RESEARCH ARTICLE

Student perception and post-exam analysis of one best MCQs and one correct MCQs: A comparative study

Mohammad Idrees Adhi,[1] Syed Moyn Aly[2]

Abstract

Objective: To find differences between One-Correct and One-Best multiple-choice questions with relation to student scores, post-exam item analyses results and student perception.

Methods: This comparative cross-sectional study was conducted at the Dow University of Health Sciences, Karachi, from November 2010 to April 2011, and comprised medical students. Data was analysed using SPSS 18.

Results: Of the 207 participants, 16(7.7%) were boys and 191(92.3%) were girls. The mean score in Paper I was 18.62±4.7, while in Paper II it was 19.58±6.1. One-Best multiple-choice questions performed better than One-Correct. There was no statistically significant difference in the mean scores of the two papers or in the difficulty indices. Difficulty and discrimination indices correlated well in both papers. Cronbach's alpha of paper I was 0.584 and that of paper II was 0.696. Point-biserial values were better for paper II than for paper I. Most students expressed dissatisfaction with paper II.

Conclusion: One-Best multiple-choice questions showed better scores, higher reliability, better item performance and correlation values.

Keywords: One-Best MCQs, Type A MCQs, One-Correct MCQs, Reliability, Item analysis, Post-exam analysis, Student perception. (JPMA 68: 570; 2018)

Introduction

Summative assessment is a necessary process to assure the profession, the public and the regulatory authorities that the graduating practitioners are capable of offering the highest quality of healthcare. Therefore, assessment is a critical part of the educational and accreditation processes across the health professions.[1] With the recent concern for patient safety, the need for efficient and valid assessment tool has become more important. The incorporation of a robust system of assessment provides credibility to the pass/fail decision-making process. This provides direct evidence about the validity of the interpretations made.[2]

This research is based on Messick's framework of validity. According to Messick,[3] validity is a unitary concept defined as the evidence collected to support the interpretation of assessment results. According to Messick and Kane,[3] the contemporary view of validity suggests that all validity is construct validity which has five broad sources of evidence. One of these sources is called 'Internal Structure' dealing with statistical post-exam analysis.

In Pakistan, multiple-choice questions (MCQs) are one of the assessment tools used for assessing 'knows' and 'knows how' levels of competence.[4] The two common types of MCQs are One-Correct, which assesses recall or 'knows', and One-Best which assesses application of knowledge or 'knows how'. It is, therefore, essential that we get to know the differences in psychometric properties of these two types so that there is evidence of one aspect of validity, i.e. internal structure, in the Pakistani context.[3]

MCQs have gained acceptance as a method that can test higher cognition.[5] Experience with MCQs suggests that a candidate is unlikely to have a good overall performance unless he performs well in MCQs. Students who do badly in MCQs are unlikely to excel in other types of test.[6]

An understanding of the post-exam analysis, whereby results are analysed in order to determine the accuracy of interpretations made from them, is an essential requirement of contemporary educational practices.[3] This authenticates the decisions taken on the basis of the marks.

Item-analysis is the process of collecting, summarising and using information from students' responses to assess the quality of test items.[3,7,8] It helps in judgement about which items are of appropriate difficulty level,

[1]King Abdul Aziz Medical City & King Abdullah Specialized Children Hospital, National Guard Health Affairs, Riyadh, Saudi Arabia, [2]Department of Medical Education, Jinnah Sindh Medical University, Rafique Shaheed Road, Karachi, Pakistan.

Correspondence: Mohammad Idrees Adhi. Email: miadhi3112@gmail.com

discriminate amongst the students and demonstrate internal consistency in assessing the construct. Point bi-serial can be calculated to provide supportive information about how well an item differentiates the students. On the basis of item-analysis, results can be made more defensible, and test items can be revised and improved for future use on a scientific basis. Feedback can also be provided to item developers.[9]

One of the major concerns in the construction of the test items is ensuring the reliability of results. This type of item-analysis determines test homogeneity of the construct being assessed. The more well-constructed the test items are, the more likely they would measure the same construct, thus ensuring internal consistency and a high Cronbach alpha value.[10] Difficulty index (P) refers to the percentage of the total number of students who answer an item correctly.[11] Discrimination index (DI) provides information about how well an item is able to discriminate among the students.[5]

There is hardly any literature comparing the psychometric properties of One-Correct MCQs with those of One-Best MCQs. Some of the earliest works are by Norcini, Baranowski, Swanson, Grosso and Webster in which they compare the psychometric properties of MCQs with patient management problems (PMPs).[12] In 1995, Downing, Baranowski, Grosso and Norcini[13] provided validity evidence for MCQs. They reported that MCQs had a higher criterion-validity than the multiple true-false variety.

