BJA

British Journal of Anaesthesia, 118 (2): 207–14 (2017)

doi: 10.1093/bja/aew412
Clinical Practice

Making robust assessments of specialist trainees' workplace performance

J. M. Weller[1,2,*], D. J. Castanelli[3,4], Y. Chen[1] and B. Jolly[5]

[1]Centre for Medical and Health Sciences Education, School of Medicine, University of Auckland, New Zealand, [2]Department of Anaesthesia, Auckland City Hospital, New Zealand, [3]Department of Anaesthesia and Perioperative Medicine, Monash Health, Victoria, Australia, [4]Department of Anaesthesia and Perioperative Medicine, Monash University, Clayton, Victoria, Australia and [5]Medical Education Unit, School of Medicine and Public Health, Faculty of Health and Medicine, University of Newcastle, New South Wales, Australia

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

Abstract

Background. Workplace-based assessments should provide a reliable measure of trainee performance, but have met with mixed success. We proposed that using an entrustability scale, where supervisors scored trainees on the level of supervision required for the case would improve the utility of compulsory mini-clinical evaluation exercise (CEX) assessments in a large anaesthesia training program.
Methods. We analysed mini-CEX scores from all Australian and New Zealand College of Anaesthetists trainees submitted to an online database over a 12-month period. Supervisors' scores were adjusted for the expected supervision requirement for the case for trainees at different stages of training. We used generalisability theory to determine score reliability.
Results. 7808 assessments were available for analysis. Supervision requirements decreased significantly ($P < 0.05$) with increased duration and level of training, supporting validity. We found moderate reliability ($G > 0.7$) with a feasible number of assessments. Adjusting scores against the expected supervision requirement considerably improved reliability, with $G > 0.8$ achieved with only nine assessments. Three per cent of trainees generated average mini-CEX scores below the expected standard.
Conclusions. Using an entrustment scoring system, where supervisors score trainees on the level of supervision required, mini-CEX scores demonstrated moderate reliability within a feasible number of assessments, and evidence of validity. When scores were adjusted against an expected standard, underperforming trainees could be identified, and reliability much improved. Taken together with other evidence on trainee ability, the mini-CEX is of sufficient reliability for inclusion in high stakes decisions on trainee progression towards independent specialist practice.

Key words: educational assessment; educational measurement; medical education, graduate; reliability; workplace

The move to competency-based medical education demands some measure of a trainee's ability to work independently and provide safe, effective and efficient care.[1 2] Workplace-based assessments (WBAs) have been introduced widely in specialist training programs, after the description of the mini-clinical evaluation exercise (mini-CEX) by Norcini in 1995 to partly address this need.[3] A further potential benefit of WBAs is that when supervisors stand back and observe trainees, and use a structured format for feedback, the quality of the feedback improves.[4–8]

Editorial decision: November 8, 2016; **accepted:** November 15, 2016

Editor's Key Points

- Structured, formal evaluations are becoming embedded throughout medical specialist training.
- An entrustment process occurs in which an expert supervisor fosters growing independence of practice, but objective measures are needed.
- This study found that mini-clinical evaluation exercise assessments are valid, reliable and can identify underperformers.
- These findings from Australia and New Zealand need confirmation in other settings.

However, WBA implementation has met with mixed fortune. At worst, WBAs have been described as an unreliable tick box exercise in compliance, an unhelpful administrative hurdle, and of little value to trainees or supervisors.[9 10] Because of perceived lack of value in formal assessment decisions, and the potential negative effect on feedback if perceived as summative, some institutions have moved to a formative-only stance on WBA.[11] We consider this is not utilizing WBA to its full potential.

The anaesthesia curriculum can be described in terms of the work that needs to be done, and entrustment decisions made on areas of work that can be safely entrusted to the trainee. These areas of work have been called Entrustable Professional Activities.[12] Clinical supervisors habitually make judgements on the extent to which they can leave the care of their cases in the hands of a trainee. Entrustment scales have been proposed as a way of capturing this expert judgement in WBA[13] and improving the reliability of clinical supervisor ratings. Changing the WBA scoring system to reflect this entrustment decision could generate reliable assessments, that could indeed be used to make defensible decisions on trainees' ability to progress through the training scheme to independent practice.[8 14 15]

