Chapter 2.2.5

ASSESSMENT IN RURAL MEDICAL EDUCATION

Tim Wilkinson
University of Otago, New Zealand

Principles of good assessment

Good assessment is good assessment regardless of whether it happens in rural settings or elsewhere. Most of this chapter therefore covers principles that apply in any setting. However, there are some situations that require special consideration in rural settings and these are discussed as they arise, and near the end of this chapter.

The best plans for the best learning course can all be easily undone if the associated assessment is not well planned. It is well known that assessment drives learning and sometimes this is seen as a negative phenomenon, as though it were somehow the fault of learners. But assessment guides learning and helps learners get to the next stage of their course – so naturally they will guide their learning to the assessments. If assessment is guiding learning in the wrong direction, it is the fault of the assessment, not the learners.

One of the most common pitfalls in designing an assessment system is to think about it too late. Sometimes we see a lot of thought being put into the learning programme and then at the last hour, someone asks the question, what assessment will we have? There can then be a tendency to think of familiar assessment tools, and decide to use them: "We'll have an MCQ test and that should sort it". Instead, assessment should be integral to any thinking about learning and particularly integral to thinking about the goals of the learning programme.

Decide the purpose

Before we think about an assessment programme, one of the most important questions to answer is why are you assessing? Determining the purpose of assessment may seem a simple question and one that need not be worried about too much, but all too often it is given insufficient thought and many problems can ensue from this.

One of the most common reasons for assessing is to guide learning. It is to help learners understand where their strengths and weaknesses lie and hopefully therefore to guide them to those gaps that they need to attend to. However other purposes of assessment can be to inform decisions on progression – particularly whether a learner is ready to proceed to the next stage of the course or to graduate from the course. Assessment can also be used as a form of course evaluation – the faculty can learn how well the course is working by looking at those areas of the course that learners seem to have learnt well and those areas that they seem to struggle with.

Another common purpose of assessment is to rank learners. Ranking is not an end in itself but can be needed if there is competition for limited places in the next stage of a programme – e.g. to select people for places in medical school or for a job where these are scarce and applicants are plentiful. Another purpose could be to motivate learners. Some mistakenly think that ranking is also useful to motivate learners yet there is good evidence that ranking does not achieve this goal (1). However, having clear expectations and clear feedback about how learners are achieving against these expectations can be highly motivating.

It is clear that some of these purposes may conflict with one another. For example if the assessment is designed to rank, then there is the risk it may demotivate. If the assessment is to inform decisions about progression, there may be less information available to guide learning. It is hard to design an assessment system to meet all these purposes. Choices need to be made about those purposes that are most important for the learning programme. Once those choices have been made, and an assessment system designed, we should not be disappointed if other purposes are not being met.

It cannot be overstressed how important it is to have a written statement about the purpose of the assessments. While they may differ according to the stage of the course, having them stated clearly goes a long way to informing subsequent actions in the design of an assessment system.

Educational impact: Decide what learning you value

The observation that assessment drives learning has sometimes been seen as a negative phenomenon, such as when learners must make a choice between preparation for assessment rather than to provide good care for patients. This negative effect only occurs when assessments are not aligned to what we value. If our assessments drive learners away from activities that we think are important, then it is our problem to solve, not the learners'. If we only assess the things that we value, then the learning behaviours should move in the desired direction (2). The effect of assessments on learning is referred to as its educational impact and, arguably, is the most important attribute of an assessment programme.

Generally we value assessments that reflect an ability to do the job well. This is performance ('does do') and contrasts with competence ('can do') (3). Although performance is the 'final common pathway' that encompasses competence, there are times when competence first needs to be shown before a person can be permitted to perform (4). This value on performance over competence has driven moves to place more assessment within the workplace.

Reliability and validity

We cannot consider assessment without clarifying reliability and validity. Good assessments are reliable and valid. However, it is also worth noting that very few assessment tools are both reliable and valid – at least not if they're trying to measure something important. To explain this further, we need to understand what these concepts are.

Reliability refers to the reproducibility of an assessment. If we were to do the assessment again, would a learner get a similar result; if the assessment were to be assessed by a different examiner, would the learner get a similar result? Ideally the answer should be yes in both cases. ***Validity*** refers to the extent to which an assessment is measuring what it's intended to measure.

Reliability can be reported numerically – often between 0 and 1. As such it can sometimes be likened to a correlation coefficient where '1' indicates complete agreement or reproducibility and '0' indicates no reproducibility or that the result is

random. In contrast, validity is expressed in words and relates to purpose. There can be various forms of validity – but common ones are how much the assessment matches real-life tasks; how much the assessment is able to detect those learners who are below the expected standard from those who are above it; how much the assessment fails the right learners and passes the right learners.

