Composite reliability of workplace-based assessment for international medical graduates

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

The known Workplace-based assessment (WBA) of the performance of doctors has gained increasing attention. The reliability of individual assessment tools has been reported in previous studies.

The new We analysed the composite reliability of a toolbox of WBA instruments in assessing international medical graduates (IMGs). For five case-based discussions, 12 Mini-Clinical Examination Exercises and six multisource feedback assessments, the composite reliability coefficient was 0.899 (standard error of measurement, 0.125).

The implications The reliability of WBA for assessing the performance of IMGs is excellent. WBA can also be used for performance assessment in other settings.

Abstract

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

Design and setting: Between June 2010 and April 2015, 142 IMGs were assessed by 99 calibrated assessors; each cohort was assessed at their workplace over 6 months. The IMGs completed 970 case-based discussions (CBDs), 1741 Mini-Clinical Examination Exercises (mini-CEX) and 1020 multisource feedback (MSF) sessions.

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

Main outcome measures: The reliability of the three WBA tools; the composite reliability of the tools as a group.

Results: The composite reliability of our WBA toolbox program was good: the composite reliability coefficient for five CBDs and 12 mini-CEX was 0.895 (standard error of measurement, 0.138). When the six MSF results were included, the composite reliability coefficient was 0.899 (standard error of measurement, 0.125).

Conclusions: WBA is a reliable method for assessing IMGs when multiple tools and assessors are used over a period of time. This form of assessment meets the criteria for "good assessment" (reliability ≥ 0.8) and can be applied in other settings.

The purpose of this article is to report the value of workplace-based assessment (WBA) for evaluating international medical graduates (IMGs). Most countries have systems for assessing IMGs. Fundamental to these systems are robust assessment procedures that assess their fitness to practise, and they typically include written multiple choice question tests and objective structured clinical examinations.[1,2] The virtue of standardised tools is that the assessment is similar for all candidates. Despite having been validated,[3] however, they do not assess proficiency in actual practice. 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, for both patients and society.[5-7] Many postgraduate training bodies are implementing WBA strategies,[8] and several undergraduate programs are already using some of its tools, particularly the Mini-Clinical Evaluation Exercise (mini-CEX), case-based discussions (CBDs), multisource 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] 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 assessment (for feedback and training), these tools have been used in programmatic assessment (in which multiple assessment tools are used to comprehensively assess a doctor or student in a well designed program) and can be used for summative purposes (to determine whether a candidate has passed or failed a course or program).

We hypothesise 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,[10] and our earlier study found that it is also cost-effective.[11] Although feedback from supervisors and staff indicate that WBA candidates are ready to work at a satisfactory level, there has been no reliability study of WBA for IMG assessment.

Moreover, studies of the reliability of WBA instruments typically focus on a single instrument, 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 undertaken by Moonen-van Loon and colleagues in a recent study of domestic graduates.[12] They found that combining the information from several methods meant that smaller samples were adequate (ie, fewer individual tests of each type).

The question therefore arises: what is the composite reliability of WBA when used for high stakes (ie, critical) assessment of IMGs?

[1]Centre for Medical Professional Development, John Hunter Hospital, Newcastle, NSW. [2]Maastricht University, Maastricht, The Netherlands. kichu.nair@newcastle.edu.au
doi: 10.5694/mja16.00069 See Editorial, p. 209

Our study estimated the composite reliability of a WBA program in Australia. Although trainees receive supervisor reports during most training programs, this has been found to "under-call under-performance", as the reports are prepared by a supervisor who is also the assessor (both coach and referee).[13] Since this was a routine assessment and many of the IMGs had completed different assessment forms, we only analysed the newer tools: mini-CEX, CBDs and MSF.[4,8]

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) in an examination centre.[14]

In 2010, we established a program to assess these doctors with 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 WBA in our program, the candidates had to pass the English and multiple choice question examinations, and be employed for the duration of the program (6 months). If the candidate passed our assessment, they were eligible for AMC certification. Our assessment program is accredited by the AMC.[15]

Data collection

Data were collected from June 2010 to April 2015. During this 5-year period, IMGs employed in Hunter New England Health, both in urban and rural areas, completed 970 CBDs, 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 a broad range of 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).

