**Russell Jones**
BSc, DipEd, BEd(Hons), PhD, is Director of Assessment, The Royal Australian College of General Practitioners, and Professor, Centre for Medical and Health Science Education, Monash University, Victoria. russell.jones@racgp.org.au

# Accurately assessing candidates for general practice

**Background**
Specialist medical examinations, such as those within general practice, are essential for identifying candidates who are able to progress to independent specialist practice. The rationale for their use is that underlying general practice related knowledge, ability, and skill of a candidate can be determined through the use of a valid, reliable, fair, practical and generalisable examination.

**Objective**
This article discusses a method for viewing all aspects of an examination for the purpose of minimising or eliminating error.

**Discussion**
An assumption is made that performance on the examination is a predictor of the underlying ability of a candidate. Although examinations are invaluable tools, they are only indicators of candidate competence or candidate mastery. Decision making based on examination results may be adversely affected if error enters examination content, processes and procedures. This is particularly the case for candidates whose examination scores fall around the pass mark. Potential strategies for minimising the 'band of uncertainty' for these candidates in The Royal Australian College of General Practitioners Fellowship examination are discussed.

**Examination results may be used as an indicator of the ability of an examinee (or candidate) to work as a general practitioner. An assumption is made that performance on the examination is a predictor of the underlying ability of a candidate. If the examination is valid, reliable, fair, practical and generalisible of true general practice knowledge, ability and skill, then this is an appropriate assumption to make.**

The primary purpose of most medical specialist examinations is to identify those candidates who have adequate competence or an acceptable level of mastery of the medical speciality and those who do not. Although there are two possible results, pass or fail, there also exist two possible levels of candidate competence or mastery: adequate or inadequate. A 2 x 2 matrix can be constructed showing this relationship using candidate competence/mastery on one axis and examination result on the other (*Figure 1*). All candidates will fall into one of the four possible quadrants. If the examination is sound, then most candidates will fall into either:

- the top right quadrant where candidates with adequate competence/mastery will pass the examination, or
- the bottom left quadrant where candidates with inadequate competence/mastery will fail.

Placement of candidates into either of these two quadrants is an appropriate decision. However, it is possible for some candidates to fall into one of two other possible quadrants:

- the top left quadrant where candidates with adequate competence/mastery will fail the examination, or
- the bottom right quadrant where candidates with inadequate competence/mastery will pass the examination.

Neither of these decisions is appropriate and, ideally, no candidate should fall into either of these two quadrants. However, all examinations, particularly those that include a pass/fail mark, are really only indirect indicators of true examinee ability. All examinations are influenced to some extent by error, which acts to interfere with the estimation of true candidate ability. This error will influence the likelihood of candidates falling into each of the four

quadrants. The greater the error, the greater the likelihood that a candidate may fall into either the top left or bottom right quadrants. This is akin to the concept familiar to all general practitioners of receiving pathology test results which might result in false positives or false negatives. The extent and frequency of these false pathology results depend on the sensitivity and specificity of the assay.

An analogy may be drawn between placement of candidates within the top left or bottom right quadrants and type 1 and 2 errors. A type 1 error is rejecting a true null hypothesis and a type 2 error is accepting a false null hypothesis.[1,2] If the proposed null hypothesis for an examination is that candidates with adequate competence/mastery will pass, a type 1 error may be said to occur if a candidate with adequate competence/mastery fails. Similarly, a type 2 error may be said to occur if a candidate without adequate competence/mastery passes an examination.

The distribution of candidate results for a hypothetical examination might appear as shown in *Figure 2*. Identifying candidates who clearly should pass is usually straightforward and may be represented by the area on the right of the graph. Similarly, identifying candidates who clearly should fail may be represented by the area on the left. Problems typically arise for candidates who fall around the cut-off score (or pass mark) when a small amount of error can cause candidates to be incorrectly classified as either pass or fail. This is represented by the darkened 'band of uncertainty' in *Figure 2*. Within the 2 x 2 quadrant of *Figure 1* the primary purpose of an examination becomes the reduction of the number of candidates who fall within either of the error quadrants. When viewing a credentialing examination from the perspective illustrated in *Figure 2*, the primary purpose is to reduce the width of the band of uncertainty.

Figure 1. A 2 x 2 matrix showing the appropriateness of decisions made in the relationship between examination performance and candidate mastery/competence

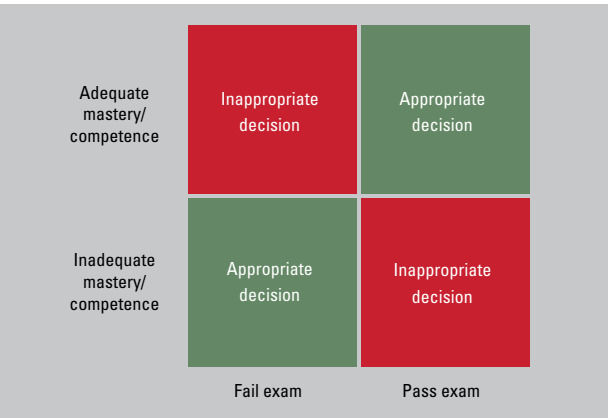
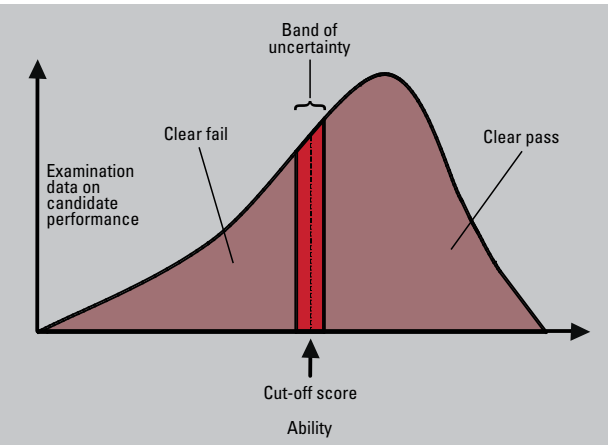



Figure 2. A hypothetical distribution of candidate results showing areas of clear pass/fail separated by a band of uncertainty. Examinations should seek to reduce the width of this band




## Sources of error

There are many potential sources of error and most arise from threats to validity or reliability.