The challenge in assessment, and a common observation within medical programmes, is that it can be hard to get assessment tools that are both reliable and valid. A supervisor's observation of a learner in action in the workplace often has high validity in terms of measuring how well that learner might undertake future work in the workplace, but the reliability may not be that high; another supervisor may give a different rating, for different reasons. In marked contrast, shoe size can often be measured very reliably – the result may well be the same if it is measured again the next day or by another person. But it has very low validity in informing how someone may work as a doctor. Although shoe size is an extreme example, and would never be used in practice for such purposes, there is the risk that we spend so much time justifying assessment tools purely on the grounds of reliability that we overlook validity. It is almost as if the test designer is thinking that if it's reliable it must be good. Of course, there is a good argument to be made that no test can be valid unless it is reliable – you may be measuring the right thing but if the result is not reproducible then it may be of little use.

These tensions have characterised assessments for many years and have led to some unfortunate decisions. It can lead to excessive attention being paid to individual assessment tools. Multiple choice question (MCQ) banks and Objective Structured Clinical Examinations (OSCEs) are good examples of this (see Appendix A). In order to achieve acceptable reliability, there is the risk that the task being asked of the learners becomes so circumscribed and defined that it becomes meaningless.

This has been supported by the observation that checklists not only do not improve reliability but may worsen validity. For example, scoring on checklists can lead to experts getting lower scores than novices (5). This reflects the efficiency by which experts undertake tasks. Experts do not need to undertake all the steps (as might be captured by a checklist) in order to achieve an accurate and efficient outcome (5,6,7).

Multiple sampling

If specification of the task and checklists don't improve reliability, then what does? In short, the best way to improve reliability is multiple sampling. Assessment is like statistical sampling – a sample is an approximation of what we want to know. The bigger the sample, or the more times we sample, the closer we get to an accurate understanding of an individual. It is difficult for one method of assessment to provide complete information; similarly it is difficult for samples at one point in time to provide complete information. This means that a variety of assessments over a variety of times, which are matched against the areas we are interested in, is much more likely to provide reliable and, as we shall learn later, valid information (8,9,10).

The value of multiple sampling has been backed up by empirical observation (10,11) and by generalisability analysis (12,13). For example, in a clinical assessment, generalisability analysis can compartmentalise how much variation in a learner's marks are due to their true ability, how much is due to the patients (case specificity), how much is due to the examiners (assessor specificity) and how much to unmeasured factors (error).

Single observations often mean a learner's true ability may form an insufficient proportion of the mark. The best way to increase the proportion of the mark that is due to a learner's true ability and decrease the proportion due to other factors, is to increase the number of observations (13,14). If there is considerable variation due to the patient, then these observations should include more patients. If it is due to the examiner, then these observations should include more examiners. This 'concentrates' the marks due to learners and 'dilutes' the marks due to the idiosyncrasies of examiners or patients. In contrast, if the variation due to examiners is low yet the variation due to patients is high, as in long cases (6,14), then examiner training will have limited influence. Instead, multiple cases with different patients are needed. It is these analyses that have guided our thinking on how best to combine assessment results and how best to use less reliable assessment tools.

This also explains why OSCEs have high reliability. Although they were designed on the basis that reliability came from its standardisation and the use of prescribed tasks with defined checklists, in reality the reliability arises from the use of multiple observations (stations).

This has led to the renaissance of global judgements (5,15-19). In the past, unreliable results were erroneously attributed to insufficient structure in an assessment tool rather than to other sources of unreliability such as insufficient number of observations, insufficient range of instruments or insufficient number of assessors. Under this influence, checklists proliferated, a good example being supervisor report forms. It was thought that having a number of checklist items would improve reliability. In fact, it often made it worse (5,7,19). Global judgements are often just as, or more, reliable – provided these judgements are aggregated with those of others.

Programmatic assessment

We know that medical practice is complex, requires multiple skills and is multidimensional. It therefore should be no surprise to learn that there is no single assessment tool that can measure good medical practice. Instead the various elements of practice need to be assessed in different ways – which means we need multiple assessment tools. As described earlier, we also need multiple samples. Putting this together means we need judicious use of a selected group of assessment tools, carefully staged over the course of a programme of learning that in their entirety give a picture of a learner's abilities. This concept is referred to as programmatic assessment.

