Composite reliability of workplace-based assessment of international medical graduates

Composite reliability of workplace-based assessment of international medical graduates

Balakrishnan (Kichu) R Nair[1,2], Joyce MW Moonen-van Loon[3], Mulavana Parvathy[1], Brian C Jolly[2], Cees PM van der Vleuten[3]

The known Workplace-based assessment (WBA) of the performance of doctors has gained increasing attention. The reliability of individual assessment tools has previously been explored.

The new We analysed the composite reliability of a toolbox of WBA instruments for assessing international medical graduates (IMGs). A combination of five case-based discussions and 12 mini-clinical examination exercises with six multi-source feedback assessments achieved a standard error of measurement of 0.24, better than the 0.26 required for an adequate level of precision.

The implications Combining data from different WBA assessment instruments achieves acceptable reliability for assessing IMGs, provided that the panel of WBA assessment types and the assessors are carefully selected.

Abstract

Objective: The fitness to practise of international medical graduates (IMGs) is usually evaluated with standardised assessment tests. The performance rather than the competency of practising doctors should, however, be assessed, for which reason workplace-based assessment (WBA) has gained increasing attention. Our aim was to assess the composite reliability of WBA instruments for assessing IMGs.

Design and setting: Between June 2010 and April 2015, 142 IMGs were assessed by 99 calibrated assessors; each was assessed in the workplace over 6 months. The IMGs completed 970 case-based discussions (CBDs), 1741 mini-clinical examination exercises (mini-CEX), and 1020 multi-source feedback (MSF) assessments.

Participants: 103 male and 39 female candidates from 28 countries (Africa, Asia, Europe, South America, South Pacific) in urban and rural hospitals of the Hunter New England Health region.

Main outcome measures: The composite reliability across the three WBA tools, expressed as the standard error of measurement (SEM).

Results: In our WBA program, a combination of five CBD and 12 mini-CEX assessments achieved an SEM of 0.33, greater than the threshold 0.26 of a scale point. Adding six MSF results to the assessment package reduced the SEM to 0.24, which is adequately precise.

Conclusions: Combining data from different WBA assessment instruments achieves acceptable reliability for assessing IMGs, provided that the panel of WBA assessment types are carefully selected and the assessors are calibrated.

In this article, we report the value of workplace-based assessment (WBA) for evaluating international medical graduates (IMGs). Most countries have systems for assessing the fitness of IMGs to practise; fundamental to these systems are robust procedures that typically include written multiple choice question tests and objective structured clinical examinations.[1,2] The virtue of standardised tools is that the assessment is the same for all candidates. Despite being validated,[3] however, the disadvantage of standardised assessment is its questionable relevance to real world clinical practice; it has been suggested that the "standardisation of final licensing, and fitness to practise examinations may make educationalists weep with joy, but there is no clear evidence that it makes for better doctors."[4] Could we perhaps do better?

In recent years, WBA has become more prominent in medical education. Its purpose is to assess proficiency in an authentic clinical environment, principally because what doctors do is more important than what they know, both for patients and society.[5-7] Many postgraduate training bodies have implemented WBA strategies,[7,8] and several undergraduate programs are already using some of its tools, particularly the Mini-Clinical Evaluation Exercise (mini-CEX), case-based discussions (CBDs), multi-source feedback (MSF), and directly observed procedural skills (DOPS). The philosophy underpinning WBA is the assessment of several domains by multiple assessors over a period of time, with feedback built into each encounter.[9] Although trainees receive supervisor reports in most training programs, this has been found to "undercall under-performance", as the reports are prepared by a supervisor who is also the assessor (ie, both coach and referee).[10]

This form of assessment can track the progress of the trainee, for which reason WBA is described as "assessment for learning" rather than the traditional "assessment of learning".[6] Although originally developed for formative assessments (for feedback and training), these tools have been used in programmatic assessments[11] (in which multiple assessment tools are used to comprehensively assess a doctor or student in a program), and can also be used for summative purposes (to determine whether a candidate has successfully passed a course).

