International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 2 Issue 1 ǁ January. 2014ǁ PP-06-10

"Reliability of four-response type multiple choice questions of pharmacology summative tests of II M.B.B.S students"

Bhavisha N. Vegada, Bharti N. Karelia, Ajita Pillai
Department of Pharmacology, P.D.U. Govt. Medical College,Rajkot, Gujarat, India.

ABSTRACT:

Introduction
One of the major concerns in the construction of test items for an examination is ensuring the reliability of the test items. In order for assessments to be sound, they must be free of bias and distortion. Reliability and validity are two concepts that are important for defining and measuring bias and distortion. Internal consistency is an estimate of reliability based on the average correlation among items within a test and examines the degree to which the MCQs in a test measure the same characteristics or domains of knowledge.

Methods
In this study ten MCQ tests from 2008 to 2012 were selected and analyzed to obtain their mean, standard deviation, reliability coefficient and standard error of measurement. Data entry was done by using Microsoft excel 2007.

Results
Mean reliability coefficient was 0.54. Out of ten tests, two tests had low reliability, five tests had very low and three tests had questionable reliability. Mean standard deviation of MCQ tests was 3.52 with range of 3.05 to 4.01. Mean standard error of measurement was 2.37 with range of 2.24 to 2.44.

Conclusion
Reliability of all MCQ tests was low and need improvement. Standard Error of Measurement is more appropriate parameter for reliability than Reliability Coefficient.

KEYWORDS: *Reliability, Reliability coefficient, Standard error of measurement*

I. INTRODUCTION

The educational objectives in medicine as well as in other discipline are generally allotted to three 'domains'-cognitive, psychomotor and affective. Hence, medical examination should be designed to answer whether an undergraduate has achieved the above educational objectives by answering the following three questions: (1) what does he know (cognitive)? (2) what can he do (psychomotor)? And (3) what sort of person is he (affective)? Regrettably the current medical examination system still could not answer these questions faithfully. [1]

Objectivising evaluation is becoming increasingly more important in the field of education, both for summative & formative purpose, as has been again & again emphasized by guidelines published by several universities. One method of achieving this purpose is the widespread use of objective written items, and the most popular form of which is the multiple choice question (MCQ). [2]

Designing MCQ is a complex and time consuming process in a multidisciplinary integrated curriculum. MCQs are used mostly for comprehensive assessment at the end of a semester or academic session and provide feedback to the teachers on their educational action. Having constructed & assessed a test, a teacher needs to know, how good the test questions are & whether the test items were able to reflect students' performance in the course related to learning. Because of their versatile character, MCQs are the most commonly used tool for assessing the knowledge capabilities of medical students. [3] There are different types of MCQs like five-response, four-response, three-response and true/false or two-response. [4] One of the major concerns in the construction of test items for an examination is ensuring the reliability of the test items. [3]

For the assessments to be sound, it should be free from bias and distortion. Reliability and validity are two concepts that are important for defining and measuring bias and distortion. *Reliability* refers to the extent to which assessments are consistent. *Validity* refers to the accuracy of an assessment. [5] Concepts related to reliability are consistency, precision, stability, equivalence and internal consistency (Beanlander al1999 p328).

Internal consistency is an estimate of 'reliability based on the average correlation among items within a test' (Nunnally& Bernstein 1994 p251) and examines the degree to which the MCQs in a test measure the same characteristics or domains of knowledge (Bean land er al 1999, Polir&Hungler 1999). Typically, internal consistency is measured by the calculation of a reliabiliry coefficient (Cronbach 1990, Beanland et al 1999, Polit & Hungler 1999). [6] Reliability depends both on Standard Error of Measurement (SEM) and on the ability range (standard deviation, SD) of candidates taking an assessment. [7]

The present study was taken up with an objective to measure the reliability of MCQs.

II. MATERIAL& METHODS

The pattern of 1st and 2nd terminal examination of pharmacology subject at our institute consists of 80 marks theory and 50 marks practical examination. Theory examination consists of 20 multiple choice questions of 1 mark each. Year indicate when the 1st terminal examination was held and A for 1st terminal and B for 2nd terminal examination.

2.1 Data collection

MCQ items were taken from the 10 summative test papers from the year 2008-2012 (each year having two terminal examinations). A total of 200 test items were selected for the item analysis. Each MCQ consisted of a stem and four choices and the students were to select one best answer from these four choices. A correct response to an item was awarded 1 mark, while an incorrect response would result in negative 0.25 marks and a no- attempt or blank response was given no mark.

