original papers

Psychiatric Bulletin (2007), **31**, 342–344. doi: 10.1192/pb.bp.106.012906

FEMI OYEBODE, SANJU GEORGE, VEENA MATH AND SAYEED HAQUE

Inter-examiner reliability of the clinical parts of MRCPsych part II examinations

AIMS AND METHOD

The aim of the study was to investigate the interrater reliability of the clinical component of the MRCPsych part II examinations, namely the individual patient assessment and the patient management problems. In the study period, there were 1546 candidates and 773 pairs of examiners. Kappa scores for pairs of examiners in both these assessments were calculated.

RESULTS

The kappa scores for exact numerical agreement between the pairs of examiners in both individual patient assessment and patient management problems were only moderate (0.4–0.5). However, the kappa scores for agreement between pairs of examiners for the reclassified pass and fail categories were very good (0.8).

CLINICAL IMPLICATIONS

The poor reliability of the traditional long case and oral examinations in general is one of the most potent arguments against their use. Our finding suggests that the College clinical examinations are at least not problematic from this point of view, particularly if global pass or fail judgements rather than discrete scores are applied.

There are undoubted changes facing the practice of medicine. These changes embrace the shift away from the care of the individual to concerns about the health of the community, from cure of disease to preservation of health, from episodic care to continuous and comprehensive care, and from in-patient care to community or home treatment (Jones *et al*, 2001). These changes are accompanied by changes in medical curricula characterised by the integration of basic sciences and clinical medicine in undergraduate medicine, the adoption of social, psychological and humane disciplines by medicine and the move from didactic teaching to the promotion of active learning. Inevitably the assessment methods that have traditionally been used in determining clinical competence in undergraduate and postgraduate medicine are also undergoing revision. In postgraduate medicine, these changes are overseen by the Postgraduate Medical Education and Training Board and are allied to the changes in the structure of training programmes overseen by Modernising Medical Careers. In this context there are bound to be far-reaching changes to the College examinations.

The MRCPsych part II clinical examinations currently comprise the individual patient assessment (IPA), also known as the traditional long case, and the patient management problems (PMP), which is also known as the structured oral examination. The traditional long case has been established for over 150 years as a method of examining clinical skills in medicine. In recent years its value as an assessment method has come under great scrutiny. The strengths of the traditional long case include its obvious face validity, as it evaluates the performance of a doctor in an encounter with real patients whereby information is gathered and treatment plans are developed under realistic conditions. The task for the candidate is to take a history, to structure the clinical problem, synthesise the findings and formulate an appropriate management plan. For many clinicians these skills are fundamental to the practice of medicine and the authenticity of the challenge for the candidate is an intuitively correct method of assessing clinical competence. Despite these obvious strengths, the traditional long case has inherent problems. The clinical challenges posed to candidates in the long case are not identical, equal or even similar in complexity. Furthermore, it is assumed that performance on one particular type of case is predictive of performance on other types of cases, when most clinicians know that they do not necessarily perform uniformly across all patient problem types. In addition, there is concern that examiner behaviour is not reliable (the problems of interrater and intra-rater reliability; Norcini, 2002). In short, there is a conflict between validity and reliability.

The problems that arise with the individual patient assessment also pertain to the clinical oral examination, the patient management problems. In unstructured *viva voce* examinations candidates are liable to be asked whatever questions the examiner chooses and there is the risk that the examiner will concentrate on their pet interests. Furthermore, there is evidence that structured *viva voce* examinations are more reliable than unstructured examinations (Tutton & Glasgow, 2005) and the College has made the necessary changes to accommodate this. The concerns about clinical oral examinations are also pertinent to PhD *vivas* (Jackson & Tinkler, 2001; Morley *et al*, 2002) and to job selection interviews (Wiesner & Cronshaw, 1988; McDaniel *et al*, 1994).

In this study we investigate the level of agreement between the two examiners in both component parts of the MRCPsych clinical examinations. This is a measure of the inter-rater reliability.