We propose that WBA has the potential to provide more relevant assessment of IMGs. When applied to assessing their fitness to practise, WBA must be robust and validated for this purpose. Earlier studies of WBA for IMG assessment found that WBA is acceptable to the candidates, assessors, and the health care system,[12] and one study found that it is also cost-effective.[13]

Studies of the reliability of WBA instruments typically focus on single instruments, but in practice, assessment information is pooled across methods. We therefore need a multivariate estimate of the composite reliability of the WBA toolbox, as first suggested by Miller and Archer[6] and investigated by Moonen-van Loon and colleagues in a recent study of domestic graduates in the Netherlands.[14] The investigators found that combining the information from several methods meant that smaller samples were adequate (ie, fewer individual tests of each assessment type).

[1]Centre for Medical Professional Development, John Hunter Hospital, Newcastle, NSW. [2]University of Newcastle, Newcastle, NSW. [3]Maastricht University, Maastricht, The Netherlands. kichu.nair@newcastle.edu.au ▪ doi: 10.5694/mja17.00130

The question therefore arises: what is the composite reliability of WBA when used for a high stakes (ie, critical) assessment of IMGs? Our study estimated the composite reliability of an established WBA program in Australia. As this was a routine assessment and many IMGs had completed different assessment forms, we analysed only the newer tools: mini-CEX, CBDs and MSF.[8,9]

Methods

All IMGs who wish to practise in Australia (except those who qualified in the United Kingdom, the United States, Canada, Ireland, or New Zealand) must pass the Australian Medical Council (AMC) examination. This assessment consists of a multiple choice examination and an English proficiency assessment, followed by a clinical examination (16 objective structured clinical examination stations).[15]

In 2010, we established a program, accredited by the AMC, for assessing these doctors by WBA as an alternative to the AMC clinical examination. Many IMGs are accorded temporary registration that allows them to work in areas where there is a workforce shortage while waiting for the AMC clinical examination. This waiting period is often long. To be eligible for our program, the candidates needed to pass the English and AMC multiple choice question examinations, and to be employed for the duration of the program (6 months). Candidates who passed our assessment program were eligible for AMC certification.

WBA assessment framework

In accordance with AMC directions, the assessment of each IMG included a minimum 12 mini-CEX and five CBD examinations. At least six different assessors had to be involved in the assessment of an IMG; 99 assessors in total rated the CBD and mini-CEX assessments. This assessment component was supplemented by one set of MSF data.

The mini-CEX, originally developed in the United States to guide learning, assesses clinical performance in authentic clinical situations.[16] IMGs were assessed in six disciplines (medicine, surgery, women's health, paediatrics, emergency medicine, mental health) that reflected the content of the AMC examination. The assessment level was appropriate for the first postgraduate (intern) year. Each mini-CEX measures several competencies in history taking and patient examination, each rated on a scale of 1 to 9; 1–3 indicates unsatisfactory performance, 4–6 satisfactory performance, and 7–9 superior performance. The CBDs, which assess the candidate's record keeping and clinical reasoning, were scored on a similar scale.[16,17]

For completing the MSF assessment form, each IMG nominated three medical and three non-medical (eg, nurse, social worker, pharmacist) colleagues with whom they had worked extensively during the assessment period, and the IMG completed a self-assessment form. The MSF assessment form included 23 questions with statements on aspects of practice such as professionalism, communication, and requesting help when in doubt, and responses were scored on a scale of 1 to 5.[6,18] To pass the assessment, the IMG needed to achieve a satisfactory result (a rating of 4 or more) in eight of the 12 mini-CEX and four of the five CBD examinations; passing the MSF required an average score of 3 across all six assessors.

