QMJ
A peer reviewed journal
QATAR MEDICAL JOURNAL

OPEN ACCESS

RESEARCH PAPER

Analysis of use of a single best answer format in an undergraduate medical examination

Fahmi Ishaq El-Uri, Naser Malas

Address for Correspondence:
Fahmi Ishaq El-Uri
Karak Government Hospital, Mu'tah University, Mutah, Karak, Jordan
Email: fahmiuri@yahoo.com

http://dx.doi.org/10.5339/qmj.2013.1
Submitted: 15 May 2013
Accepted: 29 May 2013

Cite this article as: El-Uri FI, Malas N. Analysis of use of a single best answer format in an undergraduate medical examination, Qatar Medical Journal 2013:1 http://dx.doi.org/10.5339/qmj.2013.1

QSCIENCE.com
Member of Qatar Foundation

ABSTRACT

Examinations at the Faculty of Medicine of Mu'tah University are based on a single best answer multiple-choice questions (MCQs) format. However, the reliability of this examination format has not been determined.

Objective: Using an examination of obstetrics and gynaecology as a model, this study aims to analyze the difficulty (facility) index, the discriminatory power and reliability of the examination format.

Materials, Subjects and Methods: A prospective study on the psychometric performance was carried out for an undergraduate examination in obstetrics and gynaecology. The performance of the items was measured in terms of facility, discrimination and reliability. Two statistical tests were used to estimate the reliability of the exam: The Kuder-Richardson Formula 20 (KR-20) and the Cronbach alpha test.

Results: The items scored well in facility with a significant portion (26%) achieving a positive point biserial of over 0.3 and a Cronbach alpha score of 0.947. However, 23% of the items had a negative point biserial and the Kuder-Richardson 20 (KR20) score was only 0.599.

Conclusion: In order to improve the reliability of examinations, we recommend removing the items with negative point biserials and increasing the total number of items.

Keywords: psychometric analysis, Mu'tah University, Obstetrics and Gynaecology, examination

INTRODUCTION

The Faculty of Medicine in the University of Mu'tah, the Hashemite Kingdom of Jordan, was established in 2001. The students undergo a six-year course in medical sciences, including obstetrics and gynaecology during the last three years of clinical teaching. The school has

been using exams that are based on a single best answer of five options in order to test their students. The reasons for this involve the recognised superiority of the format in terms of its ability to probe understanding without sacrificing reliability.[1] In 2006, the faculty implemented a computer-based assessment of paper examinations. The method generates important data including difficulty (facility) index and discrimination index. The facility of a test is a measure of the number of correct responses to each item. It allows determination of how 'hard' or 'easy' the question is. The facility of a question is the most basic expression of candidate performance on a question. Assuming a normal distribution of candidate intelligence, 95% of candidates' total scores for any one exam would fall within two standard deviations of the mean score for that exam. Therefore, a question that produced a facility score outside of this range, i.e., can be answered by less than 5% or more than 95% of candidates, could be deemed 'too hard' or 'too easy', respectively.

Discrimination is measured on a per question basis using the point biserial, also known as the Pearson product-moment correlation coefficient. The point biserial figure shows the correlation between candidates' overall exam scores and an individual question score. The values range between −1 (a negative correlation) and 1 (a positive correlation). Questions can be considered excellent discriminators if they have a score ≥0.40. Good discriminators score in the range of 0.30–0.40, mid-range discriminators within a 0.10 to 0.30, and modest discriminators in the range of 0.001 to 0.099 point biserial. If questions have a score of 0.00, then they are considered non-discriminators.

However, the reliability of examinations can as well be estimated. Two statistical tests were used to estimate the reliability of the exam in this study: The Kuder-Richardson Formula 20 (KR20) and the Cronbach alpha test. The KR20 formula is a measure of internal consistency for examinations with dichotomous choices. It produces a correlation measure between 0 where a high KR20 coefficient (e.g., >0.90) is indicative of a homogeneous test. Usually, a KR20 figure of 0.8 is considered the minimal acceptable value. A figure below 0.8 could indicate that the exam was not reliable. The KR20 is influenced by difficulty, spread in scores, and length of the examination. On the other hand, the Cronbach test is commonly used as a measure of the internal consistency or reliability of a test score. It can be used for non-dichotomous (continuous) measures. Cronbach's alpha will generally increase as the inter-correlations among test items increase. Alpha can take values between negative infinity and one. Usually, a reliability of 0.70 or higher is required for the use of an examination.

