ORIGINAL ARTICLE

Correlation of MCQ and SEQ Scores in Written Undergraduate Ophthalmology Assessment

Hamid Mahmood

Abstract

Objective: To determine the psychometric worth of Multiple-Choice Questions (MCQs) and Short-Essay Questions (SEQs) used in internal summative assessment of undergraduate ophthalmology annual send-up examination and to ascertain the quality of items using subjective measures.

Study Design: Correlational analytical study.

Place and Duration of Study: Fatima Jinnah Medical College, Lahore, from January to March 2012.

Methodology: Fourth year MBBS students appearing in their written send-up annual examination in ophthalmology. The MCQ and SEQ scores were correlated for validity and reliability. Construct validity was estimated (Pearson correlation coefficient). MCQ's split-half reliability (Spearman correlation coefficient) and SEQ's inter-rater reliability (Pearson correlation coefficient) were also estimated.

Results: Moderate correlation was seen between MCQ and SEQ scores ($r=0.5$, $p < 0.01$). Split-half reliability of MCQs had moderate correlation ($r=0.4$, $p < 0.01$) while inter-rater reliability of SEQs scores showed high correlation ($r=0.9$, $p < 0.01$). Most of the MCQs (55%) and SEQ's (90%) were at recall level of Bloom's taxonomy. MCQs (30%) attained a higher level (interpretation) as compared to SEQ's (10%). None of the MCQ and SEQ attained problem solving level.

Conclusion: Using variety of tools for assessment improves both the validity and reliability enabling examiners to draw fair conclusions about a student's ability. Validity and reliability can be further improved by arranging training opportunities for item writers to achieve accurate transfer of hypothetical construct in the form of an item. This is crucial since essential and important learning outcomes can only be assessed with valid and reliable tools.

Key Words: *MCQ. SEQ. Psychometrics. Validity. Reliability.*

INTRODUCTION

There is a considered view that assessment drives learning.[1] The domains of assessment include cognition, performance of skills and attitudes.[2] The written tests like Multiple-Choice Questions (MCQs) and Short-Essay Questions (SEQs) test formats are used for the assessment of cognitive domain.[2] The MCQs are more objective and essentially select type of item response format.[3] MCQs have a cueing effect, which promotes guessing and leads to higher scores.[4] In addition, writing MCQs of higher cognitive level of problem solving is challenging.[5] On the contrary, the SEQs are more subjective and have a supply or construct type item response format, which does not have any cueing effect and can effectively assess problem solving skills.[5]

In terms of scoring system, the MCQs are easy to score both manually or by a computer.[6] The SEQ scoring is time-consuming and may have an element of bias in terms of subjective judgment which can be circumvented by using a scoring rubric.[6] MCQ and SEQ are valid and reliable tools in the assessment of cognitive domain.[7]

Department of Ophthalmology, Services Institute of Medical Sciences (SIMS), Services Hospital (SHL), Lahore.

Correspondence: Prof. Dr. Hamid Mahmood, 140-A, Shadman II, Lahore.
E-mail: hamidbut@gmail.com

Received: October 25, 2013; Accepted: December 31, 2014.

This study was done to determine the psychometric worth of MCQs and SEQs used in internal assessment i.e. annual send-up summative examination in undergraduate ophthalmology and to ascertain the quality of items used in the examination using subjective measures. The rationale was to provide evidence of psychometric quality of MCQs and SEQs prepared by the institutional faculty in order to identify gaps in practice for future faculty development workshops.

METHODOLOGY

This correlational analytical study was done at Fatima Jinnah Medical College, Lahore, Pakistan from January to March 2012. Participants of the study were fourth year MBBS students undertaking end of course send-up examination in the subject of ophthalmology after one academic year. The total class comprised of 239 students. Nine students were absent and were thus excluded. Hence a total of 230 students who appeared in the examination were included. The study was approved by the institutional review board for ethical purposes. One 45 item MCQ test paper and one 9 item SEQ paper was administered. Each MCQ began with a stem and a lead-in followed by five options. A table of specifications was used to develop/select MCQs according to the core curriculum in ophthalmology to ensure content validity. There was a maximum total of 45 marks and one hour test taking time was allowed.

Each question carried one mark for a correct answer. There was no negative marking for a wrong answer. Hand scoring of MCQ test was done.