This study represents the third phase of a program of research on the mini-CEX in the Australian and New Zealand College of Anaesthetists (ANZCA) training program. In our context, the mini-CEX comprises an holistic assessment of a trainee's performance over an entire case from planning and preparation through case management, and including communication, team collaboration and risk minimisation. In our first study[16] supervisors were asked to make judgements on trainees on a scale of unsatisfactory, satisfactory or superior performance. While the quality of supervision and feedback improved, we estimated over 50 assessments would be required to generate a reliable score for any trainee, and we did not identify any trainee whose performance was classified as unsatisfactory.[7 16] In our second study[15] supervisors were asked to judge how closely they needed to supervise the trainee for the case (i.e. from within the theatre suite, or hospital, or from an offsite location). This led to markedly improved reliability estimates. In addition, when scores were adjusted against an independently derived standard for the expected level of performance with that case a reliable estimate of trainee ability would be obtained with as few as ten mini-CEX assessments. In addition, a substantial group of trainees was identified who performed below the expected standard,[15] a capacity of the mini-CEX that we had previously identified as lacking.

In 2013, ANZCA introduced a raft of compulsory workplace-based assessments, including mini-CEX, for all anaesthesia trainees in Australia and New Zealand, using our previously tested scale based on supervision requirement (SReq) (BOX 1). We were unsure if the very positive results from our small studies involving volunteer trainees and supervisors would translate to the real world of ANZCA training, with around 1500 trainees, over 4000 potential supervisors, and compulsory mini-CEX assessments.

In this study we explored reliability and validity of the mini-CEX assessments using all such assessments submitted to the Trainee Portfolio System (TPS). As in our previous study, we were also interested in the reliability of the scores for SReq adjusted against an external standard for expected SReq (i.e. with a specific case did the trainee require more or less supervision than expected for their training level).

Our specific research questions (with the evidence they would supply) were:

1. Are mini-CEX scores for SReq strongly related to level of training? (Evidence of construct validity)
2. What is the reliability of the mini-CEX scores for SReq? (Evidence of reliability)
3. What is the reliability of the score for the observed Mini-CEX SReq when adjusted for expected SReqs for that case (Evidence on how variation from a standard might be more useful than a simple score)
4. Can we use mini-CEX scores to identify the underperforming trainee? (Evidence that, contrary to some studies' findings, WBA can detect underperformance reliably)

Methods

Ethics and consent

Ethics approval was obtained from the University of Auckland Human Participant Ethics Committee (Ref. 011108) and the Monash University Human Research Ethics Committee (CF14/1668 – 2014000796). ANZCA trainees sign a training agreement in which they consent to their training records being accessed by those with appropriate authority and to the use of de-identified TPS data for the purpose of monitoring and evaluation. Access to the Trainee Portfolio System data was through ANZCA staff and all data provided to the research team for analysis was encrypted such that no individual trainee or supervisor could be identified.

Context

The ANZCA Training Program: The ANZCA training program comprises four distinct stages which are completed in a minimum of five yr. These stages are: Introductory Training (IT), zero to six months, where trainees are under direct supervision and must pass the Initial Assessment of Anaesthetic Competence; Basic Training (BT), a further 18 months, during which trainees may undertake some work under indirect supervision, participate in the after-hours roster, and must pass the FANZCA Part 1 Exam; Advanced Trainee (AT), of two yr duration, during which time trainees must pass the FANZCA Part 2 Exam; and Provisional Fellowship Trainee (PFT) of one yr duration, where trainees may undertake a subspecialty fellowship and prepare for independent practice. Trainees enter extended training (IT-E, BT-E, AT-E, PFT-E) when they fail to complete the requirements for that stage. In addition to time and formal assessments, these requirements include specified volumes of practice, research, teaching and audit activities, and a minimum number of WBAs submitted to the TPS. These WBAs comprise Direct Observation of Procedural Skills (DOPS), case-based discussion, Multi-source Feedback and mini-CEX. This study focusses on the mini-CEX.

Mini-CEX assessments in ANZCA: The mini-CEX scale and descriptors can be viewed at http://www.anzca.edu.au/documents/mini-cex.pdf. Supervisors mark trainees on a 9-point scale with three categories: Trainee needs assessor in the theatre suite; Trainee needs assessor in the hospital; and Trainee could manage this case independently and does not require direct supervision. This scale is accompanied by descriptors for each point on the scale, for example, 1= Not comfortable leaving the trainee unsupervised in theatre for any period of time, to 9= Trainee could manage this case as a consultant. Appropriate if they don't contact supervisor. May have collegial discussion on the phone.

The ANZCA mini-CEX on-line form is completed by supervisors on cases selected by trainees and supervisors. Case selection is before the start of the case but not otherwise constrained. Access to the form is via ANZCA login for both trainee and supervisor. Completed assessments are submitted electronically to a central data base – the TPS. The Departmental Supervisor of Training (SOT) has access to the mini-CEX data and may use them to inform decisions on progression to the next stage of training. There is currently no ANZCA guideline on how mini-CEX should be used in decisions of progression to the next stage of training.