2.2 Data analysis

MCQ scores of students of different batches of last five years from 2008 to 2012 were included for analysis and data entry was done by using Microsoft excel 2007. Different statistical parameters like mean, standard deviation, reliability coefficient, standard error of measurement and confidence interval for MCQ tests were calculated.

The Equation for Reliability Coefficient is as follow: [8]

$$\text{Alpha} = [n/(n - 1)] \times [(\text{Vart} - \Sigma\text{Vari})/\text{Vart}] \quad (1)$$

Alpha = estimated reliability of the full-length test
n = number of items
Vart = variance of the whole test (standard deviation squared)
ΣVari = sum the variance for all n items

The values for reliability coefficients range from 0 to 1.0. A coefficient of 0 means no reliability and 1.0 means perfect reliability. Since all tests have some error, reliability coefficients never reach 1.0. Generally, if the reliability of a standardized test is above .80, it is said to have very good reliability; if it is below .50, it would not be considered a very reliable test. [5]

Tests with a reliability coefficient 0.90 and above were considered as excellent reliability, those between 0.80-0.90 were considered very good, those between 0.70-0.80 were good, those between 0.60-0.70 were considered low and therefore needs to be supplemented by other measures to determine grades, those between 0.50-0.60 needs revision of test and those with 0.50 and those below were considered to have questionable reliability. [9]

The Equation for Standard Error of Measurement is as follow: [10]

$$\text{SEM} = S\,(1-r)^{½} \quad (2)$$

Where, S = theStandard Deviation for the test.
r = the Reliability coefficient for the test

III. RESULTS

As shown in Table 1, mean reliability coefficient was 0.54. Out of ten tests, two tests had low reliability, five tests had very low and three tests had questionable reliability.

(Table-1)

As shown in Table 2, mean standard deviation of MCQ tests was 3.52 with range of 3.05 to 4.01. Mean standard error of measurement was 2.37 with range of 2.24 to 2.44.
(Table-2)

IV. DISCUSSION

The reliability of an examination provides useful information about its performance (and it is self-evident that an examination with a very low reliability is unlikely to be a good or an effective examination, to the point where zero reliability means that the marks from an examination are no more effective than are random numbers at distinguishing between candidates). Having said that, the mere fact that an examination has a high reliability does not ensure that it is necessarily functioning effectively, because the reliability is heavily dependent upon the ability range of the candidates who are taking it. As has already been seen: [7]

1) The very same exam can apparently drop its reliability dramatically if it is retaken but only by those who have already passed it;
2) The reliability can be artificially inflated by encouraging very weak candidates to take it, thereby increasing the SD of the marks;
3) It is almost inevitable where successive examinations are taken, as with the Part 2 Written examination of MRCP(UK) being taken after Part 1, that the SD will necessarily be lower (only able candidates passing Part 1), and that the reliability of a second examination will usually be lower than the first examination.
4) When examinations have very small numbers of candidates, as with the SCEs, there is a greater risk that the reliability will be distorted by an unusually high or low spread of candidate abilities

Reliability can always be increased by making an assessment progressively longer, thereby increasing the number of examination items, although that is expensive in time, effort and opportunity cost. [7] Our results showed that two tests have low reliability means they need to be supplemented by other measure and there are probably some items which could be improved. Five tests have very low reliability means they need revision and supplemented by other measure. Three tests have questionable reliability means thesetests should not contribute heavily to the course grade, they need revision.

Cortina et alconsider reliability coefficient at least 0.70 or above to be adequate for classroom assessment. [11] None of the test was found to fulfill this criteria ,so our MCQ tests have low reliability. High reliability means that the questions of a test tended to "pull together."Students who answered a given question correctly were more likely to answer other questions correctly. If a parallel test were developed by using similar items, the relative scores of students would show little change. Low reliability means that the questions tended to be unrelated to each other in terms of who answered them correctly. The resulting test scores reflect peculiarities of the items or the testing situation more than students' knowledge of the subject matter. [9]

Another way to express reliability is in term of the standard error of measurement. This measure provides an estimate of how much an individual' score would be expected to change on re-testing with no change in knowledge and perception with the same or an equivalent form of the test. Our result showed that standard deviation of candidate scores showed large variation (3.05-4.01) as compared to variation in standard error of measurement (2.24-2.44).

