The reliability of a portfolio of workplace-based assessments in anesthesia training

Can J Anesth/J Can Anesth (2019) 66:193–200
https://doi.org/10.1007/s12630-018-1251-7

REPORTS OF ORIGINAL INVESTIGATIONS

The reliability of a portfolio of workplace-based assessments in anesthesia training

Fiabilité d'un portfolio d'évaluations sur le lieu de travail dans la formation en anesthésie

Damian J. Castanelli, MBBS · Joyce M. W. Moonen-van Loon, PhD · Brian Jolly, PhD · Jennifer M. Weller, MD

Received: 13 February 2018 / Revised: 1 October 2018 / Accepted: 2 October 2018 / Published online: 14 November 2018

Abstract

Purpose *Competency-based anesthesia training programs require robust assessment of trainee performance and commonly combine different types of workplace-based assessment (WBA) covering multiple facets of practice. This study measured the reliability of WBAs in a large existing database and explored how they could be combined to optimize reliability for assessment decisions.*

Methods *We used generalizability theory to measure the composite reliability of four different types of WBAs used by the Australian and New Zealand College of Anaesthetists: mini-Clinical Evaluation Exercise (mini-CEX), direct observation of procedural skills (DOPS), case-based discussion (CbD), and multi-source feedback (MSF). We then modified the number and weighting of WBA combinations to optimize reliability with fewer assessments.*

Results *We analyzed 67,405 assessments from 1,837 trainees and 4,145 assessors. We assumed acceptable reliability for interim (intermediate stakes) and final (high stakes) decisions of 0.7 and 0.8, respectively. Depending on the combination of WBA types, 12 assessments allowed the 0.7 threshold to be reached where one assessment of any type has the same weighting, while 20 were required for reliability to reach 0.8. If the weighting of the assessments is optimized, acceptable reliability for interim and final decisions is possible with nine (e.g., two DOPS, three CbD, two mini-CEX, two MSF) and 15 (e.g., two DOPS, eight CbD, three mini-CEX, two MSF) assessments respectively.*

Conclusions *Reliability is an important factor to consider when designing assessments, and measuring composite reliability can allow the selection of a WBA portfolio with adequate reliability to provide evidence for defensible decisions on trainee progression.*

D. J. Castanelli, MBBS (✉)
School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia
e-mail: damian.castanelli@monash.edu

Department of Anaesthesia and Perioperative Medicine, Monash Health, Clayton, VIC, Australia

J. M. W. Moonen-van Loon, PhD
Department of Educational Development and Research, Faculty of Health, Medicine, and Life Sciences, Maastricht University, Maastricht, The Netherlands

B. Jolly, PhD
School of Medicine and Public Health, Faculty of Health and Medicine, University of Newcastle, Newcastle, NSW, Australia

J. M. Weller, MD
Centre for Medical and Health Sciences Education, School of Medicine, University of Auckland, Auckland, New Zealand

Department of Anaesthesia, Auckland City Hospital, Auckland, New Zealand

Résumé

Objectif *Les programmes de formation en anesthésie basés sur les compétences nécessitent de solides évaluations des performances des stagiaires et combinent habituellement des évaluations sur le lieu de travail (ÉLT) couvrant de nombreux aspects de la pratique. Cette étude a mesuré la fiabilité des ÉLT dans une grande base de données existante et a exploré comment elles pourraient être combinées pour accroître leur fiabilité pour des décisions sur les évaluations.*

Méthodes *Nous avons utilisé la théorie de la généralisation pour mesurer un critère composite de*

fiabilité de quatre types d'ÉLT utilisés par les collèges d'anesthésiologistes d'Australie et de Nouvelle-Zélande : un exercice de mini-évaluation clinique (mini-CEX), l'observation directe des habiletés procédurales (DOPS), une discussion de cas (CbD) et une rétroaction de multiples sources (MSF). Nous avons alors modifié le nombre et la pondération des combinaisons d'ÉLT pour optimiser la fiabilité avec moins d'évaluations.