Method

Data from four sittings of the MRCPsych examinations (Spring 2003, Autumn 2003, Spring 2004 and Autumn 2004) were available for analysis. In both the individual

original papers

patient assessment and patient management problems, examiners are required to award an independent and individual mark from 1 to 10 (1 being very poor and 10 being excellent) before entering into the discussion that precedes the award of the final jointly awarded mark. It is the individually awarded marks that are the subject of this study. Scores 1–4 designate failure and 5–10 a pass.

Data were analysed using SPSS version 12 for Windows. In this study the kappa statistic (κ) was used as the measure of strength of agreement between pairs of examiners. The generally accepted standards of strength of agreement of κ are: <0 poor, 0.1–0.2 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, 0.81–1.0 almost perfect (Landis & Koch, 1977).

Results

In the study period, there were 1546 candidates and 773 pairs of examiners. Data were not available for all candidates who sat the examinations during the study period; only for 1537 candidates who sat the individual patient assessment and 1518 who sat the patient management problems. The κ for exact agreement between the pairs of examiners for the individual patient assessment for all scores was 0.513 ($P<0.0001$), for the failing candidates only was 0.462 ($P<0.0001$) and for the passing candidates was 0.485 ($P<0.0001$). When the scores were reclassified into pass and fail categories, κ for agreement was 0.794 ($P<0.0001$). The κ for exact agreement between the pair of examiners for the patient management problems for all scores was 0.515 ($P<0.0001$), for failing candidates only it was 0.562 ($P<0.0001$) and for the passing candidates it was 0.475 ($P<0.0001$). The κ for agreement between examiners in the patient management problems when the scores were reclassified into pass and fail categories was 0.802 ($P<0.0001$).

Discussion

The kappa scores for exact numerical agreement between the pairs of examiners in both individual patient assessment and patient management problems were only moderate. However, the kappa scores for agreement between pairs of examiners for the reclassified pass and fail categories were substantial. We believe that this is the first report of the measure of agreement between pairs of examiners in a clinical examination in postgraduate psychiatry comprising the traditional long case (individual patient assessment) and the structured oral examination (patient management problems). The finding of substantial agreement between pairs of examiners on the reclassified pass or fail categories suggests that global pass or fail judgements may be more robust than the award of discrete marks. However, it is important to note that the lower kappa statistic in the comparison of individual marks is understandable and predictable because lower reliability is expected when there are a higher number of rating options. Notably when a decision is made on pass or fail ratings, there are only two categories, so agreement is statistically more likely. This probably accounts for the higher kappa statistic for pass or fail judgements. There are reports of high interrater reliability in undergraduate clinical examinations (Schwiebert & Davis, 1993; Morgan *et al*, 2001; Wass & Jolly, 2001). However, there are also reports of low reliability, particularly with global performance ratings as opposed to checklists (Daelmans *et al*, 2005). In postgraduate medical examinations, there are very few published findings. In anaesthesia, Devitt *et al* (1997) reported high interrater reliability between pairs of examiners observing performance on an anaesthesia simulator test.

The reliability of the traditional long case and oral examinations in general is one of the most potent arguments against their use. Our finding suggests that the College clinical examinations are at least not problematic from this point of view. Norcini (2002) has argued that there are at least three ways to improve the reproducibility of scores awarded by examiners in the traditional long case: employing a statistical model to remove difference among them; training examiners; or increasing the number of examiners. It is likely that the close agreement between examiners in the College examinations is due to the training provided. New examiners receive initial training before examining and all examiners are required to attend the annual board of examiners meeting where refresher training takes place. However, Norcini (2002) argues that any improvements in reproducibility of scores, that is, in the reliability of the traditional long case, will only be modest and that the largest effect is likely to be due to increasing the number of examiners.

