Vol.- 07, Issue- 01, 2016

Bangladesh Journal of Medical Education
ISSN: 2306-0654

Original Article

Post-application Quality Analysis of MCQs of Preclinical Examination Using Item Analysis

Brigadier General (Dr) Md Ahsan Habib[1], Professor Dr Humayun Kabir Talukder[2], Brigadier General (Dr) Md Mahbubur Rahman[3], Dr. Shahnila Ferdousi[4]

Abstract

Multiple choice questions (MCQs) have considerable role in the preclinical medical assessment, both formative as well as summative. This cross sectional descriptive study was conducted to observe the quality of MC items (completion type) of anatomy, biochemistry and physiology used in preclinical undergraduate medical examinations of 2012 and 2013 of a public university of Bangladesh. Each MC item had a stem and 5 options, and 1200 options were analyzed for difficulty and discrimination indices. Total 556 options were false statements (distracters) and were analyzed to observe their effectiveness as distracter. The study revealed that 18.67% of options were with appropriate difficulty (0.660.80). Highest frequency (43.5%) of difficulty indices was in easy class interval (0.911). Over all frequencies of items of three subjects in the ascending order were difficult, appropriate, marginal and easy as per their difficulty indices. Satisfactory or better discrimination indices (=0.20) were observed in 29.33% options. The mean difficulty and discrimination indices observed were respectively 0.82±0.18 (95% confidence interval [CI] 0.81 to 0.83) and 0.13±0.14 (95% CI 0.122 to 0.138). Out of the options, 6.75% had negative discrimination indices. Items with difficulty index around 0.60 had maximum discriminatory power (up to 0.68) and more difficult as well as easy items had less discriminatory ability. Out of the distracters 83.45% were observed effective and the mean effectiveness was 22.3±18.7% (95% CI 20.75% to 23.85%). The study recommended using the method and findings to improve the quality of the items leading to development of a standard Question Bank.

Key Words: Multiple choice questions, difficulty, discrimination, distracter

Introduction

The challenge for all pre- and para-clinical departments of medical colleges is to teach sufficient factual knowledge and practical skills and also to encourage students to be enquiring as well as analytical and to develop desirable professional attitudes. In teaching learning process objective of an assessment is to assure the ability of a student and to observe the effectiveness of the educational programme. Assessment is significantly effective when appropriate techniques and tools are implemented in the process of assessment. An assessor needs to be aware of the potential factors that can influence all components of the assessment cycle (Tavakol & Dennick, 2011), from question creation to the interpretation of examination scores.

The assessment methods most commonly used in undergraduate preclinical course are written, oral and practical examinations. In written examinations, questions used are mostly Short Answer (SAQs) type and Multiple Choice (MCQs) type. Oral examination is gradually moving towards more structured form from traditional unstructured one. Practical examinations are either individual full practical exercises or multiple stations of objectively structured practical examinations (OSPEs). First phase of undergraduate medical course of Bangladesh includes anatomy, biochemistry and physiology. The subjects are covered over a period of eighteen months. The learning of the students is assisted and assessed by both formative and summative tests which include written, practical and oral tests. Written tests of first phase bear 40% of the total marks of summative anatomy examination. In case of other two subjects it is 50% of total. In all the subjects (anatomy, biochemistry and physiology) written test comprises of 70% SAQ, 20% MCQ and 10% formative components.

MCQs can test well all the levels of the cognitive domain of education (Bloom, 1956). Although there is considerable criticism for its use in professional education as, it tends to assess only recall of knowledge, however, it can test any higher level of the cognitive domain and it can also discriminate well between students, if it is well constructed (McCoubrie, 2004). In the Miller's Pyramid, the MCQs with other types of written test can fit assessing the bottom two levels, which are the 'Know' (knowledge) level and the 'Knows How' (competence) level (Al-Wardy, 2010). This depends mainly on the level of competency exhibited in the construction of the MCQs items. It is important to have sufficient distracters (incorrect options) and to find a good balance between the number of correct options (key) and distracters. In addition, it is essential to construct the question so that keys are defensibly correct and distracters are defensibly incorrect (Case & Swanson, 2001). The

[1]Professor and Head, Department of Anatomy
Armed Forces Medical College, Dhaka Cantonment, Bangladesh

