British Journal of Anaesthesia **112** (6): 1083–91 (2014)
Advance Access publication 17 March 2014 · doi:10.1093/bja/aeu052

BJA

QUALITY AND PATIENT SAFETY

Can I leave the theatre? A key to more reliable workplace-based assessment

J. M. Weller[1,2*], M. Misur[2], S. Nicolson[2], J. Morris[3], S. Ure[4], J. Crossley[5] and B. Jolly[6]

[1] Centre for Medical and Health Sciences Education, Faculty of Medical and Health Sciences, University of Auckland, 2 Park Rd, Grafton, Auckland 1010, New Zealand
[2] Department of Anaesthesia, Auckland City Hospital, 2 Park Rd, Grafton, Auckland 1010, New Zealand
[3] Department of Anaesthesia, Royal Melbourne Hospital, Grattan Street, Parkville, VIC 3052, Australia
[4] Department of Anaesthesia, Wellington Hospital, Riddiford Street, Newtown, Wellington 6021, New Zealand
[5] Academic Unit of Medical Education, University of Sheffield, 85 Wilkinson Street, Sheffield S10 2GJ, UK
[6] University of Newcastle, University Drive, Callaghan, Newcastle, NSW 2308, Australia

* Corresponding author. E-mail: j.weller@auckland.ac.nz

Editor's key points

- Existing tools for work-based clinical assessment have been limited by low reliability and capability to identify poorly performing individuals.
- This paper evaluated a new scoring system for clinical assessment of trainees.
- This system combined traditional assessments with the addition of case difficulty and the level of supervision required.
- This new scoring system appears reliable, with better detection of poor performance.

Background. The value of workplace-based assessments such as the mini-clinical evaluation exercise (mini-CEX), and clinicians' confidence and engagement in the process, has been constrained by low reliability and limited capacity to identify underperforming trainees. We proposed that changing the way supervisors make judgements about trainees would improve score reliability and identification of underperformers. Anaesthetists regularly make decisions about the level of trainee independence with a case, based on how closely they need to supervise them. We therefore used this as the basis for a new scoring system.

Methods. We analysed 338 mini-CEXs where supervisors scored trainees using the conventional system, and also scored trainee independence, based on the need for direct, or more distant, supervision. As supervisory requirements depend on case difficulty, we then compared the actual trainee independence score and the expected trainee independence score obtained externally.

Results. Compared with the conventional scoring system used in previous studies, reliability was very substantially improved using a system based on a trainee's level of independence with a case. Reliability improved further when this score was corrected for case difficulty. Furthermore, the new scoring system overcame the previously identified problem of assessor leniency and identified a number of trainees performing below expectations.

Conclusions. Supervisors' judgements on trainee independence with a case, based on the need for direct or more distant supervision, can generate reliable scores of trainee ability without the need for an onerous number of assessments, identify trainees performing below expectations, and track trainee progress towards independent specialist practice.

Keywords: educational assessment; educational measurement; medical education, graduate; reliability; workplace

Accepted for publication: 30 November 2013

Anaesthesia training programmes aim to produce graduates capable of independent specialist practice. Traditional assessments have emphasized knowledge acquisition, rather than clinical ability, and workplace-based assessments (WBAs) have been introduced across many postgraduate and undergraduate programmes to address this. WBAs are now a compulsory component of many specialist training programmes,[1] many using modifications of Norcini's mini-clinical evaluation exercise[2] (mini-CEX). The Royal College of Anaesthetists' (RCA) 'Anaesthesia Clinical Evaluation Exercise' (A-CEX) is an example.[3]

Reliability and validity are of central importance in any assessment, including WBAs. While anaesthesia fellowship examinations are valid and reliable tests of knowledge, they may not be a good measure of the ability to practice as an anaesthetist. While WBAs should be a more valid measure of

this ability, previous studies suggest they have low reliability, and fail to identify the struggling trainee whom experienced clinicians have no difficulty recognizing.[4 5]

In comparison with formal examinations, a number of factors affect the reliability of WBAs. Formal examinations can be standardized for difficulty and content, but WBAs cannot—cases are unpredictable and cannot be scheduled or repeated. The examiners in formal anaesthesia examinations are trained, and agree on set standards of performance, but the 'examiners' in WBAs can include all specialist anaesthetists working in teaching hospitals, many of whom have limited training in the use of the WBA tools.

