Rural and Remote Health
FRAME
ARHEN
ISSN 1445-6354
The International Electronic Journal of Rural and Remote Health Research, Education, Practice and Policy

ORIGINAL RESEARCH

'Lost in translation'?: Developing assessment criteria that value rural practice

EJ Bell[1], G MacCarrick[2], L Parker[2], R Allen[3]

[1]*University Department of Rural Health, University of Tasmania, Hobart, Tasmania, Australia*

[2]*School of Medicine, University of Tasmania, Hobart, Tasmania, Australia*

[3]*Tasmanian Qualifications Authority, Hobart, Tasmania, Australia*

Submitted: *17 March 2005;* ***Revised:*** *19 May 2005;* ***Published:*** *22 July 2005*

Bell EJ, MacCarrick G, Parker L, Allen R
'Lost in translation'?: Developing assessment criteria that value rural practice
Rural and Remote Health **5: 420. (Online), 2005**

Available from: http://rrh.deakin.edu.au

A B S T R A C T

Introduction: Rural workforce preparation is often discussed in terms of specific interventions such as rural placements. More technical discussions of education matters seem to belong in the realm of education experts. However, this issues article argues that a focus on quality assessment techniques is important to the rural health agenda. Making connections between the medical education literature and the broader education literature, it explores elements of a qualitative decision-making model as an alternative to narrow competency-based and norm-referenced approaches. In the process it explores assessment techniques that may help educators better translate their intentions to value rural practice into the learning of students. Background: Research suggests that, in Australia at least, many university educators have different and conflicting understanding of assessment criteria. At the same time, the literature on the development of assessment criteria is relatively small in a context in which the medical education literature takes a quantitative, reliability-driven approach. This has important implications for how we ensure that rural practice is given enough emphasis at the level of education that most strongly drives student learning - assessment.

Methods: This article explores such matters by examining the steps needed to develop assessment criteria in undergraduate medical education courses. It draws on key writings from the past, as well as current debates, in the medical education and broader education literature. It focuses on the detail of assessment techniques to show how the intention to value rural practice can be 'lost in translation' with narrow norm-referenced and competency-based assessment models.

Conclusions: Rural health has a stake in technical debates about education in health sciences courses. Like other knowledge and skills, the knowledge and skills important to rural practice cannot be valued at the coalface of student learning if our assessment techniques subvert intentions. Developing the quality of assessment techniques involves scrutiny of not only the medical education

literature, but also the broader education literature, including writings about working models of criteria-and-standards-based assessment. This scrutiny suggests assessment techniques are not equal in terms of how well they translate intentions. More than that, it suggests the value to rural health education of shifting from narrow norm-referenced models to best practice in criteria-and-standards-based assessment.

Key words: assessment criteria, learning, medical education.

Introduction

Sound preparation for rural practice is about more than providing students with opportunities for rural placements[1]. It is about the knowledge and skills valued in health sciences education courses. Specifically, it is about how well assessment practices[2] in these courses translate intentions to value rural practice.

However, it seems that, in Australia at least, university educators have conflicting understandings of assessment criteria[3]. At the same time, the literature exploring specific features of assessment criteria, including how to write them, is relatively small. The medical education literature emphasises a quantitative, reliability-driven approach[4]. All this raises questions about how to ensure that knowledge and skills relevant to rural practice are given enough emphasis at the level that most strongly drives student learning - assessment. Are there key junctions where the intention to value rural practice can be 'lost in translation'? This article explores this question by weaving together select writings from the medical education literature and the broader education literature, past and present. It looks at the key steps in developing assessment criteria:

- writing course objectives
- writing the assessment plan
- defining standards
- deciding overall achievement and the rules for progression
- developing assessment instruments
- writing assessment criteria.

Methods

Course objectives

Rural health education has something to gain from clear definitions of objectives. Without assessment instruments guided by overarching curriculum objectives, assessment can hide implicit objectives (and values and priorities) 'nested' in instruments of measurement to create a 'hidden curriculum'[5,6]. Clear statements of objectives are the foundations for valid assessment programs that go beyond traditional clinical models of medical education[4,7,8].

