NAEM JOURNAL
ISSN 1997-4248
Volume 16 Issue 01
Serial 31

Evaluation of Multiple Choice Questions (MCQ) in Higher Secondary Certificate (HSC) Examination

Rowshan Ara Parvez[1]
Muhammad Kamal Uddin[2]
Md. Shahjahan[3]

Abstract

This study describes the importance of evaluation of MCQ in higher secondary certificate examination. The objective of the present study was to thoroughly evaluate MCQs of selected subjects in HSC Examinations. Qualitative & quantitative approaches have been followed and purposive sampling techniques were deployed in the study areas. Qualitative data was collected through Focus Group Discussion (FGD). For quantitative part, the primary data as well as secondary data were collected. Primary data were collected through structured questionnaires and secondary data were collected from Dhaka education board. This study revealed that, the questions that were asked in each subject in 2018 had more difficulty and the percentage of 3 functional distractors were highest for 2018 than 2016. On the other hand, the discrimination power of the majority items were good in 2016 and the number of functional distractor were highest that of than 2018. In the case of 3 functional distractors, the percentage was lowest for Civics i.e. humanities and the overall percentage of functional distractors was highest for Physics i.e. science. Internal consistency reliability was good in 2016 and correlation between creative question CQ and MCQ were high in 2018. In cognitive level analysis according to Bloom's Taxonomy, it can be seen that in most cases the experts could not agree that the item what actually measures. FGD revealed that formative assessment was absent in all colleges, those who had received training do not have sufficient knowledge to construct MCQ, however many have not received training.

Keywords: Multiple Choice Questions, Difficulty Index, Discrimination Index, Functional Distractors, Non-functioning distractors, Focus Group Discussion, Internal Consistency Reliability.

[1] Lecturer, Dept. of Psychology, Begum Badrunnessa Govt. Girls College, Bakshibazar, Dhaka-1211, Bangladesh. E-mail: rowshan.p30@gmail.com cell: 01721784530.
[2] Professor, Department of Psychology, University of Dhaka, E-mail: kamaluddin67@hotmail.com , cell- 01713456644
[3] Professor, Training Speacialist, NAEM

Introduction

The (MCQs) are widely used on classroom tests in colleges and universities and they often account for a substantial portion of a student's course grade (Mavis, Cole, & Hoppe, 2001; McDougall, 1997). A typical MCQ consists of a question, referred to as the stem, and a set of two or more options that consist of possible answers to the question. The correct or best answer is referred to as the keyed option and the remaining options are called distractors (Cizek, & O' Day, 1994). Function of a distractor is to attract students who do not know the correct answer while students who know the correct answer ignore them.

The MCQ format offers many advantages. It allows teachers to efficiently assess large number of students with minimal human intervention (McCoubrie, 2004); grading tends to be quick and easy, especially when students indicate their answers on an optically scanned MC response sheet (DiBattista & Kurzawa, 2011). Another important advantage is that a well-constructed MCQ test can yield test scores at least as reliable as those produced by a constructed-response test, while also allowing for broader coverage of the topics covered in a course (Downing, 2002; Bacon, 2003). Despite these advantages, MCQ testing is often criticized. Some authors have pointed out that MCQs are only capable of testing factual information and so are ill suited for testing higher-order cognitive skills (Walsh & Seldomridge, 2006; Masters, Hulsmeyer, Pike, Leichty, Miller, & Verst, 2001). However, it is clear that thoughtfully written MCQ items can serve to assess higher-level cognitive processes, although creating such items does require more skill than writing memory-based items (Buckles & Siegfried, 2006; Palmer & Devitt, 2007). Furthermore, MCQs can accurately discriminate between high-and-low achieving students (Downing, 2002; Schuwirth, & van der Vleuten, 2003).

Good item construction is critical since test grades affect students' educational outcomes and subsequent career paths (Tarrant, Knierim, Hayes, & Ware, 2006). However, well-constructed MCQ items are time consuming and difficult to write, and it is estimated that a good MCQ may require about one hour to construct (Farley, 1989). Teachers often spend a great deal of time constructing the stem and much less time on developing probable options. High quality MCQ requires the plausible options to be well-written (Haladyna, & Downing, 1989). Poorly constructed options or distractors frequently contain cues that allow students to guess the correct answer without the prerequisite knowledge and as a result, the item poorly discriminates between high-and-low achieving students (Downing, 2002; Biggs, 1999).

A functional or acceptable distractor is defined as an option with either a response frequency of ≥5% or a negative discriminating power. Conversely, a non-functional distractor is defined as an option with either a response frequency of <5% or a non-negative (zero or positive) discriminating power (Tarrant, Ware, & Mohammed, 2009; Ware & Vik, 2009). Di Battista and Kurzawa (2011) reported that as the number of functional distractors in MCQ items increased, discriminatory power increased and items became more difficult. They further reported that the discriminatory power of items fell off sharply as the number of distractors with a non-negative discrimination correlation increased, but discriminatory power was barely affected by the presence of distractors that were selected by less than 5% of examinees.