The SEQ paper had 9 questions with 5 marks for each answer. A test blue print was used to select the SEQs in content areas similar to selection of MCQs. An analytic rubric was used for independent marking by a pair of examiners and rater training was done. The average score of the pair was assigned as the final mark for that particular answer. The total test time was one hour and thirty minutes. The total score for the SEQ paper assessment was also a maximum of 45 marks. Absolute method of standard setting at 50% or 45/90 cumulative marks in MCQ and SEQ was set as the pass mark.[2]

Students' scores generated from administration of MCQ and SEQ sent-up examination in ophthalmology were recorded (percentage marks). Data analysis was done using SPSS version 20. Mean value ± SD was calculated. Correlation between MCQ and SEQ scores was estimated by using Pearson product moment method. Internal consistency of MCQ result was determined through split-half reliability by employing Spearman correlation coefficient. The SEQ inter-rater reliability was estimated by Pearson correlation coefficient. P-value of < 0.05 was considered statistically significant.

RESULTS

The mean percent score of the total group of 230 students was 64.90 ± 8.28, while it was 70.58 ± 10.07 in the SEQ group and 59.24 ± 9.25 in the MCQ group. Students who achieved cumulative pass score of 50% or more were 226/230 (98.3%) while 4/230 (1.7%) failed. A statistically significant moderate correlation was found out between SEQ and MCQ scores (r= 0.5, p < 0.01) (Figure 1). The cognitive level of MCQ and SEQ was judged by two independent reviewers using the modified Bloom's taxonomy. More than 55% of the MCQs and 90% of the SEQs were found to be at recall level of

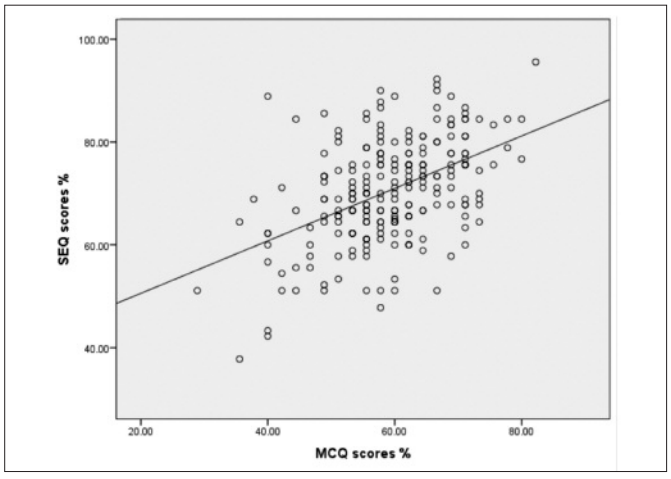

Figure 1: Scatter plot of overall student performance in written ophthalmology send-up examination (r = 0.5, p < 0.01).

Table I: Modified Bloom's taxonomy: MCQ-SEQ inter-rater agreement.

Modified Bloom's Taxonomy level	MCQ inter- rater agreement		SEQ inter-rater agreement	
	Rater 1	Rater 2	Rater 1	Rater 2
Level 1 Recall	25 (55%)	29 (64%)	8 (89%)	9 (100%)
Level 2 Interpretation	15 (33%)	16 (35%)	1 (11%)	-
Level 3 Problem solving	5 (11%)	--	-	-
Total	45 (100%)	45 (100%)	9 (100%)	9 (100%)

Table II: Rating scale used for judging MCQ item writing flaws.

Conditions required to achieve rating	Rater 1	Rater 2
1. Pass the cover test and no item writing flaws	44/45 (98%)	28/45 (62%)
2. Pass the cover test and 1 to 2 item writing flaws	1/45 (2%)	16/45 (36%)
3. Cover test dubious and no item writing flaws	-	1/45 (2%)
4. Fail the cover test and 1 to 2 item writing flaws	-	-
5. Fail the cover test and more than 2 item writing flaws	-	-
Total	45 (100%)	45 (100%)

modified Bloom's taxonomy. About 30% of MCQ's attained interpretation or level 2 as opposed to none of the SEQs having attained that level; 5 (11%) MCQs attained problem solving level (Table I). The structural aspects of construct validity of MCQ item was checked by using a rating scale and cover test by two raters (Table II).

MCQs split-half reliability was estimated by correlation of odd item scores (23 items) and even item scores (22 items) and the correlation coefficient was determined to be moderate (r=0.4, p < 0.01). SEQ inter-rater correlation coefficient (Pearson's) was found to be high (r=0.9, p < 0.01).