Based on the assumption that any test score contains an error , SEM is used to estimate a band or interval within which a person's true score would fall, that is the score (hypothetical) the student would receive if there were no error of measurement.[12] The smaller the SEM is; the narrower the interval. Narrow intervals are more precise, containing less error, than larger intervals. SEM is inversely related to the Reliability Coefficient. [13]

For example, in our study 2008-A exams ,Mean Observed Score was 8.11 and SEM was 2.33.Wecan say with 95% Confidence ,true score of students of this batch lies in an interval within two SEM of the observed score.(between3.45 and 12.77). An alternative interpretation states that 95 times out of 100 times the students' score on a retest would be between 3.45 and 12.77.

As shown in Table 1, reliability coefficient of first terminal examinations of year 2009 and 2011 were 0.51 and 0.63 respectively but standard error of measurement of these examinations was same 2.44. So for reliability of examinations don't consider only reliability coefficient but standard error of measurement is also important stastical parameter. SEM is more appropriate for reliability than reliability coefficient. [7]

Our results suggest that our MCQ tests have not very reliable tests and need to improve.Reliability also shows problems when numbers of candidates in examinations are low and sampling error affect the range of candidate ability. SEM is not subject to such problems; it is therefore a better measure of the quality of an assessment and is recommended for routine use. [7]

V. CONCLUSION

Reliability of all MCQ tests was low and need improvement. Standard Error of Measurement is more appropriate parameter for reliability than Reliability Coefficient.

REFERENCES

[1] T. Ho, W. Yip, and J. Tay, *The use of multiple choice questions in medical examination: An evaluation of scoring and analysis of results,* Singapore Medical Journal, *22(6),* 1981, 361-367.

[2] N. Ananthakrishnan, Item analysis-validation and banking of MCQs, in N. Ananthakrishnan, K. Sethuraman, S. kumar, (Ed.), Medical Education principles and practice, 2(JIPMER, Pondicherry)131-137.

[3] N. Mitra, H. Nagaraja, G. ponndurai, et al, *The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests,*IeJSME, *3(1),* 2009, 2-7.

[4] Understanding item analysis reports, [online] available at: http://www.washington.edu/oea/service/scanning-scoring/item_analysis.html [accessed December 20, 2012].

[5] Classroom assessment, [online] available at: http://www.fcit.usf.edu/assessment/basic/basic.html [accessed December 20,2012].

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

[7] J. Tighe, I. McManus, N. Dewhurst, L. Chis, andJ. Mucklow, *The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations,* BMC Medical Education, 10-40.

[8] Introduction to reliability, [online] Avialable at: http://www.ncsu.deu/jlnietfe/EDP560_notes_files/reliability.pdf [accesssed January 10, 2013].

[9] Understanding item analysis reports, [online] Available at: http://www.washington.edu/oea/service/scanningscoring/scanning/itemanalysis [accessed January 10, 2013].

[10] Standard error of measurement, [online] Available at: http://web.sau.edu/WaterStreetMaryA/NEW%20intro%20to%20tests%20&%20measures%20website_files/standard_error_of_measurement.htm [accessed January 10, 2013].

[11] J. Cortina, *What is Coefficient Alpha?An Examination of Theory and Applications,* Journal of Applied Psychology, 78(1), 1993, 98-104.

[12] Test reliability, [online] Available at: http://www.indians.edu/best/testreliability [accessed January 10, 2013.

[13] L. Harvill, *Standard error of measurement,* Education measurement : issues & practice, *summer*, 33-41.

Table-1 :***Reliability coefficient of ten MCQ tests***

Year	*Examination*	*Reliability Coefficient*
2008	A	0.50
	B	0.54
2009	A	0.51
	B	0.60
2010	A	0.66
	B	0.38
2011	A	0.63
	B	0.53
2012	A	0.51
	B	0.49
Mean		**0.54**

Table-2*: Standard deviation, standard error of measurement and confidence interval of ten MCQ tests*

Year	*Examination*	*Mean*	*Standard deviation (SD)*	*Standard Error of Measurement (SEM)*	*Confidence interval at 95% (CI) (Mean ± 2SEM)*
2008	A	08.11	3.30	2.33	3.45-12.77
	B	09.49	3.57	2.42	4.65-14.33
2009	A	08.74	3.49	2.44	3.86-13.62
	B	09.68	3.77	2.38	4.92-14.44
2010	A	10.46	3.98	2.32	5.82-15.10
	B	08.68	3.05	2.40	3.88-13.48
2011	A	09.71	4.01	2.44	4.83-14.59
	B	08.02	3.40	2.33	3.36-12.68
2012	A	10.28	3.20	2.24	5.80-14.76
	B	07.26	3.39	2.42	2.42-12.10
Mean		**09.04**	**3.52**	**2.37**	