Essentially, these objectives are global statements of 'knowledges and skills'[6]. This term is often associated with a style of old-fashioned global judgment by a preceptor making inferences about latent traits. In contrast, 'competency' is often positioned as a focus upon the evidence of performance or features of student work. However, the distinction is somewhat bogus. There is always an element of inference in assessment of non-trivial skills, such as those associated with medicine in general and rural practice in particular.

While the primary objectives or macro knowledges and skills[6] important to a medical education course can be defined using different taxonomies[9-11], the focus should be on whether the broad groupings systematically reflect the critical elements in professional practice.

Rural practice adds distinct elements to this total set, not simply particular emphases or contexts: different ways of handling patient care, undertaking teamwork and

communication, as well as resource and self-management skills[12-15]. This is not to make a claim for a 'rural' bowel cancer, but to observe that attending to someone with bowel cancer when the nearest hospital is a very long way away will require different skills of a GP[12].

The existence of distinctive research literature, organisations for rural health and rural practice, and rural medical training programs suggest there may be a distinct discipline of 'rural medicine'[12]. This issues a challenge to develop the appropriate emphasis on rural practice in medical courses. Designing courses that help meet this challenge involves 'developing a hierarchy downwards by specifying meanings for a primary criterion' as part of an iterative process of 'interpretation and elaboration' where 'each lower level amplifies the level immediately above it'[16]. In other words, using models for developing objectives from professional practice[17], under each primary objective, course developers can include broad-brush descriptions of the secondary objectives or micro knowledge and skills that are later developed into detailed schemes of assessment criteria for each assessment instrument.

The assessment plan

The assessment plan specifies the assessment instruments (eg multiple choice, portfolios, simulated patient situations), their use (formative or summative), any remedial action and special consideration and a 'blueprint'[10] linking assessment to its explicit purposes. A blueprint is a matrix with the generic 'competencies' on one axis and, on the other, the information about the assessment instruments or items that will be used to assess each of those competencies[18,19]. It is an important opportunity for educators to audit how rural practice is valued in assessment.

Developing the assessment plan involves defining the standards, deciding how to arrive at an overall result for each student, and specifying minimum acceptable standards for course progression.

Defining standards

Medical education literature includes reference to a standard as 'a special score that serves as a boundary between those who perform well enough and those who do not'[20]. However, every symbol used to discriminate one student's achievement from another represents a standard, explicit or implicit.

These tell an important story about the assessment ethos of a course and suggest ways in which rural workforce preparation can be 'lost in translation'. For example, numerical marks have a long history of providing convenient representations of judgments, but their association with objectivity and precision is more chimera than reality[21]. In contrast, competency-based assessment has a history of being industry-driven and defined without the involvement of educators, which may have left it without a firm base in education theory and practice[22,23]. There is a tension between the dichotomy of 'competent' and 'not [yet] competent', and ideas about competency as development along a continuum, often associated with Rasch[24] and, ultimately, Glaser[25]. The underlying issue is that if competence is an ability (distributed from a little to lots) then having one standard (competent) requires a line where 'not quite enough' becomes 'just about enough', leaving misclassified some of those above the cut and some of those below. To extend the earlier point about the inferential nature of assessment, the information we have—about a student's performance in particular situations—is an estimate of the ability/competence we are interested in. Where an estimate of a scale is cut into categories, the more categories there are the greater the chance of misclassification, but the less serious or misleading the misclassification is likely to be. The discussion about deciding overall achievement will explore why, when there are only two categories, there are fewer misclassifications but the consequences are more misleading.

Accordingly, there is evidence that standards-referenced discriminations along some finer rating scale than simply 'competent/not competent' (provided there are not too many ratings), are not only more precise, but provide richer

information - a developmental framework for communication for learner and teacher[26]. Certainly, research into competency-based education has called for a movement away from 'yes/no' atomistic checklists and toward integrated assessment of more global job functions involving some kind of grading[23].

We suggest the use of a form of standards-referencing involving bands or levels defined along achievement continua. The descriptions of these levels (labelled with letter grades) would then be definitions of thresholds along developmental continua[27]. This approach to defining standards aligns with broader education literature. More than that, it may better serve efforts to provide learning-rich programs that value the complex multi-disciplinary skills associated with rural practice.