In our previous study of the mini-CEX (modified for anaesthesia), more than 60 assessments were required to reach a level of reliability sufficient to make defensible decisions on trainee progression. Moreover, no trainee received an unsatisfactory grade in any of the 331 assessments we studied.[4] While interviews with trainees and supervisors strongly supported the value of mini-CEX to improve supervision and feedback, we found many anaesthetic supervisors lacked confidence in their ability to judge trainees against a scoring system that used the term 'expected level of performance', and were also reluctant to tell a trainee their performance was unsatisfactory.[5] WBAs depend on willing supervisors, but where WBAs are seen as unreliable, supervisors will disengage from the exercise, and decisions on trainee progression made on the basis of unreliable assessments will be open to challenge.

Data from studies of the way experts make judgements in complex settings, including medical ones,[6 7] suggest that a scoring system reflecting the way clinicians usually make judgements about trainees would reduce disagreement between them, and increase score precision. Anaesthesia supervisors are accustomed to judging the need for direct, indirect, or more distant supervision required by a trainee managing a particular case. We therefore developed a scoring system based on the extent to which the supervisor trusted the trainee to independently manage a case, with descriptors reflecting the need for close or more distant supervision (e.g. going to the theatre tearoom, the hospital cafeteria, being out of the hospital). We called this the 'trainee independence score'. To overcome the observed reluctance of supervisors to award scores of unsatisfactory or below standard,[5] we used non-pejorative descriptors, that is, the amount of supervision required.

Our primary hypothesis was that supervisors' scores would be more reliable when scoring trainee independence with the case than when using the conventional system scoring trainees below, at, or above expectations for stage of training.

The extent to which a trainee can independently manage a case depends on two factors—the ability of the trainee and case difficulty. Correcting for the latter required an external standard stating the extent to which the trainee should be able to manage independently a particular type of case at their stage of training, or from the supervisor's perspective, the need for direct, indirect, or distant supervision for such a case. Comparing the expected supervisory requirements with the actual supervisory requirements for a particular case allows calculation of the 'corrected trainee independence' score.

Our secondary hypotheses were: that the corrected trainee independence score would be more reliable than the (uncorrected) trainee independence score; and that the corrected independence score would identify more trainees performing below expectations than the conventional system.

Methods

The National Multi-region Ethics Committee considered the project fell under the category of quality assurance, where we were evaluating a development within an existing programme of assessment and ethics approval was not required. To ensure confidentiality, all trainee, case, and assessor data were de-identified on submission to the centralized database.

The context

This study took place in the anaesthetic departments of three major teaching hospitals, two in New Zealand and one in Australia before the introduction of compulsory mini-CEX assessments for the Australian and New Zealand College of Anaesthetists (ANZCA) training programme in 2013. The ANZCA training programme requires progression through five levels (basic trainee year 1–2, advanced trainee year 1–3).

The online mini-CEX form

We changed the scoring systems in the original version of our online mini-CEX form to address the identified issues of assessor variability and leniency.[4 5] We asked anaesthesia supervisors to rate the following: each of the 10 domains of practice against a scoring system of developing autonomy; overall level of independence with the case; and overall performance against that expected for stage of training. We used a nine-point scale for all scoring systems, divided into three categories, with three points in each category, each with descriptors. A word version of the online mini-CEX form used in this study is shown, with descriptors, in the Appendix.

Participants

Assessments were voluntary and all trainees and all supervisors in the three departments at the time of the study were eligible to submit assessments. Mini-CEX assessment data were submitted online in real time to a single database.

Sample size

We aimed to collect a minimum of 300 assessments, including a large and representative sample of trainees and supervisors from across the regions. Each of these factors is important for the precision and generalizability of the reliability estimates.[8]

Generating scores for expected level of independence for the case

We convened a panel of three experienced supervisors of training (SOTs). SOTs are appropriately trained specialist anaesthetists, officially appointed by ANZCA and responsible for training in ANZCA-accredited departments. They oversee each trainee's clinical performance and WBAs, perform regular clinical placement reviews, and confirm progression of trainees through the training programme. We provided the three SOTs

with case details of all the submitted mini-CEX assessments in this study. These details included patient age, gender, ASA physical status classification, surgical complexity score, surgical subspeciality, and name of the operative procedure. We removed all information relating to the trainee, hospital, date, and assessor and trainee scores. For each case, the three SOTs independently judged the expected independence score for a trainee in each of the 5 yr of the ANZCA training programme, on the basis of the level of supervision expected. We called this the 'expected trainee independence score'.