DISCUSSION

End-of-term send-up test is good assessment strategy regarding preparation for the university professional certification examination since the marks attained are included in the cumulative university examination scoring system in a ratio of 10%; hence the students take it seriously and consider it good preparatory practice. The notion of assessment for learning has a formative element as students' result is discussed with them, making it different from the end-of-course exit examination, which is assessment of learning.[9] Formative assessment significantly contributes in making learning more meaningful by helping students identify their deficiencies and also by providing a direction for corrective measures, it steers learning towards the desired direction. Besides, it has a strong role in motivating and empowering students to become self-regulated learners as well.[5]

An assessment has to be valid and reliable.[10] Validity is a property of the scores generated from student responses and not of the assessment instrument. Rather it is the meaningful interpretation of scores. The hypothetical construct of a task or domain relates to the level or construct of cognition on an ascending scale from recall, understanding and problem solving.[10] All

validity is said to be construct validity having two important aspects; Construct-Under-Representation (CUR) and Construct-Irrelevant-Variance (CIV).[11] CUR is avoided by adequate sampling of content by examination blue printing and assigning appropriate level of cognition to the items according to Bloom's taxonomy. The items should be written to the highest level according to the learning objectives of the desired domain.[11,12] Item Writing Flaws (IWF) are an important cause of CIV and can be avoided by proper item writing according to the guidelines and best practices. IWF make the items too easy or too difficult. Too easy items unduly favour the weak students and too difficult items put the good student at a disadvantage. In addition, due attention should be paid to language, cultural and social appropriateness.[11,12]

The scores produced through MCQ and SEQ were correlated for validity and reliability evidence in terms of their meaningful interpretation. A statistically significant moderate correlation was seen between SEQ and MCQ scores of the total group of students ($r= 0.5$, $p < 0.01$). The moderate correlation is also suggestive of the fact that the two assessment methods are different but inter-related. A strong correlation between MCQs and SEQs has been reported ($r=0.6$, $p < 0.01$) in other studies.[13] It can also be inferred from the moderate to strong correlation between MCQ and SEQ that the students who perform well in MCQ also perform better in SEQ also.[14,15] It is also pertinent to note that a correlation coefficient of 0.6 shows a strong relationship in the context of construct validity as opposed to interpretation of reliability coefficients where 0.7 to 0.8 is also seen as moderate correlation.[10] On the contrary, the correlation between the written tests and clinical tests like OSCE are surprisingly low showing that the assessment methods are markedly different in terms of construct validity.[16]

It was further seen in this study that the desired level of cognitive assessment in terms of construct validity did not seem to be attained as evaluated by two independent raters in terms of modified Bloom's taxonomy. More than 55% of the MCQs and 90% of the SEQs were found to be at recall level of modified Bloom's taxonomy. Interestingly, about 30% of MCQs attained interpretation or level 2. This could be due to better faculty understanding of MCQ item construction in terms of the vignette or clinical scenario based stem. It has been suggested that MCQs can be the preferred tool to assess problem solving skills.[7] In a study by Baig *et al*. in the basic sciences, 76% MCQ were at level 1, 24% at level 2, zero at level 3.[17] In another study of clinical subjects, 60% of MCQs were found to be at level 1, 6% at level 2 and 28% at level 3.[18] Tarrant showed in a study that 90% of MCQs were at the recall level.[19] Lower level MCQ are easy to make and need less time.[20] Lower level MCQ items have more IWF which in turn lowers the construct validity. This can be minimized by writing higher level MCQ's as they are found to have less IWF.[21] Lower level MCQ's lowers the quality of items and hence affects the difficulty and discrimination indices. Lower difficulty and poor discrimination favours poorly performing students. Higher difficulty and poor discrimination negatively affects good students.[21]

Besides lowering the overall validity, items of lower quality also affect the reliability of the assessment as validity is the upper limit of reliability.[10] Reliability relates to the consistency or repeatability of measurement and is important regarding fairness and hence defensibility of results.[1] Various aspects of reliability have to be interpreted in the context of the desired dimension of reliability and pattern of judgment of assessors.[22] The desired reliability of MCQ is 0.8 or higher.[1,10] In this study, the split-half reliability of MCQ seems to be moderate ($r= 0.4$, $p < 0.01$). This also calls for regular use of Cronbach's alpha if item is deleted to objectively determine the role of scores produced through each item on overall internal consistency and most importantly to improve the construct of the items responsible for bringing down the Cronbach's alpha.

Although the reported reliability of open-ended questions is low, it can be improved by use of analytic rubrics and rater training.[6] The inter-rater reliability correlation coefficient of the SEQs test scores in this study is quite high ($r = 0.9$, $p < 0.01$) due to use of analytic rubrics and rater training. It is reported that the marking accuracy increases as the length of essay answers is reduced.[23] The structuring of an essay question also improves the internal consistency from 0.31 ($p > 0.05$) to 0.69 ($p < 0.05$).[24] The SEQ's behaved as trimmed down version of essay question that the faculty is so used to write and students are accustomed to respond.

