Reliability of the Evaluation of Students' Answers to Essay-type Questions

T Anatol, S Hariharan

ABSTRACT

Introduction: *This paper seeks to quantify the reliability of the assessment of students' answers to essay-type questions, in an attempt to define the role of such questions in University examinations.*
Methods: *The marks awarded for essay-type questions during three consecutive final undergraduate examinations in surgery were analyzed. The mean scores, 95% confidence intervals and the standard error of the mean were calculated to determine the distribution of the marks. Statistical analysis was used to determine the correlation of the marks awarded for the same answer by different markers and deduce the dependability of this method of testing.*
Results: *The marks awarded to 233 answer papers were available for analysis. The marks awarded by each pair of examiners for student answers to individual questions coincided on only 46.3% of occasions, but varied within just ± 5% on 90.7% of occasions. Use of the kappa index to determine the agreement between markers produced a value of just 0.385, well short of the ideal of 1.0. Assessment of the overall reliability of this type of examination by Cronbach's reliability coefficent gave a value of 0.672.*
Conclusion: *There was a significant variation among markers in the evaluation of answers to essay-type questions. However, the overall test reliability was acceptable enough to justify continuation of this type of assessment as a supplement to other methods.*

Confiabilidad de la Evaluación de las Respuestas de los Estudiantes a las Preguntas de Ensayo

T Anatol, S Hariharan

RESUMEN

Introducción: *Este trabajo busca cuantificar la confiabilidad de la evaluación de las respuestas de los estudiantes a las preguntas de ensayo, en un intento por definir el papel de este tipo de preguntas en los exámenes de la Universidad.*
Métodos: *Se analizaron las notas otorgadas en cirugía a las preguntas de ensayo durante los tres exámenes finales consecutivos de pregrado. Se calcularon los puntajes promedio, intervalos de confianza de 95%, y el error estándar de la media, con el fin de determinar la distribución de las notas. Se usó el análisis estadístico para determinar la correlación de las notas dadas a las mismas respuestas por diferentes evaluadores, y para deducir la confiabilidad de este método de evaluación.*
Resultados: *Las notas otorgadas a 233 pruebas respondidas fueron puestas a disposición para su análisis. Las notas dadas por cada par de examinadores a las respuestas de los estudiantes a las preguntas individuales, coincidieron sólo en 46.3% de las ocasiones, pero variaron en justamente ± 5% en 90.7% de las ocasiones. El uso del índice de Kappa para determinar el acuerdo entre evaluadores, produjo un valor de sólo 0.385, bien lejos del ideal 1.0. La evaluación de la confiabilidad general de este tipo de examen, mediante el coeficiente de confiabilidad de Cronbach, arrojó un valor de 0.672.*
Conclusión: *Hubo una variación significativa entre los evaluadores a la hora de calificar las respuestas a las preguntas de ensayo. Sin embargo, la confiabilidad de la prueba en general fue suficientemente aceptable para justificar que se continúe con este tipo de evaluación como un complemento de otros métodos.*

West Indian Med J 2009; 58 (1) : 13

From: Department of Clinical Surgical Sciences, Faculty of Medical Sciences, Eric Williams Medical Sciences Complex, Trinidad and Tobago

Correspondence: Dr T Anatol, Department of Clinical Surgical Sciences, Faculty of Medical Sciences, Eric Williams Medical Sciences Complex, Mt Hope, Trinidad and Tobago. Fax: (868) 663- 4319, e-mail: trevana@wow.net

INTRODUCTION

Effective clinical reasoning calls for the unique integration of several types of learning. These include previously acquired familiarity with the basic medical sciences, information obtained from the study of clinical situations and experience gained on the wards. The synthesis of this knowledge is then applied to diagnostic and management decisions. Relevant knowledge regarding clinical skills is now commonly assessed by Multiple Choice Questions (MCQ) and Objective Structured Clinical Examinations (OSCE) but patient management problems for answering in an essay format are still considered helpful in evaluating the ability to make appropriate judgments (1).