Due to the young age of the Faculty of Medicine at Mu'tah University and the lack of data on the appropriateness of this examination system in Jordan and the type of questions, an assessment of the performance and reliability of the examination format in the context of Jordan is essential. An examination of obstetrics and gynaecology was used as a model.

SUBJECTS AND METHODS

Fifty-six medical students attempted an examination in obstetrics and gynaecology as part of their sixth-year assessments. The exam consisted of 100 items in single best answer format. There were five choices per stem, and the students were asked to select the single most correct answer. The items were based exactly on the lectures the students were given and were written by three specialists. The responses were marked and counted by an optical reader (Digital Scanner AX1011-AXM980) and analyzed by HODA Tool Reader 2.0 program (Axiome Alpha SA, Peseux, Switzerland).

The results were analysed in terms of facility, discrimination and reliability. The reliability score was based on both Cronbach alpha and Kuder-Richardson 20 (KR20).

RESULTS

Facility

The facility of a test is a measure of the number of correct responses to each item. It allows determination of how 'hard' or 'easy' the question is. A facility score of 0% would be show the 'hardest' question, with no correct answers, and a score of 100% would indicate the 'easiest' question. As illustrated in Table 1, the results

Table 1. Facility test of the exam.

Facility (%)	Number of questions
100	1
90–99.9	11
80–89	10
70–79	11
60–69	16
50–59	13
40–49	6
30–39	13
20–29	19
10–19	1
0.01–9	0
0	0

showed an almost equal distribution of test facility between 20 – 99.9%. One question was answered by less than 20% of the students, and another question was answered correctly by all students. None of the questions were left unanswered by the students.

Discrimination

Discrimination is measured on a per-question basis using the point biserial, also known as the Pearson product-moment correlation coefficient. The point biserial figure shows the extent to which two sets of data (in this case, candidates' overall exam scores and individual question score) correlate. The range of values produced by the point biserial is between −1 and 1, depending on whether the data have a negative correlation, no or modest correlation, or a positive correlation.

Nine items achieved point biserial figures of over 0.4. Seventeen items achieved a point biserial between 0.3 and 0.39. Therefore, 26 items achieved a point biserial of 0.3 or over. Thirty-six items achieved a point biserial between 0.1 and 0.29. Thirteen items achieved a point biserial between 0.001 and 0.099. Two items had a point biserial score of zero. Collectively, 77 questions were zero and above and, hence, negative point biserials were obtained by 23 items.

Reliability

The Cronbach alpha score was 0.947, the standard measurement of error was 1.529 and the KR20 was 0.599.

DISCUSSION

One of the easiest psychometric measures to understand and rate is facility, which is merely a count of correct and incorrect answer frequency. Questions that are 'too hard' or 'too easy' should be flagged for review. Overall, there is a good spread of difficulty, with admirably even coverage, in the analyzed exam. Both features are highly appropriate and useful for testing candidates. A single question was determined to be 'too easy' with a facility of 100% indicating all candidates answered correctly. Nonetheless, if the question covered a core topic that should be known, this can be defended as appropriate. No questions were deemed 'too hard', although this is difficult to rate for this question format due to guessing factor.

There are a good number of highly discriminatory questions in this examination. Nine questions (9% of the exam) achieved excellent point biserial figures of over 0.40. In addition, seventeen questions were very good discriminators achieving point biserial of over 0.30 but less than 0.40. These results are for a cause for optimism with a total 26% of the examination questions performing an excellent discriminatory job. Mid-range discriminatory questions also form a solid core with 36 questions being well discriminatory (0.10 to 0.30 point biserial). Overall, a total of 62 questions (62% of the examination) had a point biserial 0.10 or above. They reflect the capability of the-best-of-five format to create more sophisticated questions than the true/false format. As a result, better discrimination can be expected, as the guessing factor is reduced significantly with five choices, compared to the 50 – 50 chance of the true/false format. While not impossible, point biserials of over 0.40 are very rare in the true/false MCQ format. In addition, while 5 – 7% of true/false MCQ format will reach the level of very good discriminatory performance, it is still below the level of the-best-of-five format.