Résultats *Nous avons analysé 67 405 évaluations de 1 837 stagiaires et 4 145 assesseurs. Nous avons supposé une fiabilité acceptable pour les décisions intérimaires (enjeux intermédiaires) et définitives (enjeux élevés) à, respectivement, 0,7 et 0,8. Selon la combinaison des types d'ÉLT, 12 évaluations ont permis d'atteindre le seuil de 0,7 lorsqu'une évaluation de chaque type a le même poids, alors qu'il en a fallu 20 pour que la fiabilité atteigne 0,8. Si la pondération des évaluations est optimisée, la fiabilité acceptable pour les décisions intérimaires et finales est possible avec, respectivement, neuf évaluations (p. ex., deux DOPS, trois CbD, deux mini-CEX, deux MSF) et quinze évaluations (p. ex. deux DOPS, huit CbD, trois mini-CEX, deux MSF).*

Conclusions *La fiabilité est un facteur important dont il faut tenir compte quand on conçoit les évaluations et la mesure d'une fiabilité composite permet la sélection d'un éventail d'ÉLT avec une fiabilité adéquate pour l'obtention de données probantes et la défense de décisions sur les progrès des stagiaires.*

Many anesthesia training programs no longer rely exclusively on knowledge examinations to assess their trainees but define the work they expect graduates to be able to do and then assess this in the workplace.[1-3] They incorporate workplace-based assessments (WBAs) of various types with other assessments into a programmatic approach to assessment,[4,5] which allows assessment of competencies over time and across different assessment methods. These assessments are then used to determine progression through training and ultimately to specialist practice.[1,4,5]

We know from multiple examples in the literature that individual types of WBAs can be used reliably.[6] Specifically in anesthesia, our group has previously investigated the psychometric properties of the mini-Clinical Evaluation Exercise (mini-CEX) in the Australian and New Zealand context.[7-9] With the use of an entrustment scale, acceptable reliability is achievable with a feasible number of mini-CEX.[7] Similarly, an evaluation of the direct observation of procedural skills (DOPS) assessment used in Australian and New Zealand anesthesia training also found acceptable inter-rater reliability.[10]

Training programs now commonly use a portfolio of different types of WBAs to assess across the different domains of professional practice. Combining WBAs has proved reliable in a study of Dutch postgraduate medical education.[11] Nevertheless, as far as we know, researchers and training programs in anesthesia are yet to investigate the reliability of combining different types of WBAs. Without this information, we cannot be confident of the reliability of the numbers and weightings of different types of WBAs we use to decide if trainees should progress through the training program.

Our aim in this study was to investigate the composite reliability of a portfolio of WBAs, and the predicted reliability of different combinations of WBAs, using an existing assessment database of WBA assessments maintained by the Australian and New Zealand College of Anaesthetists (ANZCA).

Our specific questions were:

1. What is the reliability of the scores associated with individual WBA types?
2. What numbers and combinations of WBAs would be required for scores to provide a level of reliability acceptable for interim and final decisions on trainee progression?

Methods

We obtained ethics approval from the Monash Health Human Research Ethics Committee (Ref. 16015L) and the University of Auckland Human Participant Ethics Committee (Ref. 017408).

Context

The ANZCA implemented a competency-based curriculum in December 2012 across Australia and New Zealand, which included a portfolio of WBAs. The intended purpose of the WBAs was two-fold—to improve the quality of feedback and trainee learning, and to assess the competence of trainees. ANZCA training is centrally administered, and all 158 training departments in Australia and New Zealand use a single assessment system. Any anesthesiologist and trainee can complete a WBA in any ANZCA accredited hospital.