This traditional approach to student assessment using open-ended questions based on clinical scenarios has been occasionally challenged (2). Although this form of testing has the attraction of limiting the provision of cues to students, it also has the drawbacks of being resource and labour intensive, as well as of being open to accusations of selective sampling, marker subjectivity and uneven reliability (3–5). Although reliability may be improved by multiple assessments using different assessors, several related effects may still compromise the impartiality of the marking process. These include variations in severity among different markers, of the same marker at different times and in the relative difficulty of the questions set (6).

The Faculty of Medical Sciences of the University of the West Indies currently uses the combination of an essay-type question paper, a clinical examination and a *viva voce* examination in the 'exit' assessment of clinical knowledge in some subject areas. The present study attempts to explore the reliability of this form of testing by studying the marks awarded to students for the essay components of three recently conducted examinations.

SUBJECTS AND METHODS

The marks awarded by each examiner for three consecutive written examinations in undergraduate surgery were used for the analysis. Each of the three examination papers contained eight essay-type questions, and the students had to answer all eight questions without any choice at selection or omission. Markers rated batches of papers independently and each paper was marked by a pair of markers. All the markers were experienced clinicians who had been involved in examining final-year medical students for a number of years. They had progressed from being observers to being actual examiners at the clinical and *viva voce* examinations, before being invited to mark the answer sheets of essay-type question papers.

A band-type marking system was used for the examinations, with five points ranging between 40 and 75%. The marks were assigned as follows: 40 = very poor (irretrievable), 45 = below average (retrievable), 50 = average (pass), 55 = above average, 60 = good, 65 = very good, 70 = honours and 75 = distinction. No intermediate marks could be assigned. Each marker was given specific instructions about the marking system being used.

A total of 13 examiners overall assessed these papers, each answer paper being marked by a pair of examiners. Measures of the method of evaluation under analysis included the concordance, inter-rater reliability and internal consistency. Concordance assessed the rate of coincidence of marks given for the same answer by different markers. Inter-rater reliability defined the correlation between the marks assigned to the same answer by different markers. Internal consistency reflected the correlation between the marking of a set of answers as a determinant of the evenness of the assessment.

To assess the inter-marker concordance for each examination, the precise agreement between each marker (within the pair) for each question was analyzed. So too was the agreement within one band (*ie* within a range of ± 5%). The mean score for each question awarded by each marker along with 95% confidence intervals and the standard error of mean were also calculated. Inter-rater reliability was determined by the use of kappa statistics. The internal consistency was derived by using Cronbach's reliability coefficient to estimate the correlation between the marks awarded to each answer. The ideal test would have a reliability of 1.0. The closer the score approached 1, the greater the reliability.

The Statistical Package for Social Sciences (SPSS) version -12 (Chicago, IL, USA) and Microsoft Excel 2000 (Seattle, WA, USA) software programmes were used for the analyses.

RESULTS

A total of 233 answer papers from three consecutive batches of undergraduate surgery examinations were included for analysis. Batch I had 106, Batch II 35 and Batch III 92 papers.

The mean percentage agreement between marker-pairs for all the questions in the three batches of papers studied was 46.3% (range 35.8 to 65.7%). This percentage increased to 90.7% (range 82.1 to 100%) when concordance within 1 band (± 5%) was estimated. The details are depicted in Table 1. The box-plot (Figure) shows the distribution pattern of the marks awarded by the individual examiners.

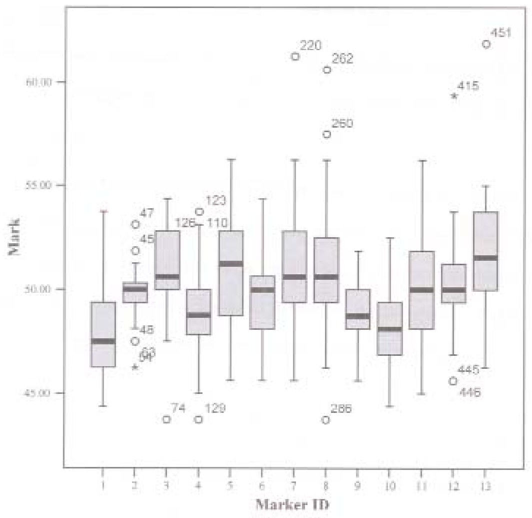
The median score awarded to each question was 50%. The assessment of each marker was compared to the median of the marks awarded by the whole cohort of markers. An average of 53.6% of examiners awarded marks at the overall median level (range 35.4 to 78.8%), 24.8% (range 12 to 47.2%) awarded marks below the overall median and 21.6% (range 5.4 to 38.9%) awarded marks above the overall median. These details are set out in Table 2.