Generating scores for corrected level of independence for a trainee

To determine if the trainee required more or less supervision than expected for a case, we then calculated the difference between the independence score awarded by the supervisor who observed the case in theatre, and the level of independence expected of the trainee (from the mean SOT score). We called this the 'corrected trainee independence score'.

Scores generated for analysis

We thus generated four scores for each individual assessment for analysis:

(i) Overall performance for current level of training, with three points within each category of the nine-point scale: below (1–3); at (4–6); and above (7–9) expected level.
(ii) Composite mean score for the 10 domains of clinical practice, with three points within each category of the nine-point scale: required supervisor input for safe practice (1–3); generally autonomous, required some input (4–6); autonomous (7–9).
(iii) Trainee independence score with three points within each category of the nine-point scale: supervisor required in theatre (1–3); supervisor required in hospital (4–6); supervisor not required in hospital (7–9).
(iv) Corrected trainee independence score: difference between observed trainee independence score and expected trainee independence score (−8 to +8).

Analysis

To ensure that judgements from the three SOTs were meaningful and consistent, we calculated single and average intraclass correlation coefficients across SOTs for scores at each of the five levels of training.

The score on an assessment is, ideally, a true measure of trainee ability. However, the score inevitably contains errors due to a number of factors; variations between assessors (assessor scoring the same thing differently); variations between test items or cases; positive or negative interactions between assessors and trainees; interactions between assessors and particular test items or cases (e.g. pet topics); and interactions between trainees and particular items or cases. Generalizability theory allows the contribution of these different sources of error to be estimated and can be used to generate an estimate of reliability taking all these into account (*G* coefficient).[8 9] We used generalizability theory (MinQUE procedure, in SPSS GLM section) to calculate the impact on score variance of: trainee ability; assessor stringency (strictness, rigor); assessor subjectivity (across trainees); and residual case-to-case variation (which combines a number of factors including the case itself). D-studies were used to estimate the reliability of varying combinations of numbers of assessors and cases per trainee.

Results

We collected 338 assessments from 84 different assessors on 80 trainees from the three hospitals between September 2010 and May 2012. Individual trainees were assessed between one and 15 times on a mini-CEX by different assessors from the assessor pool. Fifty-six of the 80 trainees had at least two assessments, and 42 had at least three.

Intraclass correlation coefficients across three SOTs judgements on the expected level of trainee independence with the 338 cases for trainees ranged from 0.74 to 0.86 across the five levels of training ($P<0.001$ for each level) (Supplementary Appendix), thus establishing this as a reliable external standard.

Generalizability analysis: contribution of different factors to score variance

In a perfect assessment of trainee ability, all score-to-score variance would be due to trainee ability. In the real world, other factors such as assessor stringency and trainees' case-to-case variation in performance affect the scores.

Table 1 Variance components in generalizability analysis for different scales. Overall, overall performance in this case for current level of training; Composite, composite score of progression towards autonomy for the mean of 10 components of performance in this case; Independence, overall independence score in this case; Corrected, observed independence score for level of independence minus expected independence score for level of training in that case, which could have a potential range of −8 to +8. Var, variance

Variance component	Overall estimate (%)	Composite estimate (%)	Independence estimate (%)	Corrected estimate (%)	Factor interpretation
Var_trainee	0.112 (9%)	0.251 (22%)	0.992 (21%)	1.44 (28%)	Trainee ability
Var_assessor	0.340 (29%)	0.452 (40%)	0.847 (18%)	0.91 (18%)	Assessor stringency
Var_trainee × assessor	0.303 (26%)	0.248 (22%)	1.157 (25%)	0.08 (1%)	Assessor trainee-related subjectivity
Var_residual	0.431 (36%)	0.177 (16%)	1.661 (36%)	2.76 (54%)	Residual case-to-case variation

Variance estimates from the generalizability analyses using the four scoring systems are shown in Table 1.

The most obvious observation from the table is that all three scoring systems aligned to trainee independence reflect trainee ability better than the 'overall performance' scoring system that is aligned to 'expectation'. (For the estimate of variance, the percentage variance figures provide an easier comparison than the absolute variance figures, which depend on how much of the range of the scale is used.) In addition, the independence scoring systems (corrected and uncorrected) are less subject to variable supervisor stringency than the composite scores. Finally, the corrected independence score eliminates most of that part of the supervisors' variability in marking that can be linked directly to bias towards or against a particular trainee.