[2]Professor, Curriculum Development and Evaluation, Centre for Medical Education (CME)

[3]Chief Health Officer, Dhaka South City Corporation

Associate Professor, Department of Biochemistry, Dhaka Medical College

Address of correspondence: Brig. General (Dr) Md Ahsan Habib
Professor and Head, Department of Anatomy
Armed Forces Medical College, Dhaka Cantonment, Bangladesh
email: doctorahsan@gmail.com

appropriate analysis of students' responses to an assessment is an essential step in improving the quality of the assessment itself as well as students' learning. There are formal processes used to analyze assessment results and the standard statistical methods associated with analyzing the validity and reliability of an assessment (Crisp, & Palmer, 2007; Cupic, *et al.*, 2009). The most standard ways of reviewing validation of a test are pre-validation, post-validation and key-validation (Sood, *et al.*, 1995; Participants Workbook, 2011). Pre-validation is done before the conduction of examination. A committee of three or four experts checks the relevancy of the contents and construction of each item. Post-validation is the statistical analysis of items to check that they are effectively evaluating students' learning and is called item analysis. There are several indices to judge and improve the quality of MCQs as an assessment tool (Nnodim, 1992; Considine,*et al.*, 2005). Difficulty index or facility index, discrimination index and distracter effectiveness are the most import among them (Sim & Rasiah, 2006; Barman,*et al.*, 2010; Tavakol & Dennick, 2011).

No single assessment instrument is perfect and no single instrument can test all aspects of medical competence and performance. Each instrument has its strengths and weaknesses (Schuwirth, & van der Vleuten, 2003; Swanwick, 2010). Some of the assessment tools are inherently subjective, while the rest of them may be applied with a greater degree of objectivity. No method of assessment can, however, be intrinsically flawless (Downing, 2005; Stagnaro-Green & Downing 2006). Evaluation of the quality of 40 New England Journal of Medicine (NEJM) MCQs of weekly continuing medical education (CME) programme, using the standard evidence-based principles of effective item writing showed that each multiple choice item reviewed had at least three item flaws, with a mean of 5.1 and a range of 3 to 7 (Stagnaro-Green & Downing 2006). The flawed multiple-choice test items, which violate well established and evidence-based principles of effective item writing, misled some students and unduly penalize some examinees (Downing, 2005). So, it is perceivable that an assessment system based on available tools and techniques of assessment shall be evaluated so that it becomes highly objective, reliable and valid.

This cross sectional comparative study was conducted to observe the quality of multiple choice test items of Anatomy, Biochemistry and Physiology of First Professional MBBS Examinations of one of the public Universities of Bangladesh using item analysis.

Method

The study was conducted at Armed Forces Medical College (AFMC) of Dhaka Cantonment, Bangladesh. Answer script of MCQ part of written test of three subjects of first professional MBBS examinations of Bangladesh University of Professionals of July 2012 & July 2013 (regular batches) were collected and studied.

In the studied professional examinations there were two papers (I & II) in written test of each subject. Each paper had 20 MCQ true/false completion type items. Each item had a stem and 5 options. Students had to select each of the options as true/false. So, in each paper 20 items had 100 options (20x5). Two hundred options of each of the subjects were analyzed; that involved analyses of 600 (200x3subjects) options in each professional examination. As total population of two professional examinations (regular examinations of 2012 & 2013) were studied, the procedure involved analyses of (600x2) 1200 options (sampling unit) for difficulty indices as well as same number (1200) for discrimination indices. Total 556 options were false statements (distracter) and were analyzed for distracter effectiveness.

Data of each paper was managed separately. Calculation and classification were completely done on a Microsoft excel template prepared by principal author. For calculation of difficulty index (p) and discrimination index (DI) respectively p= [(H+L) / N] and DI = 2 x [(H-L) / N] formulae were used. There, H was correct response in high ability group, L was correct response in low ability group and N was total response (Guilbert, 1981; Chauhan, *et al.*, 2013). Calculated variables were transferred to SPSS version 19 for statistical analysis and observed for correlation. For calculation of indices 27% high performer and 27% of low performer was considered. A distracter was considered effective in this study when it was 5% or more. The statistical analyses were performed as applicable. The p values less than 0.05 were considered statistically significant.

