TITLE

Oral assessment and high stakes postgraduate medical examinations: The issue of validity, reliability and fairness

AUTHORS

Muhammed Ashraf Memon, MBBS, MA Clin Ed, DCH, FRACS, FRCSI, FRCSEd, FRCSEng[1,2,3,4]

Gordon Rowland Joughin, BA, BSW, DipTertEd, MEd, PhD[5]

Breda Memon, RGN, LLB, PGCEdu[6]

TITLE/INSTITUTIONS/DEPARTMENTS

[1]Consultant Surgeon, Department of Surgery, Ipswich Hospital, Queensland, Australia

[2]Associate Professor, Department of Surgery, The University of Queensland, Herston, Queensland, Australia

[3]Associate Professor, Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Queensland, Australia

Visiting Professor, Faculty of Health Sciences, Bolton University, Bolton, Lancashire, UK

[5]Senior Lecturer, Educational Development, Centre for Educational Development and Interactive Resources (CEDIR), NSW, Australia

[6]RN, Department of Surgery, Ipswich Hospital, Queensland, Australia

CORRESPONDENCE/REPRINTS

M. A. Memon, FRCS, Department of Surgery, Ipswich Hospital, Chelsmford Avenue, Ipswich, Queensland, Australia

Tel: +61 7 38101140, Fax: +61 7 38101592, Mobile: +61 448614170

Email: mmemon@yahoo.com

KEY WORDS

Oral assessment; Postgraduate medical examination; Validity; Reliability; Fairness

INTRODUCTION

Although oral assessment has a long history in the certification process of medical specialists, there remain concerns about its use. In recent years the validity, reliability and fairness of oral examinations have been questioned and they have been dropped from many postgraduate medical examinations[1-2]. This paper reviews the practice of oral examination in the context of the core assessment constructs of validity, reliability and fairness by: (a) delineating these constructs and noting the issues that arise in relation to them; (b) examining the issues that arise in respect to a number of core dimensions of oral assessment; and (c) reviewing specific research into oral assessment in postgraduate medical education with an emphasis on the core constructs and dimensions that highlight outstanding issues. While oral assessment is a common practice, it is clearly not unproblematic. One hopes that through the careful selection and training of examiners, the use of an examination blueprint, the development of clear questions and criteria for marking, continuing discussion of techniques and problems and the debate about equal opportunities and the possible discriminatory outcomes of the examination towards international medical graduates will help ensure that the validity, reliability and fairness conditions of oral examinations are met at the highest level on the part of examination authorities and examination panel members. The paper concludes by proposing 15 conditions under which oral assessment is valid, reliable and fair, and by emphasising matters that warrant urgent attention from professional bodies using oral assessment for certification of medical specialists.

WHY ASSESS ORALLY?

Oral assessment has been seen to have a number of particular benefits in higher education[3,4]. The rationale for its use in certification focuses on the following: (a) it mirrors the oral form of communication that dominates professional practice, (b) it can test the limits of a candidate's knowledge and understanding and (c) it is thought to be a particularly effective way of assessing cognitive processes[5]; interpersonal competence and intra-personal qualities. On the other hand the disadvantages of oral assessment are many and include: (a) it is resource intensive; (b) it involves making judgements on the basis of limited evidence; (c) in the case of appeals, it may involve justifying marks without written evidence; (d) it is usually an unfamiliar (and potentially difficult) form of assessment for examiner and examinee alike; (e) it may be difficult to distinguish between what a candidate says and how they say it (discourse); and (f) it can induce performance-inhibiting stress in some candidates[4-6].

Most importantly, oral assessment raises questions about three fundamental qualities that should adhere to any form of assessment:

(i) Assessment should explicitly focus on, and accurately measure, those attributes (knowledge, skills and values) that are requisite to professional practice, while excluding unrelated qualities (such as language skills that exceed the requirements of future practice).

(ii) A candidate's performance should depend on what they know and are able to do and on attitudes or values that are essential to work in the profession — it should not depend, or be influenced by, the particular person or persons examining them, the context of the examination, the particular questions they are asked, or any aspects unrelated to their actual capabilities.

(iii) It should go without saying that judgement of a candidate's performance should not be influenced by such factors as age, gender, race, or socio-economic status.