D-studies: required number of assessors and cases per trainee

In Tables 2–5, we show estimates of reliability generated from the generalizability analysis (called decision or 'D-studies'). While the *G* coefficient is a measure of overall reliability of the assessment, taking into account variance due to trainee, case, and assessor and the interaction between these factors, the D-studies, allow us to estimate reliability for a trainee with different numbers of assessors and cases in different configurations (e.g. more assessor *vs* more cases). The level of reliability required will depend on the purpose for which an assessment is being used and the consequences. Most educational measurement professionals suggest a reliability of at least 0.9 for very high stakes assessments such as licensure examinations, levels at or above 0.8 for the end of year assessments, and for assessments with lower consequences, such as formative or summative assessments administered by local faculty, one would expect reliability to be in the range of 0.7–0.8.[10] We would consider mini-CEX decreases most logically into the last category.

Reviewing the data in Tables 2–5, we noted mini-CEX assessment formats where the reliability coefficient is 0.7 or greater. From this, we can see that when supervisors use the

Table 2 Overall performance for current level of training: generalizability analysis D-studies showing reliability estimates for different numbers of cases and assessors

Number of assessors	Cases per assessor				
	1	2	3	4	5
1	0.09	0.12	0.13	0.13	0.13
2	0.17	0.21	0.22	0.23	0.24
3	0.24	0.28	0.30	0.31	0.32
4	0.30	0.34	0.36	0.37	0.38
5	0.34	0.40	0.42	0.43	0.44
6	0.39	0.44	0.46	0.47	0.48
7	0.42	0.48	0.50	0.51	0.52
8	0.46	0.51	0.53	0.55	0.55
9	0.49	0.54	0.56	0.57	0.58
10	0.51	0.57	0.59	0.60	0.61

Table 3 Composite score of progression towards autonomy: generalizability analysis D-studies showing reliability estimates for different numbers of cases and assessors

Number of assessors	Cases per assessor				
	1	2	3	4	5
1	0.22	0.24	0.25	0.25	0.25
2	0.36	0.39	0.40	0.40	0.41
3	0.46	0.49	0.50	0.50	0.51
4	0.53	0.56	0.57	0.57	0.58
5	0.59	0.61	0.62	0.63	0.63
6	0.63	0.66	0.67	0.67	0.67
7	0.67	0.69	0.70	0.70	0.71
8	0.70	0.72	0.73	0.73	0.73
9	0.72	0.74	0.75	0.75	0.75
10	0.74	0.76	0.77	0.77	0.77

Table 4 Overall independence score: generalizability analysis D-studies showing reliability estimates for different numbers of cases and assessors

Number of assessors	Cases per assessor				
	1	2	3	4	5
1	0.21	0.26	0.28	0.29	0.30
2	0.35	0.41	0.44	0.45	0.46
3	0.45	0.51	0.54	0.55	0.56
4	0.52	0.58	0.61	0.62	0.63
5	0.58	0.64	0.66	0.67	0.68
6	0.62	0.68	0.70	0.71	0.72
7	0.65	0.71	0.73	0.74	0.75
8	0.68	0.74	0.76	0.77	0.77
9	0.71	0.76	0.78	0.79	0.79
10	0.73	0.78	0.80	0.80	0.81

Table 5 Corrected independence score: generalizability analysis D-studies showing reliability estimates for different numbers of cases and assessors

Number of assessors	Cases per assessor				
	1	2	3	4	5
1	0.27	0.37	0.43	0.47	0.49
2	0.42	0.54	0.60	0.64	0.66
3	0.52	0.64	0.70	0.73	0.75
4	0.59	0.70	0.75	0.78	0.80
5	0.64	0.75	0.79	0.82	0.83
6	0.69	0.78	0.82	0.84	0.85
7	0.72	0.81	0.84	0.86	0.87
8	0.74	0.83	0.86	0.88	0.89
9	0.77	0.84	0.87	0.89	0.90
10	0.78	0.86	0.88	0.90	0.91

independence scoring system, there is a very marked reduction in the number of cases and assessors required to reach a reliability coefficient of 0.7. The overall performance scoring system fails to reach a reliability of 0.7 with 10 assessors each scoring the trainee in five cases, that is, 50 cases in total (Table 2); the composite scoring system (Table 3) and the independence scoring system (Table 4) each require between eight and nine cases to reach a reliability of 0.7. For the corrected independence scoring system to reach a reliability of 0.7, it requires just seven cases if each case is assessed by a different assessor. Assessor numbers can be reduced to four if each assesses two cases, or three if each assesses three cases. A reliability coefficient over 0.8 is attained with seven assessors each assessing two cases (Table 5).