It appears from the above literature that there are many factors to consider when evaluating the quality of MCQs. One might examine the extent to which items conform to widely accepted item-writing guidelines (Haladyna, Downing, & Rodriguez, 2002). Another way to evaluate the

quality of MCQ items involves analyzing the responses that examinees make. MCQ testing has been a part of public examinations in Bangladesh. The first MCQ tests were included at the SSC Examinations. About 35 - 40% marks are currently allocated for MCQs on public examinations namely on JSC and SSC. These tests are developed by teachers teaching the courses. However, a few teachers have proper knowledge and adequate training in developing high-quality MCQs. Teachers either develop MCQs themselves or rely on MCQ test-banks as a source of questions, both of which may result in question quality which is less than optimal.

Statement of the Problem

HSC Examination is one of the most important public examinations in Bangladesh. It is very important to construct questions for higher secondary certificate examination. Many believe that quality of education has drastically fallen and one of the reasons is MCQs they claim. There are often substantial deficiencies in MCQ tests prepared by course teachers (Mehrens & Lehmann, 1991).

The present study aims to address the following questions concerning MCQs:

1. Are they reliable?
2. Are they valid?
3. Are they acceptable or optimum in difficulty level?
4. Are they able to mat up discrimination the key option?
5. Are distractors functional?

Rationale of the Study

Higher Secondary education is important to build up a career of the students. Government of Bangladesh has taken keen interest to improve the quality of higher secondary education. Quality at all levels of education has been a matter of growing concern for teacher's educators and policy makers in Bangladesh. It has been a talk of the town today that quality is gradually falling. Many believe that quality of education has drastically fallen and one of the reasons is MCQs, they claim. However, there is no known research in Bangladesh that has empirically addressed the issue. MCQs is a very popular type of assessment which is sometimes considered as easy tools for assessment and proper evaluation of the efficiency of the distractors will help the students to removes non-functioning distractors. The main purpose of this study, therefore was to examine the quality of whole test and individual items of MCQs on HSC Examination. Specifically, we analyzed MCQs on HSC Examination in 2018 and 2016 focusing on three key characteristics of individual MCQ: difficulty, discrimination, and effectiveness of the distractors.

Objectives of the study

The general objective of the present study was to thoroughly evaluate MCQs of selected subjects in HSC Examinations. The specific objectives were as follows:

1. To evaluate the effectiveness of MCQ items.
2. To examine reliability of MCQs.
3. To explore association between MCQs and CQs.
4. To assess reflection of curriculum intentions in MCQs.

5. To investigate whether criteria or guidelines were followed in preparing MCQs.
6. To suggest ways and means to improve quality of MCQs in future examinations.

Scope and limitation of the study

The present study examined the quality of multiple choice questions in HSC examination. There are 10 education boards (including Technical & Madrasha Boards) in our country. We selected only Board of Intermediate and Secondary Education, Dhaka among them. We got access to Board of Intermediate and Secondary Education, Dhaka for collecting secondary data. We also got the scope to look into the activities of Board of Intermediate and Secondary Education, Dhaka and also knew the scenario of the quality of Multiple Choice Questions and the numbers which are achieved by the students. In this research, only two years' multiple choice questions were selected. It is not enough for evaluating the quality of multiple choice questions at HSC level. It is so much time consuming that's why we could not select more years for this research purpose. Time and budget was not enough to do this research,

Methodology of Study and Data collection

Design

This study followed a mixed method research design involving both qualitative and quantitative approaches. For qualitative part, we conducted three Focus Group Discussion (FGDs) enquiring perception and experiences of teachers of urban, semi-urban, and rural colleges regarding Higher Secondary Certificate Examinations. For quantitative part, we collected primary as well as secondary data. To collect primary data, we used structured questionnaires which verified whether MCQs of HSC in 2016 and 2018 Examinations were prepared in accordance with Bloom's Taxonomy. Secondary data were collected from the board based on the following inclusion criteria.

Sampling

Complete responses of randomly selected 1000 students each in Bangla, ICT (Compulsory), Physics (Science), Civics (Humanities), and Management (Business) for 2016 and 2018 HSC Examinations were collected from Dhaka Education Board.

Inclusion criteria

1. Examinees who answered all the MCQs in selected subjects
2. Examinees representing male and female
3. Examinees representing urban, semi-urban, and rural institutions

Data Analysis

Apart from descriptive statistics (e.g., mean, standard deviation etc.), we conducted the following:

1. Item Difficulty Index
2. Item Discrimination Index
3. Distractor Functionality Index

4. Internal Consistency Reliability (Cronbach Alpha) of each set of questions.
5. Pearson product moment correlation between total scores on MCQs and CQs.

FGDs and workshop to assess reflection of curriculum intentions and to ascertain whether criteria or guidelines were followed in preparing MCQs.