Hingorjo and Jaleel[14] published post-exam analysis of One-Best MCQs in which they found items with average difficulty to have high discrimination. Baig, Ali, Ali and Huda[15] compared MCQs with Short Essay Questions similar to Mahmood H.[16] Abdul Ghani et al.[17] reported the effect of faculty development on quality of MCQs which they measured via item analysis and Cronbach's Alpha. Karelia[18] described the P and DI of One-Best MCQs in pharmacology and found insignificant correlation between these two. Mitra, Nagaraja, Ponnudurai and Judson[6] also correlated P with DI and found insignificant correlations. Taib and Yusoff[19] compared item analysis results of One-Best MCQs with those of long case and reported that One-Best performed better.

This research may be considered as the first step in comparing the psychometric properties of the two tools. The current study was planned to find out the differences between One-Correct and One-Best MCQs with relation to student scores, post-exam item analyses results and student satisfaction.

Subjects and Methods

This comparative cross-sectional study was conducted at the Department of Ophthalmology at the Dow University of Health Sciences, Karachi, Pakistan, from November 2010 to April 2011, and comprised medical students. Non-probability, convenience sampling was used after getting approval from the institutional review board.

Fourth-year medical students, who were at the end of their clerkship in ophthalmology, were targeted. At the end of every ophthalmology rotation, the principal researcher gathered them in a classroom and explained the research process, its purpose and the formative nature of the test to be administered to them. He then requested the students to volunteer and assured them of complete anonymity. Verbal consent was then taken from the students to use their exam result for research purposes. These students were given a test comprising 100 items. Only those students who had attended two months of clerkship in the ophthalmology department, one month in the third year and one month in the fourth year were included in the research. These students were explained that the test was of two hours. There would be two papers, paper I and II, each comprising 50 items. Paper I would consist of One-Correct type and paper II of One-Best type.

A blueprint of the entire ophthalmology content was first developed to optimise content validity. The topics were then divided equally between the two papers. All the items of both the papers were written by the principal researcher. Both MCQ types were written based on guidelines given in the National Board of Medical Examiners (NBME) item writing manual. Item writing flaws (IWFs) identified in the manual were avoided. A break of 15 minutes was given between the two papers to avoid fatigue. The researcher ensured that the students did not get any chance to discuss the items during the break. For every cohort, the sequence of MCQs was changed and the scenarios tweaked in order to give them a fresh look.

At the end of both papers, a feedback form was distributed. This questionnaire asked students about their general feelings regarding items in both papers, their opinion about the difficulty level of items, the impact that these items might have on their learning process and how comfortable they would be if these items were included in their professional examinations.

At the end of exams, response sheets were scanned and data was saved in Microsoft (MS) Excel. Item analyses for P and DI were done using MS Excel. SPSS 18 was used for data analysis. Statistical analysis was carried out on the test scores of Paper I and Paper II by applying t-test.

p<0.05 was considered statistically significant. Point bi-serial was calculated by Pearson correlation coefficient, which was also used to study the relationship between P and DI of Paper I and Paper II.

Results

Of the 207 participants, 16(7.7%) were boys and 191(92.3%) girls. The mean score in Paper I was 18.62±4.7, while in Paper II it was 19.58±6.1. The researchers reject the hypothesis that there will be a significant difference in the marks of papers I and II (p = 0.075).