Data source

Our analysis used the scores entered into the TPS database from the mini-CEX assessments by supervisors from all ANZCA trainees in Australia and New Zealand. The mini-CEX form included a rating for case complexity based on case co-morbidities, age and surgical procedure. This was rated by the supervisor and provided an overall measure of case difficulty. Other self-populating fields included trainee characteristics, and identifier codes for the trainee and supervisor.

Generating scores for expected SReq for the case

The supervision required in a case depends on case difficulty. A senior trainee would be expected to require the supervisor in the theatre suite for complex cardiac surgery, thus scoring 1-3 on the SReq scale. The same trainee would however be expected to manage a straightforward patient requiring cholecystectomy with the supervisor at a distance (i.e. 7-9). To interpret the SReq score and decide if a particular trainee was performing at, above or below expectations for that case, we adjusted their raw SReq score against a standard for expected SReq. We had previously generated expected SReqs for a series of cases for trainees at different levels of training.[15] Three SOTs independently judged the expected supervisory requirement for each of the 338 cases in that study. The ICC values for the judgements produced for each level of training ranged from 0.74 to 0.86. We sorted these 338 cases by procedure, and excluded surgical procedures with fewer than three cases in that category. We were left with 31 different surgical procedures where we could generate stable scores for expected levels of supervision for use in the current study.

Generating scores for observed minus expected (O-E) SReq

From the TPS data base of all mini-CEX assessments, we identified a subset of assessments involving the 31 surgical procedures for which we had generated the standard of performance. To determine if the trainee required more or less supervision than expected for a case we calculated the difference between the SReq score awarded by the supervisor in theatre, and the expected SReq. We called this the [O-E] SReq score. The scale limits were -8 and +8, and the score was zero when the trainee's requirement for supervision was at the expected standard.

Statistical analysis

To determine if SReq decreased with increasing duration of training, we calculated mean scores and 95% confidence intervals around these scores for SReq for trainees at each level of ANZCA training. These confidence intervals were calculated using the standard error of measurement (SEM) defined as $\surd(\sigma^A/N^A+\sigma^{AT}/N^A+\sigma^R/N^A.\ N^C)$ and used as 1.96xSEM.

We used generalizability theory to estimate score reliability. The score variance on a mini-CEX assessment would ideally be largely due to trainee ability, but scores will also be influenced by other factors. Generalizability theory quantifies the variance components for all the sources of error in a score: 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 variance itself). A G co-efficient of 0.8 or above is considered acceptable for a high stakes assessment decision. Generalizability theory can be used to predict score reliability (G coefficient) using different combinations of case and assessor numbers, taking all these into account in what are called Decision studies or D studies.[17] [18] We used MinQUE variance component procedure, in SPSS version 23 General Linear Model section, to account for unbalanced study design where only one assessor rates each case and all cases are unique. The MINQUE comprises a system of multiple generated linear equations in which variables are the ratios of the sought variance components to the residual variance, and the residual variance itself. The system of linear equations is then solved to obtain the MINQUE estimates. In this analysis, in simplified terms, we identified the following as predicting the overall variance of the data

$$\text{Total Variance} = \sigma^T+\sigma^A+\sigma^{AT}+\sigma^R$$

Where $\sigma^{T,}$ σ^A σ^{AT} and σ^R are the variance components for Trainee, Assessor, Assessor x Trainee interaction and Residual error respectively. The data were unbalanced and repeated for trainees and assessors but each mini-CEX was a separate case so it was not possible to identify individual components of variance for cases, case x trainee interaction and assessor x case interaction. These components are all part of the residual error. In a naturalistic environment it is not possible to estimate these components without trainees repeating the same case and/or multiple assessors viewing the same case.[19] However because individual trainees were sampled across more than one occasion (case), it was possible to estimate reduction in the contribution of residual error because of repeated measures across different cases and use the variance component for the residual error sampling distribution in the estimation of the G score.

Hence, data from trainees with only one mini-CEX assessment were removed from the analysis. Variance components were calculated for SReq scores and for [O-E] SReq scores. D-studies were used to estimate the reliability and precision of scores when derived from varying combinations of numbers of assessors and cases per trainee for both scoring systems.