The difficulty (or facility) of an item was determined by computing P. The P denotes the percentage or proportion of students answering an item correctly. The recommended range of P is 0.30 to 0.70 (De Champlain, 2010; Oermann & Gaberson, 2013).

$$P = \frac{Noofstudentsansweringtheitemcorrectly}{Tota\,\mathbf{h}\ oofstudents} \text{ .....(1)}$$

Item discrimination was determined by computing Upper-Lower Index (ULIC), point-biserial correlation (PBC) between item and total, and corrected point-biserial correlation (CPBC). The difference between PBC and CPBC is that the correlation in latter case is calculated between the item and the total excluding the item in question. This correction is made because total scores that have the item in question embedded within them will have a spuriously higher relationship (i.e., correlation). The following cut-offs were used in the present study where CPBC>.35 (excellent), .25<= CPBC<.35 (good), .15<= CPBC<.25 (fair), 0.0<= CPBC<.15 (poor), and CPBC< 0.0 (Miskeyed or Other Flaws).

$$ULI_C = P_{Upper} - P_{Lower} \text{ .....(2)}$$

The upper group is made up of the 27% who are the best performers (have the highest overall test scores) and the lower group is made up of the 27% who are the poorest performers (have the lowest overall test scores) (Cureton, 1937). The difficulty of an item, as computed by using Equation 1, for upper group is subtracted from that for lower group determining ULIC.

When performance on the item is dichotomous, the product-moment correlation is called the point-biserial correlation, PB. When the dichotomous variable creates a distinction between those examinees who answered correctly and those who did not answer correctly, the formula is:

$$\boldsymbol{B}_{C} = \frac{M_C - M}{S}\sqrt{\frac{P_C}{1-P_C}} \text{ .....(3)}$$

Where *Mc* is the mean score on the criterion of the examinees who answered the item correctly; *M* and *S* are the mean and standard deviation on the criterion (total test score) of all the examinees; and *Pc* is the proportion of the examinees who answered the item correctly.

As for the correct answer, a dichotomous variable distinguishes between whether or not that distractor was selected (Millman & Green, 1989, p. 362). Distractor Discrimination Index was calculated in three different ways as follows:

$$ULI_D = P_{Upper} - P_{Lower} \text{ ..... (4)}$$

Here ULI_D denotes upper lower index for distractor and P denotes the percentage or proportion of students selected a particular distractor. This is equivalent to Equation 2. The second index is equivalent to Equation 3 and was calculated as follows:

$$\boldsymbol{B}_{D} = \frac{M_D - M}{S}\sqrt{\frac{P_D}{1-P_D}} \quad \text{.................................} \ (5)$$

Where PBD is the point biserial correlation for distractor, MD is the mean score on the criterion of the examinees who selected the distractor and PD is the proportion of examinees selecting the distractor. For most measurement purposes, an item which is deemed acceptable should have a positive PBC and negative PBD each (Millman & Green, 1989, p. 362).

When PBC is computed for the discrimination of the item as a whole, it forms a distinction between those examinees who answered correctly (i.e., chose the right answer) and those who did not answer correctly (i.e., chose one of the distractors). On the other hand, when PBD is computed, a distinction is made between those examinees who chose the distractor (i.e., answered incorrectly) and those who did not choose the distractor (some of whom answered correctly and some of whom did not). This distinction does not differentiate among examinees who answered correctly in contrast to those who did not.

A correct usage of PB for distractors is to change the two groups of examinees who are to be compared. One group should be those who chose the distractor, and the other group should be those who chose the correct answer. The dichotomous variable that would be formed would make a meaningful distinction between those examinees who answered incorrectly and chose this specific distractor and those examinees who chose the correct answer. If we call this index PBDC (for the comparison between Distractor and Correct choices), the formula for calculating it would be:

$$\boldsymbol{B}_{DC} = \frac{M_D - M_{DC}}{S_{DC}}\sqrt{\frac{P_D}{P_C}} \quad \text{..} (5)$$

Where MD is the mean score on the criterion of the examinees who selected the distractor, MDC is the mean score on the criterion of examinees who chose the distractor or the correct answer; SDC is the standard deviation on the criterion of examinees who chose the distractor or the correct answer, and PD and Pc are the proportion of examinees who selected the distractor and the correct answer, respectively. According to Attali and Freankel (2000), this modified version protects against type II error, i.e., from incorrectly rejecting a distractor whose MD is lower than MDC, but not lowers than M.

Distractor Functionality Index (DFI) or in other words Distractor Effectiveness (DE) was determined by using two indexes. A functional or acceptable distractor is defined as an option with a response frequency of ≥5% and a negative discriminating power (PBDC). Conversely, a non-functional distractor is defined as an option with either a response frequency of <5% or a non-negative (zero or positive) discriminating power (Tarrant, Ware, & Mohammed, 2009; Ware & Vik, 2009).

Findings, Interpretation and Presentation

The present study evaluated MCQs of Bangla, ICT, Physics, Civics, and Management of HSC examinations in 2016 and 2018. Please note that each question comprised of 4 options where one of them was the correct answer or key and remaining three were distractors. The number of questions given and analyzed in each subject is shown in Table 1.