The mean values of 'P' in Paper I and Paper II were 0.37±0.19 and 0.39±0.13, respectively. Moreover, 19(38%) items in Paper I and 12(24%) items in Paper II were marked as 'difficult' (P<0.30). Besides, 30(60%) items in Paper I and 37(74%) items in Paper II were marked as 'moderate' (P>0.30 and P<0.80, respectively). Both Paper I and Paper II had 1(2%) easy item (P>0.80). No statistically significant difference was found in P between the two papers (t= -0.619, p= 0.537) (Appendix A).

Mean DI values in Papers I and II were 0.14±0.09 and 0.30±0.11, respectively. Also, 32(64%) items in Paper I and 9(18%) items in Paper II had DI of less than 0.20 and were marked as poorly discriminating. Moreover, 11(22%) items in Paper I and 14(28%) items in Paper II had DI between 0.20 and 0.29 and were marked as acceptable. Paper II had 17(34%) items as compared to only 2(4%) good discriminating items in Paper I (DI between 0.30 and 0.4). There was 1(2%) excellent item in Paper I (DI>0.4), and 10(20%) excellent discriminating items in paper II. There

Appendix A

Guidelines for Difficulty index	
1. Difficult -	0.0 to 0.3
2. Moderate -	0.3 to 0.8
3. Easy -	0.8 and above
Guidelines for Discrimination index	
1. Poor =	0.2 and below
2. Acceptable =	0.21 - 0.29
3. Good =	0.3 - 0.4
4. Excellent =	0.4 and above

Appendix B

Categories for Point-biserial

A. Insignificant = Regardless of the Correlation value (r2), the p-value is = or > 0.05 (poor)
B. Significant = p-value is < 0.05
   i. r2 = 0.19 and lower (poor)
   ii. r2 = between 0.20 and 0.29 (satisfactory)
   iii. r2 = 0.30 and higher (good)

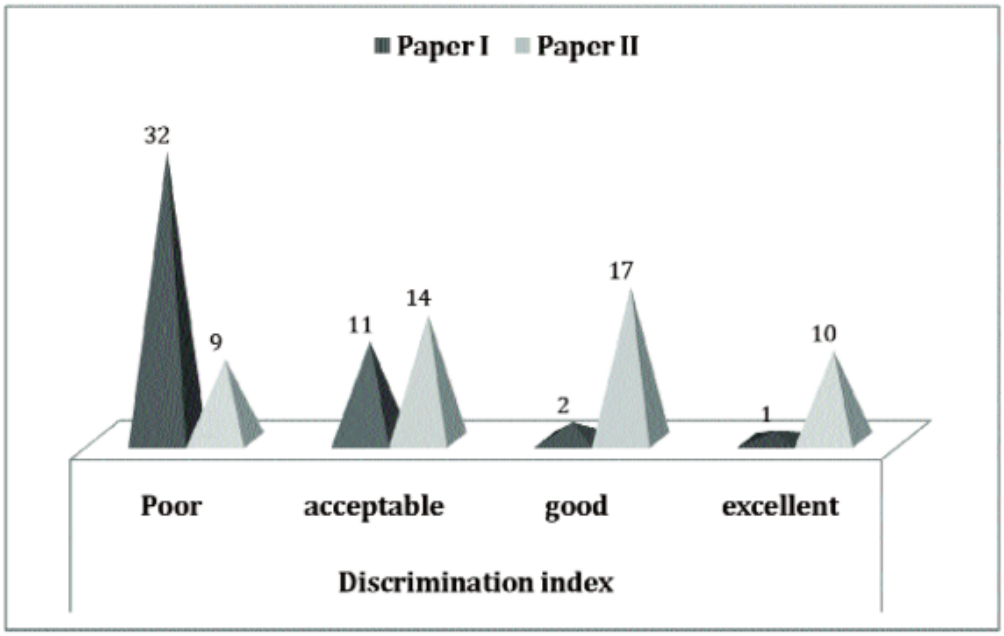

Figure-1: Comparison of discrimination index in Papers I & II.

Table: Proportion of responses from feedback questionnaire.