Each training department has one or more ANZCA-appointed supervisors of training who undertake clinical placement reviews at six-monthly intervals, or earlier if

trainees move hospital, where they make interim decisions on trainee progress within stages of training. Toward the end of each of the four stages of training, they must decide if the trainee has met the performance level required to progress to the next stage. Educational supervisors in training programs across the world frequently make interim and final decisions such as these.

To address the different domains of practice, the ANZCA WBA portfolio comprises four different types of WBA: the mini-CEX, DOPS, case-based discussion (CbD) and multi-source feedback (MSF). The mini-CEX is used to assess part or all of an observed case (e.g., providing anesthesia care for a patient for a surgical operation). The DOPS is used to assess an observed technical procedure and uses a generic form for all procedures.[12] The DOPS and mini-CEX use an entrustability scale where supervisors are asked to make a judgement on whether the trainee needed the supervisor in the room, nearby in the hospital, or required only distant supervision for that case or procedure. The CbD form, used to assess clinical reasoning, requires a trainee to present a previous case to a supervisor for discussion. The rating scale reflects the degree of input required from the supervisor to develop a shared understanding of the issues involved in the case and the justification for how it was managed. The MSF is completed by anesthesiologists and other work colleagues, including surgeons, nurses, and anesthesia assistants. The scale requires these colleagues to score the trainee performance against expectations for their level of training and trainees receive a final score based on the interpretation of the collated ratings and comments by the training supervisor. This final score is used in this study.

The ANZCA determines the minimum number of each assessment type required during each stage of training, and this applies to all training sites. The length of each stage of training and the minimum total number of each assessment type required are presented in Table 1. Trainees are required to complete a minimum number of assessments every three months to ensure regularity of observation and feedback and to avoid clustering assessments in proximity to decision points. Each three months, trainees must complete at least two DOPS and two mini-CEX in introductory training, one CbD, two DOPS, and two mini-CEX in basic training, and one CbD, one DOPS, and two mini-CEX in advanced training. In addition, trainees must complete a minimum of six mini-CEX and four DOPS in introductory training. Either trainees or assessors may initiate WBAs, and trainees are encouraged to select their own cases for assessment. For MSF, trainees select their assessors.

The ANZCA maintains an electronic database in which all the WBAs are recorded. Anesthesiologists and trainees access the forms online using a secure login. Training supervisors have access to all of a trainee's WBAs and they are available to inform decisions on progression in training. Copies of the WBA forms are available at http://www.anzca.edu.au/training/2013-training-program/forms.

Data

We obtained de-identified data for all trainees from the introduction of WBAs on December 1, 2012, until the access date of May 23, 2016. This dataset includes trainees who transitioned to the new training program as well as trainees commencing at any point during that period and hence includes trainees at all levels with training of varying durations. Each WBA has a global score, which we used for the analysis. We excluded data from trainees with only a single assessment for any WBA type, with the exception of MSF, which is a composite of assessments from a range of individuals.

Analysis

Reliability of a measurement is an estimate of its consistency over different occasions, and reflects the degree of confidence with which we can claim that the measured value reflects the true value.[13] Generalizability

Table 1 Minimum total number of workplace-based assessments required during each stage of ANZCA training, with minimum duration of each stage

	CbD	DOPS	Mini-CEX	MSF
Introductory training (six months)	-	4	6	1
Basic training (18 months)	6	12	12	1
Advanced training (24 months)	8	8	16	1
Provisional fellowship (12 months)	2	-	-	1
Total (60 months)	16	24	34	5*

*One MSF is also required while training in intensive care medicine

ANZCA = Australian and New Zealand College of Anaesthetists; CbD = case-based discussion; CEX = Clinical Evaluation Exercise; DOPS = direct observation of procedural skills; MSF = multi-source feedback

Table 2 Total number of workplace-based assessments and related numbers of trainees and the mean scores (on a 1-9 scale), standard deviations, and reliability for each assessment type from December 1, 2012 until May 23, 2016