Wakeford[7] claims effective assessment must reflect truthfully some combination of an individual's abilities, achievement, skills and potential and be valid, reliable and fair. A clear understanding of these constructs is central to the proper use of oral assessment for certification purposes. Furthermore the assessment tool must also be feasible which in turn will achieve patient safety, improved training, trainee feedback, increase public confidence, and achieve accreditation and certification[8]. The Quality Assurance Agency (QAA)[9] Code of Practice states that all assessment must be explicit, equitable, valid and fair.

Assessment must be explicit. QAA use the term explicit (or transparent) to mean that all parties concerned in the assessment must be aware of what is being assessed and how it is being assessed.

Assessment must be equitable. The use of this term insists that the assessment must not unfairly advantage nor disadvantage any particular group or individual.

Assessment must be valid. The QAA use this term to mean appropriate. Others including Reznick[10] define validity as "whether we are measuring what we think we are measuring". Validity is assessed qualitatively and can be sub-divided into several sub-types. The readers can find the definition of these types of validity in any standard educational text and on-line.

Assessment must be reliable. Reliability is defined by the precision of the assessment tool; it is a quantifiable attribute that may be calculated mathematically. Reliability draws our attention to the 'error' that is inherent in measurement. Reliability relates to the question: "If the test were to be given on two separate days to the same individual and if there were no intervening changes or learning; to what extent would the results be identical?"[10]. Reliability is affected by (a) adequate test length; (b) intra-observer error; (c) inter-observer error; (d) inter-rater error and (e) inadequacies of test items[11].

Reliability becomes a key concern in oral assessment because such assessment often entails subjective judgements, different examiners, different sub-sets of questions being asked (often as 'follow-up' questions), examiners whose judgement may not have been firmly established through appropriate training and experience in this form of assessment, and assessment conducted at different sites. This has important implications, e.g. in borderline cases, measurement error may allow a competent candidate to fail and an incompetent candidate to pass.

Assessment must be fair. Fairness mean that candidates "of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership"[12]. Fairness entails both the absence of bias within the test and assessment processes that give all candidates an equal opportunity to demonstrate their "standing on the construct the test is intended to measure".

A possible, and potentially highly significant, source of unfairness can arise when a candidate is examined in what is not their first language (language discourse). Here,

fairness requires caution regarding the level of language expertise required. This should be commensurate with the construct being tested. If a higher level of language is required, the test's validity is under threat, since it becomes a test of linguistic ability that is not required by the construct. Where that construct is professional practice, it would be requiring a level of linguistic skill that exceeds what practice demands.
An "ideal assessment" must, therefore, be explicit, equitable, valid, fair and reliable but it must also be "easily understood by a range of trainers and trainees, not time consuming and easy to apply"[13].

DIMENSIONS OF ORAL ASSESSMENT
While validity, reliability and fairness are the overriding concerns in any assessment, reference to the six 'dimensions of oral assessment' identified by Joughin[14] — primary content type, interaction, authenticity, structure, examiners and orality — may help identify other aspects of oral assessment that are critical to postgraduate medical education.
Primary content type is concerned with the object of assessment. The two principal issues regarding content in oral assessment are (i) to ensure that oral assessment is used to assess those aspects of a candidate's performance that are best assessed using this medium[15-17] and (ii) that there is the utmost clarity regarding the role of the candidate's communicative ability in the process. Where communication and/or language ability is not being assessed, it must not be allowed to influence the examiners' judgement. Where it is being assessed, this needs to be made an explicit focus of assessment, with clear criteria and standards applied to the judgement process.

Interaction represents one of oral assessment's main strengths since it allows examiners to probe a candidate's reasoning, ethics and level of knowledge[16]. However, interaction also highlights the nature of oral assessment as an interpersonal event; it gives rise to the possibility that the social interaction entailed in this form of assessment may distort communication and affect both a candidate's performance and how that performance is perceived by the examiners[3].

Authenticity concerns the extent to which an examination reflects the context of professional practice. While it is true that most professional interactions occur orally, there is considerable disparity between professional interaction with a patient, colleague, or consultant, and interaction with a pair of experienced examiners meeting with the specific purpose of judging one's performance.