An analogy with programmatic assessment is a family photograph album. In this analogy, the album may contain photos of some family members that are well constructed but poorly focused, there may be others that are finely focused but only of half the family. Others might not have everyone smiling. However, taken in its entirety, the album gives a reasonable portrayal of the family. Likewise, a programme of assessment might include some tools that are highly valid but only moderately reliable, other tools might be very reliable but restricted in the range of attributes they can assess. But taken together, we can gain a reasonable picture of a learner. Multiple snapshots, even if some are not totally in focus, give a better picture than one poorly aimed photograph.

This means we should not discard tools on the basis of unreliability if they have high face validity and a positive educational impact (8,9) – but we should also not place over-reliance on them on their own. Instead we should build in more observations and other assessments before making decisions.

Programmatic assessment therefore creates an unusual win-win situation: multiple assessments by multiple observers over multiple time periods improve reliability. In addition, multiple tools over times by multiple observers also contribute to validity. The key to getting programmatic assessment right is in the selection of the right mix of assessment tools. This is where 'blueprinting' is important.

Blueprinting

Because it is not feasible to assess everything of value, we have to take a sampling approach. We often need just enough of some assessment tools, not too many of another and careful attention to a few selected ones.

Mapping our assessments to our curriculum to determine the content validity is called blueprinting, of which there are two types.

The first type is a macro-level blueprint that aims to look at each of the components of the curriculum and to decide which assessment tool or tools would be ideally suited to sample that area of interest. In Table 1, the components of the curriculum that need to be assessed are given in the columns and the possible assessment tools in the rows. The tools are then mapped to the components to demonstrate which tools could assess which components. The goal is to choose the fewest number of assessment tools that ensure all the areas of interest are assessable.

Table 1:

Developing a macro-level blueprint

	What is being assessed		
Method	**Content area 1 (Knowledge)**	**Content area 2 (Knowledge and clinical skills)**	**Content area 3 (Clinical skills)**
MCQ	✓	✓	
Clinical		✓	✓

After completing this for the whole curriculum, we usually find that no single tool, or indeed, even two or three tools, can assess everything. For example, knowledge might best be assessed by a written test and multiple-choice exam, clinical examination skills by observing a learner–patient interaction over 10–15 min (e.g. by mini-CEX (20,21,13)), integrative skills by long case (14,22) or case-based discussion (23-25), teamwork by multisource feedback (26,27,28) (see Table 2). As clinical expertise is a multifaceted entity, the need for more than one tool should come as no surprise.

The second type of blueprinting is a micro-level blueprint and is used to decide how to sample appropriately for any particular tool. For example, in creating a multiple-choice exam, once again it is not possible to assess everything. But sampling in a balanced and measured way across the knowledge areas of interest ensures that all aspects are assessable, if not actually assessed, within a particular exam.

Table 2:
Developing a micro-level blueprint

	What is being assessed		
MCQ	Content area 1	Content area 2	Total
Acute patient care	20	3	23
Chronic patient care	20	4	24
Emergency care	40	13	53
Total	**80**	**20**	**100**

The outcome of good blueprinting is a parsimonious choice of assessments that reflects the balance of elements within the training programme and a good 'sampling' across all domains of interest.

Assessment in the exam room or in the workplace

Sometimes it is asked whether assessment should be in the workplace or in the exam room. By now it should be apparent that the answer to this question is likely to be that we should have both central and workplace-based assessments. No assessment method, on its own, can assess everything of interest. However, there are strengths and weaknesses for each location of assessment.

Assessments that are centralised, such as multiple-choice tests or traditional clinical assessments, have the advantage of economies of scale – one exam can be delivered to many learners. They are, therefore, better than workplace-based assessments at assessing some things, such as a core foundation of knowledge. They have symbolism in being seen as a standardised major event, or even as a rite of passage, but they have emerged from a model that says reliability can only come from rigid standardisation and control (29).

A disadvantage of centralised assessments is that the range of problems that can be assessed is limited. Clinical examinations, for example, are mostly restricted to patients with stable signs or symptoms. Furthermore, despite efforts to reduce unreliability due to examiners, overall reliability will not be improved if a limited number of patients is used (14).

A final disadvantage of centralised assessments is that, one-off events, in artificial situations, with a limited range of problems, with a limited number of examiners and with some unreliability, all contribute to learner stress. This is because learners (and examiners) see that competency alone is no guarantee of success (29).