Table 1

No of questions analyzed and given in HSC examinations

Code	Subject	2016		2018	
		Analysed	Given	Analysed	Given
101	Bangla	29	40	25	30
275	ICT	34	35	25	25
174	Physics	32	35	25	25
269	Civics	40	40	30	30
277	Management	38	40	30	30

Note: Based on Primary Data

While computing difficulty, discrimination and distractor functionality indexes, we analyzed only those questions having one correct option and excluded those having none or more than one correct option. The distributions of difficulty, discrimination, distractor functionality indices are presented in Tables and Figures.

Table 2

Range, Mean, Frequency, and Percentage of P values falling in each category of difficulty for each subject by year

Subject	Year	Range	Mean	*P* < .30 = Difficult		.30 < = *P* <.70 = Moderate		*P* >.70= Easy		**Total**
				f	%	f	%	f	%	
Bangla	2016	.18 – .80	.55	02	6.90	22	75.86	05	17.24	29
	2018	.19 – .81	.49	06	24.00	15	60.00	04	16.00	25
ICT	2016	.13 – .98	.50	08	23.53	18	52.94	08	23.53	34
	2018	.19 – .78	.41	08	32.00	16	64.00	01	4.00	25
Physics	2016	.12 – .81	.50	03	9.38	26	81.25	03	9.38	32
	2018	.10 – .70	.38	08	32.00	17	68.00	00	0.00	25
Civics	2016	.15 – .95	.61	04	10.00	22	55.00	14	35.00	40
	2018	.05 – .97	.62	03	10.00	15	50.00	12	40.00	30
Management	2016	.00 – .99	.71	02	5.26	14	36.84	22	57.89	38
	2018	.10 – .97	.59	03	10.00	16	53.33	11	36.67	30

Note: Based on Primary Data

Table 2 shows that the majority of the items were in moderate difficulty (about 50% to 80%) across subject and year. Comparing across 3 streams namely science, humanities, and business, percentage of easy questions were highest for Management (36.67% to 57.89%), followed by Civics (35% to 40%), and lowest for Physics (0.00% to 9.38%). Comparing by year, 2018 contained

higher percentage of difficult questions than 2016; and physics was the most difficult in 2018. The results are presented in Figure 1 for easy understanding.

Table 3

Range, Mean, Frequency, and Percentage of CPBC falling in each category of discrimination for each subject by year

Subject	Year	Range	Mean	$^{C}PB_{C}$ >.35 = Excellent		.25<= $^{C}PB_{C}$ <.35 = Good		.15<= $^{C}PB_{C}$ <.25 = Fair		0.00<= $^{C}PB_{C}$ <.15 = Poor		$^{C}PB_{C}$ < 0.00 = Flaws		Total
				f	%	*f*	%	*f*	%	*f*	%	*f*	%	
Bangla	2016	-.21 – .50	.32	13	44.83	9	31.03	3	10.34	3	10.34	1	3.45	29
	2018	-.10 – .30	.18	1	4.00	7	28.00	8	32.00	5	20.00	4	16.00	25
ICT	2016	.00 – .45	.21	3	8.82	10	29.41	8	23.53	12	35.29	1	2.94	34
	2018	-.02 – .38	.18	2	8.00	6	24.00	8	32.00	7	28.00	2	8.00	25
Physics	2016	.00 – .56	.37	18	56.25	5	15.63	7	21.88	1	3.13	1	3.13	32
	2018	.04 – .41	.23	3	12.00	9	36.00	6	24.00	7	28.00	0	0.00	25
Civics	2016	-.09 – .38	.25	4	10.00	23	57.50	9	22.50	2	5.00	2	5.00	40
	2018	-.12 – .43	.24	5	16.67	10	33.33	11	36.67	2	6.67	2	6.67	30
Mana	2016	-.22 – .40	.23	6	15.79	12	31.58	15	39.47	2	5.26	3	7.89	38
gement	2018	-.30 – .37	.14	2	6.67	3	10.00	11	36.67	10	33.33	4	13.33	30

Note: Based on Primary Data

Table 3 shows that the majority of the items were in fair to excellent DI (about 50% to 90%) across subject and year. Comparing across 3 streams namely science, humanities and business, percentage of poor and flaws questions that means low/poor discriminating power were highest for Management (46.66%) in 2018, followed by ICT (36% to 38%), Bangla (36%) in 2018 and lowest for Physics (6.26%) in 2016. Comparing by year, 2016 contained higher percentage of good and excellent DI than 2018 and Physics was the excellent discriminatory power in 2016. The results are presented in Figure-2 for easy understanding.

