Medical education

Reliability of physiology MCQs' examinations at the Faculty of Medicine, University of Khartoum

*Afraa M.M. Musa

Background:

One of the major challenges that face exam constructors is generating highly reliable exams. An assessment cannot be viewed as valid unless it is reliable. Test reliability, which is the best single measure of test accuracy, is the extent to which test results are consistent, stable, reproducible and free of error variance. Reliability is influenced by internal factors related to exam construction, as well as external factors which depend on the situation of test administration.

Objective:

To estimate the reliability of multiple choice questions (MCQs) of physiology exams as part of an overall quality assessment at the Faculty of Medicine, University of Khartoum.

Methods:

Reliability influential factors related to exam construction and administration were controlled and catered for by departmental and administrative staff according to the exam regulations of the faculty. Remark software was used for post-examination analysis of scores of ten consecutive summative physiology MCQ exams at the Faculty of Medicine, University of Khartoum. The number of the examinees who sat for each of the ten exams ranged from 332–359. In addition to reliability coefficients, item difficulty index (DIF I) and point-biserial correlation coefficient **(r_{pbis})** as a measure of item discrimination ability, were calculated as part of item analysis results.

Results:

The study revealed high exams' reliability (0.84-0.95) as measured by different formulas [Kuder-Richardson Formulas (KR-20, KR-21) and Cronbach's-Alpha], and low standard error of measurement/SEM (3.07-3.80). Factors which contributed to the high reliability of our ten exams were: their high discrimination power (0.32-0.47), their recommended mean difficulty (48.62-65.67%), and the relatively large numbers of items (60–80) per each exam.

Conclusion:

The high exams reliability of this study was an indicator of the precise control of external and internal factors influencing reliability. The most important contributing factor was the proper construction of exams with high quality items; in addition to careful exam administration and meticulous scoring system.

***Corresponding author:** *Department of Physiology, Faculty of Medicine, University of Khartoum, Sudan*
Email: aframusa@gmail.com

Introduction:

Multiple choice questions (MCQs) are used extensively to assess knowledge capabilities of medical students and they often account for a substantial portion of their course grades. It is an efficient form of the written assessment as it can have broad coverage of content in a relatively short time and can be graded by computers [1, 2]. These factors help in standardizing exam administration to large numbers of trainees[3].

There are two main factors that cause errors in measurements: external and internal factors [4]. One of the key goals of assessment in medical education is the minimization of all errors influencing a test in order to produce an observed score which approaches a learner's 'true' score, as reliably and validly as possible. In order to achieve this, assessors need to be aware of the potential biases that can influence all components of the assessment cycle from question creation to the interpretation of exam scores[5].

Item analysis uses statistics and expert judgment to evaluate tests based on the quality of individual items and entire sets of items, as well as the relationship of each item to other items. It is a valuable integral component of course assessment performed after the examination to provide information regarding reliability and validity of a test by calculating many exam quality indicators. It gives some idea of how well the examination has performed relative to its purposes and, thus, how future learning can be supported and directed [1, 6, 7]. Remark software provides detailed statistical analysis of students' scores, exam reliability measurements, item difficulty index (DIF I), point-biserial correlation (as a measure of item discrimination) and a detailed distracter analysis [8].

Test Reliability, which is the best single measure of test accuracy, is the extent to which test results are consistent, stable, and free of error variance. It is also defined as the extent to which a test provides the same ranking of examinees when it is re-administered[5].There are numerous indices that may be used to assess the internal consistency of an assessment. The most widely used measure is Cronbach's Alpha (or Coefficient Alpha) which was first named as alpha by Lee Cronbach in 1951[9], best measures surveys or attitude data[8, 10]. The Kuder Richardson-20 (KR-20) developed by Kuder and Richardson in 1937 [11], is a special case of Cronbach's alpha, and is the best indicator of how well the exam measures a single cognitive factor (subject)[8]. As KR-20 is specifically used for ordinal dichotomies or binary variables (e.g. items scored as right or wrong), it is considered the most appropriate index of test reliability for multiple-choice examinations [8, 12].

Intrinsic to the validity of any assessment is analysis of the scores to quantify their reproducibility. An assessment cannot be viewed as valid unless it is reliable [12, 13]. The 'utility index' described by Cees van der Vleuten [14] serves as an excellent framework for assessment design and evaluation. It describes five criteria for determining the usefulness of a particular method of assessment: reliability, validity, impact on future learning and practice, acceptability to learners and faculty, and costs [13]. Face-content and construct validity are usually ensured by expert staff who reviews the exam before administration. Furthermore, looking at the results of examinees performance using item analysis programs will assist departments and faculties to make judgments about validity and reliability of their assessment tools and improve the quality of their assessment programs.