Workplace-based assessments are not the panacea we are looking for either, but have some important advantages (30). They often have high validity because they look at performance rather than competence, and therefore can capture information on actually doing the job, such as caring for acutely ill patients or teamwork. They can be made reliable by having enough of them and by considering the results in their entirety (20). They contribute to aligning assessments to things that we value, thereby being more likely to have a positive educational impact.

However, there are two challenges: one relates to feasibility and the other relates to the effect on the assessor–trainee relationship. Asking supervisors to assess trainees more often in the workplace is to ask them to observe the trainee doing the job. For some, this may be an increase in work. For others, this may be part of normal practice (31). For the trainee, it is likely to be highly valued as it is a rich source of information from which to provide feedback. For the patient, it is likely to be valued as a tangible marker of quality improvement (29).

The effect of workplace-based assessments on the supervisor–trainee relationship is challenging, particularly in rural settings where there may be only one supervisor and one learner. On the one hand, we want this relationship to be one of support and trust where trainees can display their weaknesses so that they can improve. When it comes to summative assessments, however, there is a tension, as now we want trainees to display their strengths. If we are not clear which relationship applies at a particular time then confusion can arise. This highlights the importance of differentiating between formative and summative assessment.

Formative and summative assessment

Differentiating between formative and summative assessment has particular relevance to rural settings where a supervisor may be asked to undertake both forms of assessment. Confusing these concepts can lead to unfortunate consequences.

While deciding on the purpose of assessment is the most crucial decision to make, the purpose of an assessment programme in general is to gather high-quality evidence to make well-informed decisions. The decisions that need to be made can usually be divided into two main categories: (i) decisions on what to learn and on areas to improve; and (ii) decisions on progress. The first decision is usually made *with* the learner and is the basis of formative assessment. Formative assessments are used to guide learning. The second is usually made *for* the learner and is the basis of summative assessment (29). Summative assessments contribute to making high-stakes decisions.

The crucial distinction here is that formative assessments aim to help find weaknesses and guide learning. In such assessments, a learner will be forthcoming in acknowledging areas that need improvement. In contrast, summative assessments expect learners to display their strengths. If a learner believes an assessment is summative to inform high-stakes decisions, they will wish to conceal weaknesses. Problems can therefore arise if the purpose is not clearly stated at the outset. Learners may conceal weaknesses in formative assessments if they believe they could be used for summative purposes.

The relevance of this is that we can easily send confusing messages to our learners if we are not clear which assessments we are undertaking. In our teaching, we are keen to help learners identify their weaknesses so we can guide them in ways that might fill those learning gaps. We probe their abilities, seek clarification of their concepts and check their understanding. In turn, we encourage learners to identify their weaknesses through self-assessment and reflection. All these activities are highly useful in effective learning and are to be encouraged.

We may also use formative assessments to help find these weaknesses. If learners believe their supervisors are continually undertaking summative assessments of their abilities, then there will be the temptation for them to conceal their weaknesses. They may find teaching sessions stressful, like they are an exam. They may wish to be taught only on things that they know they are good at and may be reluctant to make explicit to their supervisors, the results of their self-assessments of areas of weakness.

Being clear about whether an assessment is formative or summative can send a mixed message that can undermine good learning. This can be a particular risk if learners are in settings where there is only one supervisor and that supervisor is responsible for both the teaching and the summative assessments – such as may occur in rural settings.

Some argue that within programmatic assessment, such formative and summative distinctions are less critical as each assessment episode is such a small part of the greater whole. All assessment therefore can be both summative and formative. This can apply particularly to some professional behaviours – some behaviours are unacceptable, regardless of whether the learner thinks they're being assessed or not. Provided learners are aware which of the assessments contribute to summative decisions and which are used purely to guide learning, then most of these pitfalls can be avoided. There are other solutions to this, which we shall come to later.

In summary, there are many aspects, and principles, of assessment that are identical, regardless of whether the programme is based in a rural setting or not. These are being clear of the purpose; the usefulness of multiple assessment tools, over multiple time periods; using multiple raters to contribute to both reliability and validity; the importance of a blueprint; and being clear about the formative or summative nature of assessments.

However there are some situations that require special consideration.

Special situations

The following are some specific issues that deserve some discussion. None is unique to rural and remote settings but can be more problematic in those contexts.

Multi-site medical schools

Most medical schools are making increasing use of a variety of sites. This is an appropriate response to the observation that health care is also delivered in a variety of sites. Some of these sites can be quite distant from the main medical school, while some medical schools are completely dispersed to the extent that there is no 'main' medical school.

