high and low students for these 2 items.

Previous studies have indicated similar results . A study by Patel and Mahajan showed 80% of the items fall in the acceptable range.

In a study by (Gajjar S, Sharma R, Kumar P, Rana M) on 148 MBBS students in a 50 items test, 24 (48%) items found in the category of (Good to Excellent) having DIF I (31-60%) and 15 items had (DI>=0.25). The mean DE was 88.6% which was acceptable. The NFDs were only 11.4%. Mean DI was 0.14 fall in range of DI<0.15 and 10(6.75%) items with negative DI indicated that there might be low preparation of students or some issues with framing few MCQs.

In a study by Shafizan Sabri, when a test containing 41 items were taken on 16 music students, the item analysis result concluded that 44% percent of the total test items exceed the difficulty index of 0.8 showed that items were easy. Also, 59% percent of items were found in acceptable range of discrimination index.

In another study by (Rao C, Prasad KHL, Sanjitha K, Permi H, Shetty J, 2017), 85% items were in acceptable range, 5% items were too easy while 10% were too difficult. The DI of 60% items were excellent, 10% items were good , 15% were acceptable and another 15% items were poor.

A study by (Quaigrain K and Arhin AK, 2017), showed that the internal consistency reliability of the test was 0.77using KR_{21}formula, the Mean Difficulty index and Mean Discrimination index was found to be 58.46% and 0.22 respectively which were quite similar to results of this study.

A study of item analysis on medical students by Christian DS, Prajapati AC, Rana MB, Dave VR, 2017) indicated that reliability of test Kuder-Richardson$_{21}$ ranged from 0.29 to 0.52. Also, 79 out 200 MCQs had Discrimination index (DI<0.15 i.e. poor) and 61 items had (DI>=0.35 i.e. excellent), 47 had DI between (0.15 and 0.24 i.e. marginal items) and 13 had (DI between 0.24 and 0.34 i.e. good items). Also, 17.3% distractors were not selected by anyone.

Another study conducted by (Namdeo SK and Sahoo B), 2016 showed that 14(56%) items were in acceptable range, 32% were too easy and 8% were too difficult. Also, the Discrimination index DI of 48% items was excellent, for 12% items DI was good and for 32% items DI was poor. 53.4% NFDs were present in 22 items, 12% items had no NFDs, whereas 32%, 40%, 16% items had 1,2,3 NFDs respectively.

The results from a study by (Mukherjee P and Lahiri SK, 2015) found mean difficulty index p and discrimination index DI as 61.92% and 0.31 respectively. Also, 46.67% items were called as *ideal* having p-value from 20%-90% and DI>=0.3. The internal consistency reliability of the test KR_{20} was 0.9 almost similar to the results of present study.

Conclusion

The difficulty index DIF I of 2(10%) items lying in the range of DI(0-30%) indicates that two items were too hard. Another 6(30%) items were found too easy where difficulty index (DI>=70%) and rest of the 12(60%) items in the range of DI(30%-70%) were found in acceptable range. The internal consistency reliability $KR_{21}$of the test was 0.72 and Standard deviation was 3.68. The mean score was 10.75.

4(20%) items having Discrimination Index DI<0.20 were found as poor items and required to be written differently or discard completely. 4(20%) items found with DI in the range of0.20 to 0.24 were marginal/acceptable items and need little revision. Also, 8(40%) items were found in the range of DI (.25 to .35) and 4(20%) items having DI>=0.35 indicated that the items were excellent.

An ideal item will be one which has average difficulty index between 31% and 70% and high discrimination (DI>0.35). Items analyzed in the test were neither too easy nor too hard (mean difficulty index = 59.72 and discrimination index was 0.25), which is in acceptable range. It was observed that the majority of items fulfilled the criteria of acceptable difficulty and good discrimination. The results of this study will initiate a different method of selecting questions in MCQ test and review the assessment strategy in curriculum development.

The results of the study will assist faculty and curriculum designers to design an objective type question paper i.e. to write/revise questions and distractors to improve the performance of students and strengthening learning process in students. Such studies if conducted after each examination, will be helpful to assess the level of knowledge understanding in learners about the courses.

References

1. Thompson B. &Levitov J.E (1985). Using microcomputers to score and evaluate test items. Collegiate Microcomputers,3, 163-167.
2. Kelly TL , 1939, the selection of upper and lower groups for validation of test items. Journal of Educational Psychology. Vol 30, pp 17-24.
3. Hotiu A. The Relationship between Item Difficulty and Discrimination indices in Multiple choice Tests in a Physical Science Course Florida: Florida Atlantic University; 2006.
4. Ebel RL, Essentials of educational measurement 1st Edition, Printice Hall. 1972.
5. Mehrens, W. A., & Lehman, I. J. (1991). Measurement and evaluation in education and psychology (4th ed.). Belmont, CA: Wadsworth/Thomson Learning.
6. Sabri, Shafizan (2013). 'Item Analysis Of Student Comprehensive Test For Research In Teaching Beginner String Ensemble Using Model Based Teaching Among Music Students In Public Universities' In International Journal of Education and Research Vol. 1 No. 12
7. Patel KA, Mahajan NR. Itemized analysis of questions of multiple choice question exam. Int J Sci Res 2013; 2:279-80.
8. 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 Gujarat, Indian J Community Med 2014;39:17-20 .
9. Rao C, Kishan Prasad HL, Sanjitha K, Permi H, Shetty J. Item Analysis of multiple choice questions : Assessing an assessment tool in medical students. Int J EducPsychol Res 2016;2.201-4.
10. Quaigrain K, Arhin AK. Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation.. Retrieved from http://dx.doi.org/10.1080/2331186x.2017.1301013
11. Christian DS, Prajapati AC, Rana BM, Dave VR. Evaluation of multiple choice questions using item analysis tool : a study from a medical institute of Ahmadabad, Gujarat. Int J Community Med and Public Health. 2017 Jun;4(6): 1876-1881 http://www.ijcmph.com
12. Namdeo SK, Sahoo B. Item Analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. Int J Res Med Sci. 2016 May; 4(5): 1716-1719 www.msjonline.org
13. Mukherjee P, Lahiri SK. Analysis of Multiple Choice Questions (MCQs) : Item and Test Statistics from an assessment ina medical college of Kolkota, West Bengal. IOSR Journal of Dental and Medical Sciences (IOSR-JDMS). Volume 14, Issue 12 Ver. VI (Dec. 2015), PP 47-52.
14. Sharma M. Development Software for conducting on Demand Examination and Question Banking for Open Learning System in Network Environment : Case Study – CIC Programme of IGNOU. International Conference on Collaborative Networked Learning 1998, pp 91-103.