Precision and spread of results for the four scoring systems

Figure 1 shows the spread of results for the different scoring systems and different levels of training and the resulting confidence limits of scores around thresholds of satisfactory/unsatisfactory performance. Assessors scored across the range of the scale for the independence scoring system, while for the conventional overall scoring system, scoring range was restricted, with most trainees scoring above expectations and only one trainee approaching the 95% confidence interval (CI) around the threshold.

In addition to reporting on score reliability with the generalizability coefficients, we also looked at the precision of scores. Precision is a measure of the range within which the real score actually lies—an indicator of score accuracy. Precision is calculated from the various sources of error as standard error of measurement (SEM) and relates to the score scale. CIs of 95% (observed score ± 2 SEMs) indicate that we can be 95% confident that the real score decreases within this range. In our data, for the corrected trainee independence score, we set the threshold score as zero (i.e. no difference between observed and expected). We identified six trainees who fell more than 2 SEMs (95% CI) below zero, and could therefore be

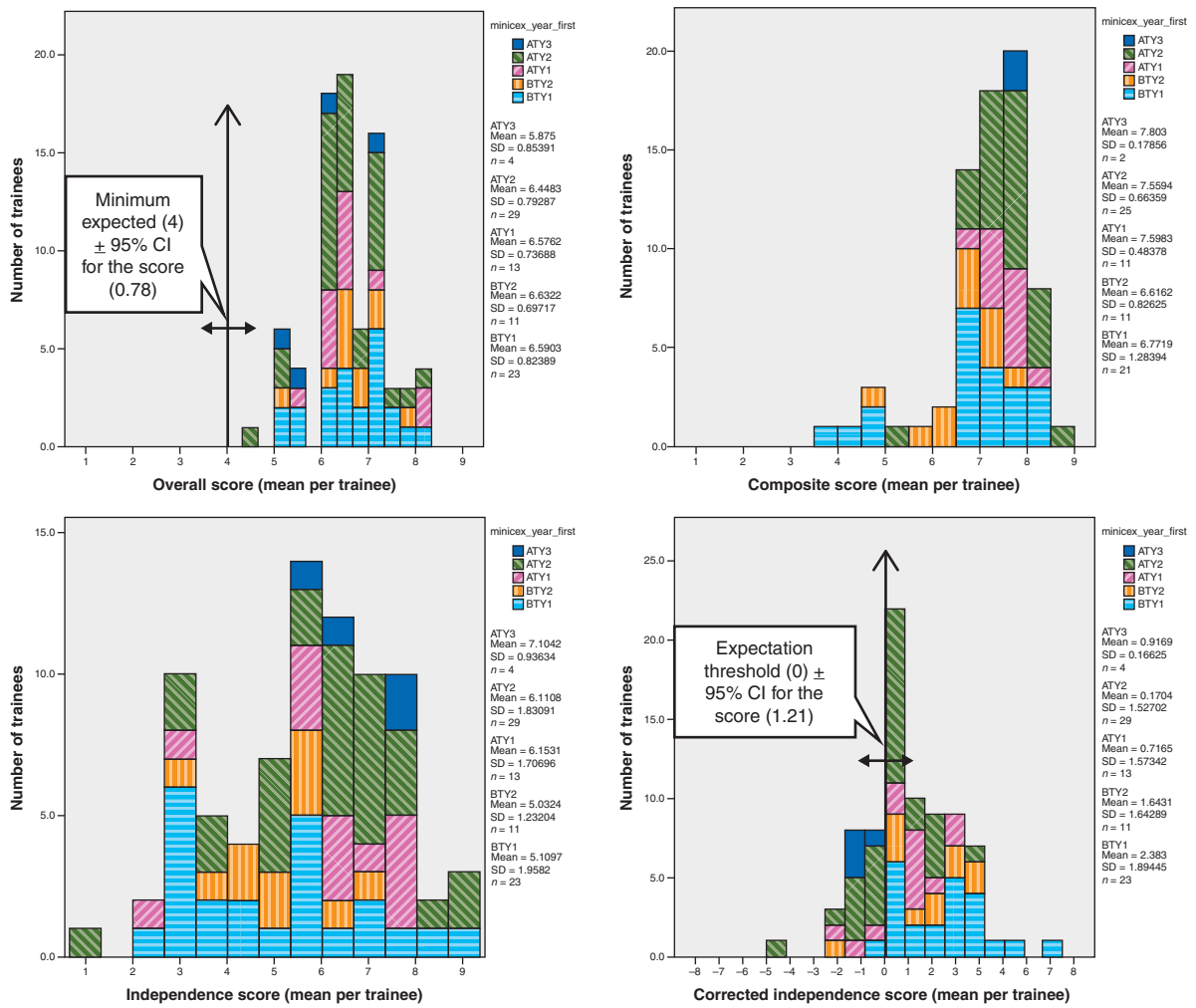

Fig 1 Spread of trainees' mean scores for each of the four scales—subdivided by training year, with CI around threshold score. *n*, number of trainees at each level (total 80 trainees, incomplete data for composite scale).

considered as underperforming; 42 trainees were within 2 SEMs of zero [0 (1.21)] and could be considered as performing within expectations; and 32 trainees were more than 2 SEMs above zero, and were clearly doing well. Applying the conventional scoring system to this same group of trainees, where the threshold for satisfactory performance is set at 4; 95% CI$=\pm 2$ SEMs$=\pm 0.78$, the lowest scoring trainee fell in the uncertain range raising some doubts about their performance, and all others scored above this, that is, at or above standard expected for that year of training. (CI calculations based on 15 cases—five assessors each observing three cases.)

The spread of trainee scores at different training levels for the independence scoring systems suggests trainees were selecting cases at the cusp of their abilities. In a *post hoc* analysis using the variable 'case complexity' as a proxy for case difficulty, we did in fact find a weak but significant correlation with trainee seniority (Spearman's $\rho=0.150$, $P<0.01$). Of note, trainees identified as less independent than expected with the case included trainees from all five levels of the programme.

Discussion

Compared with the conventional scoring system, supervisors' scores for mini-CEX assessments were considerably more reliable when scoring trainee independence with the case on the basis of need for direct, indirect, or distant supervision. A reliability coefficient of 0.7 could be attained with only nine assessments with the trainee independence score, whereas this was not attained with 50 assessments in the conventional scoring system.