The inter-rater reliability as determined by use of kappa statistics gave the coefficient of agreement between different markers as only 0.385 (SE: 0.297). The

Table 1: Percentage agreement among marker pairs

Batch (n)	Agreement*	Q1 (%)	Q2 (%)	Q3 (%)	Q4 (%)	Q5 (%)	Q6 (%)	Q7 (%)	Q8 (%)	Overall (%)
I	exact	35.8	35.8	38.7	39.6	47.2	42.5	51.9	45.3	42.0
(106)	± 5	85.6	82.1	91.5	88.7	89.6	83.0	86.8	93.4	87.6
II	exact	37.1	37.1	45.7	45.7	51.4	65.7	42.9	60.0	48.2
(35)	± 5	88.6	85.7	(100)	94.3	91.4	(100)	88.6	97.1	93.2
III	exact	44.6	63.0	46.7	54.3	55.4	51.1	47.8	39.1	50.3
(92)	± 5	88.0	95.7	87.0	94.6	96.7	96.7	94.6	92.4	93.3
All	exact	39.5	46.8	42.9	46.4	51.1	49.4	48.9	45.1	46.3
(233)	± 5	87.1	88.0	91.0	91.8	92.7	91.0	90.1	93.6	90.7

Batch = each year of examination
Q = Question number
* exact: marks within the same band ± 5 marks within one band



* and ○ show outliers

Figure: Distribution of marks awarded by examiners

internal test consistency derived by Cronbach's alpha showed a value of 0.672 for all three examinations.

DISCUSSION

The present study showed a less than 50% concordance rate between examiners assessing essay-type questions, although this rate increased to over 90% when variation within one band (± 5%) was allowed. Kappa analysis confirmed a poor correlation of 0.38 between markers evaluating the questions. Yet, the overall test reliability was acceptable, averaging 0.67 for the three examinations.

Essay-type questions require students to construct their own responses. This tests the ability to not only recall but also to organize and apply knowledge. This form of appraisal is particularly attractive for a predominantly 'problem-based' curriculum (7), such as is used at the St Augustine campus of the University of the West Indies.

A previous study in the United Kingdom (UK) indicated that essays were being used for summative assessments

Table 2: Relationship between marker ratings

Marker	Scripts marked	Questions (n)	Mean (SEM)	Scores at median[†] n (%)	Scores < median n (%)	Scores > median n (%)
1	36	288	48.1 (0.36)	102 (35.4)	136 (47.2)	50 (17.4)
2	36	288	49.8 (0.22)	227 (78.8)	35 (12.2)	26 (9.0)
3	23*	184	50.9 (0.5)	110 (59.8)	22 (12.0)	52 (28.3)
4	40	320	49 (0.38)	174 (54.4)	100 (31.3)	46 (14.4)
5	40	320	50.9 (0.42)	140 (43.8)	79 (24.7)	101 (31.6)
6	40	320	49.7 (0.33)	171 (53.4)	86 (26.9)	63 (19.7)
7	39	312	51.2 (0.46)	157 (50.3)	51 (16.3)	104 (33.3)
8	38	304	50.9 (0.51)	140 (46.1)	61 (20.1)	103 (33.9)
9	39	312	49.1 (0.23)	221 (70.8)	74 (23.7)	17 (5.4)
10	38	304	48.2 (0.31)	149 (49.0)	133 (43.8)	22 (7.2)
11	40	320	50.1 (0.4)	191 (59.7)	64 (20.0)	65 (20.3)
12	39	312	50.2 (0.38)	183 (58.7)	61 (19.6)	68 (21.8)
13	18*	144	52.0 (0.76)	53 (36.8)	35 (24.3)	56 (38.9)