	Strongly agree	Agree	Unsure	Disagree	Strongly disagree
Questions were according to content taught to students	0	28 (13.6%)	66 (32.1%)	36 (17.3%)	77 (37%)
	Yes	**Unsure**	**No**		
Were items in Paper II more difficult than those in Paper I?	179 (86.4%)	13 (6.2%)	15 (7.4%)	-	-
Will items as in Paper II have an impact on your learning strategies?	112 (54.3%)	44 (21.09%)	51 (24.7%)	-	-
Would you be comfortable if Paper II type items were included in the professional exams?	74 (36%)	33 (16%)	100 (48%)	-	-

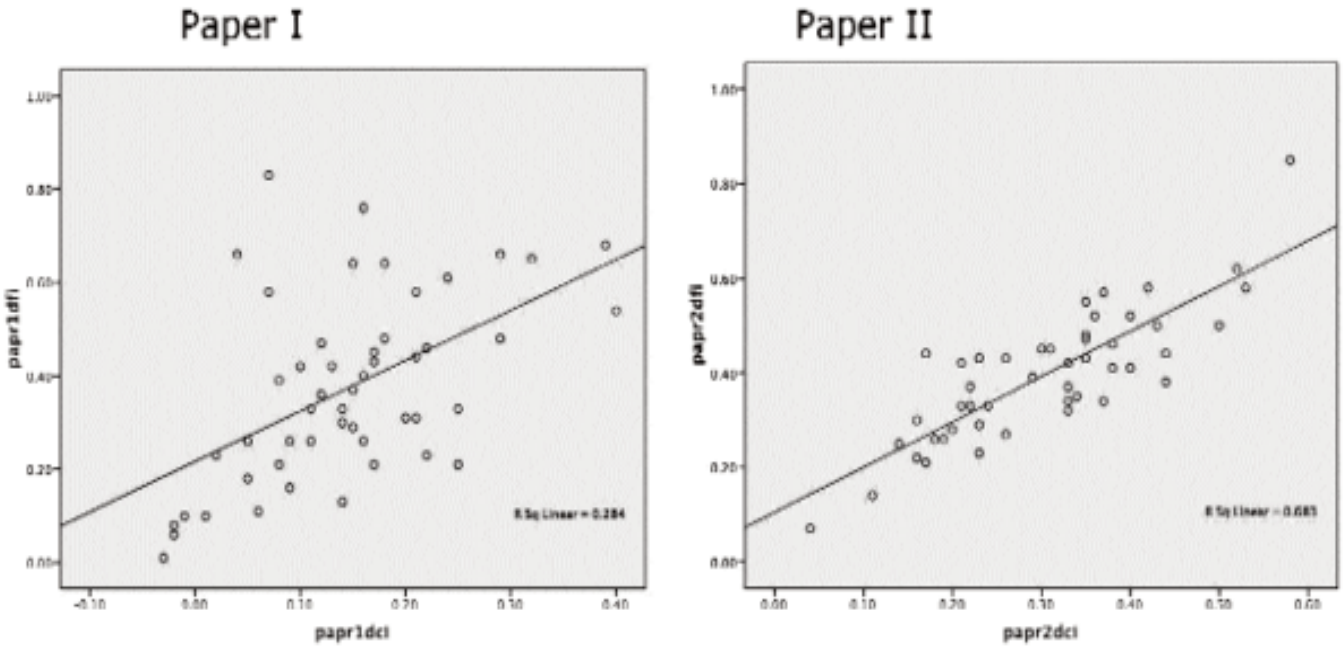

Figure-2: Difficult items are poor discriminators and moderate items are good discriminators. This relationship is positive and linear and is statistically highly significant (p= 0.000) in both Paper I, and Paper II. The r2 = 0.533 for Paper I and 0.836 for Paper II.

were no negatively discrimination items in Paper II. Paper I had 4(8%) negative discriminators. There was a statistically significant difference in the discriminating indices between the two papers (t= -7.732, p = 0.000) (Figure-1).

Point bi-serial had more satisfactory and good items in paper II than in paper I. The number of items which had insignificant correlations (i.e. no discriminatory abilities) was 15(30%) in paper I and 8(16%) in paper II (Appendix B).

Pearson correlations (r2) of P and DI of both Paper I and Paper II (r = 0.533 and 0.836, respectively) were highly significant (p= 0.000) (Figure-2).