Structure refers to the extent to which oral assessment uses a predetermined, organised body of questions or sequence of events[17]. Maximum structure increases reliability, but limits interaction. Minimum structure raises concerns about both validity and reliability. Most well designed oral assessment allows for probing within carefully specified parameters.

Examiners - Oral assessment frequently involves more than one examiner — a small panel of two or more is not unusual — and examiners are often drawn from the field of practice. Because oral assessment is time intensive, significant numbers of examiners may be involved in a single assessment process. These factors give rise to a number of issues, including inter-examiner reliability noted above, the use of

examiners who are specialists in their field but not specialists in assessment, and the use of examiners who will often have a limited experience in this form of assessment. This creates, on the one hand, the need for acute awareness of inter- and intra-examiner reliability issues, and on the other, the need for thorough training of examiners.

Orality refers to the extent to which the assessment is conducted orally. Assessment can be purely oral or, as is often the case, it can be combined with other processes, including written papers or observed performance. Purely oral processes may limit the reliability of the assessment, if only because multiple forms of assessment are likely to increase reliability. Apart from the linguistic issues, oral communication can introduce some problematic factors, including its potentially 'agonistic' (or argumentative) tone, its more personal nature, and the intensified self-awareness candidates may experience[14,18].

ORAL ASSESSMENT IN POSTGRADUATE MEDICAL EXAMINATIONS

The literature on oral assessment in postgraduate medical education is not extensive. Hutchinson et al's meta-analysis[19] could locate only 55 articles on validity or reliability in relation to certification processes in general, though their analysis does not specify which of the processes they describe were oral. Their findings are relevant since they highlight the relative paucity of published material on validation of assessment for postgraduate medical education considering the influence such high stakes examinations have on doctors' career progression and employment opportunities. Their results showed that two crucial forms of validation, namely consequential validity, and construct validity, were missing from the majority of the papers. The authors conclude that only general or family practice has demonstrated that it is prepared to be thorough in developing an oral assessment instrument and pilot testing, with other specialties lacking such commitment.

Wakefield et al[20] noted a range of problems that can arise during oral examinations and errors that can occur in subsequently judging candidates' performances. The former include dysfunctional starts, difficulty in covering the ground fast enough and slowly spoken candidates. The latter include disagreements with co-examiners about grades, allowing first impressions to be overly influential, treating candidates "like themselves preferentially", and being influenced by a candidate's appearance. They proceed to describe the extensive training programme that is required to avoid these problems.

Roberts et al[21] provide an intensive consideration of language issues in the oral component of the Membership of the Royal College of General Practitioner's MRCGP examination in a study conducted with the support of the Royal College of General Practitioners. The study was based on a linguistic and interactional analysis of "awkward moments" — points at which examiners identified tension or poor communication between examiners and candidates — using videotaped examinations. The authors conclude that significant difficulties arise because of the different levels of discourse involved in the examination. Candidates for whom English is a second language may have difficulty recognising which type of discourse (i.e. personal, professional or institutional) is being called for by a question and in moving from one level of discourse to another. They conclude that "*the oral examination seems to assess candidates' professional discourse but does so through institutional discourse*

or a hybrid of all three discourses. This can lead to misunderstandings, mismatches, and cross purposes reinforced by the difficulty of managing any oral examination or selection interview where both time pressure and the social pressures of face to face interaction must be taken into account." They note that while this creates difficulties for all candidates, doctors from ethnic minorities who have trained overseas may have particular difficulties and that the potential for discrimination which arises from this needs to be addressed in the interests of reliability, fairness and social justice.

Wass et al[22] focused on reliability in another study supported by the RCGP. In a study of 896 candidates and 141 examiners, they used "generalisability theory" to examine the basis of variance in candidates' scores and to measure inter-case reliability, pass/fail decision reliability, and standard error of measurement. This study is of major importance since it not only highlights case specificity as the chief threat to reliability but quantifies the change that would be required to establish reliability in this instance. Thus an inter-case reliability coefficient of 0.65 when two examiners are used in each of two exams covering a total of 10 topics becomes 0.74 if the examiners operate singly, enabling four exams covering a total of 20 topics to be held. They conclude that "provided an adequate length of testing time is given and sufficient independent judgements are made on a wide range of topics, orals can be made psychometrically acceptable". The authors emphasize the need for (a) the introduction of standardised questions to improve inter-case reliability; (b) increased examiner training in equal opportunities and (c) clarity of question setting. These authors have shown that examiner performance contributes 27% variance seen in the oral examination and have emphasized the need of similar studies to be undertaken by institutions using oral examination in high stakes assessment.