The G coefficients in the decision tables (D studies) for these data were derived by the following formula $G = \sigma^{T}/(\sigma^{T} + \sigma^{A}/N^{A} + \sigma^{AT}/N^{A} + \sigma^{R}/N^{A}.\ N^{C})$, where G = generalizability coefficient; N^{A} is the number of assessors used, N^{C} is the number of cases used to assess the trainee.

Results

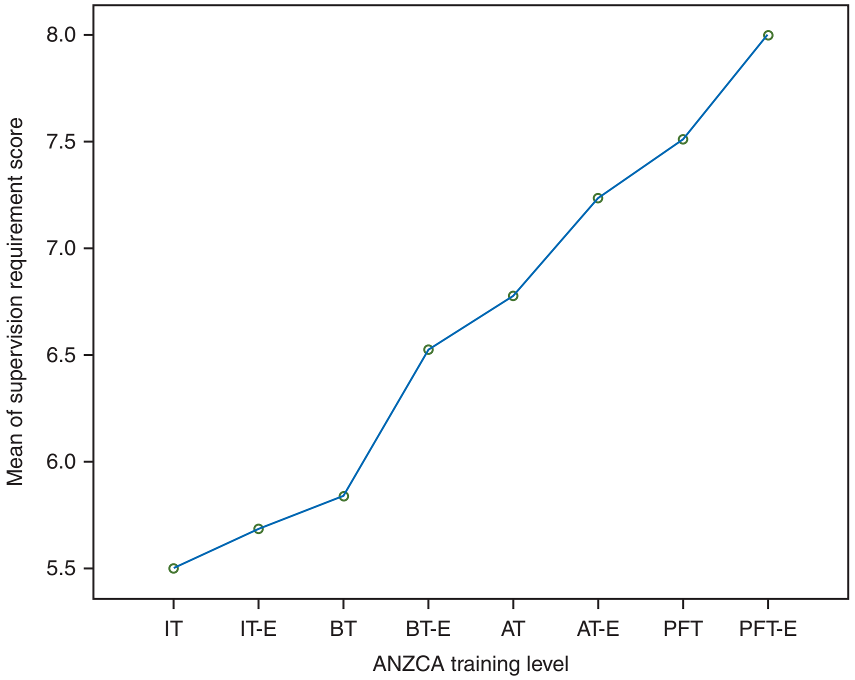
There were a total of 7808 mini-CEX assessments in the TPS data base. Scores for SReq demonstrated a spread across the full range of scores, with a mean score of 6.43 on the 9 point scale. Scores for SReq plotted against ANZCA training level increased progressively with the duration of training (Fig. 1). Excluding trainees in extended training (IT-E, BT- E, AT-E), we found significant differences ($P < 0.05$) between each ANZCA training level. (Table of 95% confidence intervals around scores is available as supplementary material.)

The progressive increase in SReq scores with increasing clinical experience supports the validity of the assessment, as more senior trainees move towards independent specialist practice. Of note, for trainees in extended training (IT-E, BT-E, AT-E, PT-E), who were held back at that level of training for various reasons (e.g. failure to pass an examination, failure to meet other training requirements) mean SReq scores were higher than those at the same training level who were not in extended training (Fig. 1). For this group, our results indicate the SReq scores, a measure of trainee ability to manage the work independently, continue to improve with time in training, independent of trainees' ability to meet the other requirements of the ANZCA training program.

We found a moderate correlation of 0.371 ($P < 0.01$) between the case complexity score (overall case difficulty), and the ANZCA training level of the trainee. This means that in general, more senior trainees chose more complex cases for the mini-CEX assessment.

We identified 60 cases from trainees with only one mini-CEX assessment in the TPS, leaving 7748 assessments for analysis. These 7748 cases involving 1149 trainees and 2401 assessors were included in the generalizability analysis of SReq scores. For the [O-E] SReq scores, there were 4147 cases, involving 1092 trainees and 1835 assessors from the TPS that matched the 31 common surgical procedures for which we had generated a score for expected SReq. Distribution of trainees and assessments across stages of training is shown in Table 1.

The variance components from the generalizability analysis of the SReq scores and the [O-E] SReq scores are shown in Table 2. The most obvious observation is the higher variance



IT=introductory trainee (first six months)

BT=basic trainee (18-24 months)

AT=advanced trainee (24-56 months)

PFT=provisional fellow trainee

E=extended time (trainee has not met all requriments to pass into next stage)

N=7808

Fig 1. Mean of supervision requirement score for trainees at each level of ANZCA training.