Data collection

Data were collected from June 2010 to April 2015. During this 5-year period, IMGs employed in urban and rural areas of Hunter New England Health completed 970 CBD, 1741 mini-CEX, and 1020 MSF assessments, managed and administered by the Centre for Medical Professional Development Unit in Newcastle. There were 103 male and 39 female candidates from 28 countries (Afghanistan, Argentina, Bangladesh, Belgium, Burma, China, Egypt, Fiji, Germany, India, Indonesia, Iran, Iraq, Italy, Jordan, Kenya, Malta, Malaysia, Nepal, the Netherlands, Norway, Pakistan, Papua New Guinea, Romania, Sierra Leone, South Africa, Sudan and Ukraine). Each IMG completed their assessments over a 6-month period.

Over the 5-year study period, more than half the assessors attended at least one follow-up recalibration and feedback session. An independent group consisting of clinical academics, educationalists and administrators oversaw the governance of the program and continuously reviewed its quality. The assessment forms were sent, collected and analysed by the Centre for Medical Professional Development; the data were stored at a secure site at the John Hunter Hospital by the WBA program coordinator.

We analysed the average overall score of the mini-CEX and CBD assessments and the average scores of all scored items in the MSF assessments. When including MSF assessments in the WBA toolbox, the scores were linearly transformed from the 1–5 scale to a score on a 1–9 scale by multiplying the average score by 2 and subtracting 1. We did not analyse the MSF self-assessment results, as they were intended to assist self-reflection by the candidates, not to evaluate their performance. Reports from supervisors were not included in our analysis because they were found to be unreliable.[10]

Data analysis

All mini-CEX, CBD and MSF assessments over 6 months for a candidate were extracted. The secured records were analysed in SPSS 23 (IBM). For each assessment, we calculated the average score in order to determine the individual reliability of the various WBA tools, as well as the composite reliability of the tools as a group.

Reliability analysis assesses the reproducibility or consistency of WBA scores, providing an indication of how well we can discriminate between the levels of performance of IMGs, and of our confidence about their having achieved a passing score.

We employed generalisability theory,[12,14] an approach we have applied in previous studies. This model takes into account different sources of variance, and is therefore considered a useful framework for estimating the reliability of complex performance assessments.[19,20] In its simplest form, generalisability theory estimates the relative sizes of the variance components of factors affecting the measurement. Some variance components are desirable (eg, systematic variation between candidates), while others introduce undesirable variance, typically reflecting differences between assessors, cases, and other independent variables.

The variance components can be used to estimate reliability coefficients and the size of the total error. The reliability coefficient lies in the range 0 to 1; when providing a high stakes assessment based on a combination of several low stakes assessments, a reliability coefficient of 0.8 is generally regarded as acceptable.[21]

Total error can also be expressed as the standard error of measurement (SEM), which can be used to estimate confidence intervals for the original scores. A small SEM indicates that the estimate of a candidate's performance is more precise. Although reliability coefficients and the SEM are related algebraically, a large reliability coefficient is not necessarily associated with a small SEM.

The SEM is the more useful index for expressing reliability because one can define the confidence interval for a candidate's performance on the original scoring scale.[22,23] In the context of our investigation, a high stakes assessment that determined whether a candidate should be permitted independent clinical practice, we needed to reliably assess within one point on the 9-point scoring scale; that is, a confidence interval of 0.5 points around each score. For a 95% confidence interval, we divide 0.5 by the corresponding *z*-score (1.96) to calculate our SEM benchmark of 0.26.

The separate univariate variance components of each WBA instrument and the covariance between the instruments can be used to estimate the composite reliability of all instruments in a multivariate toolbox.[14] By varying the number of assessments of each type included and by differentially weighting the results of the individual assessment methods, a range of estimates of the composite reliability can be calculated. We therefore investigated which weightings of the individual assessment methods resulted in the optimal composite reliability.