Deciding overall achievement and the rules for progression

The aggregation of data across different assessment instruments is another step in which rural health education can be lost in translation. Different aggregation methods can have an effect great enough to suggest that the final results obtained are less about student achievement and more an artefact of the method used[5].

Medical education literature suggests that numerical assessments should not be first combined into a single result, instead basing decisions about overall achievement on some more informative profile[28]. This is supported by arguments in the education literature that any kind of aggregation rule involving summing raw, scaled, or weighted scores carries the usually wrong assumption 'that all possible combinations of marks which give rise to the same total represent equivalent achievements and are therefore equally valuable educationally'[21].

Medical education literature includes recognition that norm-referencing (deciding standards on the basis of cuts applied to a distribution of marks on the basis of some standard distribution, without reference to explicit criteria) produces results that are a poor basis for intelligent aggregation. Norm-referencing is considered unacceptable for clinical competency tests because it ranks candidates, rather than giving any sense whether the pass mark really does separate clinical competence from incompetence[19]. The UK General Medical Council's Performance Procedures seem to avoid the use of numerical rules for considering a large amount of information, preferring instead to make overall judgements using 'principles of triangulation' of evidence[29]. While numerical approaches to deciding 'pass/fail' standards are common in the health sciences, absolute standards are increasingly being associated with the gold standard[5]. The decision about where to draw the line between adequate and inadequate overall performance is described, by at least one medical education leader, as being about finding a systematic means of gathering value judgments to find a consensus (of which the pass score is considered an expression)[20].

Yet the medical education literature is full of highly technical discussions about how to aggregate results[5,28,30,31]. There is, for example, Angoff's method, which involves using averages of judges' estimates of what percentage of hypothetical borderline candidates will respond correctly to an item[20]. More crudely, the decision about who should pass and fail has sometimes involved defining 'the cut-off score at the mean minus the standard deviation of test scores'[5]. Both approaches assume that assessment is essentially unidimensional, measuring a single well-defined construct. However, while BMI may well be essentially unidimensional, complex knowledge and skills, such as those associated with rural practice, are not.

Nor does a set of decision-making rules applied to competency-based assessment represent a much better model for translating intentions. Where a decision about overall competence is based on a series of judgments, even if the chance of each individual judgment being wrong is small, the chance of the set including at least one wrong judgment is high. While the analogous issue seems reasonably well known in hypothesis testing, it seems to be rarely acknowledged in competency-based assessment, where a competent person is someone judged as competent in every respect, and a non-

competent person is someone judged as non-competent at least once. When there are several judgments, we are likely to accept as competent a person who is not (at least one of the decisions was wrong) and we may reject as not competent a person who is (the single not-competent decision was wrong). A thousand kilometres away from the nearest hospital that may matter even more than it does in a well-serviced city.

A set of rules for managing criterion-referenced letter grades can avoid some of these problems. Suppose standards-referenced letter grades awarded for four assessment instruments had to be combined for each student. An overall high distinction or HD might be defined as at least two HD and not more than one result at credit or C and below. An overall pass might be defined as at least a C for a key assessment instrument and no lower than a pass for two other instruments. Such decision-making models can also include 'trade-offs' (eg a P on one particular instrument and a HD on another will be considered equivalent to two Cs). Trade-offs allow a superior performance on one task to compensate (to some carefully considered limited extent) for a lower performance on another. While trade-offs should not be used where the two tasks are completely disparate and both are essential, they provide a means of taking into account both the uncertainty involved in the judgments and the fact that a (non-trivial) assessment task draws on more than a single dimension of performance. Unlike the use of numerical means, where a very poor score can be compensated for by a very high score on something else, trade-offs have the strength that they are not fully compensatory.

Such decision-making models for deriving overall judgments are the basis for working models of criteria-and-standards-based assessment[11,32] that give educators better opportunities to translate their intentions into the fine-grained detail of assessment practices.