Table 1. Number of trainees and numbers of mini-CEX assessments for each stage of ANZCA training. Numbers in brackets relate to the subset of assessments adjusted against an external standard – (Observed minus expected supervision requirement score)

Stage of training	Numbers of trainees	Number of assessments	Average number of assessments per trainee
Introductory Trainee (IT)	52 (51)	312 (161)	6 (3.2)
Introductory Trainee Extended (IT-E)	22 (20)	125 (70)	5.7 (3.5)
Basic Trainee (BT)	375 (363)	2755 (1513)	7.3 (4.2)
Basic Trainee Extended (BT-E)	79 (71)	475 (238)	6 (3.4)
Advanced Trainee (AT)	475 (449)	3298 (1717)	6.9 (3.8)
Advanced Trainee Extended (AT-E)	34 (33)	214 (125)	6.3 (3.8)
Provisional Fellow (PFT)	110 (103)	564 (320)	5.1 (3.1)
Provisional Fellow Extended (PFT-E)	2 (2)	5 (3)	2.5 (1.5)
Totals	1149 (1092)	7748 (4147)	6.7 (3.8)

Table 2. Variance estimates for Supervision Requirement (SReq) scores for the 7748 cases from the TPS data base (score range one to nine) and variance estimates for [O-E] SReq for 4147 assessments (score range minus six to plus eight)

Variance estimates	SReq scores		[O-E]SReq scores	
Component	Estimate	%	Estimate	%
Variance (trainee ability)	0.526	17.99	1.526	30.48
Variance (assessor stringency)	0.631	21.58	1.073	21.43
Variance(assessor subjectivity)	0.307	10.50	0.499	9.97
Variance (residual)	1.459	49.90	1.908	38.11
Variance (total)	2.924	99.97%	5.006	99.99%

Table 3. D-Study using Supervision Requirement scores showing the estimated Generalisability (G) co-efficients for different numbers of assessors and cases per assessor. A G-coefficient ≥0.8 is considered highly reliable. N = 7748

	Cases per assessor				
Assessors	1	2	3	4	5
1	0.18	0.24	0.27	0.29	0.30
2	0.31	0.39	0.42	0.45	0.46
3	0.40	0.49	0.53	0.55	0.56
4	0.47	0.56	0.60	0.62	0.63
5	0.52	0.61	0.65	0.67	0.68
6	0.57	0.65	0.69	0.71	0.72
7	0.61	0.69	0.72	0.74	0.75
8	0.64	0.72	0.75	0.76	0.77
9	0.66	0.74	0.77	0.78	0.79
10	0.69	0.76	0.79	**0.80**	**0.81**

attributed to trainee ability in the [O-E] SReq, which is what we are in fact hoping to see.

Table 3 shows the D study for the SReq score. While moderate reliability (G > 0.7) is achieved with eight assessors each doing two cases, high reliability (G ≥ 0.8) would require ten assessors each doing four cases. Table 4 shows the D study for the [O-E] SReq score. Here we see that a G co-efficient G ≥0.8 can be achieved with considerably fewer assessments, for example nine assessors each doing one case, or seven assessors each doing two cases.

Table 4. D-Study using observed minus expected supervision requirement score showing the estimated generalisability (G) co-efficients for different numbers of assessors and cases per assessor. A G-coefficient ≥0.8 is considered highly reliable. N = 4147

	Cases per assessor				
Assessors	1	2	3	4	5
1	0.30	0.38	0.41	0.43	0.44
2	0.47	0.55	0.58	0.60	0.61
3	0.57	0.64	0.67	0.69	0.70
4	0.64	0.71	0.73	0.75	0.76
5	0.69	0.75	0.78	0.79	**0.80**
6	0.72	0.78	**0.81**	**0.82**	**0.82**
7	0.75	**0.81**	**0.83**	**0.84**	**0.85**
8	0.78	**0.83**	**0.85**	**0.86**	**0.86**
9	**0.80**	**0.84**	**0.86**	**0.87**	**0.88**
10	**0.81**	**0.86**	**0.87**	**0.88**	**0.89**

The distribution of [O-E] SReq scores around zero (at expected level), is demonstrated in Figure 2.

There were 811 cases where the trainees' [O-E] SReq mini-CEX score for that particular case fell below the expected level (zero), 460 trainees who had at least one [O-E] SReq mini-CEX below expected level of performance, and 123 trainees (3%) whose average [O-E] SReq score for their combined mini-CEX assessments fell below zero (IT = 2, BT = 16, AT = 104, PFT = 1). For these 123 trainees, the number of mini-CEX assessments for each trainee ranged from two to 19, with a median of three, and average scores ranged from minus 3.34 to minus 0.01.