Reliability analysis

The numbers of assessments and assessors varied between IMGs, and each assessor assessed a different set of IMGs. The facet (ie, source of variation) of average assessment scores (*i*) is therefore nested within the facet of IMGs (*p*), leading to the generalisability design *i*:*p*. For each WBA tool, we estimated variance components by analysis of variance with type I sums of squares (ANOVA SS1). The absolute error variance for the decision study on the separate WBA instruments is calculated by dividing the estimate of the variance component σ^2 (*i*:*p*) by the harmonic mean for each instrument. The harmonic mean was employed because the number of assessment scores differed between IMGs, and because the harmonic mean tends to reduce the effect of large outliers (ie, a single IMG with many assessments).[24]

In multivariate generalisability theory, the composite reliability of all instruments as a toolbox is calculated in a decision (D) study. For the D-study, each assessment score (*i*) is a score on exactly one assessment instrument, and the corresponding multivariate model is $i^{\circ}:p^{\bullet}$; that is, the facet of IMGs (*p*) is crossed with the fixed multivariate variables (assessment instruments) and nested within the independent facet of assessment scores (*i*). The composite universe score and absolute error variances are determined by a weighted sum of the universe scores and absolute error variances of the individual assessment instruments. Multivariable optimisation of the weights can be applied to obtain an optimal composite reliability coefficient.[14]

Ethics approval

Ethics approval for collecting and analysing the data was obtained from the Hunter New England Health Human Research Ethics Committee in 2010 (reference, AU201607-03 AU). All IMG candidates and assessors provided consent for analysing their de-identified data.

Results

Box 1 summarises the numbers of assessments and of IMGs tested during the study period, together with their mean scores (on a 1–9 scale, with standard deviations). As many IMGs undertook more than the required number of assessments, harmonic means of the numbers of assessments of each type were calculated for our analyses.

1 Numbers of assessments and of international medical graduates tested during the study period, June 2010 – April 2015, and summary of the test scores

	CBD	Mini-CEX	MSF
Number of assessments	970	1741	1020
Number of international medical graduates	142	142	142
Mean number of assessments per graduate	6.8	12.3	7.2
Mean test score	6.0	5.8	7.7
Standard deviation	0.7	0.6	0.5
Harmonic mean number of assessments	6.7	12.2	6.7

CBD = case-based discussion; mini-CEX = mini-clinical evaluation exercise; MSF = multi-source feedback. ◆

Reliability of the individual WBA instruments

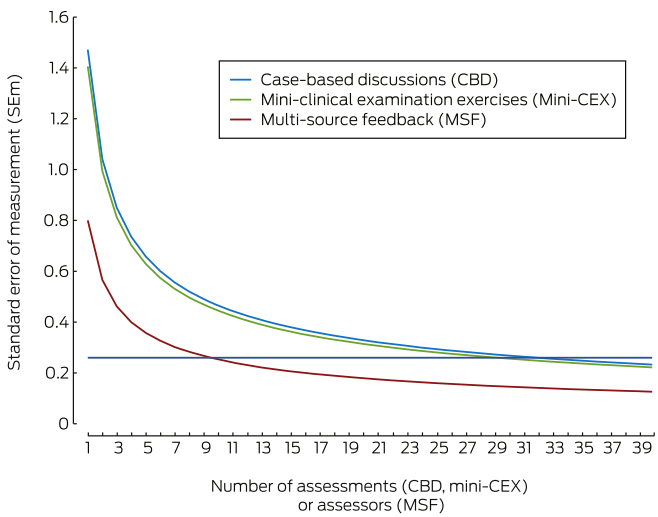
Box 2 depicts the SEM according to the number of assessments (CBD and mini-CEX) or assessors (per occasion of MSF). The data were derived from the regular variance components for the error variance associated with individual assessment tools. For an SEM of 0.26, the minimum numbers of assessments for each assessment type, if used alone, were 32 CBDs, 30 mini-CEXs and 10 MSFs.

Composite reliability of the WBA toolbox

We performed two composite reliability studies: one that excluded and one that included the MSF assessments. The rationale was that the CBD and mini-CEX assessments are similarly based on single observations by single assessors, whereas the MSF comprises a round of assessments of the performance of the IMG over a longer period of time.