Reliability was further increased when the trainee independence score was adjusted for case difficulty against an external standard. Furthermore, using the corrected trainee independence score, we identified a number of trainees requiring closer supervision than expected for their year of training. Using the conventional system, not one of these trainees was identified as performing below expected standards.

This confirms our proposition that anaesthesia supervisors can make reliable judgements when asked to judge if the trainee required direct, indirect, or distant supervision with the case, but have difficulty when asked to judge what is expected of a trainee at different stages of their training. A number of practical implications arise from our findings. First, using conventional scoring systems, the number of mini-CEXs required for a reliable estimate of trainee ability (over 50 cases) is well beyond the limits of feasibility, in contrast to the new scoring system where fewer than 10 mini-CEX assessments are sufficient. This suggests that the mini-CEX, and possibly other WBAs using similar scoring systems, can be defensibly used for high stakes decisions on trainee progress. Secondly, the new system allows tracking of trainee progression over time towards the final goal of independent specialist practice and allows early identification and remediation of trainees not tracking along the expected curve.

Reliability estimates for mini-CEX assessments vary considerably, but high assessor variability is a problem across studies,[11] and this does not seem amenable to training.[12] In our previous study, we found that to achieve the minimum acceptable reliability of 0.7, three assessors would need to each score the trainee in 20 cases (60 observations),[4] and no trainee was awarded an unsatisfactory score in any of the 331 assessments. We generated similar findings with the conventional scoring system in our current study, thus confirming our previous findings, and the shortcomings of the conventional scoring system.

The use of a corrected trainee independence score, comparing scores from departmental anaesthesia supervisors against an externally derived standard, is novel. While the trainee independence score awarded by the departmental supervisor is reliable, it depends on case difficulty and thus requires interpretation either by the trainee's SOT or against an external standard. We used descriptions of the 338 cases in our study to generate this expected standard, using the mean of three SOTs judgements for expected levels of supervision required. Future research will use this data set to develop and validate a general formula for estimating the expected level of supervision for a trainee for any case based on ASA status, surgical complexity, patient age, and subspeciality.

As trainees progress through the anaesthesia training programme, they will take on increasingly challenging cases. Their need for supervision may thus remain stable, expectations will increase, and thus the independence score may remain stable over time. We did in fact find that trainees were assessed in more challenging cases as they progressed through the training scheme. While the independence score, corrected or otherwise, may remain stable, the difficulty of the cases that the trainee can manage would be expected to follow an increasing trajectory.

