Reliability of the MRCP(UK) Part I Examination, 1984–2001

I C McManus,[1,2] J Mooney-Somers,[2] J E Dacre[2] & J A Vale[2] on behalf of the MRCP(UK) Part I Examining Board and the, Federation of Royal Colleges of Physicians, MRCP(UK) Central Office

Objectives To assess the reliability of the MRCP(UK) Part I Examination over the period 1984–2001, and to assess how the reliability is related to the difficulty of the examination (mean mark) and to the spread of the candidates' marks (standard deviation).

Methods Retrospective analysis of the reliability (KR20) of the MRCP(UK) examination recorded in examination records for the 54 diets between 1984 and 2001.

Results The reliability of the examination showed a mean value of 0·865 (SD 0·018, range 0·83–0·89). There were fluctuations in the reliability over time, and multiple regression showed that reliability was higher when the mean mark was relatively high, and when the standard deviation of the marks was high.

Conclusions The reliability of the MRCP(UK) Examination was maintained over the period 1984–2001. As theory predicted, the reliability was related to the average mark and to the spread of marks.

Keywords Clinical competence, *standards; education, medical, graduate, *standards, *methods; educational measurement, *standards; Great Britain; reproducibilty of results.

Medical Education 2003;37:609–611

Introduction

Written examinations form a central part of post-graduate medical training and assessment, with multiple choice questions (MCQs) an internationally popular format, in part because of ease and objectivity of marking. MCQ-based examinations are often used as the first component of a multistage examination, particularly to ensure that candidates have an adequate knowledge base, prior to entering subsequent clinical examinations. For an examination to be robust, it must be valid, reliable and practical. A valid examination measures what it is supposed to measure; a reliable examination measures the same features on different occasions; a practical examination is easy to organize and efficient to run.

Post-graduate medical examinations in the UK have in the past been criticised for not providing sufficient information about their statistical underpinning[1] and very few UK examinations presently publish information on their reliability.[2] Here we assess the reliability of the MRCP(UK) Part I Examination, and the factors affecting it.

The basic format of the MRCP (UK) Part I Examination was unchanged from 1984 to 2001, consisting of 60 five-part multiple true–false questions,[3] a total of 300 items overall, covering a broad range of general (internal) medicine and related basic medical science.

Methods

In the Examination, items could be answered true, false, or don't know, with negative marking for items answered incorrectly. Reliability was calculated using the KR20 coefficient[4] which is equivalent to coefficient alpha, and can be regarded as a generalized split-half reliability.[5] For each Examination the overall difficulty of the Examination was also calculated, expressed as the mean percentage correct (i.e. corrected for guessing, so that candidates guessing at random would on average score zero). The standard deviation (SD) of candidates' corrected scores was also calculated.

[1]Department of Psychology, University College London, UK
[2]MRCP(UK) Central Office, London, UK

Correspondence: Prof I C McManus, Department of Psychology, University College London, London, WC1E 6BT UK

Key learning points

The reliability of the MRCP(UK) examination was high (average 0·865) over the 54 diets held between 1984 and 2001.

Reliability was higher when the average mark was relatively high (close to 50%) and when the spread of marks was greatest (due to few questions being too easy or too hard).

The effects of the average mark and the spread of marks are as psychometric theory predicts.

Results

Figure 1(a) shows the variation in the reliability of the general medicine examination over the 54 diets held across 18 years (mean KR20 = 0·865, SD = 0·018, 90% range 0·83–0·89). Figures 1(b and c) show the fluctuations in the mean and the SD of the corrected percentage mark. Multiple regression shows both mean and SD are independent predictors of the KR20 (SD: beta = 0·727, $P < 0·001$; mean: beta = 0·613, $P < 0·001$; $R = 0·942$), the SD accounting for more of the variance than the mean. The pass mark and the number of candidates taking the Examination showed no separate association with KR20 once mean and SD had been taken into account. Figure 1(d) shows the joint regression of KR20 against mean and SD of corrected scores.

Comment

The reliability of the MRCP(UK) Part I Examination, calculated as the KR20 score, has been related to the distribution of marks obtained by candidates in 54 separate diets of the Examination over an 18-year period during which the basic structure of the Examination was unchanged. Reliability was maintained at an acceptable level over the period 1984–2001 but was higher, as classical psychometric theory predicts, when the average mark was higher and when the spread of candidates' marks was greater.

Two academic changes were introduced to the MRCP(UK) Part I Examination in the early 1990s. Firstly, from the last diet of 1993, the four MCQs on paediatric topics were removed from the Examination.