When investigating combinations of CBD and mini-CEX assessments, the reliability threshold of an SEM of 0.26 could be obtained by combinations, for example, of 15 CBD and 16 mini-CEX or of 20 mini-CEX and 11 CBD assessments (Box 3). Most IMGs underwent 12 mini-CEX and five CBD assessments during the 6-month

2 The reliability of the individual workplace-based assessment instruments, as indicated by the standard error of measurement (SEM)*



* The maximum acceptable SEM level (0.26) is indicated by the horizontal line. ◆

3 Composite reliability for combinations of mini-clinical evaluation exercises and case-based discussion assessments, with optimised weights

		Number of mini-clinical evaluation exercises												
		10	11	12	13	14	15	16	17	18	19	20	21	22
Number of case-based discussions	10	0.321	0.313	0.306	0.299	0.292	0.286	0.280	0.275	0.270	0.265	0.260	0.256	0.252
	11	0.314	0.306	0.299	0.293	0.287	0.281	0.275	0.270	0.265	0.261	0.256	0.252	0.248
	12	0.307	0.300	0.293	0.287	0.281	0.276	0.271	0.266	0.261	0.257	0.253	0.249	0.245
	13	0.301	0.294	0.288	0.282	0.276	0.271	0.266	0.262	0.257	0.253	0.249	0.245	0.242
	14	0.295	0.288	0.282	0.277	0.272	0.267	0.262	0.258	0.253	0.249	0.246	0.242	0.238
	15	0.289	0.283	0.277	0.272	0.267	0.262	0.258	0.254	0.250	0.246	0.242	0.239	0.235
	16	0.283	0.278	0.273	0.268	0.263	0.258	0.254	0.250	0.246	0.243	0.239	0.236	0.232
	17	0.278	0.273	0.268	0.263	0.259	0.254	0.250	0.247	0.243	0.239	0.236	0.233	0.230
	18	0.273	0.268	0.264	0.259	0.255	0.251	0.247	0.243	0.240	0.236	0.233	0.230	0.227
	19	0.269	0.264	0.260	0.255	0.251	0.247	0.243	0.240	0.236	0.233	0.230	0.227	0.224
	20	0.264	0.260	0.256	0.251	0.248	0.244	0.240	0.237	0.233	0.230	0.227	0.224	0.222
	21	0.260	0.256	0.252	0.248	0.244	0.241	0.237	0.234	0.231	0.228	0.225	0.222	0.219
	22	0.256	0.252	0.248	0.244	0.241	0.237	0.234	0.231	0.228	0.225	0.222	0.219	0.217

Shaded cells: standard error of measurement < 0.26 (threshold for acceptability). ◆

training period, yielding an SEM of 0.35 (with optimised weighting: 0.33), exceeding the upper limit of 0.26.

If six MSF assessments on one occasion were added to the five CBD and 12 mini-CEX assessments, the SEM improved to 0.24. By applying the harmonic means in Box 1 — that is, assuming that the IMGs underwent seven CBD, 12 mini-CEX and one set of seven MSF assessments with optimised weighting[14] — a satisfactory SEM of 0.23 was achieved (Box 4).

A composite reliability coefficient of 0.8 could be achieved with a combination of 10 CBD assessments, 12 mini-CEX assessments, and 18 assessors per MSF, provided the weighting of the MSF assessments was much greater (0.72) than that for the other assessment types (each 0.14) (data not shown). The resulting SEM of 0.16 is more than adequate for assessment purposes.

Discussion

We found that a multivariate assessment toolbox can achieve a satisfactory level of precision (SEM < 0.26) with a practicable number of individual assessments. Moreover, combining different assessment methods that examine a broader range of attributes than each method alone achieves greater precision. In addition, a reliability coefficient of 0.8 can be achieved with 40 separate assessments (10 CBD, 12 mini-CEX, and 18 MSF assessments). While this number is quite high and may cause assessment fatigue for both trainees and assessors, the workload associated with the CBD and mini-CEX components is only marginally greater than the current assessment regimen. The MSF workload is shared by a large number of assessors, half of whom (the non-medical colleagues) are not involved in the other components.