In total, 99 assessors rated the CBD and mini-CEX assessments. The MSF assessors were nominated by the IMGs, and the assessment forms were sent, collected and analysed by the central office; the forms were de-identified when results were provided to the candidates. Over the 5-year study period, more than half the assessors attended follow-up re-calibration and feedback sessions. Ongoing review of the quality of the program was undertaken by an independent group consisting of clinical academics, educationalists and administrators who oversaw the governance of the program. All candidates attended similar calibration sessions of about 3 hours each. Several different assessors assessed each IMG during the 6-month period. All results were recorded on the assessment forms and sent directly to the central office. The data were stored at a secure site.

WBA instruments (Appendix)

The assessment consisted of 12 mini-CEX examinations, five CBD examinations and one set of MSF data, and each candidate was assessed by at least six assessors. The mini-CEX assessments in medicine, surgery, women's health, paediatrics, emergency medicine and mental health were blue-printed (designed) to reflect the AMC examination. The assessment level was appropriate for the first postgraduate (intern) year.

The mini-CEX, originally developed in the US to guide learning, is used to assess clinical performance in authentic clinical situations.[16] The IMG was assessed in six disciplines and various competencies, and scored on a scale of 1 to 9; 1–3 corresponds to unsatisfactory performance, 4–6 to satisfactory performance, and 7–9 to superior performance. Case complexity and global rating were marked during the constructive feedback. The CBDs, which assess the candidate's record-keeping and clinical reasoning, were scored on a similar scale.[17,18] To pass, the IMG had to achieve a satisfactory result in eight of 12 mini-CEX and four of five CBDs, and to pass the MSF (with the average score of 3).

For the MSF, the IMG nominated three medical and three non-medical colleagues (eg, nurse, social worker, pharmacist) with whom they had worked extensively during the assessment period to complete an assessment form. The IMG also completed a self-assessment form. An MSF assessment form consisted of 23 questions with statements on aspects such as professionalism, communication, and requesting help when in doubt, and were scored on a 1–5 scale.[6,19]

We used the overall score of the mini-CEX and CBD assessments and the average scores of all scored items in the MSF assessments. When including the MSF assessments in the WBA toolbox, the scores were linearly transformed to a 1–9 score by multiplying the average score by 2 and subtracting 1.

We did not include the self-assessment results from the candidates in the MSF data, as this item was for their own reflection and not for evaluation of performance by external assessors. Reports from supervisors were not included in the analysis, as they have been found to be unreliable.[13]

Data analysis

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

Reliability analysis

Reliability analysis assesses the reproducibility or consistency of WBA scores, and therefore provides an indication on how well we can differentiate between the levels of performance (scores) of the IMGs. Generalisability theory takes into account different sources of variance and is therefore considered a useful framework for estimating the reliability of complex performance assessments.[20] It generates a reliability coefficient with a range of 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]

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 using 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 $\sigma^2(i{:}p)$ by the harmonic mean for each instrument. The harmonic mean was preferred to the arithmetic mean 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).[22]

Distinct from the separate univariate reliability of each WBA instrument, the composite reliability of all instruments as a toolbox is calculated using a D-study in multivariate generalisability theory.[22] Each assessment score (i) is a score for exactly one assessment instrument, and the corresponding multivariate model is therefore $i^{\circ}:p^{\bullet}$; ie, 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. The weights can be optimised by multivariable optimisation to obtain an optimal composite reliability coefficient.[12]

Ethics approval

Ethics approval to collect and analyse 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 to use their de-identified data.

Results

Box 1 summarises the number of assessments and the number of IMGs tested during the study period, with mean scores (on a 1–9 scale), standard deviations, and harmonic means for each of the assessment types (average number of assessments).

Reliability of the individual WBA instruments

Box 2 presents the reliability coefficients according to the number of assessments (CBD and mini-CEX) or assessors (one occasion of MSF). The data were derived from the regular variance components for the true and error variance associated with individual assessment tools. The minimum number of assessments needed for a reliability coefficient of 0.8 was 12 for CBDs, nine for mini-CEXs and ten for MSFs.

Composite reliability of the WBA toolbox

As 5-point scale was used for the MSF assessments, but 9-point scales for the CBD and mini-CEX assessments, we performed two composite reliability studies: one that excluded the MSF assessments, and one that included them after linearly transforming their scores to a 9-point scale.