Result

The mean marks (±SD) obtained by the students in the MCQ tests of anatomy, biochemistry and physiology of the examinations studied were 16.31±1.35, 16.65±1.21 & 16.02±1.21 respectively out of total 20 marks. Frequency distribution of difficulty and discrimination indices as per their class intervals is presented in the table 1 and 2 respectively. The mean difficulty and discrimination indices of all the samples were respectively 0.82±0.18 (95% confidence interval 0.81 to 0.83) and 0.13±0.14 (95% confidence interval 0.122 to 0.138). In anatomy, biochemistry and physiology respectively 6.5%, 6.5% and 7% options (samples) had negative discrimination indices. Relative state of distracters' functionality is shown in figure 1 and the table 3 has depicted the central tendency and dispersion of variables observed.

There was statistically significant difference between the frequencies of poor or worst and satisfactory or better discrimination indices of three subjects ($\div^2$=8.45, df=2, p=0.0146). Biochemistry has relatively less number of items with satisfactory or better discrimination. The mean discrimination index of biochemistry was also statistically significant less than those of anatomy (p = 0.002) and physiology (p = 0.028).

Pearson correlation between difficulty and discrimination indices showed that discrimination index correlate poorly with difficulty index (r = -0.453). The correlation was significant at 0.000 level (2-tailed). Negative correlation signifies that with increasing difficulty index values, there is decrease in discrimination index. When the data was entered in Microsoft illustration a dome shaped relationship was displayed (Figure 4.9). Initially, the discrimination power increased with the level of difficulty of the items, until it reached a plateau (discrimination index of about 0.28 with the maximum 0.65) with difficulty indices of about 0.50 (extending up to 0.70) and then began to decline with further increase in difficulty indices. The discrimination power of the items with difficulty indices 0.10 and 1.0 were 0 (zero).

Table 1: Frequency distribution of different classes of difficulty index of MCQs of two first professional examinations (n=400 for each subject)

Subject	Easy (0.91-1) f (%))	Marginal (0.81-0.90) f (%)	Appropriate (0.66-0.80) f (%)	Difficult (00-0.65) f (%)
Anatomy	173 (43.25)	86 (21.50)	73 (18.25)	68 (17.00)
Biochemistry	190 (47.50)	91 (22.75)	65 (16.25)	54 (13.50)
Physiology	159 (39.25)	90 (22.50)	86 (21.50)	65 (16.25)
Grand Total	522 (43.50)	267 (22.25)	224 (18.67)	187 (15.58)

Table 2: Frequency distribution of different classes of discrimination index of MCQs of two first professional examinations (n=400 for each subject)

Subject	Excellent (0.40-01) f (%)	Good (0.30-0.39) f (%)	Satisfactory (0.20-0.29) f (%)	Poor (00-0.19) f (%)	Worst (-01--0.01) f (%)
Anatomy	35 (08.75)	23 (05.75)	73 (18.25)	243 (60.75)	26 (06.50)
Biochemistry	17 (04.25)	21 (05.25)	58 (14.50)	278 (69.50)	26 (06.50)
Physiology	17 (04.25)	30 (07.50)	78 (19.50)	247 (61.75)	28 (07.00)
All Subjects	69 (05.75)	74 (06.16)	209 (17.42)	768 (64.00)	80 (06.67)

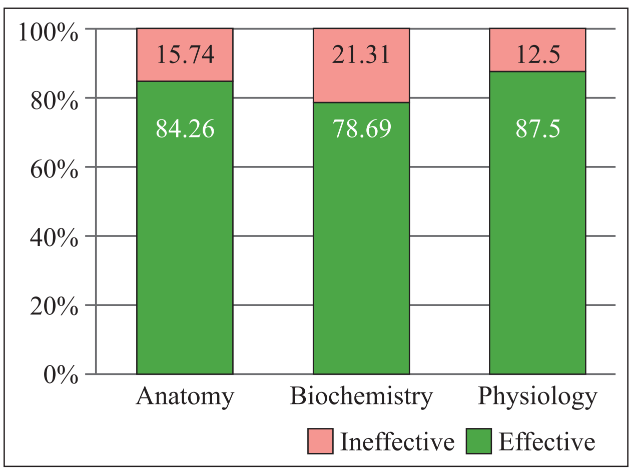

***Figure 1:** Relative frequencies of distracters by their effectiveness*
($n_{anatomy}$=197, $n_{biochemistry}$=183 and $n_{physiology}$=176