Table 5 shows the estimated 95% confidence intervals around [O-E] SReq scores for a trainee with different numbers of assessors and cases per assessor. These Confidence Intervals are calculated using the Standard Error of Measurement (SEM) defined as $\sqrt{}(s^A/N^A + s^{AT}/N^A + s^R /N^A. N^C)$ and used as 1.96xSEM. For example, we can be 95% confident that a trainee with an[O-E] SR score of 1.18 points or more below the expected standard (on the 9-point scale) from 14 assessments (7 assessors each doing 2 cases) is under-performing. Doing more cases would generate somewhat more certainty around decisions on progression. For a trainee requiring substantially more supervision for a case than expected, this could be detected after a very small number of assessments. For example, we could be 95% confident that a trainee with [O-E] SR scores 2 or more points

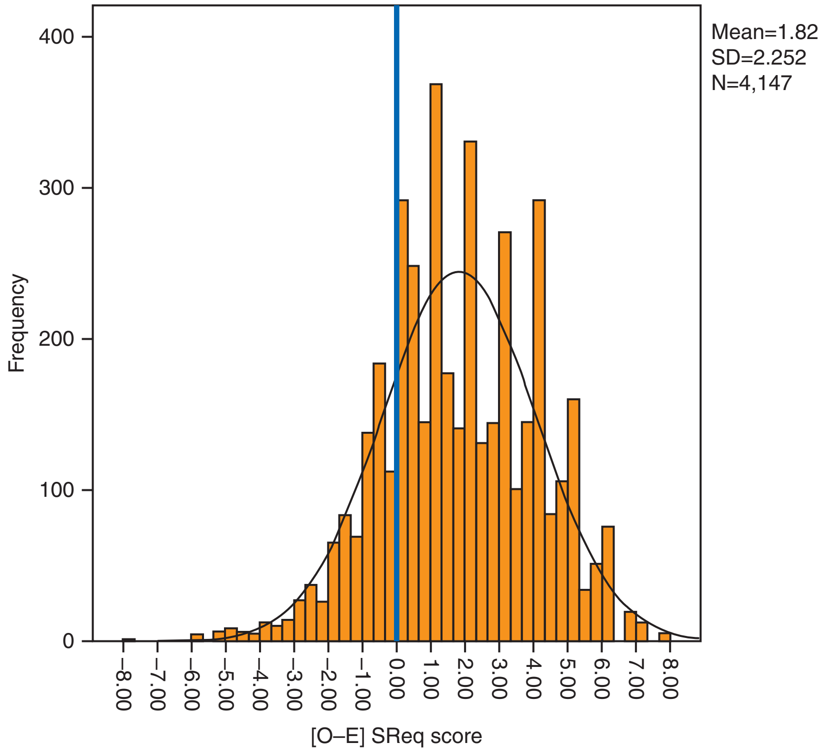


Fig 2. Distribution of [O-E] SR scores around zero (cases where the trainee had done only one mini-CEX are excluded) showing a distribution of cases falling at, above or below expected levels of performance. N = 4147.

below the expected standard over as few as four cases is in difficulty, and should be flagged for additional assistance. On the other hand, for trainees performing well above the expected standards, fewer assessments would be required to confidently make a decision on progression.

Discussion

In this study of 7748 mini-CEX assessments, we showed that when supervisors make judgements based on how much supervision their trainee required for the case, these judgements can provide moderately reliable scores. Adjusting supervisor scores against a standard for expected SReq for common cases considerably increased reliability, as indicated by the increase in score variance attributable to trainee ability. We could determine if a trainee was performing at, above or below expectations with a high degree of reliability (G >0.8) within a feasible number of assessments. This high level of reliability is sufficient for high stakes decisions on trainee progression. Furthermore, we were able to identify underperformers; 3% of trainees had average mini-CEX scores falling below expectations. Trainees performing well below expectations could be identified with only four mini-CEX assessments, providing an opportunity for early remediation. Furthermore, we provided evidence to support the validity of our scoring system for mini-CEX. Trainees required progressively less supervision with increasing duration of training, with significant differences in SReq between ANZCA training levels.

This provides clear support for using a scoring system that is intuitive to supervisors.[14]

There are relatively few publications reporting the reliability of WBA assessments using entrustment scales. Crossley and colleagues[20] compared scores from 2000 WBA assessments of UK Foundation doctors, where assessors scored trainees using both a traditional scale of performing at, above or below expectations to a scale and a scale aligned with supervisory requirement and demonstrated greater reliability with the scale aligned to SReq. George and colleagues[21] scored 31 surgical residents performing operative procedures using a resident operative autonomy scale and demonstrated good inter-rater reliability, validity and feasibility. Like ours, the results of these studies support the use of entrustment scales for WBA.