Yaphe and Street[23] have researched how examiners actually make decisions within the MRCGP framework. According to their findings, examiners make a strong initial judgement based on the "first impressions" formed by the candidate's initial response. Further exploratory questions lead to them forming a "provisional grade", with other questions being asked to confirm this "(final) grade", what the authors refer to as a three-stage process. A further noteworthy finding of this study concerns the role of candidates' personal qualities. The authors noted that "candidates who were able to give a confident, fluent, articulate and comprehensive answer scored well", while candidates who scored poorly showed a "lack of coherence in answering questions, providing narrow, inflexible or superficial answers based on limited experience or failure to develop issues raised". They point out that such qualities as fluency and creativity (personal attributes) lie outside the construct being examined. Surprisingly enough some examiners were influenced by these personal attributes which may have an impact on the final grade of the candidate. The authors concluded that *"examiners can learn to focus on professional and ethical issues and not to be distracted by personal attributes, which have no relevance to the grading criteria"*. It is hoped that this study will have a direct impact on the further development of the oral component of the MRCGP examination.

THE CONDITIONS UNDER WHICH ORAL ASSESSMENT IS VALID, RELIABLE AND FAIR

The above studies highlight the complexity of oral assessment as an examination format and draw attention to the elements of assessment that require special attention when it is conducted orally. When the analytical reports of examination practice are

considered in the context of the two sets of constructs introduced at the beginning of this paper — validity, reliability and fairness, and the dimensions of oral assessment — we can begin to move towards a description of good practice in oral examinations for specialty certification. We therefore propose the following conditions under which oral examination is valid, reliable and fair.

Validity conditions

1. Examination items focus on the capabilities required for professional practice that are best assessed orally, namely clinical-reasoning and decision making.
2. The specific capabilities for professional practice are established by a representative group of practitioners. The content of the examination is determined by a panel of experts based on these capabilities.
3. Examination items are within the scope of professional practice.
4. Where language capabilities are examined, this is done explicitly and at the level required of professional practice.

Reliability conditions

5. An adequate sampling of questions are asked in order to provide sufficient coverage of the depth and breadth of practice and to ensure inter-item variability is at an acceptable level.
6. Examiners are formally trained in oral examination issues and methods.
7. Inter-examiner variations are monitored. Discrepancies are addressed.
8. Items and implementation processes are standardised.
9. Statistical methods are used to establish and monitor reliability.

Fairness conditions

10. Consideration of bias is recognised by administrators as an essential element of good examination practice.
11. Examination items are scrutinised by a representative panel to detect item bias.
12. Result patterns are monitored to identify differential response levels from identifiable sub-groups.
13. Examinations are designed to minimise threats to their validity and reliability due to language differences of candidates.
14. The language ability required in an examination should be commensurate with that required by professional practice.
15. Where systematically lower or higher scores for particular groups of examinees occur, the possibility of bias should be considered.

CONCLUSIONS

Ensuring that the validity, reliability and fairness conditions of oral examinations are met calls for a high level of professional practice on the part of examination authorities and examination panel members. Given what is at stake in terms of the professional future of candidates and the well-being of patients, nothing less than this is acceptable. It is clear from the empirical studies cited above and from the issues associated with the core constructs of oral assessment that valid, reliable, and fair oral examinations on a large scale are not easy to design and implement. To do this requires time, resources, expertise, and ongoing research, as well as commitment. It is also clear from the studies cited above that, while the complexities of assessment have been recognised to a significant extent by the Royal College of General Practitioners, this is not the case in relation to most specialist training. Although summative

assessment procedures for the award of certificates of specialist training have been in operation for some time now, issues of validity, reliability and fairness do not seem to have been systematically addressed. Given the nature of educational assessment noted above, it is likely that significant issues concerning validity, reliability and fairness remain to be identified, and addressed, in many fields of speciality examination. One area that clearly warrants further study is the performance of international medical graduates for whom English is not a first language. Such candidates may be disadvantaged through bias towards more fluent, articulate candidates, and through difficulties in managing unfamiliar movement between discourse types within the examination.