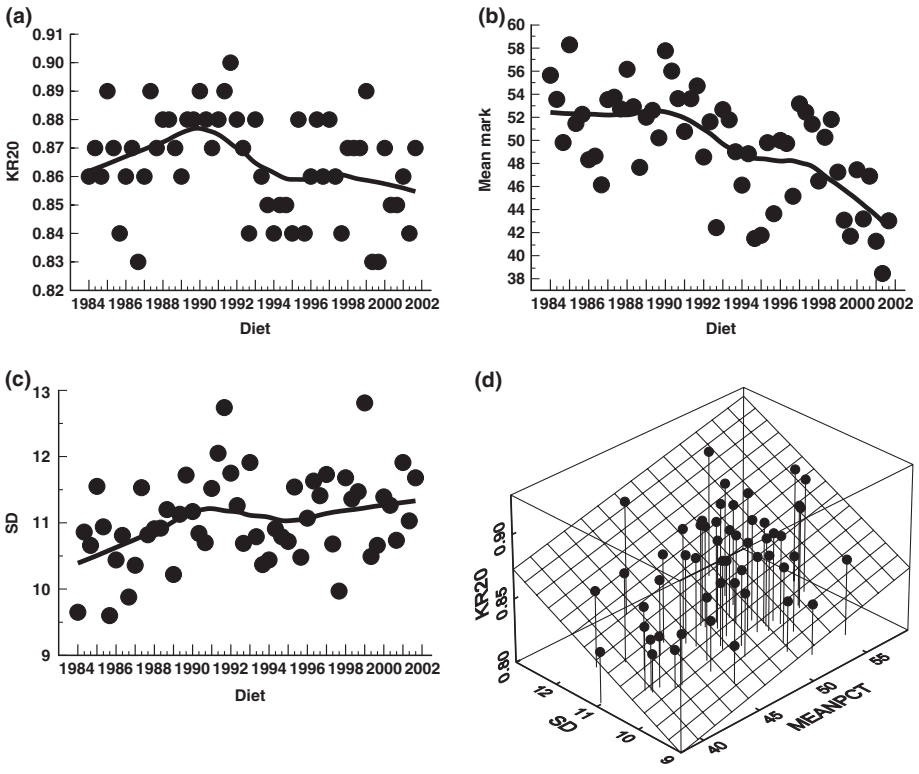

Figure 1 (a) Reliability (KR20), (b) mean corrected percentage score and (c) SD of corrected percentage score from 1984 to 2001 for the MRCP(UK) Part I Examination. Lines are fitted using a lowess technique. (d) Three-dimensional plot of KR20 for the Examination (vertical) against mean corrected percentage score and SD of corrected percentage score, with fitted regression surface.

Thereafter, no paediatric topics were tested in the MRCP(UK) Part I Examination. A separate MRCP(UK) paediatric paper was set for paediatric candidates from the last diet of 1993 and this Examination became the MRCPCH Part I Examination in 1999, which is now administered by the Royal College of Paediatrics and Child Health. Secondly, in the early 1990s the MRCP(UK) Part I Examining Board included more basic science MCQs in the Examination. These questions were not answered by a substantial number of candidates. Nevertheless, the overall reliability of the examination was maintained at an acceptable level.

Theory suggests that if an examination becomes too difficult then its reliability will decrease (until in the limiting case all questions are impossible for all candidates, who resort to mere guessing, and the reliability is zero). That relationship is clearly demonstrated here empirically, and suggests that the introduction of more basic science questions (which the majority of candidates found to be challenging) had more impact than the removal of paediatric questions, which general (internal) medicine candidates welcomed. Likewise, when candidates show a wide spread of marks (i.e. a high SD), then reliability will be higher, as true variance in candidate ability is relatively greater than measurement error; again the effect is shown empirically in these data. Fluctuations in the reliability of the MRCP(UK) Part I Examination are to a large extent dependent upon the standard deviation, low values reflecting an excess of questions which either are too easy or too difficult to assess useful variance in candidate ability, or show poor discrimination. The SD of the examination increased between 1984 and 2001, which maintained the level of the KR20 despite an overall fall in the mean mark.

Reliability is a prerequisite of any examination. The reliability of the MRCP (UK) Part I Examination has been maintained at an acceptable level over the period 1984–2001. The fluctuations that have occurred in the reliability are shown to relate to the difficulty (mean) and question mix (SD) of the Examination as theory predicts.

Acknowledgements

We thank Rosalind Barrington-Leigh, Kate Beaumont and Jennifer Mollon for their help with this study. JAV was the Chairman of the MRCP(UK) Part I Examining Board, JED was Secretary to the Board, ICM was a member of the Board, and JMS was a Research Assistant in the MRCP(UK) Central Office at the time the research was carried out and the paper was written. Data collection was carried out by JMS, statistical analysis was by ICM, the first draft of the paper was written by ICM and all authors contributed to the final draft.

References

1 Anonymous. Examining the Royal Colleges' examiners [editorial]. *Lancet* 1990;**335**:443–5.
2 Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Med Educ* 2002;**36**:73–91.
3 Albanese MA, Sabers DL. Multiple true-false items: a study of interitem correlations, scoring alternatives, and reliability estimation. *J Educ Measurement* 1988;**25**:111–23.
4 Fleming PR, Sanderson PH, Stokes JF, Walton HJ. *Examinations in Medicine*. Edinburgh: Churchill Livingstone 1976.
5 Ghiselli EE, Campbell JP, Zedeck S. *Measurement Theory for the Behavioral Sciences*. San Francisco: W H Freeman 1981.

Received 25 *May* 2002; *editorial comments to authors* 10 *October* 2002 *and* 29 *November* 2002; *accepted for publication* 27 *January* 2003