Increasingly learners may spend a large part, or all, of their time at more distant sites. Furthermore, many medical schools are offering programmes in different sites that also differ in curricular structure. For example, longitudinal placements and immersion programmes. For learners in distant sites, it is appropriate that their assessments also occur in these sites. Assessment is sometimes the 'glue' that binds these different curricula together because all learners are aiming for the same destination, despite taking different paths. The important consideration for assessment in multi-site medical schools is for the assessments to be equivalent in all sites. Some, but not all, could also be identical or simultaneous.

The two key considerations in relation to equivalence are fairness and standard setting.

Fairness

Fairness refers to treating learners equally. An assessment is unfair if some learners are given more information about that assessment than others. This can occur, for example, if some learners are told the content of an assessment while others are not. This is a particular risk for assessments that are not simultaneous as learners who have undertaken the assessment first may inform other learners who are yet to take that assessment.

In relation to assessment, fairness does not refer to whether different groups of learners have had equivalent opportunities to learn the required attributes. If that is an issue, then fairness needs to be addressed at the curriculum level, not compensated for at the assessment level.

There are four ways to ensure fairness across distant sites. The first is to offer the same exam at the same time to all learners. This is feasible if the time zone difference between sites is not large and if the nature of the assessment is appropriate. A written test, for example, can easily occur identically and simultaneously.

The second is to quarantine those learners who have undertaken the assessment, from those who are yet to be assessed, so that they are unable to communicate with those other learners. This is feasible if the time for such isolation from the other learners is a few minutes or hours. This often occurs in an OSCE where different groups of learners sequentially rotate through the same set of stations (32).

The third solution is to have different sets of the assessment that are known to be equivalent. In this case, there could be several blueprinted sets of each assessment where it is known that each set covers the same range of areas of interest but each contains different questions. Calibrating standards between these sets of assessments needs careful consideration and is discussed shortly. Examples of this are seen in multiple choice examinations where the same number of questions is allocated to each subject area, but the questions within each subject area may be different for different groups of learners.

The fourth solution is to have assessments that are so generic that knowing the actual content offers no advantage. For example if the attribute to be assessed is history taking, then provided there is a sufficient number and range of patients, observing history taking with one set of patients may well be comparable to observing history taking in another set. The key prerequisite to this is knowing that there is a sufficient number and range of patients. As discussed earlier, generalisability theory (12,13) can help here to determine the number of patients to be seen before the score that a learner obtains is known to reflect the learner's true ability. This is the basis of the mini-CEX (see Appendix A) where it has been shown that sufficient reliability can be obtained after 8-12 observations and sufficient validity can be obtained if those observations are on a blueprinted range of patient problems (20,21,33). Each mini-CEX encounter however is unique to each learner. These same principles apply to many other workplace-based assessments.

Whatever methods are used, learner perception of fairness also needs to be considered. High-stakes exams are powerful mills for creating anxiety and this is only worsened if learners believe that they are at a disadvantage compared with other learners. Sharing information about the processes the institution has in place to ensure fairness and providing some data to confirm the effectiveness of these processes are often key components in an assessment programme.

Standard setting

It is beyond the scope of this chapter to discuss the various methods of calibrating standards for the various assessment tools. However some general comments may be helpful.

Setting standards requires a process by which different groups of examiners develop a shared understanding of what is expected of learners. This can be achieved face-to-face, by calibration exercises, or mathematically. As has already been emphasised, this should always be prefaced by all parties being clear about the purpose of the assessment. Face-to-face discussions among examiners, whereby examples of the range of learner performances are discussed and agreement reached, can work well provided there is good facilitation and strong personalities are not allowed to dominate. Calibration exercises are often helped by providing exemplars of various standards, for example of written work of past learners or videotaped recordings of learner performances. The discussions that ensue from observing these exemplars can be very useful.

The collective view of examiners can also be collated mathematically by aggregating their opinions. This occurs for example in the various Angoff procedures for some written assessments (34) or the borderline group (35) or borderline regression (36) procedures for some clinical assessments.

More sophisticated standard setting can occur using Item Response Theory whereby question difficulty can be calculated separately from learner ability (37). Aggregating questions to create tests of equivalent difficulty can be used for different groups of learners, while the aggregated questions based on a learner's ability can be used to decide pass-fail decisions.