Table 4

Frequency and Percentage of distractors falling in each category of Distractor Functionality Index for each subject by year

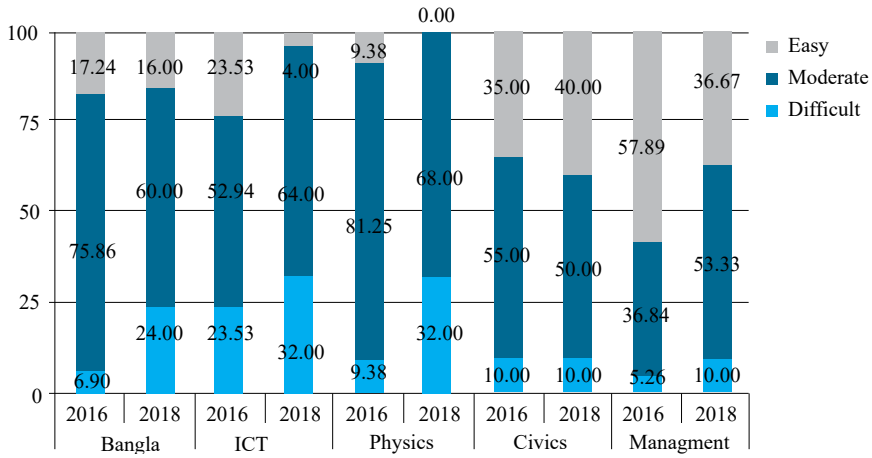
Subject	Year	Items with 3 FD DE = 100%		Items with 2 FD DE = 66.67%		Items with 1FD DE = 33.33%		Items with 0 FD DE = 0%		Total
		f	%	*f*	%	*f*	%	*f*	%	
Bangla	2016	21	72.41	6	20.69	1	3.35	1	3.45	29 (100%)
	2018	14	56.00	11	44.00	0	.00	0	.00	25 (100%)
ICT	2016	26	76.47	5	14.71	1	2.94	2	5.88	34 (100%)
	2018	23	92.00	1	4.00	1	4.00	0	.00	25 (100%)
Physics	2016	24	75.00	8	25.00		00		00	32 (100%)
	2018	20	80.00	3	12.00	1	4.00	1	4.00	25 (100%)
Civics	2016	15	37.50	14	35.00	9	22.50	2	5.00	40 (100%)
	2018	10	33.33	12	40.00	5	16.67	3	10.00	30 (100%)
Manage	2016	7	18.42	14	36.84	11	28.95	6	15.79	38 (100%)
ment	2018	15	50	05	16.67	08	26.67	02	6.67	30 (100%)

Note: Based on Primary Data

Table 4 shows that the majority of the items were 3 functional distractors that means 100% distractor efficiency (about 50% to 92%) across subject and year. Comparing among 3 streams namely science, humanities and business, percentage of 3 functional distractors were lowest for Civics (33.33% to 37.50%), followed by Management (18.42%) in 2016 and highest for ICT (92%) in 2018. Percentage of 0 functional distractors were highest for Management in 2016 and lowest for ICT (0%) in 2018, followed by Physics (0%) in 2016. The results are presented in figure-3 for better understanding.

Figure 4.1

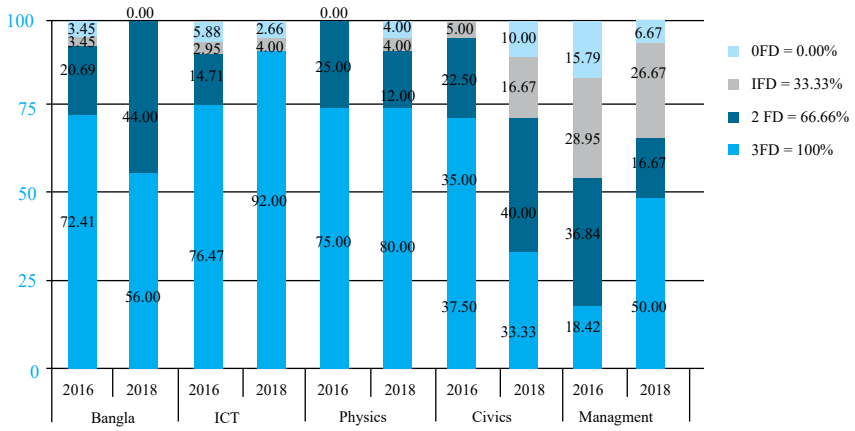
Percentage of questions falling in each category of difficulty index



Note: Based on Primary Data

Figure 4.2

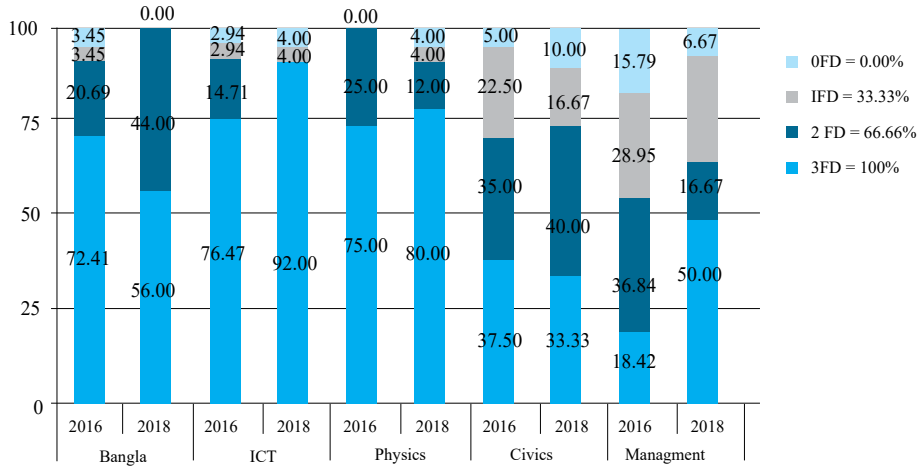
Percentage of questions falling in each category of discrimination index