Table 3: Central tendency and dispersion of variables *(n=1200 for indices and n= for distracters)*

Variable	Mean	Median	Mode	SD	Range
Difficulty index	0.82	0.88	0.98	0.18	0.92
Discrimination index	0.13	0.09	0.00	0.14	0.92
Distracter effectiveness	22.3%	18.0%	0.00%	18.7%	86.4%

☆ SD = Standard deviation

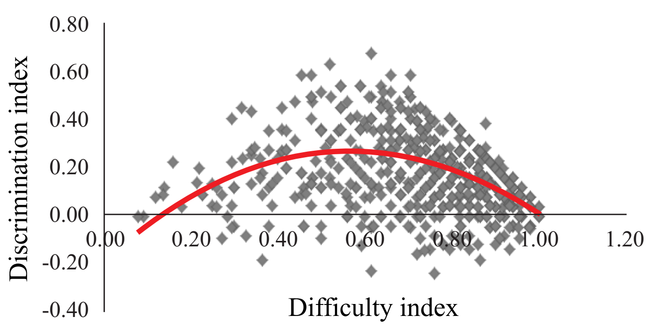

***Figure 2:** Relation between difficulty and discrimination indices as reflected on Microsoft illustration. There is a dome shaped regression line (n=1200)*

Discussion

Assessment in medical education is a subject of criticism since Abraham Flexner shaped the medical education. During the year 1953 to 1955 Queen's University of Belfast tried to establish reliability of marking in final examination of medicine (Bull, 1956). With the publication of the report the subject became concern of educationist. This inculcated the need to move towards more objective form of tests particularly in view of heterogeneous nature of response of examiners and to gain the trust of students. But soon MCQ formats were blamed to penalize the higher able students while rewarding the less able group (Alker, *et al.*, 1969). Discussion and criticism continued for MCQ (Anderson, 1979) and against MCQ (Pickering, 1979). Advantages and disadvantages were widely studied (Chandratilake, *et al.*, 2011) and educationists tried to find out the ways of constructing better MCOs (Harden, 1979; Campbell, 2011).

Multiple choice questions (MCQs) in various formats are widely used tools throughout the world in medical education. Of course educationists prefer to shift from multiple true/false (MTF) format to a single best answer type or an extended matching format with the argument that these are better for assessment of higher order knowledge and clinical competencies (Al-Wardy, 2010) and better meet the expectation (Mobalegh & Barati, 2012). In Bangladesh the tool (MTF) is used in undergraduate course as well as in postgraduate medical education.

In the present study majority of the items (43.5%) of all the tests together were easy type (difficulty indices >0.90). Less than one-fifth (18.67%) of the items had appropriate difficulty indices (indices ranged 0.66 to 0.80). When analyzed separately no statistically significant difference between the frequency distribution of three subjects (anatomy, biochemistry and physiology) was observed (p=0.26). The mean difficulty index of all the items (n=1200) studied was 0.82 (95% confidence interval [CI] 0.81 to 0.83). This mean value is located in the marginal class interval (0.81 to 0.90). Observations of a recent study (Caballero, 2013) conducted at Nova Southeastern University College of Pharmacy (USA) on the items used in 3 course ending tests of patho-physiology revealed results close to this study. There was no option of penalization (negative marking) for wrong answer in the tests included in this study. A study conducted at University of Benin (Nnodim, 1992) covering the Items of the subjects anatomy, biochemistry and physiology reported that the majority (68.3%) of MCQ completion type items were in the appropriate class interval (indices range: 0.30 to 0.75). The author conducted the study on the test where there was option of negative marking for wrong answer (+1 for each correct and -1 for each wrong answer).