Supporting high quality published research into examination practices and outcomes, and acting on the findings of such research, may serve to allay concerns about the transparency and fairness of these examinations, especially where the failure rates for identifiable sub-groups of candidates has been significantly higher than for other candidates. Other Royal Colleges (or equivalent) may do well to follow the practice set by the Royal College of General Practitioners in producing substantive evidence that their examination processes are fair and just, especially when it comes to assessing international medical graduates.

REFERENCES

1. Spike, N., Jolly, B. (2003). Are orals worth talking about? Med Educ, 37, 92-93.

2. Cunnington, J. P., Hanna, E., Turnhbull, J., Kaigas, T. B., Norman, G. R. (1997). Defensible assessment of the competency of the practicing physician. Acad Med, 72, 9-12.

3. Abrahamson, S. (1983). The oral examination: the case for and the case against. In J.S. Loyd & D.G. Langsley (Eds), Evaluating the skills of medical specialists. Chicago: American Board of Medical Specialists (pp.121-124).

4. Habeshaw, S., Habeshaw, T., Gibbs, G. (1994). 53 Interesting ways to assess your students (3rd Ed), Technical & Educational Services Ltd, Bristol.

5. Eraut, M., Cole, G. (1993). Assessment of competence in higher level occupations. Competence and Assesment, 21, 10-14.

6. Joughin, G., Collom, G. (2003). Oral assessment. Biomedical Scientist, 47, 1078.

7. Wakeford R. Principles of Student Assessment. In: Fry H, Ketteridge S, Marshall S (2nd ed). A Handbook for Teaching & Learning in Higher Education. Routledge-Falmer, Oxon, pp 42-61.

8. Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Assessment of technical surgical skills. Eur J Surg 2002; 168: 139-144.

9. QAA: The Quality Assurance Agency for Higher Education: Code of practice for the assurance of academic quality and standards in higher education – Section 6:

Assessment of students 2000, http://www.qaa.ac.uk/academicinfrastructure/codeof practice/section6/COP_AOS.pdf

10. Reznick RK. Testing and Teaching Surgical Skills. Am J Surg 1993; 165: 358-361.

11. American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCIME). Standards for educational and psychological testing. American Educational Research Association 1999, Washington, DC.

12. Pitts D, Rowley DI, Sher JL. Assessment of Performance in Orthopaedic Training. J Bone Joint Surg 2005; 87-B: 1187-1191.

13. Joughin, G. (1999). Dimensions of oral assessment and student approaches to learning. In S. Brown & A. Glasner (Eds), Assessment matters. pp 146-156. Buckingham: The Society for Research into Higher Education & Open University Press.

14. Levine, H. G., McGuire, C. H. (1970). The validity and reliability of oral examinations in assessing cognitive skills in medicine. J Educ Meas, 7, 63-74.

15. Lunz, M. E., Stahl, J. A. (1993). Impact of examiners on candidate score: An introduction to the use of multifacet Rasch model analysis for oral examination. Teach Learn Med, 5, 174-181.

16. Muzzin, L. J., Hart, L. (1985). Oral examinations. In V Neufeld & GR Norman (Eds), Assessing Clinical Competence (pp 71-93). New York: Springer.

17. Ong, W. J. (1982/2002). Orality and Literacy (pp 204). London: Routledge.

18. Hutchinson, L., Aitken, P., Hayes, T. (2002). Are medical postgraduate certification processes valid? A systematic review of the published evidence. Med Educ, 36, 73-91.

19. Wakefield, R., Southgate, L., Wass, V. (1995). Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP. B Med J, 311, 931-935.

20. Roberts, C., Sarangi, S., Southgate, L., Wakeford, R., Wass, V. (2000). Oral examinations-equal opportunities, ethnicity, and fairness in the MRCGP. Br Med J, 320, 370-375.

21. Wass, V., Wakeford, R., Neighbour, R., Van der Vleuten, C. (2003). Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. Med Educ, 37, 126-131.

22. Yaphe, J., Street, S (2003). How do examiners decide?: a qualitative study of the process of decision making in the oral examination component of the MRCGP examination. Med Edu, 37, 764-771.