Note: Based on Primary Data

Figure 4.3

Percentage of questions falling in each category of distractor functionality index



Note: Based on Primary Data

Table 5

Frequency and Percentage of distractor across items falling in functional (FD) and non-functional (NFD) categories

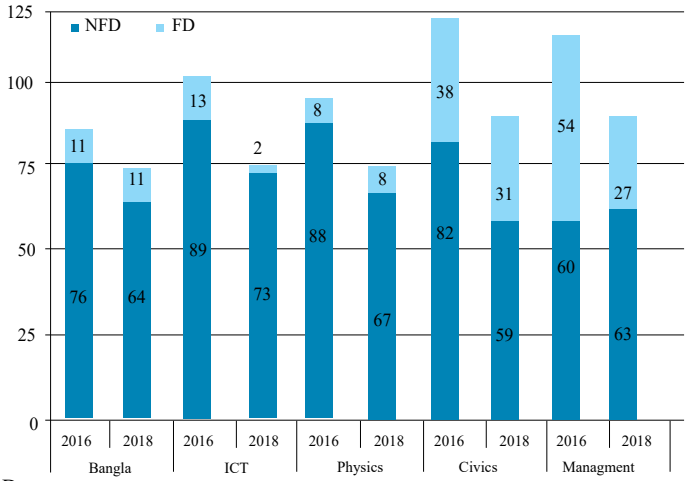
Subject		FD	FD %	NFD	NFD %	Total
Bangla	2016	76	87.36	11	12.64	87
	2018	64	85.33	11	14.67	75
ICT	2016	89	87.25	13	12.75	102
	2018	73	97.33	2	2.67	75
Physics	2016	88	91.67	8	8.33	96
	2018	67	89.33	8	10.67	75
Civics	2016	82	68.33	38	31.67	120
	2018	59	65.56	31	34.44	90
Management	2016	60	52.63	54	47.37	114
	2018	63	70.00	27	30.00	90

Note: Based on Primary Data

Table 5 shows that the majority of the items were in functional distractor (about 52.63% to 97.33%) across subject and year. Comparing across 3 streams namely science, humanities and business, percentage of functional distractor were highest for Bangla, ICT and Physics (about 85.33% to 97.33%) and lowest for Management (52.63%) in 2016. Comparing by year, 2016 contained higher percentage of functional distractor than 2018. The results are presented in fugure-4 for easy understanding.

Figure 5.1

Frequency of distractor across items falling in functional (FD) and non-functional (NFD) categories



Note: Based on Primary Data

Table 6

Internal consistency reliability and Pearson correlation between total score on MCQ and CQ

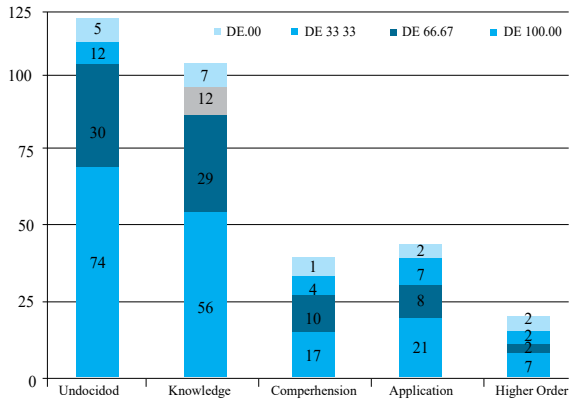
Subject	Year	Cronbach Alpha	Remark	*r* (*N* = 1000)
Bangla	2016	0.81 (n = 29)	**Good**	0.55
	2018	0.59 (n = 29)	Poor	0.41
ICT	2016	0.69 (n = 29)	Satisfactory	0.46
	2018	0.60 (n = 29)	Poor	0.49
Physics	2016	0.86 (n = 29)	**Good**	0.48
	2018	0.69 (n = 29)	Satisfactory	0.62
Civics	2016	0.78 (n = 29)	**Good**	0.42
	2018	0.74 78 (n = 29)	**Good**	0.45
Management	2016	0.73 78 (n = 29)	**Good**	0.37
	2018	0.52 (n = 29)	Poor	0.34

Note: Based on Primary Data

Table 6 shows that internal consistency reliability of the 50% items were in good across subject and year. Comparing by year, 2016 contained higher internal consistency reliability than 2018. Apart table-6 also shows that the correlation between total score on MCQ and CQ were in moderate correlation but it was supposed to be high. Comparing by year, 2018 contained higher moderate correlation than 2016.