This study analyzed 1200 samples (items) and observed that 29.33% of those items were with satisfactory or better discrimination indices (indices=0.20) and 70.67% were with poor or even worse discrimination indices (indices<0.20). Out of the total items 6.75% of items had negative discrimination indices. The mean discrimination index of all the items studied was 0.13 (95% CI 0.122 to 0.138). This mean value is located in the poor discrimination class interval (0.00 to 0.19). Study of Caballero J *et al.* (2013) reported that the mean discrimination index for true-false items was 0.18 (95% CI: 0.10 to 0.26) and the mean value was 0.26 (95% CI: 0.22 to 0.29) for multiple true false completion items. As per the classification of present study the first value is in poor discrimination group and the second value in satisfactory discrimination group. A study of University Sains Malaysia (Barman, 2010) observed that 66% of 100 MCQs items each with 5 true-false responses were with <0.20 (poor and worst) discrimination indices. But that of University of Malaya, Singapore (Sim, & Rasiah, 2006) reported means of six sessions ranged from 0.21 to 0.33 (value range: -0.41 to 0.71). The study observed that about two-thirds of those 'very easy' and 'very difficult' items had poor (=20%) or even negative discrimination. The present study was based on summative examination held at the end of extensive course. Effort was taken to bring all the students to an expected level of performance. This might be the cause of observing mean discrimination index of the studied items <0.02.

A dome shaped relationship line of difficulty and discrimination indices observed in this study. The items with difficulty index around 0.60 showed maximum discrimination power (up to 0.68). Sim and Rasiah (2006) studied the relationship between item difficulty index and discrimination index values of three separate parts of MCQ examination papers (n=250 test items). In all the cases, dome-shaped relation line was observed. Initially, the discrimination power increased with the level of difficulty of the items, until that reached a plateau (discrimination index of 51% to 71%) with moderately difficult items (difficulty index of 40% to 74%), and then began to decline with further increase in difficulty (difficulty index <25%). Approximate dome shaped curve relation curve was also observed by Cupic and Basic (2009) while analyzing 140 MCQ items. Mitra *et al*. (2009) conducted a study involving 120 MCQ test items of single best answer type. Pearson correlation between difficulty and discrimination indices showed that discrimination index correlate poorly with difficulty index (r = -0.325).

The tendency of wide dispersion of difficulty and discrimination indices was observed in the reported study. The ranges of values of difficulty indices and discrimination indices of present study were respectively 0.08 to 1 and -0.24 to 0.68. Similar trends were also observed in recent studies (Sim & Rasiah, 2006; Barman, *et al.*, 2010). Students' opportunity to answer the questions out of guessing is mostly blamed for this type of disperse data (Barman, *et al.*, 2010; Chandratilake, *et al.*, 2011; Mitra, *et al.*, 2009). This idea is reinforced by the presence of bulk of negative discrimination indices. In this study 6.67% of the discrimination indices were negative. The guessing opportunity is further encouraged by non existence of negative marking for wrong answer. In this study tendency of answering all the questions was marked. Chauhan et al. (2013) observed that option of negative marking reduced the dispersion of values. Of course different opinion is there that negative mark does not necessarily prevent students from guessing (Sim & Rasiah, 2006; Chandratilake, *et al.*, 2011). Remarkable existence of negative discrimination (10% or even more) indices are observed in various study on post-application item analyses (Ware, & Vik, 2009; Sayyah, *et al.*, 2012). This negative discrimination may also be an indication of defective item construction or even wrong key (Participants Workbook, 2011).

Distracters are unquestionably wrong answer that shall be plausible to those who have not mastered the knowledge on the subject that the MCQ is designed to measure; but clearly incorrect to those who possess the knowledge on that subject required to answer that MCQ. The term is more applicable to the single best answer type of MCQ than to MCQ true false completion type. In the later case each option works as an individual unit. On the other hand in single best answer type the whole item is a composite unit. So dispute is more regarding number of distracters in a single best answer item (Tarrant, 2009) and its functionality (Hingorjo & Jaleel, 2012).

In the present study, respectively 49.25%, 45.75% and 44% of options were distracters in Anatomy, Biochemistry and Physiology. The number was relatively less in Biochemistry and Physiology. But the differences were not statistically significant (p value 0.32 and 0.14 respectively). No strict guidelines exist about how many true and false items appear in an item, but expecting a balance between the number of true and false options per set shall be reasonable (Schuwirth, & van der Vleuten, 2003; Haladyna, 2004). Of those 15.74% of anatomy, 21.31% of biochemistry and 12.5% of physiology distracters were ineffective. Existence of nonfunctioning distracter is not an uncommon finding of researchers. Hingorjo and Jaleel (2012) observed 47 (23.5%) non-functional distracters out of total 200 distracters studied. In another study revealed 137 (38.4%) ineffective distracters out of 450 (Sarin, *et al.*, 1998). Out of those, 94 (20.8%) distracters were with '0' distracter effectiveness. Tarrant *et al*. (2009) analyzed 1542 distracters and found 52.2% of all distracters as functioning and of those 10.2% distracters were with '0' distracter effectiveness. The distracters that are not chosen or are consistently chosen by only a few participants are obviously ineffective and should be omitted or replaced. If a distracter is chosen more often than the correct answer, this may indicate poor instructions or a misleading question or the key is incorrect. In this study 54 distracters (9.7%) were observed to have effectiveness more than 50%.