Several authors have proposed modifications to the long case to make it fit for purpose. These discussions have developed because of anxiety that the objective structured clinical examination (OSCE) assesses breadth of skill but at the expense of depth (Wass & van der Vleuten, 2004). The OSCE is able to assess many competencies but because of the format, very limited time is available for the assessment of these competencies. In the MRCPsych part I OSCE there are currently twelve 7-min 'stations'. This illustrates the problem well; there is extensive coverage of novel clinical areas such as information-giving to patients, carers, doctors and other healthcare professionals, yet, the actual time allocated to the evaluation of these competences is arguably limited. Furthermore, in psychiatry, there is the risk that the OSCE promotes a disjointed acquisition of clinical skills and that the capacity to integrate a case in all its fullness may be lost in the process. In real-life situations, the discrete clinical competencies are deployed in order to serve the interest of an individual patient, and to do this satisfactorily the various aspects of the case need to be integrated into a meaningful whole. The proposals to improve the intercase reliability of the long case include:

- increasing the number of patient encounters that candidates have, perhaps up to ten (Wass *et al*, 2001)
- increasing the number of examiners
- observing the clinical encounter as well as the candidate's presentation.

https://doi.org/10.1192/pb.bp.106.012906 Published online by Cambridge University Press

original papers

McKinley *et al* (2005) suggest an innovative solution to the problem of increasing the number of patient encounters; they advocate sequential testing. This involves all students being directly observed in four consultations by a different pair of examiners for each case. Each consultation lasts 30 min. Those considered to be unlikely to fail are excused further testing; the rest, approximately a quarter of the class, are observed consulting with four more patients by another four pairs of examiners. In this system, failing candidates are examined on eight cases by eight pairs of examiners.

These proposals are resource intensive and probably impractical. In the current MRCPsych part II, the time required for each candidate to examine ten patients will amount to at least 10 h. In an examination dealing with approximately 1000 candidates annually, this will be an impossible task. The same problem pertains to any extension in the number of examiners or the introduction of wholly observed clinical examinations. The issue is whether these proposals will produce significant or merely marginal improvements, and ultimately whether they will be cost-effective.

There is evidence that the desire to create assessment methods that rely on standardised and objectified tasks in a controlled environment is returning full circle to the assessment of candidates in the real world of patients and the workplace (van der Vleuten & Schuwirth, 2005). The concern about the variance introduced by real patients and the emphasis on the desirability of standardised patients has lessened with the use of the Mini-Clinical Evaluation Exercise (mini-CEX) in workplace-based assessments, with limited observations of candidate encounters with real patients (Norcini *et al*, 1995). However, it is doubtful that the mini-CEX can be successfully applied to psychiatry without modification. What is now also clear is that the reliability of clinical examinations is not dependent solely on objectification or standardisation, but also on careful sampling across clinical content domains which needs substantial hours of testing time (Petrusa, 2002). The reliability estimates for the long case depending on hours of testing are reported as 0.60 for 1 h, 0.75 for 2 h, 0.86 for 4 h and 0.90 for 8 h. These estimates are comparable for multiple choice question papers, oral examination and OSCE (van der Vleuten & Schuwirth, 2005).

It is clear that the proposals to improve the traditional long case are unlikely to be efficient or cheap. However, the energy going into the process suggests an awakening to the potential risks of relying merely on tests of competence such as the OSCE. At present it is uncertain how far workplace-based assessments of clinical performance using instruments such as the mini-CEX can adequately replace the traditional long case and oral examination. Our findings show that there is a good measure of agreement between pairs of examiners in these examinations, particularly for global pass or fail judgements. In this transitional period, as assessments of clinical competence and performance evolve, whatever programme of assessments is developed and adopted, the value of the traditional long case and structured oral examination need to be carefully considered. It is probably true to say that the unique contribution of the long case in particular is unlikely to be surpassed by simulated patients or standardised and objectified assessments.

Declaration of interest

F.O. was Chief Examiner, Royal College of Psychiatrists 2002–2005.

References

DAELMANS, H. E., VAN DER HEM-STOKROOS, H. H., HOOGENBOOM, R. J., *et al* (2005) Global clinical performance rating, reliability and validity in an undergraduate clerkship. *Netherlands Journal of Medicine*, **63**, 279–284.