Estimates of trainee performance that include 95% confidence intervals are helpful. For example (Table 5), a trainee with an[O-E] SReq score of 1.18 points or more below the expected standard (on the 9-point scale) from 14 assessments (seven assessors each doing two cases) is likely to be under-performing. Doing more cases would generate somewhat more certainty around decisions on progression. For a trainee requiring substantially more supervision for a case than expected, this could be picked up after a very small number of assessments. For example we could be confident that a trainee with [O-E] SReq scores two or more points below the expected standard over as few as four cases is in difficulty, and should be flagged for additional assistance. On the other hand, for trainees performing well above the expected standards, fewer assessments would be required to confidently make a decision on progression.

We believe that ours is the first large study to incorporate a standard setting exercise into the interpretation of mini-CEX assessments and has implications for all specialist training programs.

Our D studies provide guidance to the ANZCA training program and individual training rotations on the different combinations of case and assessor numbers required to generate reliable scores for trainees. Of note, while a reliable score could be generated by nine assessors, each only scoring one case, nine cases may not be enough to provide a valid sample across the range of possible cases for a trainee at a particular stage in their training.

As would be expected adjusting the supervision score against an external standard for expected performance considerably improved reliability. Adjusting the SReq score for a case against an external standard takes into account case difficulty. Trainees are not penalized for selecting difficult cases. For easy cases, the expected standard will be higher. A relevant analogy would be the Olympic diving competition, where scores depend on dive difficulty and execution of the dive. Trainees may thus be encouraged to demonstrate their ability by choosing the more challenging cases for assessment. By choosing cases at the cusp of their ability, they are more likely to receive useful feedback on areas for further development, maximising the value of the mini-CEX assessment for their own learning. This is the subject of a related qualitative study.[22]

Table 5. 95% Confidence Intervals around Observed minus expected Supervision Requirement scores from the D-Study, showing CI for score estimates for different combinations of assessor numbers and cases per assessor. Scores were on a nine point scale. N = 4147

	Patients per assessor				
Assessors	1	2	3	4	5
1	3.66	3.12	2.91	2.81	2.74
2	2.59	2.20	2.06	1.98	1.94
3	2.11	1.80	1.68	1.62	1.58
4	1.83	1.56	1.46	1.40	1.37
5	1.64	1.39	1.30	1.25	1.23
6	1.49	1.27	1.19	1.15	1.12
7	1.38	1.18	1.10	1.06	1.04
8	1.29	1.10	1.03	0.99	0.97
9	1.22	1.04	0.97	0.94	0.91
10	1.16	0.99	0.92	0.89	0.87

In this study, decisions on trainee ability were made by comparing supervisor scores against an external standard. In an anaesthesia training program a number of options exist to put this into practice. Firstly, with a comprehensive set of standards for supervisory requirement in index cases and an online database of trainee assessments, an automated process could generate aggregate mini-CEX [O-E] SReq scores for trainees. The training program could use these scores in decisions on progression to the next level of training. These automated scores could usefully provide an early flag for an at risk trainee, or inform trainees on their own progress. A more considered approach to score interpretation may be preferred. At key points in progression within a training program (e.g. basic to advanced training, advanced training to provisional fellowship, graduation as a specialist) a panel of experts could review a trainee's portfolio of workplace assessments and make defensible decisions on trainee progression knowing the supervisor scores for level of supervision required in the case, the expected standard for those cases, and the number of assessments required to make a reliable decision. Taken together with formal examinations, we could thus have more confidence that we are graduating specialists who are both knowledgeable and can do the work to an acceptable standard.

In a competency-based specialist training programme where the curriculum is described in terms of EPAs,[12] progression through the curriculum depends on decisions on the degree of entrustment of the work to the trainee for each EPA. Scores from mini-CEX assessments based on the SReq score could be used to make robust decisions on the degree of entrustment of clinical work to a trainee in any particular area of practice. For example, regional anaesthesia for elective Caesarean section on a healthy parturient with uncomplicated pregnancy, with the supervisor in the hospital could be considered as an early

Box 1. Mini-CEX scale and descriptors.