The reliability threshold of 0.8 could be attained by a combination of five CBD and five mini-CEX assessments, but also with three CBD and six mini-CEX assessments (Box 3). As described in the introduction, IMGs generally undergo 12 mini-CEX and five CBD assessments during the 6 months; this combination had a reliability coefficient of 0.886 and a standard error of measurement (SEM) of 0.144. The SEM estimates how average scores per assessment of an IMG were distributed around their "true" score (ie, performance level). However, the dataset indicated that typically more assessments of all types were undertaken than required, leading to a reliability coefficient of 0.895 and an SEM of 0.138 when using harmonic means of test numbers and optimised weights (Box 1).

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	141	141
Mean number of assessments per graduate	6.8	12.3	7.2
Harmonic mean number of assessments	6.7	12.2	6.7
Mean test score	6.0	5.8	7.1
Standard deviation	0.7	0.6	0.7

CBD = case-based discussion; mini-CEX = Mini-Clinical Evaluation Exercise; MSF = multisource feedback. ◆

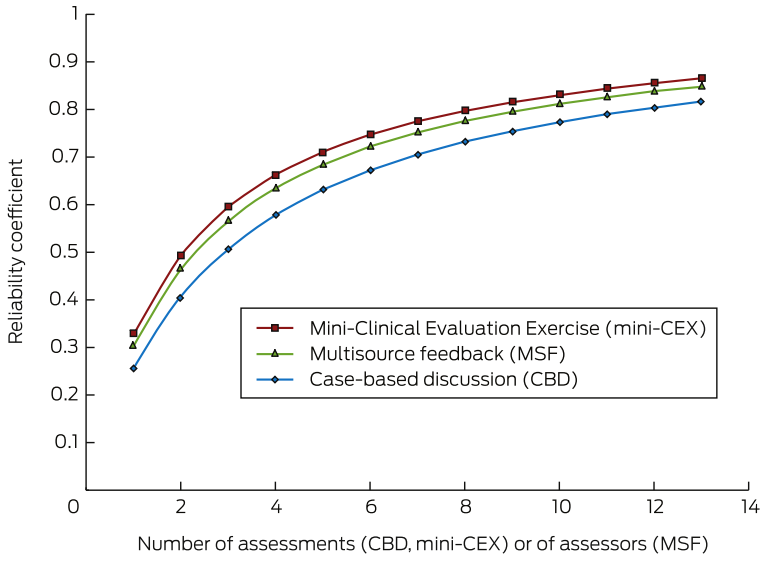
Adding the six MSF assessments to the five CBD and 12 mini-CEX assessments slightly increased the reliability coefficient from 0.886 to 0.890, with an SEM of 0.131. Using harmonic means (Box 1) and optimised weights, we obtained a reliability coefficient of 0.899 and an SEM of 0.125 (Box 4).

Discussion

An assessment instrument for evaluating performance in a high stakes setting should have a reliability coefficient of at least 0.8. Our study found that our WBA program meets this criterion. The composite reliability we found is as good as or even better than that of most standardised assessments.[23] Our previous studies have found the WBA program has good acceptability, educational impact, and validity.[10] Taken together, our program therefore satisfies the criteria for a "good assessment" program.[9]

Further, when the components were used as part of a WBA toolbox, we achieved good reliability with fewer individual assessments.[12] This may lead to changes in the procedure, reducing the workload for IMGs and assessors. It should be noted that all instruments in

2 The reliability of the individual workplace-based assessment instruments



3 Composite reliability when combining different numbers of Mini-Clinical Evaluation Exercises and case-based discussion assessments, with optimised weights

		Number of Mini-Clinical Evaluation Exercises												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Number of case-based discussions	1	0.463	0.577	0.650	0.701	0.739	0.769	0.792	0.812	0.827	0.841	0.852	0.862	0.871
	2	0.549	0.633	0.690	0.731	0.763	0.788	0.808	0.824	0.838	0.850	0.860	0.869	0.877
	3	0.610	0.676	0.722	0.756	0.782	0.803	0.821	0.835	0.848	0.858	0.867	0.876	0.883
	4	0.657	0.709	0.747	0.776	0.798	0.817	0.832	0.845	0.856	0.865	0.874	0.881	0.888
	5	0.693	0.736	0.768	0.792	0.812	0.828	0.842	0.853	0.863	0.872	0.880	0.886	0.892
	6	0.722	0.758	0.785	0.807	0.824	0.838	0.850	0.861	0.870	0.878	0.885	0.891	0.897
	7	0.746	0.777	0.800	0.819	0.834	0.847	0.858	0.868	0.876	0.883	0.889	0.895	0.900
	8	0.767	0.793	0.813	0.830	0.843	0.855	0.865	0.874	0.881	0.888	0.894	0.899	0.904
	9	0.784	0.807	0.825	0.839	0.852	0.862	0.871	0.879	0.886	0.892	0.898	0.903	0.907
	10	0.799	0.819	0.835	0.848	0.859	0.869	0.877	0.884	0.890	0.896	0.901	0.906	0.910
	11	0.812	0.829	0.844	0.856	0.866	0.874	0.882	0.889	0.895	0.900	0.905	0.909	0.913
	12	0.823	0.839	0.852	0.862	0.872	0.880	0.887	0.893	0.898	0.903	0.908	0.912	0.916
	13	0.833	0.847	0.859	0.869	0.877	0.885	0.891	0.897	0.902	0.907	0.911	0.915	0.918