Developing assessment instruments

In the medical education literature at least, it is well-recognised that purpose must drive every aspect of assessment design, and that assessment instruments should be developed to integrate different secondary objectives (or their equivalents)[10,18,33]. However, it may be that intentions to value rural practice can be lost in translation at another level - the design features of assessment instruments.

The medical education literature suggests why this might be so. It is because certain kinds of assessment have come to be rather mechanically associated with certain goals of education[34]. This process seems to have begun with a movement away from a traditional, content-knowledge-driven, decontextualised model of medical education, to a 'real world' problem-solving model closer to clinical practice. Instruments such as multiple choice questions, essays and orals tend to be associated with the former, less 'authentic' education model[19,33-35].

Not surprisingly then, the medical education literature suggests a strong recognition that educators should use a range of assessment instruments[19,36,37]. Yet, to the extent that educators assume an automatic match of an assessment genre and a particular assessment goal, educators may not look closely at the specific design features of assessment instruments in each genre. Having problem-based learning does not automatically confer validity or even authenticity. Neither does having portfolio assessment[34]. What matters most for student learning is the quality, fitness-for-purpose, and authenticity of the assessment instrument[34] and its substantive content[37]. If an instrument is meant to be assessing clinical reasoning, the design features of the instrument should reflect what is known about the way clinical reasoning works in practice and how to assess it[36,38]. Without a design-features approach to assessment instruments we are at risk of having appearance without substance. Without that substance, rural medicine is less effectively translated into students' learning experiences.

Design features include the conditions of assessment, the language of the instrument (such as the purposeful use of specialist language), the layout and presentation of the instrument, and so on[39].

A design-features approach to developing assessment instruments does more than help assessment reflect curriculum intentions. It can help educators capitalise on the reality that assessment drives student learning[34].

This focus upon validity need not reject the best the medical education literature offers about reliability. Obviously we want reproducible judgments that are not mere artefacts of who is making the assessment judgment, the particular items selected, the day on which a test was held, and so on. The advice is to sample student performance adequately and use different examiners for high stakes results[19].

Developing an assessment scheme that uses criteria-based assessment

Finally, rural health education can be lost in translation at the level of elaborating the micro knowledges and skills into terms that are useable in the assessment scheme or what medical education literature calls the 'answer key'[37]. The assessment scheme can make the assessment instrument work like a mirror reflecting back to learners their strengths and weaknesses.

This last ideal, common in the education literature, recalls the French philosopher Foucault's analysis[40] of techniques for regulating and shaping the self in modern society. Feedback - what is said, when it said and how is said - has long been known to have a powerful shaping influence on the development of the practitioner as member of the profession[41]. That is, the self reflected back to the student through the assessment criteria and their application to student performances may be part of the way health sciences education develops professionals who are more (or less) compatible with rural practice.

In the criteria-and-standards-based approach explored in this article, an assessment scheme takes the form of a grid. The list of knowledges and skills developed from the secondary objectives will be in the first column and the letter grades in the first row. The descriptors for each grade are then produced.

Such descriptors should not be couched in abstract terms such as 'good' or 'outstanding'. They are not descriptions of the reactions of an assessor to a piece of work or rules for the award of results, as has been observed in university assessment criteria[3]. They are not generic descriptions applied indiscriminately across different assessment situations. Nor should they take the form of fragmented checkbox lists of competencies[10,35,42,43]. As experience with the Objective Structured Clinical Examination suggests, detailed checklists can sacrifice validity for reliability at the individual item level, the important and complex at the expense of the trivial, and the generic at the expense of the contextual[18,43]. Consequently, some have argued for a combined approach: using checklists for more practical and technical skills, and global ratings for more complex skills such as communication and diagnostic problem-solving[18].

The descriptors describe for each possible grade the point-at-able features of student work, developed from considering, among other things, what an exemplar student response looks like. They are about what students must do, not what they must not do, to achieve a particular grade[6]. Some will need to be sufficiently detailed to help educators make sound judgements about technical clinical skills, others will need to be broad enough to allow global judgments of complex multi-faceted constructs, such as problem-solving.

Good descriptors will help the educator provide clear feedback to students[44] about the strengths and weaknesses of their performance. Descriptors of a particular grade account for the wide range of possible student responses, and recognise that essentially the same achievement can often be demonstrated in different ways, without affecting the substantive validity of the assessment judgment[6].