	CbD	DOPS	Mini-CEX	MSF	Total
Number of WBAs	11,125	23,670	29,124	3,486	67,405
Number of trainees	1,411	1,745	1,771	1,639	1,837
Number of assessors	2,895	3,518	3,765	-	4,145
Mean score	7.14	6.64	6.45	7.38	-
SD	0.71	0.82	0.93	0.91	-
Average number completed per trainee	7.88	13.56	16.45	2.13	-
Reliability based on average number	0.73	0.59	0.80	0.55	0.87
% Variance Trainee	25%	10%	19%	37%	-
Number for reliability of 0.7	7	23	10	5	
Number for reliability of 0.8	12	38	17	7	

CbD = case-based discussion; CEX = Clinical Evaluation Exercise; DOPS = direct observation of procedural skills; MSF = multi-source feedback

theory is commonly employed to calculate reliability in rater-based judgements such as WBA.[13] Generalizability theory apportions the variance in scores to the different factors affecting the measurement, which allows the degree to which the score reflects trainee performance to be estimated.

We used the MINQUE procedure in IBM SPSS Statistics for Windows (IBM Corp., Armonk, NY, USA) to generate variance components for trainee performance and error.[14] These variance components were then used to calculate the reliability coefficient for each WBA type using generalizability theory. Next, we used multivariate generalizability theory to estimate the reliability of combinations of WBAs using the individual WBA variance components and the covariance between them.[11] The contribution of an individual WBA type to the combined reliability is proportional to the variance attributable to the trainee and the number of assessments of that type used. We combined different numbers of each WBA to determine which combinations produced acceptable reliability while minimizing the number of assessments required.

A further modification that can affect reliability is to vary the weight applied to each assessment type. This is analogous to the judgements made when allocating different weighting to components of an examination. We varied each weighting in turn, starting with the assessment type with the highest trainee variance, to obtain the optimum composite reliability. As that assessment types' contribution to the composite reliability was observed to approach the optimum, it was fixed and the weighting of the next was then adjusted in the same way, until each weighting had been determined.

When deciding what level of reliability was acceptable, we applied the principle that the consequences of the decision inform the quality of the evidence required.[4] When generalizability theory is used, a reliability coefficient of 0.8 is generally the minimum expected for assessments leading to final (high stakes) decisions, whereas a lower reliability coefficient of 0.7 is often accepted for interim (intermediate stakes) decisions.[13,15-18] These values are lower than those traditionally used in classical test theory as generalizability theory accommodates more potential sources of error.[19]

Results

After excluding 402 single assessments, there were 67,405 assessments from 1,837 trainees and 4,145 assessors included in the analysis. The number of assessments, trainees, and assessors are summarized in Table 2 together with the mean scores, standard deviations, and calculated reliability of each individual type of WBA. Ideally, the proportion of variance attributable to the trainee should be high, as the trainee's performance is the object of measurement in WBA. The calculated trainee variance ranged from a low of 10% with DOPS to a high of 37% with MSF. (Table 2) The number of assessments of each individual assessment type required to reach the 0.7 and 0.8 reliability thresholds is shown in Table 2.

The optimal combinations of assessments that would result in a composite reliability of 0.7, sufficient to inform intermediate stakes decisions such as interim decisions during training, or 0.8 to inform final decisions, are presented in Table 3. Depending on the combination of

Table 3 Examples of combinations of workplace-based assessments, with each assessment equally weighted, for composite reliability of 0.7 or 0.8, which minimize the number of assessments required

CbD	DOPS	Mini-CEX	MSF	Total	Composite reliability coefficient
3	3	7	0	13	0.71
4	3	6	0	13	0.71
5	2	5	0	12	0.70
3	3	6	1	13	0.71
5	2	4	1	12	0.70
4	2	5	1	12	0.70
5	5	11	1	22	0.81
8	4	8	1	21	0.80
8	7	7	1	23	0.80
4	3	11	2	20	0.80
7	7	7	2	23	0.80
8	4	7	2	21	0.80