### Validity

Validity is the extent to which an examination measures what it is supposed to measure. In the case of general practice, this is a candidate's knowledge, skill, ability, attitude, practice and behaviour in general practice. There are numerous forms of validity:

- content validity is the extent to which an examination assesses an adequately representative sample of the content it purposes to measure
- criterion related validity is the extent to which examination scores are related to one or more external measures (ie. criteria)
- predictive validity measures the strength of the relationship between the examination result and prediction of future performance
- concurrent validity considers the strength of the relationship between performance on one examination and performance on another and is designed to assess the same underlying abilities (eg. the strength of the correlation between a multiple choice and short answer examination if both are purported to assess knowledge of general practice)
- construct validity is the degree to which an examination measures the underlying 'construct' the examination is designed to measure, ie. how closely does the examination measure 'real world' performance?
- face validity is what an examination appears to assess at first glance. Although it may be argued that face validity is not of equal importance to other forms of validity, it is essential in order that an examination be perceived as appropriate by examiners, candidates and others interested in the examination process. Maintaining credibility of the credentialing examination among the public and medical communities is extremely important.[3]

Most recently, the effect on learning of an examination has been introduced as a measure of validity, ie. does the examination have a desirable effect on learning? Examinations may introduce invalidity if their procedures, structure or content encourage inappropriate learning.

## Reliability

Reliability is the extent to which an examination or examination process is consistent over time, on different occasions, with different examiners or candidates, or using different questions. There are several forms of reliability:

- alternate forms, or test-retest, reliability is the extent to which an examination produces consistent results over several administrations to the same candidates
- intra-rater reliability is the extent to which an examination produces consistent results using the same examiner
- inter-rater reliability is the extent to which different examiners award the same mark or score for the same individual during an examination, and
- internal consistency is the extent to which the different questions comprising an examination consistently measure the same attribute.

Those readers seeking a comprehensive discussion about validity and reliability are referred to the excellent discussions by Messick,[4] and Feldt and Brennan.[5] Those after a more focused discussion of reliability in relation to general practice examinations are referred to the article by Hays, Fabb and van der Vleuten.[6]

## Minimising the band of uncertainty

A key role for The Royal Australian College of General Practitioners (RACGP) is the summative assessment of doctors wishing to attain Fellowship. This is achieved through the Fellowship examination. In an era of increasing accountability it is essential that the band of uncertainty is reduced and the likelihood of a candidate falling into the top left or bottom right quadrants in *Figure 1* is minimised.

One of the most effective ways, and one which is rarely used in most medical specialist examinations, is to obtain additional information from each candidate whose performance falls within the band of uncertainty. In essence, this is analogous to events in every day general practice. General practitioners frequently encounter patients whose diagnoses are uncertain and therefore seek additional information, such as tests or more detailed physical examination, until they are confident of being able to make an accurate diagnosis. Similarly, if those candidates whose performance falls into the band of uncertainty undergo further 'tests', this will provide additional information that can be used to accurately determine their placement as either a true pass or true fail.

There are a range of approaches that can be used to obtain further useful information about a candidate whose performance falls into the band of uncertainty. Reliability of an assessment is known to increase with an increase in the number of questions. A straightforward method to obtain additional information would be to provide candidates with a supplementary examination in a format similar to the original examination. Using the pathology analogy described earlier, this might be considered akin to repeating a pathology test in an attempt to confirm a positive or negative result. However, the opportunity for additional testing allows an assessment program to utilise alternative assessment formats. These offer an alternative, potentially rich source of information. Again using the pathology analogy, this might be considered akin to a GP requesting an alternative test; one which has greater specificity. Such assessments could include viva-voce, videotape review or an examiner observing a candidate's clinical practice. There is also the potential to utilise information obtained from in-training assessments. Such an approach would succeed in minimising the band of uncertainty, as well as reducing the incidence of unnecessary resits and the incidence of false positive results.

Conflict of interest: none declared.

## References

1. Dawson B, Trapp RG. Basic and clinical biostatistics. New York: McGraw-Hill, 2004.
2. Myles PS, Gin T. Statistical methods for anaesthesia and intensive care. Oxford: Butterworth Heinemann, 2000.
3. Norcini JJ. Examining the examinations for licensure and certification in medicine. JAMA 1994;272:713–4.
4. Messick S. Validity. In: Linn RL, editor. Educational measurement. 3rd edn. National Council on Measurement in Education and American Council on Education, 1989;13–103.
5. Feldt LS, Brennan RL. Reliability. In: Linn RL, editor. Educational measurement. 3rd edn. National Council on Measurement in Education and American Council on Education, 1989;105–46.
6. Hays RR, Fabb WE, van der Vleuten C. Longitudinal reliability of the RACGP Fellowship examination. Med Educ 1995;29:317–21.

AFP CORRESPONDENCE afp@racgp.org.au