Cronbach's alpha for Paper II (0.696) was higher than that for Paper I (0.548).

Distractor analysis showed that 21(42%) items had non-functional distractors (NFDs) in Paper I as compared to only 11(22%) with NFDs in Paper II. It is possible that some of these flaws were due to the wrong key being marked.

Only 28(13%) students were satisfied about items in Paper II reflecting the content taught. Besides, 179(86.4%) students found Paper II items to be more difficult. also, 112(54.3%) students thought that the items in Paper II would have impact on their learning. Only 74(36%) students thought they would be comfortable if One-Best MCQs were included in their final professional exams (Table).

Discussion

This study concentrated on the post-exam analyses of two types of MCQ examinations and compared the obtained data. Paper I consisted of One-Correct MCQs whereas paper II consisted of One-Best MCQs. Results show that, generally, One-Best MCQs had higher reliability, discrimination and correlation values than One-Correct MCQs, thus providing clear evidence about higher validity of the former tool. Results also showed that most students were not very comfortable with these vignette-based MCQs.

Students scored higher in Paper II, which was able to spread out the students more than Paper I. This indicates the overall discriminating ability of the second paper, a fact supported by point-biserial values. The difference in maximum scores is negligible. A plausible reason for the higher mean score obtained in the One-Best format exam is that the items were closer to the kind of cases students saw in rotation, and the teaching and learning was more towards application of knowledge, clinical decision and problem-solving. Therefore, the constructs taught and assessed were probably aligned.

Our study demonstrates a significant and linear relationship between P and DI. This relationship is more prominent and stronger in Paper II. This stronger relationship may be because of higher quality of items in Paper II.

Cronbach's alpha, indicating internal consistency, of Papers II was found to be higher than that of paper I (0.696 vs 0.548). This shows that the results of paper II were more trustworthy for decision-making than those of paper I. This adds to the evidence of validity in favour of One-Best MCQs.[3] Axelson and Creiter clearly state that a Cronbach's alpha value between 0.7 and 0.79 is acceptable for lower stakes exams, e.g. a formative test,[20] as in this case. Norcini et al. also found that One-Best MCQs had a reliability of at least 0.72 or above.[12] Tan and McAleer[21] also reported true/false MCQs to have lower reliability than the One-Best variety.

This study concurred with the results found in Caroll's[11] early work where P and DI of One-Correct and One-Best were compared. The difference in P was found to be insignificant in both studies in contrast to DI which was found to have a significant difference. This adds value to One-Best MCQs since one main purpose of assessment is to differentiate among the various groups of learners based on their competencies. Tan and McAleer[21] also reported higher DI for One-Best.

Downing reported Haladyna's work and classified items based in DI and P. The classification showed that item difficulty and discrimination were often reciprocally related. Questions with high and low difficulty indices (i.e. easy ones and difficult ones) generally, but not always, show low discrimination values.[3] If an item is easy most of the students get it right and thus this item is unable to segregate the average ones from the below average. Similarly, if an item is difficult, only the top students may answer it correctly thereby lumping the average and below average. Questions with moderate difficulty index are, by and large, the best discriminators. The present results are in line with these principles highlighted in the work by Carroll.[11] Sim and Rasiah.[22]

Student perceptions are important because they provide insight into the factors that hinder learning as novices in clinical practice and can suggest approaches for improvement.[23] Students were wary of the One-Best items and perceived them as a threat. They had never before been exposed to this format and hence saw it as a challenge if it came in their final examinations. Students as well as faculty need to be trained in this format. Further, once they understand that One-Best MCQs check knowledge application and are the same types being used in examinations of College of Physicians and Surgeons and US Medical Licensure Examination (USMLE), they would probably be more willing to accept it.

There are a number of limitations of this study. The number of MCQs analysed was just 100 from a single discipline. The test was formative; it is likely that students did not prepare for it and hence item analysis may not reflect the true level of student knowledge. Data was taken from only one institution. Moreover, the research is being published after a delay of more than five years because of administrative reasons in the university. Only one expert was involved in developing the blueprint and the items when, in fact, a group of experts should have been involved.