*Marker 3 marked only 2, and marker 13 only 1, out of the 3 batches of papers
[†] The median score for each question was 50%

by 81% of the schools in the first-year, 71% in the second-year but by only 29% in the final-year (5). In another study which elicited responses from 70% of 126 medical schools accredited in the USA, half of the respondents reported the use of essay questions in the first two years, but by the fourth and final years this figure had fallen to 32% (8).

Multiple markers for essay-type answers can obviously improve reliability (9). In fact, almost half of the schools responding to the UK questionnaire made extensive use of a second marker for written assessments (5). This is standard practice at the University of the West Indies. However, when different people rate essays, inter-marker reliability becomes a legitimate concern. Even with rigorous training, differences in the background and experience of the markers can lead to subtle but important differences in grading. The present study showed a high variability in the inter-marker ratings. Differences in background and experience may have an effect in such a situation.

Real consistency in examination markings can be expected only if markers are experts *ie* highly knowledgeable in the domain which they are marking. The major errors are variation in marker severity or leniency and a lack of inter-marker reliability (10).

Well-designed scoring rubrics may respond to the concern of inter-marker reliability by establishing a description of scoring criteria in advance (11). In the Department of Clinical Sciences, although precise written guidelines about the implications of each marking band are given to the examiners, rubrics are not provided and the assessment of the content of the answer is eventually subjective. This may contribute to inter-marker variability.

In summary, essay-type questions in this setting produce considerable inter-marker variability which possibly indicates bias in the method of assessment. However, because the general reliability of this method of examination is supported by its internal consistency, complete abolition of this form of assessment does not appear justified. It may be appropriate to retain this modality in combination with other forms of assessment, particularly for postgraduate programmes, where expert markers are the norm, or perhaps for use as a formative type of assessment in the penultimate undergraduate year.

ACKNOWLEDGEMENT

The authors wish to thank Ms Shelly-Anne Hunte for her help in data collection and analysis.

REFERENCES

1. Newble D, van der Vleuten C, Norman G. Assessing clinical reasoning. In: Clinical Reasoning in the Health Professions. Higgs J, Jones M, Editors. 2000, 2nd Ed, Butterworth Heinemann Ltd: Oxford, UK. 156–165.
2. Neufeld VR, Norman GR: Assessing Clinical Competence. 1985, New York: Springer
3. Day SC, Norcini JJ, Diserens D, Cebul RD, Schwartz JS, Beck LH et al. The validity of an essay test of clinical judgement. Acad Med 1990, **65 (Suppl 9):** S39–40.
4. Lowry S. Assessment of students. BMJ 1993; **306:** 51–4.
5. Fowell SL, Maudsley G, Maguire P, Leinster SJ, Bligh J. Student assessment in undergraduate medical education in the United Kingdom, 1998. Med Educ 2000, **34 (Suppl 1):**1–49.
6. Barrett S. Question choice: Does marker variability make examinations a lottery? HERDSA Annual International Conference, July 12–15, 1999, Melbourne: 1–17.
7. Palmer D, Rideout E. Essays in Evaluation Methods: A Resource Handbook. McMaster University Program for Educational Development, 1995.
8. Mavis B, Cole B, Hoppe R. A survey of student assessment in US medical schools: the balance of breadth versus fidelity. Teach Learn Med 2001, **13:** 74–9.
9. Rudner L. Reducing Errors Due to the Use of Judges. Practical Assessment, Research and Evaluation 1992, **3:** 1–4.
10. Barrett S. The impact of training on marker variability. Int Educ J 2001, **2:** 49–58.
11. Moskal B, Leydens J. Scoring Rubric Development: Validity and Reliability. Practical Assessment, Research and Evaluation 2000; **7:** 1–11.