This study highlights the importance of validity, more specifically in terms of construct validity. Construct under-representation and construct irrelevant variance have to be rigorously addressed.[11] The faculty must follow the item writing guidelines according to best evidence based practices. Item review committees should ensure quality improvement in curriculum delivery in an assessment process.[25]

The study was done at FJMC only and not at other institutions. The results of only one subject namely ophthalmology are being studied in the written component only. For the sake of a complete assessment the domains of skill and attitudes also need to be assessed by tools like Objective Structured Clinical Examination (OSCE). The other examination subjects of fourth year are not included. Item analysis or Cronbach's alpha if item deleted were not performed for the MCQ's during this study.

CONCLUSION

The construct validity of MCQ and SEQ can be improved by arranging training opportunities for item writers. The role of item review committees for quality assurance is also mandatory. A combination of different tools to measure the cognitive domain helps in evaluating examinees' ability from different perspectives. The learning outcomes can only be properly assessed with valid and reliable tools.

Disclosure: This article was written for partial fulfillment for award of MCPS-Health Professions Education and there is no conflict of interest.

REFERENCES

1. Gronlund NE, Linn RL, editors. Measurement and assessment in teaching. Columbus, Ohio: Merrill, Prentice Hall; 2000.
2. Mc Aleer. Choosing assessment instruments. In: Dent JA, Harden RM, editors. A practical guide for medical teachers. London: *Churchill Livingstone*, 2009; p. 322-3.
3. Schuwirth LW, Vleuten CP. Different assessment methods: What can be said about their strengths and weaknesses. *Med Educ* 2004; **38**:974-9.
4. Schuwirth LWT, van der Vleuten CPM, Donkers HHLM. A closer look at cueing effect of multiple-choice qestions. *Med Educ* 1996, **30**:44-9.
5. Schuwirth LW, Van Der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No.57. *Med Teach* 2011; **33**:783-97.
6. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 2nd ed. Philadelphia: National Board of Medical Examiners; 1998.
7. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions?: *BMC Med Edu* 2007; **7**:49.
8. Buckwalter JA, Schumacher R, Albright JP. Use of an educational taxonomy for evaluation of cognitive performance. *J Med Educ* 1981; **56**:115-21.
9. Larsen DP, Butler AC, Roediger HK. Test-enhanced learning in medical education. *Med Educ* 2008; **42**:959-66.
10. Cook DA, Beckman TJ. Current concepts for validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006; **119**:166.
11. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ* 2004; **38**:327-33.
12. Haldayana TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines. *Appl Measurement Educ* 2002; **15**:307-33.
13. Mujeeb AM, Pardeshi ML, Ghongane BB. Comparative assessment of multiple choice questions versus short essay questions in pharmacology examination. *Indian J Med Sci* 2010; **64**:118-24.
14. Oyebola DD, Adewoye OE, Iyaniwura JO, Alada AR, Fasanmade AA, Raji YA. Comparative study of students performance in physiology assessed by multiple choice and short essay question. *Afr J Med Sci* 2000; **29**:201-5.
15. McCloskey DI, Holland RA. A comparison of student performances in answering essay- type and multiple choice questions. *Med Educ* 1976; **10**:382-5.
16. Wass V, Vleuten CV, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001; **357**:945-9.
17. Baig M, Ali SK, Ali S, Huda N. Evaluation of multiple choice and short essay question items in basic medical sciences. *Pak J Med Sci* 2014; **30**:3-6.
18. Khan MZ, Aljarallah BM. MEQs and MCQ as a tool for assessing the cognitive skills of undergraduate medical students. *Int J Heal Sci* 2011; **5**:45-50.
19. Tarrant M, Ware J. Impact of item-writing flaws in multiple choice questions on student achievement in high-stakes nursing assessments. *Med Educ* 2008; **42**:198-206.
20. Tarrant M, Ware JA. Framework for improving the quality of multiple-choice assessments. *Nurse Educ* 2012; **37**:98-104.
21. Tarrant M, Kneirim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple choice questions used in high stakes nursing assessments. *Nurse Educ Today* 2006; **26**:662-71.
22. Wass V, McGibbon D, Vleuten CV. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability. *Med Educ* 2001; **35**:326-30.
23. Smith J, Neely D, Hirschtick R. Achieving inter rater reliability in evaluation of written documentation. *Med Educ* 2009; **43**: 485-6.
24. Verma M, Chhatwal J, Singh T. Reliability of essay type questions-the effect of structuring. *Assess Educ Principles Policy Pract* 1997; **2**:265-70.
25. Jozefowicz RF, Koeppen BM, Case SP, Galbraith RM, Swanson DP, Glew RH. The quality of in-house medical school examinations. *Acad Med* 2002; **77**:156-61.