The emphasis in the criteria-and-standards approach on validity at the broad brush and fine detail levels, expressed in multi-dimensional assessment tools[32], is one key reason this is compatible with the fundamental aim of rural workforce preparation to produce well-rounded professionals.

Conclusions

Valuing rural practice at the coalface of student learning is about more than having good intentions. It is about effectively translating intentions into the assessment practices that drive student learning. The knowledges and skills important to rural practice cannot be effectively valued in student learning if assessment techniques subvert intentions. Ensuring that this does not happen involves, at every stage of developing assessment, attention to how different approaches work, or don't work, to translate intentions. This article has scrutinised the development of assessment criteria to show that rural health education has a stake in more technical debates about the quality of education practices. It has emphasised the value to rural health education of shifting from narrow norm-referenced and competency-based assessment to best practice in criteria-and-standards-based assessment. High quality rural health education requires an engagement with not only the medical education literature, but also the broader education literature, including writings about working models of criteria-and-standards-based assessment, wherever they may be found.

References

1. Wilkinson D, Laven G, Pratt N, Beilby J. Impact of undergraduate and postgraduate rural training, and medical school entry criteria on rural practice among Australian general practitioners: national study of 2414 doctors. *Medical Education* 2003; **37:** 809.

2. Newble D. Assessment of clinical competence. *British Journal of Anaesthesia* 2000; **84:** 432-433. (Editorial)

3. Barrie S, Brew B, McCulloch M. *Qualitatively different conceptions of criteria used to assess student learning*. (Online) 1999. Available: http://www.aare.edu.au/99pap/bre99209.htm (Accessed 4 October 2004).

4. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Medical Education* 2002; **36:** 972-978.

5. Muijtjens A, Hoogenboom R, Verwijnen G, Van der Vleuten C. Relative or absolute standards in assessing medical knowledge using progress tests. *Advances in Health Sciences Education* 1998; **3:** 81-87.

6. Bell E, Dixon S. *Developing and using sound criteria-based assessment methods: A workshop resource for university teachers*. Brisbane: Queensland Board of Senior Secondary School Studies, 2001.

7. Hudson J, Tonkin A. Evaluating the impact of moving from discipline-based to integrated assessment. *Medical Education* 2004; **38:** 832.

8. Hall J. Assessment dispersion matrices. *Medical Education* 2001; **35:** 345-347.

9. Bloom B. *Taxonomy of Educational Objectives*. London: Longman, 1956.

10. Crossley J, Humphris G, Jolly B. Assessing health professionals. *Medical Education* 2002; **36:** 800-804.

11. Pitman J DR. *Criteria-based assessment: the Queensland experience*. Brisbane: Queensland Board of Senior Secondary School Studies, 1998.

12. Smith J, Hays R, Sen Gupta T. Is rural medicine a separate discipline? *Australian Journal of Rural Health* 2004; **12:** 67-72.

13. Kenny A, Duckett S. Educating for rural nursing practice. *Journal of Advanced Nursing* 2003; **44:** 613-622.

14. Britt H, Miller G, Valenti L. *It's different in the bush: a comparison of general practice activity in metropolitan and rural area of Australia 1998-2000*. Sydney: Australian Institute of Health and Welfare, 2001.

15. Hays R, Sen Gupta T. Ruralising medical curricula: the importance of context in problem design. *Australian Journal of Rural Health* 2003; **11:** 15-17.

16. Anon. *General principles for organising criteria*. Assessment Unit discussion papers: discussion paper 9: Review of school-based assessment (ROSBA). Brisbane, Queensland: Queensland Studies Authority. (Online) 1985-1987. Available: http://www.qsa.qld.edu.au/publications/yrs11_12/assessment/rosba009.pdf (Accessed 15 October 2004).

17. Kristina T, Majoor G, van der Vleuten C. Defining generic objectives for community-based education in undergraduate medical programmes. *Medical Education* 2004; **38:** 510-521.

18. Newble D. Techniques for measuring clinical competence. *Medical Education* 2004; **38:** 199-203.

19. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *The Lancet* 2001; **357:** 945-949.