4 Result of the D-study with equal and optimised weights* for the different workplace-based assessment tools, using the harmonic means of numbers of assessments

	CBD/Mini-CEX		CBD/Mini-CEX/MSF	
	Equal weights	Optimised weights	Equal weights	Optimised weights
Weights	0.50, 0.50	0.33, 0.67	0.333, 0.333, 0.333	0.20, 0.30, 0.50
Universe score	0.17	0.16	0.12	0.11
Error score	0.12	0.11	0.07	0.05
Reliability coefficient	0.58	0.60	0.65	0.67
SEM	0.35	0.33	0.26	0.23

CBD = case-based discussion; mini-CEX = Mini-Clinical Evaluation Exercise; MSF = multi-source feedback; SEM = standard error of measurement.
* That is, weights that minimise the SEM. ◆

Each instrument in the toolbox meets the standards of the AMC. They focus on different aspects of performance, but have comparable assessment scales and are applied by calibrated assessors. These characteristics allow for the combination of the WBAs in a single toolbox. Of the optimal weights for the individual instruments used in the aggregation for the composite score, the greatest weight is clearly that for the MSF, consistent with feedback from assessors; that is, the MSF makes the greatest contribution to the reliability of the toolbox. Content validity is another advantage of our program: in the AMC examination, standardised patients are employed over a period of 180 minutes, whereas the WBA is based on interactions with genuine patients over 180 days.

Our study has limitations, in that data were collected over 6 months. It has been argued that both classical test theory and generalisability theory may be compromised by repeated measures over a long period of time;[25] this would especially apply to our scale, which is based on "satisfactory performance". However, the process we are assessing is considerably shorter than most specialty training programs in which these analyses have been employed.[8,26] Moreover, these techniques are currently the best available for investigating the psychometric properties of WBA. Potential modifications of WBA tools, such as using scales with fixed reference points (eg, the standard of performance at the completion of training[25] or the amount of supervision the trainee requires[26]), may improve the psychometric quality of these instruments.

While attempting to concurrently achieve a reliability coefficient of 0.8 and an SEM below 0.26, we moved our chief focus from cohort-focused reliability coefficient values to the margin of error per individual assessed. Independently of the reliability coefficient, the SEM is the feature that drives reliable (confidence interval-based) discrimination between individuals and between an individual's score and standard or cut-off scores.

Assessment fatigue is a major problem in clinical assessment, and any assessment program should aim to optimise the demands on

assessors' time.[27] Combining different assessment instruments leads to fewer assessments per instrument being required for high stakes judgements. Our WBA program was acceptable to the IMGs because of the educational value provided by the immediate constructive feedback as described in our qualitative study.[12,28]

Verdicts about an assessment program should be based on the reliability, validity, acceptance, cost, and educational impact of the program. We have previously reported that this program is valid, has a satisfactory educational impact, and is acceptable to trainees, health services and assessors,[12] as well as being cost-effective.[13]

The performance of doctors (what they actually do) has a greater impact on patient care than competency (what they can do under examination conditions). WBA is relatively new in medicine; several lessons have already been learned, but many questions remain to be answered. The strength of WBA is that it can assess professionalism, decision making, and time management, as well as clinical skills.

The consensus statement from the 2011 Ottawa Conference on Assessment and Clinical Competence indicated that the outstanding problem for WBA is establishing sufficient reliability when combining the individual tools.[29] While our current WBA model is useful, its reliability can be improved by fine-tuning the combination of individual tools. This is especially important in the case of doctors from different training systems. WBA programs including multiple tools provide a reliable approach to assessing IMGs, and it can be delivered as a blue-printed program that assures the breadth and depth of assessment. Similar programs could significantly improve the clinical performance of IMGs and thereby patient outcomes. However, we do not know whether the long term outcomes for candidates examined by WBA differ from those of IMGs who passed the traditional examination; comparative investigations of the two pathways would be desirable.