Off-site placements with few staff and learners

Teachers or supervisors often wear different 'hats'. Sometimes they are a mentor, sometimes they provide pastoral care and at other times they must be an examiner. These roles can become confused at the best of times, but this is a particular risk when there is a small number of staff and learners. The tensions that can arise when the formative and summative functions of assessment are insufficiently distinct have been outlined earlier. It can also be easy to think a poor learner did not do so well on an assessment because the teaching wasn't up to scratch. Or conversely place the blame for a poor outcome on the capable learner rather than the poor teaching. Remote settings also risks blurring the boundaries between teacher as 'friend' or even teacher as 'accommodation host' and teacher as 'judge'. This can be particularly problematic if the small number of staff and learners occurs within a longitudinal placement as the personal relationships that form can become stronger. Such relationships are often beneficial for learning but can exacerbate the boundary issues when it comes to assessment.

There are two possible solutions: the first is to ensure the supervisor is only one of many persons who assess the trainee. This is the ideal solution as having a variety of assessors not only preserves the supervisor–trainee relationship of support, but also improves reliability and validity of the assessments (8,10). The second is for trainee and supervisor to be absolutely clear when assessments are being used formatively (to find weaknesses and to guide learning) and when they are being used summatively (to find strengths and to contribute to decisions on progress) (29). This would require the supervisor to explicitly state when a defined period of observation forms part of the summative assessment and then to state when that period of observation finishes. This is also where many of the workplace based assessment tools can be useful as recording the observation on a specific form can be a tangible indicator that a summative assessment is taking place. The important point to remember here however is that aggregation of a number of these observations (preferably by a number of examiners) will be needed before sufficient reliability and validity is achieved to make summative decisions (13).

Assessment in interprofessional learning

Like all assessment, there needs to be clarity of purpose. In interprofessional learning, what is the assessment for and what is it aiming to assess? In many cases the answer may well include wanting to assess abilities to collaborate, understand the roles of others, work within teams, negotiate care with other health professionals and resolve contrasting views. Herein lies the main paradox of the assessment of interprofessional learning. Assessment is traditionally at the level of the individual whereas interprofessional learning is at the level of the group. Resolution of this paradox will require development of assessment tools that are aimed at a group, not just an individual.

Self-assessment

Insight can be defined as when a person's self-assessment matches external assessments. Accuracy of self-assessment is therefore an important skill. However, it can only help insight if it is then compared against an external measure, appropriately debriefed and subsequently acted upon. Self-assessment is therefore often better regarded as a learning tool than an assessment tool. However, some judgements are possible, such as the accuracy of the self-assessment or the appropriateness of the actions that arise from it.

Peer assessment

As for all assessment, defining the purpose of peer assessment is the first step. Is it formative or summative? Is it aimed at identifying outlying behaviour (the 'bad apples') or is it aimed at providing feedback to all learners so all can improve (moving the bell shaped curve)? Peer assessment can fill important gaps as peers are often in a unique situation to observe some behaviours that are not easily observed by others. However when learners are asked to be assessors, they also need some assessor training. Specifically they need to know what they are rating, why they are rating it, what happens to the results, and what gap it is trying to fill. Unless these questions are answered, there is a risk that peer assessment can cause harm.

Peer assessments commonly take one of two forms: one-off reporting of aberrant behaviours (either desirable or undesirable) or collated ratings by several peers on pre-specified attributes. Because there is likely to be variability among assessors, any aggregation of ratings from many peers will be needed to form a reliable assessment. How such an aggregation is then communicated to individual learners also needs to be carefully managed, so that the ratings are interpreted constructively. One-off reporting of aberrant behaviours also requires careful management. The ground rules of whether such reports are anonymous or confidential should be established. Ideally they should not be anonymous but could be confidential. In either event, their interpretation requires careful consideration as individuals displaying similar behaviours do not always have similar underlying causes, contexts or implications.

Patient satisfaction surveys

The principles outlined above for peer assessment apply equally to patient satisfaction surveys. While it is hard to argue that the opinions of patients shouldn't be one of the ultimate indicators of clinical practice, there are also many confounding factors that will influence a patient's view of an encounter – and many of these factors are either unrelated or not within the control of the person being rated. Being very clear therefore about which aspects we are asking patients to rate and why, is the first premise. There is even more variability between rates in patient satisfaction surveys than there is from peer assessments. For these reasons, the aggregation of a large number (typically 30-50 ratings) is needed before reliable and valid results can be obtained. This is often infeasible. Fewer ratings re needed if these are done within controlled settings, such as within an OSCE (17, 38, 39, 40).

References

1. Wilkinson TJ, Wells JE, Bushnell JA. What is the educational impact of standards-based assessment in a medical degree? *Medical Education* 2007; 41(6): 565-72.
2. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; 17(3): 165-71.
3. Miller GE. The assessment of clinical skills/competence/performance. *Academic Medicine* 1990; 65(9 Suppl): S63-S7.