Limitations

While we have identified trainees whose average scores suggested they needed closer supervision than would be expected for their stage of training, we have no other comparative measure of their performance. Future studies should look for relationships with other measures of performance, including progression through the training scheme and formal assessments.

We used a modified version of the mini-CEX for our WBAs in this study. We would expect that our scoring system would produce similar results when applied to other forms of WBAs that depend on supervisors making judgements on trainees, but this remains to be tested.

The extent to which our scoring system is generalizable to medical domains beyond anaesthesia is unclear. The context for supervision, and the way supervisors make judgements on their trainees or students may be different. In particular, the extent to which well-defined criteria for case difficulty exist to enable development of an external standard is unclear.

This study was undertaken in volunteer trainees in large, well-resourced teaching hospitals. Results from compulsory mini-CEX assessments after introduction of this scoring system across all ANZCA training sites will be the subject of future research.

Conclusion

To make the best use of WBAs in anaesthesia training, our findings suggest we should ask supervisors to make judgements on the basis of the level of supervision a trainee needs when managing a case. With this approach, WBAs such as mini-CEX are a feasible and reliable option in anaesthesia specialist training programmes to make judgements on trainee progression, with greater potential to identify underperforming trainees to facilitate timely remediation.

Supplementary material

Supplementary material is available at *British Journal of Anaesthesia* online.

Authors' contributions

J.M.W.: conception and design, acquisition of data, analysis and interpretation of data, writing and revising the article, and submission of the final approved version. M.M.: conception and design, creation of online assessment form, acquisition of data, data management, and final approval of the submitted version. S.N.: conception and design, generating data for levels of performance, acquisition of assessment data, revising article, and final approval. J.M.: acquisition of data, generating data for levels of performance, revising article, and final approval of the submitted version. S.U.: acquisition of data, generating data for levels of performance, revising article, and final approval of the submitted version. J.C.: data analysis and interpretation, writing sections of manuscript, critically revising article, and final approval of the submitted version. B.J.: data analysis and interpretation, writing sections of manuscript, critically revising article, and final approval of the submitted version.

Acknowledgement

We wish to acknowledge the contribution of all the supervisors and trainees who contributed mini-CEX assessments during the study period.

Declaration of interest

J.M.W. is an Associate Board member of the *BJA*.

Funding

No external funding.

References

1 *Academy of Medical Royal Colleges. Improving Assessment*. London, UK: AoMRC Press, 2009

2 Norcini J, Blank L, Duffy F. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med* 2003; **138**: 476–81

3 Royal College of Anaesthetists. *Workplace Based Assessments*. Available from http://www.rcoa.ac.uk/training-and-the-training-programme/workplace-based-assessments-wpba (accessed October 2013)

4 Weller J, Jolly B, Merry A, *et al.* Mini-clinical evaluation exercise in anaesthesia training. *Br J Anaesth* 2009; **102**: 633–41

5 Weller J, Jones A, Merry A, Jolly B, Saunders D. Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: implications for implementation. *Br J Anaesth* 2009; **103**: 524–30

6 Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ* 2011; **45**: 560–9

7 Crossley J, Jolly B. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Med Educ* 2012; **46**: 28–37

8 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002; **36**: 972–8

9 Shavelson R, Webb N. *Generalizability Theory: A Primer*. Newbury Park, CA: Sage Publications, 1991

10 Downing S. Reliability: on the reproducibility of assessment data. *Med Educ* 2004; **38**: 1006–12

11 Pelgrim EAM, Kramer AWM, Mokkink HGA, den Elsen L, Grol RPTM, Vleuten CPM. In-training assessment using direct observation of single-patient encounters: a literature review. *Adv Health Sci Educ* 2011; **16**: 131–42

12 Cook D, Dupras D, Beckman T, Thomas K, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med* 2009; **24**: 74–9

Appendix. Mini-CEX form and descriptors for scale and domains

Anaesthesia Mini-CEX (written version of online form)

Case Details

Case Description _____
Surgical Subspecialty _____
Surgical Complexity _____
Setting _____
ASA _____
Age _____

Surgical Complexity
Minimal – e.g. cystoscopy, I&D
Moderate – e.g. lap appendicectomy, TURP, ORIF, Fem-pop bypass
High – e.g. body cavity surgery, craniotomy, and knee replacement

Progression to Autonomy	Required Supervisor input for safe practice			Generally autonomous, some guidance required			Autonomous practice			NC
Domains	1	2	3	4	5	6	7	8	9	
Patient assessment/investigation										
Preparation for anaesthesia										
Clinical planning										
Patient communication										
Staff communication										
Procedural skills										
Problem solving/decision making										
Vigilance										
Organisation/efficiency										
Professionalism										

Level of Independence	Supervisor required in the theatre suite			Supervisor required in hospital			Supervisor not required		
	1	2	3	4	5	6	7	8	9
What level of supervision did the trainee require for this case?									