Estimation of reliability as part of quality management of an assessment program is of central importance to ensure stakeholders that doctors' competencies would reach the same conclusions if it were possible to administer the same test again on the same doctor in the same circumstances[13]. So, medical educators are required to construct valid and reliable tests because reliability and validity are both needed to assure adequate measurement of the constructs intended to be measured[5].

The aim of this study was to estimate the reliability of physiology MCQs exams at the Faculty of Medicine, University of Khartoum as one of the indicators of its validity as an assessment tool.

Methods:

This study was descriptive and cross-sectional. It was carried-out on test statistics reports produced by Remark software which provides analyses of the students' responses in the form of Excel file formats. Ten physiology MCQ exams held from September 2015-September 2016 at the Faculty of Medicine, University of Khartoum were included in the study.

Each exam paper consisted of 60 - 80 (five-option) items. The number of the examinees in each exam (i.e. answer sheets analyzed/exam) ranged from 332 to 359.

To construct valid and highly reliable exams, physiology department staff followed strictly the faculty examination regulations to minimize as much as possible occurrence of external and internal errors that affect exam reliability. All staff members who participated in teaching contributed in exam construction. They constructed their items guided by proper blueprinting. The collected items were then reviewed meticulously by the departmental examination committee which was formed from senior staff members. The test papers were typed and formatted well so that the candidates could see the papers clearly. Papers were printed in the Academic Office and stored securely in the departmental examination office. The faculty Academic Office established comfortable safe and quiet environment for exam administration. This was achieved by preparing exam halls with properly spaced seating, cooling and light. Moreover, the Academic Office staff was responsible for checking examinee attendance through signing in sheets using exam numbers, confirming examinees identities, adjusting timing, and securing paper collection. In addition, faculty assigned medical staff to look into examinees' medical complaints, psychological worries, excitement and accidents.

The exam instructions in addition to being written in the front page of the paper, were made clear and understandable by the exam invigilators. Enough time (2 minutes for each item) was provided for the examinees to solve the questions. The teams of invigilators ensured that the exams were carried-out on time; answer examinee queries; and guard against plagiarism and cheating. To decrease errors in scoring, Remark software was used to analyze the students answer sheets and provide detailed statistical analysis of students' scores. In addition students answer sheets were printed clearly and students were directed to mark the exam number and the chosen answer properly so that the scanner could easily identify them.

Reliability measurement of each exam was calculated by the software and was expressed in three formulas: KR-20, KR-21 and Coefficient Alpha. KR-20 formula is calculated using the number of test items on the exam variance (standard deviation squared) of student performance on every test item and total test score [12]. KR-21 is a simpler formula and easier to compute, and is derived from the KR-20 formula. It differs in that it assumes that all test items have identical difficulty index and produces lower estimates than the KR-20 formula[15]. The correct answer was given one mark, while no negative mark was given to the wrong answer.

Beside reliability measurements, Remark software provided item analysis data including item difficulty index (DIF I), item discrimination using point-biserial correlation coefficient (r_{pbis}) and detailed distracter analysis. Data analysis was carried-out using the Statistical Package for the Social Sciences (IBM SPSS statistics) program. *P* value <0.05 was considered as significant.

This study was approved by the Research Ethics Committee of the Faculty of Medicine, University of Khartoum. As the software provided anonymous data [i.e. deals with exam numbers rather than examinees' names], there was no need to obtain informed consents from the examinees.

Interpretation of reliability indices:

A perfectly reproducible test would have a coefficient of 1.0; that is 100% of the trainees would achieve the same rank order on retesting. Reliability indices range from 0.00 to 1.00; values near 0.00 indicate measuring many unknown factors, but not what is intended to measure; while values near 1.00 indicate measuring a single factor. The desirable level of reliability is dependent on the type of examination being considered; for a multiple choice test, an internal consistency measure of over 0.90 is achievable and desirable[16]. Reliabilities as low as 0.50 are satisfactory for short tests of 10 to 15 items, but tests with more than 50 items should have reliabilities of 0.80 or higher. Traditionally, a reliability coefficient of greater than 0.8 has been

considered as an appropriate cut-off for high stakes assessments. The higher the reliability, the lower will be the amount of error variance in the test (the standard error of measurement-SEM). SEM is inversely related to the Reliability Coefficient[17]. The higher the reliability, the better the performance of the test as a whole and consequently the items within it [4, 16].

