

Multivariate analysis of factors Influencing reliability of teacher made tests

Meshkani, Z., PhD¹; Hossein Abadie F., MSc. ²

¹Associate Professor, Department of Community Medicine, Director of EDO, College of Medicine ,Tehran University of Medical Sciences.

²Director of Educational Testing Services, College of Medicine, Tehran University of Medical Sciences.

ABSTRACT

Background: According to the measurements literature reliability of the test refers to the consistency of the test results and shows whether the obtained score is stable indication of the student's performance in particular test. Reliability can be measured by different statistics formula.

Purpose: To determine the factors influenced the reliability of 392 MCQ examinations.

Methods: The correlation of reliabilities of MCQ based examination and other characteristics of tests such as length, difficult items, discrimination index, mean, standard deviation and time for answering was calculated based on the data available on examination center of Tehran University of Medical Sciences. Multivariate regression has been used for data analysis.

Results: Overall reliability of teacher made test is at satisfactory level in most cases. The mean value of reliability was 0.71 ± 0.15 . In comparing previous semester with last series of examination some improvement have been found during these years ($P=0.000$, for first semester, $P=0.002$ for second, $P=0.005$ for third and $P=0.005$ for fourth semester). Keeping other variable fixed the interaction of length of exam according to item difficulty showed significant difference on value of test reliability. Comparing difficult and easy items question with moderate difficulty index can increase reliability 8 times more than difficult and 13 times more than easy items $P=0.000$.

Conclusion: Our study showed that with documentation of tests' metric features an analysis and evaluation of tests are within reach of medical school.

Key words: RELIABILITY, TEACHER MADE TEST, RELIABILITY MEASUREMENTS

Journal of Medical Education Winter 2005 6(2); 149-152

Introduction

High quality assessment system in medical education is an obvious need for medical schools. However there is no gold standard for high quality system, however some factors are addressed by educators and researcher in field of measurement such as reliability, validity, objectivity and feasibility of exam.

According to measurements literature, "reliability of instrument in student assessments concerns the extent to which the instrument yields the same results on repeated trials, or tendency toward consistency found in repeated measurements is referred to as reliability" (1,2)

Reliability computed by any indices reflects whether the obtained score is a stable indication of the student's performance on particular test. There are three major categories of reliability for most instruments: test-retest, equivalent form, and internal consistency. A fourth category is (scorer agreement) often used with performance and

product assessment (scorer agreement is consistency of rating a performance or product among different judges who are rating the performance or product). A number of statistics formulas can be used for different instruments, but each one has some potential disadvantages. Test-retest measures consistency from one time to the next exam by using correlation coefficient formula between two exams. Kuder Richardson formula 20 and 21 (K20 or K21) measures, the consistency of the item within the test and is equivalent to Cronbach's coefficient alpha when items are scored either right or wrong. In another word the items of the test should be dichotomously scored (0 for incorrect and one for correct) for all items of the test and items are compared with each other, rather than half of the items with the other half of the items. (3).

K-R 20 assumes that difficulty index of the questions are different (4,5,6). For research purposes, a minimum reliability of 0.70 is required. Some researchers feel that the value of reliability

should be higher. A reliability of 0.70 indicates 70% consistency in the scores that are produced by the instrument. Many tests, such as achievement tests, strive for 0.90 or higher reliabilities (7,8).

Literature on reliability estimation has some consideration on a number of reasons why the reliability estimate for a measure is low. Factors that cause error in measurement and results the low reliability of test items are listed below:

Item sampling (Longer tests can provide better reliability), length of the test,(reduce the chance of guessing) time limit for the test,(increase test anxiety and effect students performance and causes poor reliability),difficulty of test item, (difficult and easy items induce error and cause low reliability), student's awareness of how they will be assessed (causes better performance of students) scoring procedure, testing condition and test taker behaviour (effects students performance) test taker (perhaps the subject is having a bad day which causes poor performance), test itself (the questions on the instrument may be unclear induce error and cause low reliability) testing conditions (there may be distractions during the testing that detract the subject) test scoring (scores may be applying different standards when evaluating the subjects' responses)(9,10,11,12). Since there are many factors influencing reliability, this study attempts to examine the effects of each factor on reliability of teacher's made test and also to identify the trend of test reliability, since the analysis of test and it's items, implemented by the Testing Services of Medical School.

Materials and Methods

Data of this study extracted from item analysis sheets of 392 MCQ (Multiple Choice Question) form examinations in college of medicine, at Tehran University of Medical Sciences during years 2001-2004. The reliabilities of the tests which were calculated by Kuder Richardson formula (kR-20) were used as dependent variable. Length of test (number of test items), number of difficult item of the test (Difficulty index, i.e. The percents of items answered correctly), discrimination index (ability of the test to distinguish between subjects who really know and those who don't), mean ,standard deviation and time limit of the test were used as independent variables.

Data of the study analyzed by SPSS (statistical software), using multivariate regression analysis for determining the effect of each factor.

Results

Descriptive statistics of the finding shows among 392 multiple-choice exams (MCQ), 204 exam were provided by basic science groups and 188 by clinical science faculties .As table one shows the mean value of reliability of test in basic science were lower than clinical science (0.7 for basic and 0.72 for clinical). The lowest value for reliability among whole data was 0.34 and highest value was 0.92 (figure1).

TABLE 1. Reliability value according to student level of study

Student level of Study	Number of Exam	Mean value Of reliability	SD.
Basic	204	0.70	0.14
Clinical	188	0.72	0.16
Total	392	0.71	0.15

The results also indicate that there was some improvement of test reliability in comparison with 5 previous semesters (table2).