CbD = case-based discussion; CEX = Clinical Evaluation Exercise; DOPS = direct observation of procedural skills; MSF = multi-source feedback

WBA types, to reach the required reliability, we require 12-13 assessments where one assessment of any type has the same weighting. For example, with equal weighting and one MSF, four CbD, two DOPS, and five mini-CEX, we achieve a reliability of 0.7 with 12 assessment events, where the weighting in the overall reliability per assessment type is 1/12, 4/12, 2/12, and 5/12 respectively. To obtain a reliability coefficient of 0.8, sufficient to inform final decisions with high stakes such as progress to the next stage of training or graduation from training, we require approximately 20-23 assessments, each with equal weighting. For example, with one MSF, eight CbD, four DOPS, and eight mini-CEX we achieve a reliability of 0.8 with 21 assessment events, where the weighting is 1/21, 8/21, 4/21, and 8/21, respectively. As the variance attributable to trainees is lowest for the DOPS, the DOPS has the least positive effect on the composite reliability and combinations with a higher proportion of DOPS require a greater number of assessments to achieve the desired reliability coefficients. Conversely, increasing the proportion of assessments with higher variance attributable to the trainee means fewer assessments are required to achieve the desired reliability coefficient.

Optimizing the weighting of each assessment type as described above reduces the total number of assessments required for both purposes. Increasing the weighting of the assessment types with greater variance attributable to the trainee minimizes the number of assessments required to reach a desired level of reliability. We can achieve a reliability of 0.7 with nine assessments and 0.8 with 15 assessments if we vary the weighting as indicated in Table 4.

Discussion

Using a large online database of 67,405 WBAs from anesthesia training, we discovered that combinations of the four different types of WBA can achieve an acceptable reliability with nine assessments for interim decisions and 15 assessments for final decisions. We found some of the WBA types generated higher reliability than others, and this influences their potential contribution to the composite reliability measure. In our case, using fewer DOPS in proportion to mini-CEX and CbD allows a higher composite reliability with fewer assessments.

In applying our findings, we think reliability needs to be balanced against feasibility, the purpose of the different WBAs, and their appropriateness at different stages of training. Feasibility reflects the capacity of a training system to support the desired number of assessments, and although we have not examined it here, we think it is likely that the number of WBAs required for sufficient reliability to support robust decisions is within the reach of most training programs. Using results such as ours would help assessment program designers to balance the cost and workload involved for assessors against the benefit of individual assessments in terms of their contribution to sampling of curriculum content and the overall reliability of decisions based on the trainee's WBA portfolio.

We found the highest trainee variance for MSF, which indicates, as a unitary measurement, it would be most reliable and make the greatest contribution to composite reliability. Nevertheless, MSF is itself a composite measure requiring a number of colleagues to contribute to a single

Table 4 Examples of combinations of workplace-based assessments and weighting of assessment type to achieve a composite reliability of 0.7 or 0.8 while minimizing the number of assessments required. The reliability of the combination of assessments without adjusting the weighting, i.e., equal weight, is provided for comparison

	CbD	DOPS	Mini-CEX	MSF	Total	Composite reliability coefficient
Number	5	2	3	0	10	
Weight	0.70	0.10	0.20	0.00		0.70
Equal weight	0.50	0.20	0.30	0.00		0.66
Number	4	2	3	1	10	
Weight	0.53	0.08	0.20	0.19		0.70
Equal weight	0.40	0.20	0.30	0.10		0.66
Number	3	2	2	2	9	
Weight	0.40	0.07	0.13	0.40		0.70
Equal weight	0.33	0.22	0.22	0.22		0.63
Number	9	3	5	0	17	
Weight	0.74	0.07	0.19	0.00		0.80
Equal weight	0.53	0.18	0.29	0.00		0.77
Number	8	2	5	1	16	
Weight	0.62	0.05	0.21	0.12		0.80
Equal weight	0.5	0.125	0.313	0.063		0.77
Number	8	2	3	2	15	
Weight	0.65	0.05	0.10	0.20		0.80
Equal weight	0.53	0.13	0.20	0.13		0.76