Acknowledgements: We thank Kathy Ingham and Lynette Gunning (Centre for Medical Professional Development, John Hunter Hospital, Newcastle) for data collection, Ian Frank (Australian Medical Council) for ongoing support and Tim Wilkinson (Christchurch Medical School) for valuable comments on the manuscript).

Competing interests: No relevant disclosures.

Revised version received 10 Feb 2017, accepted 25 Sept 2017. ■

1 Tiffin PA, Illing J, Kasim AS, McLachlan JC. Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. *BMJ* 2014; 348: g2622.

2 Takahashi SG, Rothman A, Nayer M, et al. Validation of a large-scale clinical examination for international medical graduates. *Can Fam Physician* 2012; 58: e408-e417.

3 Peile E. Selecting an internationally diverse medical workforce. *BMJ* 2014; 348: g2696.

4 Neilson R. Authors have missed gap between theory and reality. *BMJ* 2008; 337: a1783.

5 Kogan JR, Conforti LN, Iobst WF, Holmboe ES. Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Acad Med* 2014; 89: 721-727.

6 Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ* 2010; 341: c5064.

7 ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med* 2007; 82: 542-547.

8 Wilkinson JR, Crossley JG, Wragg A, et al. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008; 42: 364-373.

9 van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005; 39: 309-317.

10 Bingham CM, Crampton R. A review of prevocational medical trainee assessment in New South Wales. *Med J Aust* 2011; 195: 410-412. https://www.mja.com.au/journal/2011/195/7/review-prevocational-medical-trainee-assessment-new-south-wales

11 van der Vleuten CP, Schuwirth LW, Scheele F, et al. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol* 2010; 24: 703-719.

12 Nair BK, Parvathy MS, Wilson A, et al. Workplace-based assessment; learner and assessor perspectives. *Adv Med Educ Pract* 2015; 6: 317-321.

13 Nair BK, Searles AM, Ling RI, et al. Workplace-based assessment for international medical graduates: at what cost? *Med J Aust* 2014; 200: 41-44. https://www.mja.com.au/journal/2014/200/1/workplace-based-assessment-international-medical-graduates-what-cost

14 Moonen-van Loon JM, Overeem K, Donkers HH, et al. Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Adv Health Sci Educ Theory Pract* 2013; 18: 1087-1102.

15 Australian Medical Council. AMC clinical examination [website]. http://www.amc.org.au/assessment/clinical-exam (accessed Sept 2015).

16 Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med* 2003; 138: 476-481.

17 Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. *Med Teach* 2007; 29: 855-871.

18 Davies H, Archer J, Southgate L, Norcini J. Initial evaluation of the first year of the Foundation Assessment Programme. *Med Educ* 2009; 43: 74-81.

19 Moonen-van Loon JM, Overeem K, Govaerts MJ, et al. The reliability of multisource feedback in competency-based assessment programs: the effects of multiple occasions and assessor groups. *Acad Med* 2015; 90: 1093-1099.

20 Swanson DB. A measurement framework for performance-based tests. In: Hart IR, Harden RM, editors. Further developments in assessing clinical competence. Montreal: Can-Heal, 1987; pp 13-45.

21 Crossley J, Davies H, Humphries G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002; 36: 972-978.

22 Tighe J, McManus IC, Dewhurst NG, The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Med Educ* 2010; 10: 40.

23 Norcini JJ. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med* 1999; 74: 1088-1090.

24 Brennan RL. Generalizability theory. New York: Springer, 2001.

25 Prescott-Clements L, van der Vleuten CP, Schuwirth LW, et al. Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Med Educ* 2008; 42: 488-495.

26 Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth* 2014; 112: 1083-1091.

27 Norcini J, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995; 795-799.

28 Lefroy J, Hawarden A, Gay SP, et al. Grades in formative workplace based assessment: a study of what works for whom and why. *Med Educ* 2015; 49: 307-320.

29 Boursicot K, Etheridge L, Setna Z, et al. Performance in assessment: consensus statement and recommendations from the Ottawa conference. *Med Teach* 2011; 33: 370-383. ■