Although there is a cause for some satisfaction at the overall discriminatory performance of this examination, there are some important caveats. Too many questions discriminate modestly, not at all, or negatively. Thirteen (13%) questions were modestly discriminatory (0.001 to 0.099), and two questions did not discriminate at all (0.000-point biserial). Most worryingly of all, 23 questions (23%) had a negative point biserial. This means that candidates who overall scored well on the examination scored worse on these questions than those who scored poorly on the whole examination.

The KR20 formula was derived by two statisticians, Kuder and Richardson,[2] with the aim to produce a formula that evaluated the reliability of a test compared with all other possible tests. The most basic way to do this is to split the test into two halves, and compare candidates' performance between the two. The correlation between the two sets of results will show how reproducible the test is, i.e., the higher the correlation, the more reliable the test. It is not, however, sufficient to compare data simply from one pair of split halves. For example, if the item performance from a test containing 100 items was split into halves of items 1 – 50 and 51 – 100, this may not produce a reasonable reliability score. For instance, if the first 50 items were based on one topic and the second 50 items on another topic, a correlation between scores in these two halves would not necessarily be useful.

A more sophisticated way to determine reproducibility, therefore, would be to split the questions alternately into two halves, i.e. the first half would contain odd-numbered questions (1, 3, 5 etc.) and the second half would be composed of even-numbered questions (2, 4, 6 etc.). This would generate a more useful correlation score, as it would not be prejudiced against either topic. Kuder and Richardson,[2] however, decided that the best test of reproducibility would be to find the correlation between the scores from every possible pair of split halves available from any given test. The KR20 formula

produces the mean of all these correlations in order to produce a reliability coefficient for the whole test. The KR20 formula produces a correlation measure, which will therefore be a number between 0 and 1. A KR20 figure of 0.8 is considered the minimal acceptable value. A figure below 0.8 could indicate a variety of occurrences, namely that the paper was particularly difficult, or perhaps that it tested unknown or unexpected topics. In general, a larger number of items will produce a more reliable test.

As the KR20 evaluates the reliability of a test compared with all other possible tests it is considered superior to the older Cronbach alpha test, which estimates reliability by averaging point biserials. Unfortunately, for this examination, due to its inclusion of a number of negative point biserial question items, whereas the Cronbach alpha test is good with a score of 0.947, the KR20 scored relatively poorly at 0.599. As the standard measurement of error uses the Cronbach alpha, this is admirably low at 1.529. This suggests that despite the relatively low KR20 reliability, decision-making can be based on these examination results with good confidence—the confidence interval being 3.06% or result plus or minus 1.53%. The KR20 reliability figure could be substantially improved if the negative point biserial question items were removed.

We recommend that the examination is developed further by removing items with a negative point biserial and or increasing the number of items. It is a rule of thumb in psychometrics that the number of items needs to be quadrupled if reliability is to be doubled.[3] However, efficiency in test time is also an important consideration. In this test, we would recommend that the number of items be increased by 50% to 150 (with the time to answer increased in parallel), if there are qualms that the bank of questions cannot be substantially improved. This should increase reliability, on a modern KR20 rating, to reasonable levels, particularly if combined with some question improvement.

However, elimination of the majority of the poorly performing questions would be the superior measure to improve the reliability of the exam and most other performance measures. If some confidence can be felt that addressing the problems with the current questions, particularly those with negative discrimination, can be achieved in future sittings, a more moderate increase to 120 questions, or no increase at all, may well be justified.

The ongoing development of a relevant, modern, discriminatory and reliable test is a model for the future in the field of Undergraduate Medical Education in the Hashemite Kingdom of Jordan.

REFERENCES

1. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach*. 2004;26(8):709 – 712.
2. Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika*. 1937;2 (3):151 – 160.
3. Downing SM. Reliability: on the reproducibility of assessment data. *Med Edu*. 2004;38 (9):1006 – 1012.