Figure 6.1

Frequency of functionality of distractors across items by levels (undecided, knowledge, comprehension, application, and higher order) as judged by experts



Note: Based on Primary Data

Figure 6.1 shows that the majority of the items according to levels (undecided, knowledge, comprehension, application, and higher order) were 100% distractor efficiency. Comparing by levels, percentage of distractor efficiency were highest for undecided questions and lowest for higher order questions which was not matched suggesting by the Board.

This research team also conducted an Focus Group Discussion (FGD) as a part of qualitative portion. The FGD portion describes the perception and experiences of teachers of college. The objective of the study was to access the reflection of curriculum intentions in preparing MCQs and to ascertain whether criteria or guidelines were followed in MCQs. FGD were carried out with teachers of rural, semi-urban and urban college. Through purposive sampling they were selected. The main findings of FGDs are given below:

- Study findings revealed that formative assessment is absent among all colleges and CQ and MCQ as summative assessment to assess the progress of the students but not for the purpose of providing students feedback.
- It appears that there is no similarities among colleges in terms of examination system as well as assessment strategy. Study findings revealed that most of the teachers of rural and urban colleges mentioned that they use both CQ and MCQ test in their internal exam. On the other hand, differences were found in the discussion with the semi-urban college where they mentioned that they use only CQ test in their internal exam except final exam.
- Based on the findings, there are several strengths and some drawbacks are fowned in using MCQs as a assessment. Those strengths were easy and quick scoring, time efficient to administer, easily cover the large range of curriculum and drawbacks were easy to guess and language complexity in case of few subjects.
- Constructing effective MCQ item requires a very deep knowledge as it is a difficult task (Flavio Tangianu, Internal Medicine, S. Martino Hospital, Oristano). This study revealed that except urban college teachers, most of rural and semi-urban teacher's knowledge about using guidelines seem not sufficient.

- It is also noted that, teachers who has training do not have sufficient knowledge for constructing an effective MCQ and non-trained up teacher mentioned that when they are asked to create any MCQ question, they just construct it by their own idea regards with some instruction of other senior teacher.
- This study revealed that many factors hinders teachers to practice to construct MCQ question items. It appears that teachers feel disinterest to make a new MCQ questions due to the availability of readymade question items of the guide. Some teachers mentioned that lack of time influence them to do this because they have teaching work load and other responsibilities.
- Study findings illuminated that lack of review and validation system of test item makes teachers more incapable to construct an effective MCQ question item. Except few urban college, there is no system of review and validation of MCQ question items is found in most of the rural and urban college.
- In FGD, it is revealed that online question bank could be a scope to construct an effective MCQ questions. From this discussion it is informed that there is initiated an online question Bank in Jessore Board in where all teachers input questions after creating questions all over the year. Teachers only can use that question when the question is reviewed and validated.

Discussion, Conclusion and Recommendation

Multiple Choice Questions is one of the most reliable assessment tools which should be designed in this research following objective. If the objective is not clear, we would be unable to understand the assessment strategy and the research will be a failure. Multiple Choice Questions if properly written and well-constructed, are one of the strategies of the assessment tool that quickly assess any level of cognition according to Bloom's taxonomy. The difficulty and discrimination indices are among the tools to check whether the MCQs are well constructed or not. Another tool used for further analysis is the distractor efficiency which analyses the quality of distractors and is closely associated with difficulty and discrimination indices. A distractor used by less than 5% of students is not an effective distractor and should be either replaced or corrected as it affects the overall quality of the question (Hingorjo, R. M; & Jaleel. F).

The present study evaluated MCQs of Bangla, ICT, Physics, Civics, and Management of HSC examinations in 2016 and 2018. Each items comprised of 4 options where one of them was the correct answer or key and remaining three were distractors. Majority of the items were in moderate difficulty level across subject and year and comparing by year, 2018 contained higher percentage of difficult questions than 2016. Among 3 streams, science that means Physic greshion was the most difficult ones in 2018, since it failed to assess equally all the students.

On the other hand, according to discrimination index, across 3 streams, science that means Physics had the excellent discriminating power in 2016. That means, across each subject, through Physics was able to easily distinguish between good and poor students in 2016.

Majority of the items in our study was 100% distractor efficiency across subject and year. The number of 0 functional distractor were highest for Management in 2016 and lowest for ICT in 2018

and Physics in 2016. In other words, the number of functional distractors among 3 streams business was the lowest which reduces the difficulty level of the question and highest for ICT (compulsory) and science which increases the difficulty level of the question. A comparison between functional and non-functional distractors shows that the number of NFD was highest for Management in 2016 which raised the question of accurate and well-constructed MCQ question. This result is supported by the similar study of Hingorjo and Jaleel, which showed that the MCQs with more NFDs are easier than the MCQs with a lesser number of NFDs.