Conclusion

Students had higher scores on One-Best MCQs which demonstrated better post-exam analysis results than One-Correct MCQs; the former had higher reliability and discrimination ability than the latter, while no significant difference was found in difficulty levels of the two types. Despite their better performance, the students appeared apprehensive about and dissatisfied with One-Best MCQs.

Disclaimer: None.

Conflict of Interest: None.

Source of Funding: None.

References

1. Swanwick T. Understanding Medical Education: Evidence, Theory and Practice. 1st Edition. London Deanery, London: John Wiley & Sons, Ltd., Publication, Wiley Blackwell, 2010; 137-140.
2. Rotem A, Barrand J, Azman A. Analysis of examinations in curriculum review. Med Educ 1982; 16: 3-6.
3. Downing SM. Statistics of Testing. In: Downing SM, Yudkowsky R (Eds.). Assessment in Health Professions Education. New York: Taylor and Francis, 2009; 107- 109.
4. Miller GE. The assessment of clinical skills/ competence/ performance. Acad Med 1990; 65: s63-67.
5. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "fact recall" with "higher order questions in multiple choice examinations as predictors of clinical performance of medical students. Acad Med 1990; 65: S59-60.
6. Mitra N K, Nagaraja H S, Ponnudurai G, Judson J P. The levels of difficulty and discrimination indices in type — A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. Ie JSME 2009; 3: 2-7.
7. Bridge PD, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: Methods for developing content-valid student examinations. Med Teach 2003; 25: 414-21.
8. Shea JA, Fortna GS. Psychometric models. In: Norman G, Van der Vleuten CPM, Newble D. (Eds.). International handbook of research in Medical Education. London: Kluwer Academic Publishers, 2002; 97-100.
9. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach 2011; 33: 447-58
10. Tweed M, Wilkinson T. A randomized controlled trial comparing instructions regarding unsafe response options in a MCQ examination. Med Teach 2009; 31: 51-4.
11. Carroll RG. Evaluation of vignette-type examination items for testing medical physiology. Am J Physiol 1993; 264: S11-5.
12. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical

competence. Med Educ 1985; 19: 238-47.
13. Downing SM, Baranowski RA, Grosso LJ, and Norcini JJ. Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. Appl Meas Educ 1995; 8: 187-97.
14. Hingorjo MR, Jaleel F. Analysis of One-Best MCQs: the Difficulty Index, Discrimination Index and Distractor Efficiency. J Pak Med Assoc. 2012; 62: 142-7.
15. Baig M, Ali SK, Ali S, Huda N. Evaluation of Multiple Choice and Short Essay Question items in Basic Medical Science. Pak J Med Sci 2014? 30: 3-6.
16. Mahmood H. Correlation of MCQ and SEQ Scores in Written Undergraduate Ophthalmology Assessment. J Coll Physicians Surg Pak 2015; 25: 185-8.
17. Abdulghani HM, Ahmad F, Irshad M, Khalil MS, Al-Shaikh GK, Syed S, et al. Faculty development programs improve the quality of Multiple Choice Questions items' writing. Sci Rep. 2015; 5: 9556.
18. Karelia BN, Pillai A, Vegada BN. The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of Year II M.B.B.S students. IeJSME 2013; 7: 41-6.
19. Taib F, Yusoff MSB. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. J Taibah Uni Med Sci 2014; 9: 110-4.
20. Axelson RD, Creiter CD. Reliability. In: Downing SM, Yudkowsky R (Eds.). Assessment in Health Professions Education. New York: Taylor and Francis, 2009; 62 - 64.
21. Tan LT, McAleer JJA; Final FRCR Examination Board. The Introduction of Single Best Answer Questions as a Test of Knowledge in the Final Examination for the Fellowship of the Royal College of Radiologists in Clinical Oncology. Clin Oncol (R Coll Radiol) 2008; 20: 571-6.
22. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false - type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore 2006; 35: 67-71.
23. Dolmans DHJM, Wolfhagen IHAP, Heineman E, Scherpbier AJJA. Factors Adversely Affecting Student Learning in the Clinical Learning Environment: A Student Perspective. Educ Health (Abingdon) 2008; 21: 32.