CbD = case-based discussion; CEX = Clinical Evaluation Exercise; DOPS = direct observation of procedural skills; MSF = multi-source feedback

overall score, and it is possible that its acceptability to raters might decline if repeated at more frequent intervals. MSF also relies on observation over a period of time, which necessitates a degree of separation between occasions to allow this observation to occur. These factors may create a practical limit to the number of MSF available at the time a performance judgement is to be made. On the other hand, increasing the number of CbDs, which also had high trainee variance, may be more feasible and potentially of value for trainees across all stages of the program.

To assess multiple domains of performance, we require multiple types of assessments, which is the rationale for using different WBAs in an assessment program. When we are designing an assessment system, we should use the combined reliability of the WBA portfolio (composite reliability), rather than the reliability of the individual WBA types, to guide our choices.[20] Nevertheless, excluding a type of WBA entirely, e.g., DOPS, because it has low trainee variance, may produce a less valid assessment. We might assess more accurately, but address fewer competencies. Previous exploration of the composite reliability of a WBA portfolio is limited. A large multi-specialty Dutch study on composite reliability of WBAs (also using an entrustment scale) demonstrated similar reliability, with 16 assessments providing sufficiently reliable results for a final decision.[11] Nevertheless, that study involved 12 specialties and did not examine the influence of specialization on the results, so our study provides additional confirmation of acceptable reliability in anesthesia.

The differing trainee variance of the WBA types suggests there is scope for improvement in the design and implementation of the assessments themselves. Training supervisors assessing workplace performance may improve the rigour of the judgements they submit on the WBA forms. In a previous study, we reported variable understanding by supervisors of the use of an entrustment scale, which also suggests that professional development may be useful.[21] The scores on the WBA forms reflect the extent of supervision required for a particular case, which will, of course, vary with case difficulty and the individual trainee's familiarity with the various clinical areas of anesthesia care. In a previous study,[7] we generated a set of standards for expected supervisory scores for trainees at different stages in the training program for a range of common cases. For example, a score suggesting a junior trainee requires close supervision for a particular case may be entirely consistent with a good trajectory through the program, but a similar score for a trainee nearing the end of their training may suggest a trainee is in difficulty. We found that

adjusting supervisory requirement scores against expected supervisory requirements increased the reliability of the scores. Alternatively, having a panel of experienced educational supervisors with knowledge of supervision expectations at different stages of training look together at a trainee's WBA portfolio might also increase the robustness of decisions made.[22,23]

A recent innovation in competency-based medical education is to use Entrustable Professional Activities (EPAs) as an organising framework.[24-26] An EPA is an area of professional practice that can be entrusted to the trainee. An example in anesthesiology could be "the trainee can provide safe, effective, and efficient anesthesia care to American Society of Anesthesiologists I and II patients for low complexity surgery". A WBA portfolio reflecting this particular area of practice would provide evidence for this entrustment decision. While we do not know what effect limiting the range of cases to a circumscribed area of practice, such as an EPA, would have on reliability, we suggest our results provide a useful guide to optimizing the ratios of different types of WBAs, and the likely number required.

Strengths and limitations

A strength of our study is the use of a large and diverse, real-world dataset including anesthesia trainees at all stages and with different durations of training, spread across 158 sites in two countries. This also results in some limitations. There are likely to be geographic variations in the use of WBAs. Nevertheless, the data are not collected in a way which would allow us to account for location in our study design. As trainees entered and left the program at varying times over the period of data collection, the number of assessments available on some trainees was limited.