Internal consistency reliability determines how highly the test items are correlated and how well these items are predicted each other. In this research, the internal consistency of the 50% items were in good that means their score was above 70. The internal consistency of the MCQ items in 2018 was poor and the correlation among these items was low. On the other hand, in the case in Pearson product moment correlation, it was found that the correlation between total score on CQ and MCQ in 2018 was in moderate compared to 2016.

According to Bloom's taxonomy, in measuring the various skills of cognition, it was seen that the percentage of distractor efficiency was high for the undecided question for which the experts could agree less than 75% and low for higher order questions which was not matched suggesting by the Board.

Recommendation

Based on the findings the following suggestions are proposed:

Ongoing training on Guidelines of constructing multiple choice Test: Teachers' training is crucial. Teachers need to attend workshops and short courses on testing in order to get clear ideas about using guideline. Question Setters, moderators and invigilators along with subject teachers should be provided with the opportunity to attend training at regular intervals.

Need for the Establishment of Standardized and uniform assessment strategy: Teachers should be given clear guidelines about the assessment strategy.

Monitoring and Evaluation: Ongoing monitoring and evaluation on MCQ test item should be carried out. MCQ tests items of internal exam along with public exams need to be evaluated in the light of guidelines.

Ensure Fair environment in Exam hall: The government should take step to ensure the fair environment in the exam halls.

Online Question Bank: The Government should consider about the nationalization of Online Question Bank.

Post Test Evaluation: Post test evaluation should be done in regular basis which is not done despite the recommendations.

Conclusion

Finally, there were some limitations of this study. Data was not collected from multiple sources or participants due to time constraints as well as lack of financial resources. Therefore,

triangulation of data was lacking. There was dearth of published research in the Bangladeshi context in the relevant field. However, in several occasions in qualitative research, it is important to return to the informants twice or thrice to complete the interview in order to clarify certain aspects of the interview that were not clear. For this study it was not planned and for better understanding the data we felt its necessity.

Acknowledgement

We are grateful to the authority of National Academy for Educational Management (NAEM) for giving us space for publishing our research article in NAEM JOURNAL. We must thank the editors and concerned person for publishing our article in the NAEM JOURNAL.

References

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25,* 31-36.

Biggs, J. (1999). *Teaching for Quality Learning at University*. Buckingham, UK: Society for Research into Higher Education and Open University Press.

Christian, D. S., Prajapati, A. C., Rana, B. M., & Dave, V. R. (2017). Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat. *International Journal Of Community Medicine And Public Health, 4* (6), 1876–1881. https://doi.org/10.18203/2394-6040.ijcmph20172004

Cizek, G., J., & O' Day, D. M. (1994). Further investigations of non-functioning options in multiple-choice test items. *Educ Psychol Meas, 54* (4)*,* 861-872.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests.*The Canadian Journal for the Scholarship of Teaching and Learning, 2* (2)*,* 1-23.

Downing, S. M. (2002). Assessment of knowledge with written test forms. *Ininternational Handbook of Research in Medical Education, 2*, 647-672. Kluwer Academic Publishers

Ebel, R. L, & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th Ed). Englewood Cliffs, N. J., Prentice Hall.

Farley, J. K. (1989). The multiple-choice test: Writing the questions. *Nursing Education, 14* (6)*,* 10-12.

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice question (MCQs) from an assessment of medical student Ahmedabad, Gujrat.*Indian Journal of Community Medicine, 39,* 17-20.

Golam M., (2015). A study on educational assessment of secondary and higher secondary teachers of Bangladesh. *World Scientific News.* http://www.worldscientificnews.com/

Haladyna, T. M., & Downing, S. M. (1989).Validity of taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2* (1)*,* 51-78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *EducPsycholMeas, 53* (4)*,* 999-1010.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-344.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc 62*, 142-7.

Holbrook, J. and BEDU. (2012). *Report on the Impact of Revised SSC Examination Question Under 8 BISE*, December 2012 year

Holbrook, J. and BEDU. (2012). *Report on the Impact Study of 2017 SSC Examination Questions Under 3 Examination Boards,* March 2018.

Hafiz, L., R., Md. Mehedi, (2019). Investigating the effects of Bangladesh and Global Studies (BGS) assessment of Secondary School Certificate (SSC) examination on students learning. *British Journal of Education, 7* (2). https://eric.ed.gov/?

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40,* 25-32.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Med Teach, 26* (8)*,* 709-712.

Mukherjee, P. & Lahiri, S. K. (2015). Analysis of multiple choice questions (MCQs): Item and test statistics from an assessment in medical college of Kolkata, West Bengal. *IOSR Journal of Dental and Medical Sciences, 14* (12)*,* 47-52.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2003). Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ, 38* (9)*,* 974-979.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today,* doi:10.1016/j.nedt. 2006.07.006.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments.*Medical Education, 42,* 198-206.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC British Medical Education, 9,* 40.http://dx.doi.org/10.1186/1472-6920-9-40

Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *Nursing Education, 45,* 212-219.

Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher, 31*, 238-243. http://dx.doi.org/10.1080/01421590802155597

Submission: 31 December 2021