Results:

Analysis of the performance scores of the individual ten exams (*Table 1)* revealed the number of the examinees ranged from 332 to 359, while the number of the items ranged from 60 in mid-semester to 80 in end-of-semester exams. Individual exams showed a range of mean difficulty index from 48.62 to 65.67% and mean discrimination index (point-biserial correlation coefficient) from 0.32 to 0.47.

The analysis of difficulty and discrimination level of each of the ten exams revealed that the majority of their items were of average difficulty (i.e. DIF I=30-70%), followed by the easy (i.e. DIF I>70%), and the least were the difficult (i.e. DIF I<30%) items. In addition, these exams revealed high discrimination ability resulting from presence of a very high proportion (almost 90%) of discriminating items (i.e. r_{pbis}=>0.2), and a very minimal proportion of poor and negative discriminating items (r_{pbis}< 0.2, r_{pbis}< 0.0, respectively).

The study revealed high reliability coefficients of these exams ranging from 0.84 to 0.95 as measured by different formulas (KR-20, KR-21 and Coefficient Alpha). In addition, it showed low standard error of measurement (SEM) ranging from 3.07 to 3.80. KR-21 produced lower values compared to KR-20 and Coefficient Alpha which were almost identical (*Table 2)*. Our study showed highly significant (P<0.002), strong positive correlation (r=0.842) between exam reliability and mean discrimination index (r_{pbis}) of the ten exams.

Table 1. Exam characteristics and item analysis

Exam	Examinees Number	Items umber	Mean DIF I	Mean discrimination coefficient(r_{pbis})
Mid S2 Sep 2015	342	60	53.63	0.33
Mid S3 Sep 2015	347	60	54.99	0.34
End S2 Nov 2015	342	70	59.35	0.36
End S3 Nov 2015	346	80	50.83	0.32
Mid S3 Feb 2016	339	60	48.62	0.36
Mid S4 Feb 2016	350	60	65.67	0.40
End S3 April 2016	336	60	53.63	0.33
End S4 April 2016	353	80	62.15	0.47
Mid S2 Sep 2016	332	60	58.44	0.37
Mid S4 Sep 2016	359	60	53.80	0.34

This table summarizes the characteristics of the analyzed exams (number of examinees and number of items) and item analysis of individual exams (mean difficulty index& mean discrimination coefficient of total items).

Table 2. Reliability measurements of the ten exams

Exam	SEM	KR-20	KR-21	Coefficient Alpha
Mid S2 Sep 2015	3.41	.86	.84	.86
Mid S3 Sep 2015	3.35	.87	.85	.87
End S2 Nov 2015	3.57	.90	.88	.90
End S3 Nov 2015	3.80	.89	.86	.89
Mid S3 Feb 2016	3.32	.89	.86	.89
Mid S4 Feb 2016	3.07	.90	.88	.90
End S3 April 2016	3.35	.90	.89	.90
End S4 April 2016	3.61	.95	.94	.95
Mid S2 Sep 2016	3.30	.90	.88	.90
Mid S4 Sep 2016	3.35	.87	.84	.87

This table shows standard error of measurement (SEM) & reliability coefficients of the analyzed exams as measured by different formulas (Kuder-Richardson Formula 20, Kuder-Richardson Formula 21 and Coefficient /Cronbach Alpha).

Discussion:

The quality of the test as a whole is assessed by estimating its "internal consistency." Measuring exam reliability is essential to judge its validity; therefore, it is considered one of the indicators of quality of an assessment tool. Various factors were found to affect reliability. Some are external factors which depend on the test situations and administration, such as the room temperature, guessing answers, emotional problems, physical discomfort and lack of sleep. The others are internal factors which depend on the quality and quantity of the test, such as item sampling and the way in which the item is constructed. Scorers and scoring systems can also be a potential source of error [(5)].

Compared to some published research, our reliability measurements (0.84 to 0.95) were among the highest reported figures. On evaluation of the psychometric performance of an obstetrics and gynaecology exam performed in Mu'tah University, reliability was estimated as (0.947) using the Cronbach alpha test and only (0.599) using KR-20. This was attributed to the inclusion of 23% of items having negative point biserial. It was concluded that KR20 reliability could be substantially improved if the negative point biserial question items were removed[(18)]. Although it is generally recommended that classroom-type assessments have a reliability of at least 0.70[(19)], the Canadian study of 16 tests revealed that more than half of the tests had values of Cronbach's alpha that fell below this level. This low reliability was explained by the presence of negatively discriminating and flawed items in these tests. On the other hand, some of their tests were quite good; three of them had mean discrimination coefficients of at least 0.30 and adjusted alpha values of 0.80 or higher [(2)]. In another study, analysis of ten pharmacology summative tests revealed low mean reliability coefficient (0.54). Two tests showed low reliability (0.60-0.70), five tests very low reliability (0.50-0.60), and three tests questionable reliability (=<0.50). They admitted that their exams need to be supplemented by other measure and some items need to be improved for assessment to be reliable[(17)]. Finally, evaluation of 100 MCQs of the four options type of the final exam in internal medicine at the College of Medicine in King Khalid University, reported a KR- 20 value of (0.79) which was considered by the authors as good reliability and the student scores were believed to be reliable[(20)].