When comparing last semester (second half year 2003-2004) with other previous semesters, multivariate statistical test indicates a significant differences between last and other semesters $P=0.000$ For first, $P=0.002$ for second, $P=0.005$ for third , $P=0.005$ for fourth semester Table 2).

Keeping other variable fixed the interaction of length and difficulty-indexes of test according to item difficulty shows that items with moderate difficulty index increase reliability 8 times more than difficult and 13 times more than easy items (regression coefficient for moderate difficulty 6.4, for difficult items 0.83 and for easy items 0.43), $P=0.000$ for all level of difficulty indexes, (table 3).

Items with negative value of discrimination index have negative effects on reliability value of -0.007

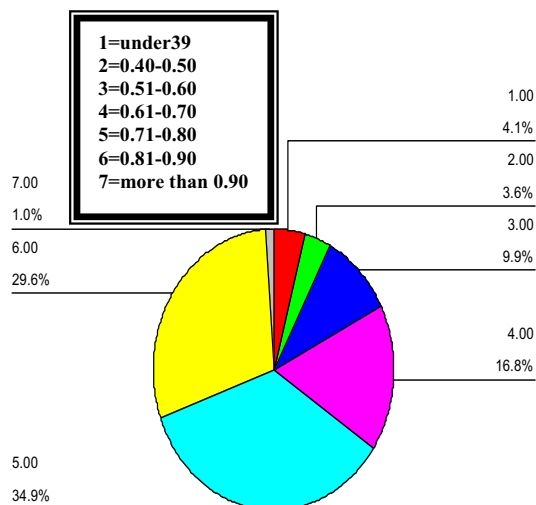


FIGURE 1. Distribution of reliability value of exams

Table 2: mean and standard deviation of reliability value according to Semester

Academic Semesters	Number Of exam	Mean of reliability	Standard Divination	Level of significance p
2001-2002-2	65	0.68	0.16	0.000
2002-2003-1	75	0.68	0.15	0.002
2002-2003-2	67	0.69	0.19	0.005
2003-2004-1	61	0.71	0.13	0.005
2003-2004-2	114	0.75	0.10	-

TABLE 3. Regression coefficient item analysis of exams

Parameter (factors)	Regression coefficient	Level significant(p)
Difficult items	0.83	0.000
Easy items	0.43	0.000
Moderate difficult	0.65	0.000
Negative Dis.-index.	- 0.007	0.005

(P=0.005) which means each items with negative value of discrimination indexes reduce 0.007 value of reliability.

Other parameters such as time allocated to exams and level of students' education (basic and clinical

level) have been omitted from the model and no statistical significant differences were found.

Discussion

For practical purposes a reliability of 0.70 may not be enough where the important decisions about the fate of individuals is made on the basis of a test score, the reliability of test should be at least 0.90 preferably 0.95 or higher ." (12). Although there has been some improvement in assessment system more consideration should be paid to this issue. Other findings of this study highly support the recommendations of the measurement literature that" test with difficult and easy items and low discrimination, influences the value of reliability. Length of examination however affects the reliability, but this study showed without the considering the quality of test items, increasing the number of questions in order to increase reliability is a big mistake (13,14,15,16).

The finding of this research has implication for Tehran University of Medical Sciences for improving the assessment system of the medical school

References

- 1- Tuckman, BW. Testing for teachers. (2nd Ed.). Sydney: Harcourt Brace Publishers; 1988.
- 2- Gay, L. Educational Evaluation and Measurement 2nd Ed. New York: Macmillan Publishing.
- 3- Linn RL, Gronlund NE. Measurement and Assessing in Teaching 7th Ed. New Jersey: Prenticehall Inc; 1995.
- 4- Roid HG, Haladyna MT. A Technology for Test Item writing. New York: Academic press; 1982.
- 5- Marso RN, Pigge FL.(1991). An analysis of teacher made tests and item construction errors. J Contemp Edu Psych 1991; 16:279-86.
- 6- Marso RN, Pigge FL. A summary of published research : Classroom teacher's knowledge and skills related to the development and use of teacher made tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco (ERIC Document Reproduction Service No. ED 346-148).
- 7- Layton JM. Validity and reliability of teacher made test. J Nurs Staff Develop 1989 ;2,(3):105-9.

- 8- Anderson TW. An introduction to multivariate statistical analysis, 3rd Ed, New Jersey: John Wiley & Son Inc;2003.pp: 126.
- 9- Osterlind S.J. 1983. Test item bias, quantitative application in the social science, Sage Publication Inc, California.
- 10- Nunnally, J. C. 1978. Psychometric theory (2nd. Ed.). New York: MacGrow- Hill.
- 13- Norman G, Swanson D, Case S.1997. Conceptual and Methodology Issues in Studies Comparing Assessment Formats, issues in comparing item formats. Teaching and learning in medicine;8(4):208-16.
- 14 -Swanson DB, Norcini JJ, Grosso LJ.1987 Assessment of clinical competence: written and computer based simulations. Assessment and Evaluation in Higher Education 1987;12(3):220 - 46.
- 15- Swanson, D.B. A measurement framework for performance-based tests. In: Hart, I and Harden, (Eds.) Further Developments in Assessing Clinical Competence. 1987, Can-Heal publications, Montreal,Canada.
- 16- Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. Medical Education 1983;17:165-71.