Conclusion

Difficulty index does not always indicate the level of difficulty or hardness of the item. Even same item shows variation in index in different context. If the contents of course are well-instructed or students are highly trained or examinees are with high-ability, the tests and their items appear easy. Same tests in reverse situation may seem to be difficult or hard. Presence of flaws in items produces impact on item difficulty and discrimination as well. The negative discrimination index may be an indicator of 'miskeys' or of potentially flawed questions or the students are misinformed. So, the main use of the findings of item analyses is to improve the quality of MCQs. This shall involve eliminating ambiguities, clarifying wordings and strengthening alternatives. Items with poor discrimination index and too low or too high difficulty index should be reviewed by the respective content experts. Suggestions of Li (2013) may be utilized as good guideline in this respect. The distracter is a part of the test item and should be useful. Distracter has been proven to affect positively test score reliability. If it is not useful, it should be removed.

References

1. TAlker, H.A., Carlson, J.A. & Hermann, M.G. (1969) Multiple-Choice Questions and Student Characteristics, *Journal of Educational Psychology*, 60, 3, pp.231-243.
2. Al-Wardy, N.M. (2010) Assessment Methods in Undergraduate Medical Education. *Sultan Qaboos University Medical Journal*, 10, 2, pp.203-209.
3. Anderson, J. (1979) For Multiple Choice Questions, *Medical Teacher*, 1, 1, pp.37-42.
4. Barman, A., Ja'afar, R., Rahim, F.A. & Noor, A.R. (2010) Psychometric Characteristics of MCQs used in Assessing Phase-II Undergraduate Medical Students of Universiti Sains Malaysia, *The Open Medical Education Journal*, 3, pp.1-4.
5. Bloom, B.S. (1956) *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*, New York: David McKay Co. Inc., pp.1-7.
6. Bull, G.M. (1956) An examination of final examination of Medicine, *Lancet*, 25 August, 2, pp.368.
7. Caballero, J., Wolowich, W.R., Benavides, S. & Marino, J. (2013) Difficulty and discrimination indices of multiple-choice examination items in a college of pharmacy therapeutics and pathophysiology course sequence, *International Journal of Pharmacy Practice*, 21, (online view before publication).
8. Campbell, D.E. (2011) How to write good multiple-choice questions, *Journal of Paediatrics and Child Health*, 47, pp.322325. (doi:10.1111/j.1440-1754.2011.02115.x).
9. Case, S.M. & Swanson, D.B. (2001) *Constructing Written Test Questions For the Basic and Clinical Sciences*, 3rd ed, Philadelphia : National Board of Medical Examiners (NBME).
10. Chandratilake, M., Davis, M. & Ponnamperuma, G. (2011) Assessment of medical knowledge: The pros and cons of using true/false multiple choice questions, *The National Medical Journal of India*, 24, 4, pp.225-228.

11. Chauhan, P.R., Ratrhod, S.P., Chauhan, B.R., Chauhan, G.R., Adhvaryu, A. & Chauhan, A.P. (2013) Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot, *Biomirror*, 1, 4, pp.1-4.

12. Considine, J., Botti, M. & Thomas, S. (2005) Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12, 1, pp.19-24.

13. Crisp, G.T. & Palmer, E.J. (2007) Engaging Academics with a Simplified Analysis of their Multiple-Choice Question (MCQ) Assessment Results, *Journal of University Teaching and Learning Practice*, 4, 2, 89-106.

14. Cupic, M., Snajder, J. & Basic, B.D. (2009) *Post-test analysis of automatically generated multiple choice exams: a case study*, Villach, Austria: ICL (Interactive Computer aided Learning) 2009 proceedings, pp.1-10.