Shaded cells: reliability coefficient ≥ 0.8 (threshold for acceptability). ◆

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	0.500, 0.500	0.270, 0.730	0.333, 0.333, 0.333	0.240, 0.638, 0.122
Universe score	0.169	0.163	0.123	0.140
Error score*	0.025	0.019	0.018	0.016
Reliability coefficient	0.869	0.895	0.870	0.899
SEM	0.160	0.138	0.136	0.125

CBD = case-based discussion. mini-CEX = Mini-Clinical Evaluation Exercise. MSF = multisource feedback. * Calculated by dividing the covariance by the harmonic mean, summed for all instruments, divided by the number of different instruments. ◆

the toolbox meet the standards set by the AMC. They focus on different aspects of performance, but have similar assessment scales, and are applied by assessors adhering to the same assessment standard after calibration. These characteristics allow for the combination of the WBAs in one toolbox, allowing composite reliability scores to be calculated. It is interesting that when we searched for optimal weights for individual instruments in the aggregation for the composite score, the mini-CEX received the most weight, perhaps because the mini-CEX has the highest individual reliability (Box 2).

Assessment fatigue is a major problem in clinical assessment, and any program should aim to optimise the use of the assessors' time.[24,25] With fewer assessments, more people are likely to implement such a program. The current program was also highly acceptable to the IMGs because of the educational value inherent in the immediate constructive feedback.[26]

We examined the performance of IMGs in Australia, but the 5-year study period and the large number of assessments included in the dataset render it sufficiently rigorous that the results can probably be extrapolated to other programs. However, the level of calibration of assessors and the structure of the assessment instruments should be similar if comparable results are to be obtained.

Evaluating the performance of doctors (what they do) is more important than assessing their competency (what they know), as their performance during training and practice is more relevant to patients and society. This is especially important in the case of doctors educated in different medical training systems. WBA programs with multiple tools provide a reliable method for assessing IMGs and can be delivered in a well organised, blueprinted program that assures the breadth and depth of the assessment. Similar programs could have a huge impact on the performance of IMGs, potentially improving patient outcomes. However, we do not know whether the long term performance of candidates who undergo WBA is different from IMGs who pass the traditional examination, and comparison of these outcomes for the two pathways would be desirable.

Most postgraduate training programs are adopting WBA components. The tools used by the assessors have individual reliabilities greater than 0.8, and our study may contribute to designing an improved portfolio of assessment, with different assessment tools for achieving more rigorous performance assessment.

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.

Received 20 Jan 2016, accepted 3 June 2016. ■

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

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

11 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

12 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.

13 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

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

15 Elkin K, Spittal MJ, Studdert DM. Risks of complaints and adverse disciplinary findings against international medical graduates in Victoria and Western Australia. *Med J Aust* 2012; 197: 448-452. https://www.mja.com.au/journal/2012/197/8/risks-complaints-and-adverse-disciplinary-findings-against-international-medical

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, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002; 36: 972-978.

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

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

24 Sabey A, Feest K, Gray S. Lessons from the UK: doctors' views of changes in postgraduate training. *Focus Health Prof Educ* 2010; 11: 42-51.

25 Nair BR, Hensley MJ, Parvathy MS, et al. A systematic approach to workplace-based assessment for international medical graduates. *Med J Aust* 2012; 196: 399-402. https://www.mja.com.au/journal/2012/196/6/systematic-approach-workplace-based-assessment-international-medical-graduates

26 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. ■