The various factors which contributed in the synthesis of high exams reliability in this study will be discussed in the following section.

In the physiology department, the examination committee used to review meticulously exam construction so as to overcome the internal factors related to test development. The process starts by drafting, checking, and then subjecting items to critical scrutiny in order to identify problems before test administration [21]. Exam constructors' expertise was important to ensure validity. Using proper assessment blueprinting, alignment of assessment methods with the intended learning outcomes (ILOs) and teaching & learning activities (TLA) was achieved. Blueprinting ensured content validity through fair and balanced selection of items which accurately covered the domains intended to be measured.

In this study, inclusion of more than 50 items in each of the ten exams (60-80 items) which covered large curricular contents played a role in increasing the reliability. The process of selecting a representative fraction of a total pool of items is referred to as item sampling. The more representative the test items to curricular contents, the more reliable will be the test. In general as the number of test items increases, sampling error will decrease and, hence, reliability will increase up to a limit. In multiple choice tests where there is the possibility of guessing, increasing the number of items will reduce errors associated with guessing [5]. On the other hand, huge number of items may cause tiredness and carelessness resulting in decreased reliability[21]. The insufficiency of time is known to decrease the reliability of the test. In this study, the sufficient time which was provided for the examinees to properly respond to each item, was one of the factors that played a role in production of high exam reliability.

In general, item statistics will be somewhat unstable for small groups of students and fifty students are considered as the minimum number required for stability. Increasing the students number to one hundred or more was noticed to improve stability of item analysis results [22]. Hence, the large number of our examinees in each exam shared in the stable performance of our items.

Providing explicitly clear exam instructions enhanced examinees understanding of what was requested exactly and assisted them in writing the answers clearly. Therefore, it is recommended by medical educators to increase exam reliability[21]. In addition, having material which was homogeneous, and known to candidates added more to the exam reliability[21,23]. Furthermore, using the digital scoring machines enabled fast, accurate, fair and objective scoring system, and detailed statistical analysis of test scores. It also minimized inter-rater variability. Identification of the candidates by numbers rather than names reduced bias and subjectivity in scoring as well, all of which contributed to the consistently high reliability.

Furthermore, faculty Academic Office together with departmental staff exerted lots of efforts to minimize external conditions that induce bias during exam administration (e.g. examinees misunderstanding or misreading test directions, cooling of exam's hall, noise level, distractions, examinees sickness, worry, excitement, accidents during examination, cheating and plagiarism,…etc.). This was achieved by establishment of well-prepared exam halls, provision of invigilator staff and medical staff.

The quality of individual items is assessed by discrimination ability which is measured by comparing students' item responses to their total test scores (point-biserial correlation coefficient)[24]. The major influential factor on reliability of test scores was the high discriminatory power of our exam items. The strongly positive significant correlation finding between mean exam discrimination ability and exam reliability was consistent with a previous study finding[2]. Hence, reliability of a test depends primarily on the discriminatory power of the test items that comprise it [24, 25]. This necessitates the construction of high discriminating ability items[2].

The reliability of a test paper is decreased if the test items are either very difficult or very easy. Medical educators recommend for optimum test reliability

to minimize the variability of item difficulty and to make the level of item difficulty somewhat easier than the halfway point between a chance percentage of correct answer (i.e. 20% in five option items) & 100% correct answers[6, 24, 26]. In the current study, the majority of items of each exam were of average difficulty while minor proportions were very easy and very difficult. This level of difficulty contributes substantially to the high reliability reported in this study. This finding corresponded well with what has been observed by other researchers that the sets of medium or average "difficulty" items are more reliable than the very easy and very difficult items, since the variability among the values of very difficult or very easy items is low[20, 27].Variation in difficulty is also related to group heterogeneity, as it is known that the more heterogeneous the group of the examinees, the higher the internal consistency of the exam[5].