Overall Performance in this Case for Current Level of Training	Below expected level			At expected level			Above expected level		
	1	2	3	4	5	6	7	8	9
Overall performance in this case for current level of training									

Supervisor's Feedback	
What did the trainee manage well in this case?	
What areas still need supervisory input for safe practice?	
What specific actions can the trainee take to improve in these areas?	

Scale descriptors:

Progression to autonomy

1-3: Required supervisor input for safe practice – Gaps in knowledge, skills or decision making that required input from supervisor to ensure safe anaesthesia care.

4-6: Generally autonomous, some guidance required – Acceptable knowledge skills or decision making for safe anaesthesia care, some guidance required.

7-9: Autonomous practice – Able to manage this aspect of the case independently at consultant level.

NC - Please select this if you feel unable to comment.

Level of independence

1-3: Supervisor required in the theatre suite

(1) Supervisor not comfortable leaving trainee unsupervised in theatre for any period of time.

(2) Supervisor comfortable to leave trainee for brief coffee break in theatre tea room. Not happy for trainee to instigate changes in management in your absence.

(3) As in 2, but comfortable staying out of theatre for a bit longer, e.g. while eating your lunch. Trainee may instigate some new actions that you have previously discussed.

4-6: Supervisor required in hospital

(4) As in 5, but supervisor feels the need to check in on the trainee at regular intervals.

(5) Supervisor happy to leave the theatre block, but remain immediately available in the hospital, e.g. not take on another case themselves. Expect trainee to notify supervisor of any significant problem or event, e.g. persistent abnormal physiological parameter, major blood loss.

(6) As in 5 but expect trainee to manage most problems initially, and call you if their initial management doesn't work.

7-9: Supervisor not required

(7) Supervisor could potentially be off-site but would want to review the trainee's management plan before the trainee started the case.

(8) Supervisor off-site. Confident that trainee can make a good assessment and plan, but want to be notified that they are doing the case.

(9) Trainee could manage this case as a consultant. Appropriate if they don't contact supervisor. May have collegial discussion on case.

Domain item descriptors

- Patient assessment *I* investigations - Elicits relevant information from history and examination of the patient, gathers information from patient notes and investigations including medication history and allergies. Appropriately orders further investigations or preop treatment.
- Preparation for anaesthesia - Prepares for anaesthetic appropriately - checks equipment and anaesthetic machine, organizes theatre and monitoring, prepares drugs, ensures appropriate personnel present.
- Clinical planning - Formulates an appropriate plan for anaesthetising the patient or managing the patient in the clinic or on the pain round.
- Patient communication - Explores patient's perspective, jargon free, open and honest, agrees management plan with patient.
- Staff communication - Works effectively and appropriately in an interprofessional team. Communicates anaesthesia plan to appropriate staff, maintains open communication with surgical team. Fosters effective team communication (open, two-way, clear, concise, closes communication loop).
- Procedural skills - Proficiency in vascular access, application of monitoring, regional technique, airway management, patient positioning.
- Problem solving/decision making - Responds appropriately and in a timely manner to changes in the patient's status or to unanticipated events. Interprets available data. Integrates information to generate differential diagnoses and management plans. Demonstrates effective management of clinical problems and complications. Performs appropriate diagnostic studies or interventions, considers risks and benefits.
- Vigilance - Demonstrates an awareness of the status of the patient (through constant clinical and electronic monitoring), the procedure and other personnel. Maintains focus on patient care and avoids distraction. Anticipates and prepares for future events.
- Organisation/efficiency - Prioritizes, is timely, succinct. Well organized workspace, efficient use of time and resources without compromising patient care. Good standard of record keeping.
- Professionalism - Shows respect, compassion and empathy for patient and establishes trust. Attends to patient's needs and comfort. Respects confidentiality. Behaves in an ethical manner, aware of legal frameworks for consent. Shows integrity. Aware of own limitations including risk of fatigue, impairment. Commitment to quality and safety (e.g. practices to reduce medical error, complies with hospital protocols).

Handling editor: J. P. Thompson