20. Norcini J. Setting standards on educational tests. *Medical Education* 2003; **37:** 464-469.

21. Sadler R. *The place of numerical marks in criteria-based assessment*. Assessment Unit Discussion Papers: Review of school-based assessment (ROSBA). Brisbane: Queensland Studies Authority. (Online) 1985-1987. Available: http://www.qsa.qld.edu.au/yrs11_12/assessment/discuss.html (Accessed 8 October 2004).

22. Pitman JBE, Fyfe I. *Assumptions and origins of competency-based assessment: new challenges for teachers*. Brisbane: Queensland Board of Senior Secondary School Studies, 2000.

23. Gonczi A. Reconceptualising competency-based education and training: with particular reference to education for occupations in Australia (PhD Thesis). Sydney: University of Technology; 1996.

24. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press, 1980.

25. Glaser R. The future of testing: a research agenda for cognitive psychology and psychometrics. *American Psychologist* 1981; **36:** 923-936.

26. Connally J, Jorgensen K, Gillis S, Griffin P. *An integrated approach to the assessment of higher order competencies*. Australian Association for Educational Research. (Online) 2002. Available: http://www.aare.edu.au/02pap/con02630.htm (Accessed 14 October 2004).

27. Griffin P. *The comfort of competence and the uncertainty of assessment*. Hong Kong Institute of Education. (Online) 2004. Available: http://www.ied.edu.hk/cric/new/principalconference/papers/c1-griffin-hkprincipal%20conference%20griffin.pdf (Accessed 14 October 2004).

28. Fowell S, Jolly B. Combining marks, scores, and grades. Reviewing common practices reveals some bad habits. *Medical Education* 2000; **34:** 785-786.

29. Southgate L, Cox J, David T et al. The General Medical Council's performance procedures: peer review of performance in the workplace. *Medical Education* 2001; **35**(1): S9-S19.

30. Wilkinson T, Newble D, Frampton C. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. *Medical Education* 2001; **35:** 1043-1049.

31. Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and empirical standard setting procedure for an OSCE. *Medical Education* 2003; **37:** 132-139.

32. Pitman J, O'Brien J, McCollow J. *High-quality assessment: we are what we believe and do*. Brisbane: Queensland Board of Senior Secondary School Studies, 1999.

33. van Luijk S, van der Vleuten C. Assessment in problem-based learning (PBL).*Annals Academy of Medicine* 2001; **30:** 347-352.

34. Schuwirth L, van der Vleuten C. Changing education, changing assessment, changing research. *Medical Education* 2004; **38:** 805-812.

35. Crossley J, Howe A, Newble D, Jolly B, Davies H. Sheffield Assessment Instrument for Letters (SAIL): performance assessment using outpatient letters. *Medical Education* 2001; **35:** 1115-1124.

36. van der Vleuten C, Newble D. How can we test clinical reasoning? *The Lancet* 1995; **345:** 1032-1034.

37. Schuwirth L, van der Vleuten C. ABC of learning and teaching in medicine: written assessment. *BMJ* 2003; **326:** 643-645.

38. Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine* 2002; **14:** 150-156.

39. Allen J, Bell E. Guidelines for Assessment Quality and Equity. In: *Queensland Board of Senior Secondary School Studies*. Brisbane, Queensland: Queensland Board of Senior Secondary School Studies, 1995.

40. Foucault M. *Discipline and punish: the birth of the prison*. 2nd edn. New York: Random House, 1995.

41. Gamble J. Modelling the invisible: the pedagogy of craft apprenticeship. *Studies in Continuing Education* 2001; **23:** 185-200.

42. Queensland Board of Senior Secondary School Studies. *General principles for organising criteria*. Assessment Unit papers: discussion paper 9. Brisbane: Queensland Board of Senior Secondary School Studies. Available: http://www.qsa.qld.edu.au/publications/yrs11_12/assessment/rosba009.pdf (Accessed 15 October 2004).

43. Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine* 1998; **73:** 993-997.

44. Ende J. Feedback in clinical medical education. *JAMA* 1983; **250:** 777-781.