What level of supervision did the trainee require for THIS patient overall	*Trainee needs assessor in the theatre suite*			*Trainee needs assessor in the hospital*			*Trainee could manage this patient independently and does not require direct supervision*		
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

1. *Not comfortable leaving the trainee unsupervised in theatre for any period of time.*
2. *Comfortable to leave trainee to go on brief coffee break in theatre tearoom. Not happy for trainee to instigate changes in management in your absences.*
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. *Happy to leave the theatre block but remain immediately available in the hospital. Feels the need to check in on the trainee at regular intervals.*
5. *Happy to leave the theatre block but remain immediately available (e.g. not take on another patient). 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. *Could potentially be off-site but would want to review the trainee's management plan before they started the patient.*
8. *Supervisor off-site. Confident that the trainee can make a good assessment and plant, but want to be notified that they are doing the patient.*
9. *Trainee could manage this patient as a consultant, Appropriate if they don't contact supervisor. May have collegial discussion on the phone.*

milestone. General anaesthesia for Caesarean section on a compromised parturient with the supervisor at a distance could be considered a late milestone in training, approaching independent specialist practice.

Limitations

The standard setting exercise was based on the spread of cases available from an earlier study of 338 cases. The surgical procedures included in the [O-E] SReq analysis were those with sufficient numbers to give reliable estimates of standard of performance, and thus were common cases. A more valid and comprehensive approach to standard setting would use a purposeful selection of surgical procedures across the whole area to be tested. This could be a task for future development. The mini-CEX is one of a number of WBAs in anaesthesia training programs. Exploring the reliability and feasibility of a combination of WBAs was beyond the scope of this study but an area for future research.

Conclusion

Using a scoring system where supervisors score trainees on the level of supervision required, moderate reliability can be achieved within a feasible number of assessments. When scores are adjusted against an expected standard, high reliability can be achieved with only nine assessments, and underperforming trainees can be identified. Taken together with other evidence on trainee ability, the mini-CEX using this entrustment scale is of sufficient reliability for inclusion in high stakes decisions on trainee progression towards independent practice.

Authors' contributions

Study design/planning: J.W., D.C., B.J. Study conduct: J.W., D.C., Y.C. Data analysis: J.W., D.C., Y.C., B.J. Writing paper: J.W., B.J. Revising paper: all authors

Supplementary material

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

Declaration of interest

None declared.

Funding

This work was supported by a grant from the Australian and New Zealand College of Anaesthetists (S24-002).

References

1. Miller G. Performance assessment. *Acad Med* 1990; **65**: S63–S7
2. ten Cate OP, Hart DMD, Ankel FMD, *et al*. Entrustment decision making in clinical training. *Acad Med* 2016; **91**: 191–8
3. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Int Med* 1995; **123**: 795–9
4. Carr S. The Foundation Programme assessment tools: an opportunity to enhance feedback to trainees? *Postgrad Med J* 2006; **82**: 576–9
5. Hauer KE. Enhancing feedback to students using the mini-CEX (Clinical Evaluation Exercise). *Acad Med* 2000; **75**: 524
6. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Med Teacher* 2007; **29**: 855–71
7. 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; **102**: 633–41
8. Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008; **42**: 364–73.
9. Bindal T, Wall D, Goodyear HM. Trainee doctors' views on workplace-based assessments: Are they just a tick box exercise? *Med Teacher* 2011; **33**: 919–27
10. Weston PSJ, Smith CA. The use of mini-CEX in UK foundation training six years following its introduction: lessons still to be learned and the benefit of formal teaching regarding its utility. *Med Teacher* 2014; **36**: 155–63
11. Rees CE, Cleland JA, Dennis A, Kelly N, Mattick K, Monrouxe LV. Supervised learning events in the Foundation Programme: a UK-wide narrative interview study. *Br Med J Open* 2014; **4**: e005980
12. ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. *Med Teacher* 2015; **37**: 983–1002
13. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: outlining their usefulness for competency-based clinical assessment. *Acad Med* 2016; **91**: 186–90
14. 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
15. Weller J, Misur M, Nicolson S, *et al*. Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth* 2014; **112**: 1083–91
16. Weller J, Jolly B, Merry A, *et al*. Mini-clinical evaluation exercise in anaesthesia training. *Br J Anaesth* 2009; **102**: 633–41
17. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002; **36**: 972–8
18. Shavelson R, Webb N, *Generalizability Theory: A Primer*. Newbury Park, CA: Sage Publications; 1991.
19. Crossley J, Russell J, Jolly B, *et al*. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ* 2007; **41**: 926–34
20. 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
21. George BC, Teitelbaum EN, Meyerson SL, *et al*. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. *J Surg Educ* 2014; **71**: e90–e6
22. Castanelli D, Jowsey T, Chen Y, Weller J. Mini-CEX in anesthesia training: perceptions of purpose, value and process. *Can J Anaesth* 2016; **63**: 1345–56

Handling editor: P. S. Myles