Generalizability theory assumes local independence of assessors and occasions.[13] Nevertheless, as the intent of workplace assessment is to improve the trainee's subsequent performance, the occasions in our study are not strictly independent. This violation of the assumption of local independence is a generally accepted psychometric limitation when studying WBAs using real-world databases.[15]

Our analysis is of anesthesia training in two countries, and the extent to which our findings are generalizable to other anesthesia training programs remains to be confirmed. Although there are obvious differences between training programs across the world, the use of a portfolio of different types of WBAs to assess the various domains of anesthesia practice is increasingly common, and we think our experience can provide insights for others as they face similar challenges in their own contexts.

In evaluating our assessment instruments, reliability is not the only consideration. Nevertheless, when we design assessment systems that use multiple assessment methods (not only combinations of WBAs, but also objective structured clinical examinations [OSCEs], multiple choice questions, etc.), we make choices that have significant consequences. When determining the balance and weighting of the various assessments, we should take their psychometric properties into account. For example, over time viva voce examinations have evolved from panel-based mechanisms to multiple mini-interviews, based on a better understanding of the interacting psychometrics of the assessors' biases, the stations, assessors, and their content.[27] A similar process of evolution should occur as we learn more about WBAs.

Conclusion

Our results demonstrate that the reliability of different types of WBAs can vary, and the choices we make when selecting the combination of WBAs will influence the reliability of the subsequent decisions we make. By analyzing a large database of WBAs, we have shown that as few as fifteen WBAs provide sufficient reliability for defensible final or high stakes decisions in an anesthesia training program, with nine WBAs sufficient for reliable interim decisions on progress through training. While other factors such as curriculum sampling and feasibility are also important, we have shown how composite reliability can provide a rational basis for the design of robust assessment in the workplace.

Conflicts of interest None declared.

Editorial responsibility This submission was handled by Dr. Gregory L. Bryson, Deputy Editor-in-Chief, *Canadian Journal of Anesthesia.*

Author contributions *Damian J. Castanelli* designed the study, contributed to data management, analysis, and interpretation, and wrote the manuscript. *Joyce M.W. Moonen-van Loon* contributed to study design, led the analysis, contributed to interpretation, and commented on each draft of the manuscript. *Brian Jolly* contributed to study design, analysis and interpretation, drafting the manuscript, and commented on each draft of the manuscript. *Jennifer M. Weller* contributed to study design, interpretation, drafting the manuscript, and commented on each draft of the manuscript.

Funding This work was supported by a grant from the Australian and New Zealand College of Anaesthetists, (ANZCA grant number: S16/043).