Downing, in his book of test development[28], proposed a convenient organizational framework for collecting and reporting all sources of validity evidence of a testing program. Following these steps of effective test development, would maximize exam validity for the intended test score interpretation. These include twelve steps starting from test planning and construction through test administration and scoring to item analysis and banking[28].

Abiding by the rules and regulations of the assessment program in the faculty is the best way to create highly reliable and valid exams in addition to keeping the process of examinations to a high standard. Post-exam analysis of students' performance would facilitate establishment of proper examination bank that would definitely contribute to future exam reliability.

Conclusion:

The high reliability observed in this study was the outcome of precise control of internal and external factors that influence reliability measures. The most important contributing factors were the generation of well-constructed items, careful exam administration and meticulous scoring system. Institutions have a responsibility to raise the awareness of staff about all the sources of potential errors of exam reliability, provide item analysis reports following every test administration and train them to utilize these valuable data in improving the quality of their assessment tools.

Acknowledgements:

The author would like to thank the academic and administrative staff of the Physiology Department, the Educational Development Center and Dean's Office at the Faculty of Medicine, University of Khartoum, for their valued support and permission to access examination procedures and department scores of ten physiology exams.

References:

1. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach*. 2005;26:709-12.
2. DiBattista D, Kurzawa L. Examination of the Quality of Multiple-Choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning*. 2011; 2:Art 4.
3. Epstein RM. Assessment in medical education. *New England Journal of Medicine*. 2007;356:387-96.
4. Anastasi A, Urbina S. Psychological tests. 7th ed. NJ: Prentice Hall: Upper Saddle River; 1997.
5. Tavakol M, Dennick R. Post-examination analysis of objective tests. *Med Teach*. 2011;33:447-58.
6. Matlock-Hetzel S. Basic concepts in item and test analysis. The Annual Meeting of the Southwest Educational Research Association; Austin, Texas: *ERIC Database*; 1997.
7. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners 1998.
8. Tucker S. Using remark statistics for test

reliability and item analysis. Baltimore: University of Maryland; 2014; Available from: tp://wwhtw.umaryland.edu/media/umb/cits/umbtestscoring_testanditemanalysis.pdf.

9. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334.
10. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*. 2003;80:99-103.
11. Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika*. 1937;2:151-60.
12. Hale CD, Astolfi D. Psychometrics: Reliability & Validity Measuring learning and performance: A primer. Florida, Saint Leo University: Charles Dennis Hale; 2011. p. 45-70.
13. Joshi H. An assessment system based on principles. Developing and maintaining an assessment system-a PMETB guide to good practice. London: Postgraduate Medical Educationand Training Board (PMETB); 2007.
14. Vleuten VD, PM C. The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*. 1996;1:41-67.
15. Lenke JM. Differences Between Kuder-Richardson Formula 20 and Formula 21 Reliability Coefficients for Short Tests with Different Item Variabilities. 1977; Available from: https://eric.ed.gov/?id=ED141411.
16. McAlpine M. A summary of methods of item analysis. University of Luton: CAA Centre; 2002.
17. Vegada BN, Karelia BN, Pillai A. Reliability of four-response type multiple choice questions of pharmacology summative tests of II MBBS students. *International Journal of Mathematics and Statistics Invention* (IJMSI). 2014;2.
18. El-Uri FI, Malas N. Analysis of use of a single best answer format in an undergraduate medical examination. *Qatar medical journal*. 2013;1:3-6.
19. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ*. 2004;38:1006-12.
20. Elfaki OA, Bahamdan KA, Al-Humayed S. Evaluating the quality of multiple-choice questions used for final exams at the Department of Internal Medicine, College of Medicine, King Khalid University. *Sudan Medical Monitor*. 2015;10:123-7.
21. Zhu J, Han L. Analysis of the main factors affecting the reliability of test papers. *Journal of Language Teaching and Research*. 2011;2:236-8.
22. Item Analysis. Michigan State University: Scoring Office Available from: https://www.msu.edu/dept/soweb/itanhand.html.
23. Symonds PM. Factors influencing test reliability. *Journal of Educational Psychology*. 1928;19:73.
24. Understanding Item Analyses. University of Washington | Seattle, WA: Office of Educational Assessment; 2018; Available from: http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/.
25. Ebel RL. The relation of item discrimination to test reliability. *Journal of educational measurement*. 1967;4:125-8.
26. Lord FM. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*. 1952;17:181-94.
27. Myers CT. The relationship between item difficulty and test validity and reliability. ETS Research Report Series. 1955.
28. Downing SM, Haladyna TM. Twelve Steps for Effective Test Development. Handbook of Test Development. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2006. p. 3-25.