15. Downing, S.M. (2005) The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education, *Advances in Health Sciences Education*, 10, pp.133-143.

16. Guilbert, J.J. (1981) *Educational handbook for health personnel*, Geneva: World Health Organization, pp.4.65-4.69.

17. Haladyna, T.M. (2004) *Developing and Validating Multiple-Choice Test Items*. 3rd ed. New Jersey : Lawrence Erlbaum Associates Publisher, pp. 67-126 & 207-274.

18. Harden, R.M. (1979) Constructing multiple choice questions of the multiple true/false type, *Medical Education, July,* 13, 4, pp.305-312.

19. Hingorjo, M.R. & Jaleel, F. (2012) Analysis of One-Best MCQs: the Difficulty Index, Discrimination Index and Distracter Efficiency, *Journal of Pakistan Medical Association*, 62, 2, pp.142-147.

20. McCoubrie, P. (2004) Improving the fairness of multiple-choice questions: a literature review, *Medical Teacher*, 26, 8, pp.709712.

21. Mitra, N.K., Nagaraja, H.S., Ponnudurai, G. & Judson, J.P. (2009) The Levels of Difficulty and Discrimination Indices in Type A Multiple Choice Questions of Pre-clinical Semester 1 Multidisciplinary Summative Tests, *International e-Journal of Science, Medicine & Education (IeJSME),* 3, 1, pp.2-7.

22. Mobalegh, A. & Barati, H. (2012) Multiple True-false (MTF) and Multiple-choice (MC) Test Formats: A Comparison between Two Versions of the Same Test Paper of Iranian NUEE, *Journal of Language Teaching and Research*, September, 3, 5, pp.1027-1037.

23. Nnodim, J.O. (1992) Multiple choice testing in anatomy, *Medical Education*, 26, 4, pp.301-309.

24. Participants Workbook of the Teacher Training Workshop (2011) *Pre-service Innovative Teaching Methodologies*. Bangalore, India: Department of Medical Education, St. John's Medical College. January, pp.79-109.

25. Pickering, G.S. (1979) Against Multiple Choice Questions, *Medical Teacher*, 1 (2), 84-86.

26. Sarin, Y.K., Khurana, M. & Natu, M.V., Thomas, A.G., Singh, T. (1998) Item Analysis of Published MCQs, *Indian Pediatrics*, 35, 11, pp.1103-1105.

27. Sayyah, M., Vakili, Z., Alavi, N.M., Bigdeli, M., Soleymani, A., Assarian, M. & Azarbad, Z. (2012) An Item Analysis of Written Multiple-Choice Questions: Kashan University of Medical Sciences, *Nursing and Midwifery Studies*, 1, 2, pp.83-87.

28. Schuwirth, L.W.T. & van der Vleuten, C.P.M. (2003) ABC of learning and teaching in medicine written assessment, *British Medical Journal (BMJ)*, 326, pp.643-645.

29. Sim, S.M. & Rasiah, R.I. (2006) Relationship between Item Difficulty and Discrimination Indices in True/False - Type Multiple Choice Questions of a Para-clinical Multidisciplinary Paper, *Annals, Academy of Medicine, Singapore (AAMS)*, 35, 11, pp.67-71.

30. Sood, R., Paul, V.K., Sahni, P., Mittal, S., Kharbanda, O.P., Adkoli, B.V., Verma, K. & Nayar, U. (1995) *Assessment in Medical Education Trends and Tools*, All India Institute of Medical Science: K.L. Wig Centre for Medical Education and Technology, pp.35-84.

31. Stagnaro-Green, A.S. & Downing S.M. (2006) Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education, *Medical Teacher*, 28, 6, pp.566568.

32. Swanwick, T. (2010) *Understanding Medical Education: Evidence, Theory and Practice*. London Deanery: Wiley Blackwell Publication, pp.195-270.

33. Tarrant, M., Ware, J. & Mohammed, A.M. (2009) An assessment of functioning and non-functioning distracters in multiple-choice questions: a descriptive analysis, *BioMed Central Medical Education*, 9, 40, pp.1-8. (doi:10.1186/1472-6920-9-40)

34. Tavakol, M. & Dennick, R. (2011) Post-examination analysis of objective tests. *Medical Teacher*, 33, 6, pp.447458.

35. 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, 3, pp.238243.