References

1. *Australian and New Zealand College of Anaesthetists.* Anaesthesia training program curriculum. Melbourne: Australian and New Zealand College of Anaesthetists, 2012. Available from URL: http://www.anzca.edu.au/documents/anaesthesia-training-program-curriculum.pdf (accessed October 2018).
2. *Van Gessel E*, *Mellin-Olsen J*, *Ostergaard HT*. *Niemi-Murola L; Education and Training Standing Committee, European Board of Anaesthesiology, Reanimation and Intensive Care.* Postgraduate training in anaesthesiology, pain and intensive care: the new European competence-based guidelines. Eur J Anaesthesiol 2012; 29: 165-8.
3. *Stodel EJ*, *Wyand A*, *Crooks S*, *Moffett S*, *Chiu M*, *Hudson CC*. Designing and implementing a competency-based training program for anesthesiology residents at the University of Ottawa. Anesthesiol Res Pract 2015; 2015: 713038.
4. *van der Vleuten CP*, *Schuwirth LW*, *Driessen EW*, *et al*. A model for programmatic assessment fit for purpose. Med Teach 2012; 34: 205-14.
5. *Bok HG*, *Teunissen PW*, *Favier RP*, *et al*. Programmatic assessment of competency-based workplace learning: when theory meets practice. BMC Med Educ 2013; 13: 123.
6. *Norcini J*, *Burch V*. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach 2007; 29: 855-71.
7. *Weller JM*, *Castanelli DJ*, *Chen Y*, *Jolly B*. Making robust assessments of specialist trainees' workplace performance. Br J Anaesth 2017; 118: 207-14.
8. *Weller JM*, *Jolly B*, *Misur MP*, *et al*. Mini-clinical evaluation exercise in anaesthesia training. Br J Anaesth 2009; 102: 633-41.
9. *Weller JM*, *Misur M*, *Nicolson S*, *et al*. Can I leave the theatre? A key to more reliable workplace-based assessment. Br J Anaesth 2014; 112: 1083-91.
10. *Watson MJ*, *Wong DM*, *Kluger R*, *et al*. Psychometric evaluation of a direct observation of procedural skills assessment tool for ultrasound-guided regional anaesthesia. Anaesthesia 2014; 69: 604-12.
11. *Moonen-van Loon JM*, *Overeem K*, *Donkers HH*, *van der Vleuten CP*, *Driessen EW*. Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. Adv Health Sci Educ Theory Pract 2013; 18: 1087-102.
12. *Wragg A*, *Wade W*, *Fuller G*, *Cowan G*, *Mills P*. Assessing the performance of specialist registrars. Clin Med (Lond) 2003; 3: 131-4.
13. *Downing SM*. Reliability: on the reproducibility of assessment data. Med Educ 2004; 38: 1006-12.
14. *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.
15. *Moonen-van Loon JM*, *Overeem K*, *Govaerts MJ*, *Verhoeven BH*, *van der Vleuten CP*, *Driessen EW*. The reliability of multisource feedback in competency-based assessment programs: the effects of multiple occasions and assessor groups. Acad Med 2015; 90: 1093-9.
16. *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.
17. *Pelgrim EA*, *Kramer AW*, *Mokkink HG*, *van den Elsen L*, *Grol RP*, *van der Vleuten CP*. In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ Theory Pract 2011; 16: 131-42.
18. *Lockyer J*, *Violato C*, *Fidler H*. A multi source feedback program for anesthesiologists. Can J Anesth 2006; 53: 33-9.
19. *Brennan RL.* Generalizability Theory and classical test theory. J Appl Meas Educ 2010; 24: 1-21.
20. *van der Vleuten CP*, *Schuwirth LW*, *Scheele F*, *Driessen EW*, *Hodges B*. The assessment of professional competence: building blocks for theory development. Best Pract Res Clin Obstet Gynaecol 2010; 24: 703-19.
21. *Castanelli DJ*, *Jowsey T*, *Chen Y*, *Weller JM*. Perceptions of purpose, value, and process of the mini-Clinical Evaluation Exercise in anesthesia training. Can J Anesth 2016; 63: 1345-56.
22. *Driessen EW*, *van Tartwijk J*, *Govaerts M*, *Teunissen P*, *van der Vleuten CP*. The use of programmatic assessment in the clinical workplace: a Maastricht case report. Med Teach 2012; 34: 226-31.
23. *Van Tartwijk J*, *Driessen EW*. Portfolios for assessment and learning: AMEE Guide no. 45. Med Teach 2009; 31: 790-801.
24. *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 Teach 2015; 37: 983-1002.
25. *Wisman-Zwarter N*, *van der Schaaf M*, *Ten Cate O*, *Jonker G*, *van Klei WA*, *Hoff RG*. Transforming the learning outcomes of anaesthesiology training into entrustable professional activities: a Delphi study. Eur J Anaesthesiol 2016; 33: 559-67.
26. *Jonker G*, *Hoff RG*, *Ten Cate OT*. A case for competency-based anaesthesiology training with entrustable professional activities: an agenda for development and research. Eur J Anaesthesiol 2015; 32: 71-6.
27. *Eva KW*, *Macala C*. Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. Med Educ 2014; 48: 604-13.