DEVITT, J. H., KURREK, M. M., COHEN, M. M., *et al* (1997) Testing the raters: inter-rater reliability of standardised anaesthesia simulator performance. *Canadian Journal of Anaesthesia*, **44**, 924–928.

JACKSON, C. & TINKLER, P. (2001) Back to basics: a consideration of the purposes of the PhD viva. *Assessment and Evaluation in Higher Education*, **26**, 355–366.

JONES, R., HIGGS, R., DE ANGELIS, C., *et al* (2001) Changing face of medical curricula. *Lancet*, **357**, 699–703.

LANDIS, J. R. & KOCH, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

McDANIEL, M. A., WHETZELl, D. L., SCHMIDT, F. L., *et al* (1994) The validity of employment interviews: a comprehensive review and meta-analysis. *Journal of Applied Psychology*, **79**, 599–616.

McKINLEY, R. K., HASTINGS, A. M. & PETERSEN, S. (2005) The long case revisited. *Medical Education*, **39**, 442–447.

MORGAN, P. J., CLEAVE-HOGG, D. & GUEST, C. B. (2001) A comparison of global ratings and checklist scores from an undergraduate assessment using an anaesthesia simulator. *Academic Medicine*, **76**, 1053–1055.

MORLEY, L., LEONARD, D. & DAVID, M. (2002) Variations in vivas: quality and equality in British PhD assessments. *Studies in Higher Education*, **27**, 263–273.

NORCINI, J. J. (2002) The death of the long case? *BMJ*, **324**, 408–409.

NORCINI, J. J., BLANK, L. L., ARNOLD, G. K., *et al* (1995) The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Annals of Internal Medicine*, **123**, 795–799.

PETRUSA, E. R. (2002) Clinical performance assessments. In *International Handbook for Research in Medical Education* (eds G. R. Norman, C. P. M. van der Vleuten & D. I. Newble) pp. 673–709. Kluwer Academic Publishers.

SCHWIEBERT, P. & DAVIS, A. (1993) Increasing inter-rater agreement on a family medicine clerkship oral examination – a pilot study. *Family Medicine*, **25**, 182–185.

TUTTON, P. J. M. & GLASGOW, E. F. (2005) Reliability and predictive capacity of examinations in anatomy and improvement in the reliability of *viva voce* (oral) examinations by the use of a structured rating system. *Clinical Anatomy*, **2**, 29–34.

VAN DER VLEUTEN, C. P. M. & SCHUWIRTH, L. W. T. (2005) Assessing professional competence: from methods to programmes. *Medical Education*, **39**, 309–317.

WASS, V. & JOLLY, B. (2001) Does observation add to the validity of the long case? *Medical Education*, **35**, 729–734.

WASS, V. & VAN DER VLEUTEN, C. P. M. (2004) The long case. *Medical Education*, **38**, 1176–1180.

WASS, V., JONES, R. & VAN DER VLEUTEN (2001) Standardised or real patients to test clinical competence? The long case revisited. *Medical Education*, **35**, 321–325.

WIESNER, W. H. & CRONSHAW, S. F. (1988) A meta-analytic investigation of the impact of the interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology*, **61**, 275–290.

***Femi Oyebode** Professor of Psychiatry, Department of Psychiatry, University of Birmingham, Queen Elizabeth Psychiatric Hospital, Birmingham B15 2QZ, email: femi.oyebode@bsmht.nhs.uk, **Sanju George** Consultant Psychiatrist, Birmingham and Solihull Mental Health NHS Trust, **Veena Math** Specialist Registrar, Greater Glasgow Primary Care NHS Trust, **Sayeed Haque** Research Fellow, Department of Psychiatry, University of Birmingham, Queen Elizabeth Psychiatric Hospital, Birmingham

https://doi.org/10.1192/pb.bp.106.012906 Published online by Cambridge University Press