4. Rethans JJ, Norcini JJ, Baron-Maldonado M, Blackmore D, Jolly BC, LaDuca T, et al. The relationship between competence and performance: Implications for assessing practice performance. *Medical Education* 2002; 36(10): 901-9.

5. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Academic Medicine* 1999; 74(10): 1129-34.

6. Norman G. The long case versus objective structured clinical examinations. *BMJ* 2002; 324(7340): 748-9.

7. van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education* 1991; 25(2): 110-8.

8. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: From methods to programmes. *Medical Education* 2005; 39(3): 309-17.

9. Schuwirth LWT, Vleuten CPM. A plea for new psychometric models in educational assessment. *Medical Education* 2006; 40(4): 296-300.

10. Wilkinson TJ, Frampton CM. Comprehensive undergraduate medical assessments improve prediction of clinical performance. *Medical Education* 2004; 38(10): 1111-6.

11. Verhoeven BH, Hamers JG, Scherpbier AJ, Hoogenboom RJ, van der Vleuten CP. The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. *Medical Education* 2000; 34(7): 525-9.

12. Brennan RL, Johnson EG. Generalizability of performance assessments. *Educational Measurement: Issues and Practice* 1995; 14(4): 9-12, 27.

13. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: A key to unlock professional assessment. *Medical Education* 2002; 36(10): 972-8.

14. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. *Medical Education* 2008; 42(9): 887-93.

15. Cohen R, Rothman AI, Poldre P, Ross J. Validity and generalizability of global ratings in an objective structured clinical examination. *Acad Med* 1991; 66(9): 545-8.

16. Swartz MH, Colliver JA, Bardes CL, Charon R, Fried ED, Moroff S. Global ratings of videotaped performance versus global ratings of actions recorded on checklists: A criterion for performance assessment with standardized patients. *Academic Medicine* 1999; 74(9): 1028-32.

17. Wilkinson TJ, Fontaine S. Patients' global ratings of student competence. Unreliable contamination or gold standard? *Medical Education* 2002; 36(12): 1117-21.

18. Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. *Medical Education* 1987; 21(6): 477-81.

19. Wilkinson TJ, Frampton CM, Thompson-Fawcett MW, Egan AG. Objectivity in objective structured clinical examinations: Checklists are no substitute for examiner commitment. *Academic Medicine* 2003; 78(2): 219-23.

20. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine* 2003; 138(6): 476-81.

21. Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Medical Education* 2006; 40(10): 950-6.

22. Norcini J. The validity of long cases. *Medical Education* 2001; 35(8): 720-1.

23. Chana N. *A practical guide to case based discussion*. London: London Deanery; 2008.

24. Munger BS. Oral examinations. In: Mancall EL, Bashook PG (eds.) *Recertification: new evaluation methods and strategies*. Evanston: American Board of Medical Specialties; 1995. p39-42.

25. Jennett P, Affleck L. Chart audit and chart stimulated recall as methods of needs assessment in continuing professional health education. *Journal of Continuing Education in the Health Professions* 1998; 18(3): 163-71.

26. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004; 328(7450): 1240-3.

27. Norcini JJ. Peer assessment of competence. *Medical Education* 2003; 37(6): 539-43.

28. Wilkinson TJ, Wade WB, Knock LD. A blueprint to assess professionalism: Results of a systematic review. *Academic Medicine* 2009; 84(5): 551-8.

29. Wilkinson TJ. Assessment of clinical performance – gathering evidence. *Internal Medicine Journal* 2007; 37(9): 631-6.

30. Norcini JJ. Current perspectives in assessment: The assessment of performance at work. *Medical Education* 2005; 39(9): 880-9.

31. Daelmans HE, Hoogenboom RJ, Scherpbier AJ, Stehouwer CD, van der Vleuten CP. Effects of an in-training assessment programme on supervision of and feedback on competencies in an undergraduate Internal Medicine clerkship. *Medical Teacher* 2005; 27(2): 158-63.

32. Wilkinson TJ, Newble DI, Wilson PD, Carter JM, Helms RM. Development of a three-centre simultaneous objective structured clinical examination. *Medical Education* 2000; 34(10): 798-807.

33. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine* 2002; 77(9): 900-4.

34. Ben-David MF. AMEE guide No 18: Standard setting in student assessment. *Medical Teacher* 2000; 22(2): 120-30.

35. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: Use of global ratings of borderline performance to determine the passing score. *Medical Education* 2001; 35(11): 1043-9.

36. Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. *Medical Education* 2003; 37(2): 132-9.

37. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Medical Education. 2009; 44(1): 109-17.

38. Weaver MJ, Ow CL, Walker DJ, Degenhard EF. A questionnaire for patients' evaluations of their physicians' humanistic behaviors. *Journal of General Internal Medicine* 1993; 8(3): 135-9.

39. Tamblyn R, Benaroya S, Snell L, McLeod P, Schnarch B, Abrahamowicz M. The feasibility and value of using patient satisfaction ratings to evaluate internal medicine residents. *Journal of General Internal Medicine* 1994; 9(3): 146-52.

40. ACGME Outcomes Project. *Toolbox of Assessment Methods.* Accreditation Council for Graduate Medical Education, American Board of Medical Specialties. Version 1.1, 2000. www.acgme.org/outcome/assess/toolbox.asp (accessed 11 April 2005).

41. Marshall VC, Alexander HG, Buick SA, Epstein J, Frank IB, Makeham MAB, et al (eds.) *AMC guidelines for assessing medical competence: Applied knowledge testing in multiple-choice question format.* Canberra: Australian Medical Council; 2011.

42. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education* 1985; 19(3): 238-47.

43. Jolly BC, Grant J, Adams E, Gale R, Jackson G, Johnson N, et al. *The good assessment guide: A practical guide to assessment and appraisal for higher specialist training*. London: Joint Centre for Education in Medicine; 1997.

44. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education* 2004; 38(2): 199-203.

Further reading

1. Smith JD, Prideaux D, Wolfe C, Wilkinson TJ, Sen Gupta T, De Witt DE, et al. Developing the accredited postgraduate assessment program for Fellowship of the Australian College of Rural and Remote Medicine. *Rural and Remote Health* 2007; 7(3): 805.
2. Wilkinson JR, Wade WB. Working smarter: New methods of performance assessment for trainees. *The Bulletin of the Royal College of Pathologists.* 2005 Oct; 132: 12-5.
3. Wilkinson TJ, Smith JD, Margolis SA, Sen Gupta T, Prideaux DJ. Structured assessment using multiple patient scenarios by videoconference in rural settings. *Medical Education* 2008; 42(5): 480-7.

Appendix A

Glossary of selected assessment tools

Mini-CEX
The mini-CEX is a 15-30 minute observed snapshot of a doctor/patient interaction. It is conducted within actual patient care settings, using real patients, but has a structured marking sheet that covers pre-defined generic areas. Validity derives from using authentic interactions and reliability is achieved by ensuring a sufficient number of encounters are aggregated (20,21,33).

Case-based discussion
This is also known as 'chart stimulated recall' (24,25). The trainee selects two case records from patients they have recently seen and in whose notes they have made an entry. The assessor will select one of these for the case-based discussion.

Written Multiple Choice Question examination
The MCQ examination is a set of questions that have been chosen to assess a representative range of topics from a defined curriculum by asking candidates to choose the best answer to a question from some offered options (40,42,43).

Multisource feedback
This is the systematic collection and feedback of data on an individual's performance, acquired from a number of stakeholders. In the past, this has sometimes been referred to as the 360-degree assessment. The areas assessed are often related to professional skills and behaviours (26,27,28).

Patient satisfaction survey
This is a collation of questionnaire-based opinions of patients about the nominated person's abilities in specified areas (17,38,39,40).

Objective Structured Clinical Examination (OSCE)
A set of assessment stations designed to cover a relevant range of areas of interest (44). The number of stations varies but can be as few as six, or as many as 30. Each station lasts 5-20 minutes.

Long case
A candidate takes a history and performs a physical examination on a patient, synthesises the findings, creates a management plan and discusses these with the examiners (6,14,22,43).

Short case
This is an observed interaction between candidate and patient. It usually observes clinical examination skills within a 10-15 minute encounter (14,43).

This article is a chapter from the ***WONCA Rural Medical Education Guidebook***.
It is available from www.globalfamilydoctor.com.

Published by:
WONCA Working Party on Rural Practice
World Organization of Family Doctors (WONCA)
12A-05 Chartered Square Building
152 North Sathon Road
Silom, Bangrak
Bangkok 10500
THAILAND

manager@wonca.net

Suggested citation: Wilkinson T. Assessment in Rural Medical Education. In Chater AB, Rourke J, Couper ID, Strasser RP, Reid S (eds.) *WONCA Rural Medical Education Guidebook.* World Organization of Family Doctors (WONCA): WONCA Working Party on Rural Practice, 2014. www.globalfamilyldoctor.com (accessed [date]).