Developing and validating a tool for measuring the educational environment in clinical anesthesia

Élaboration et validation d'un outil de mesure de l'environnement éducatif en anesthésie clinique

Navdeep S. Sidhu, MBChB, FANZCA, MClined, FAcadMED · Eleri Clissold, MBBS

Received: 13 November 2017 / Revised: 30 April 2018 / Accepted: 14 May 2018 / Published online: 10 July 2018
© Canadian Anesthesiologists' Society 2018

Abstract

Purpose We aimed to develop a contemporary measure for anesthesia teaching and learning in the operating theatre that was applicable to a variety of training jurisdictions, the Measure for the Anaesthesia Theatre Educational Environment (MATE).

Methods A systematic review of the literature and modified Delphi approach was used to identify items for content validity. Reliability and exploratory factor analyses were conducted after a pilot survey of trainees to show construct validity, with removal of redundant items. Item domains were identified through a global assessment of factor structure accuracy and relation to real-world constructs.

Results Literature review generated an initial 73-item list. A modified Delphi approach with 24 experts identified 44 relevant items. The pilot survey generated 390 responses. Reliability analysis, exploratory factor analysis, and global assessment refined the measure to 33 items. Four domains were identified according to factor structure: teaching preparation and practice, assessment and feedback, procedures and responsibility, and overall atmosphere.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12630-018-1185-0>) contains supplementary material, which is available to authorized users.

N. S. Sidhu, MBChB, FANZCA, MClined, FAcadMED (✉)
Department of Anaesthesia and Perioperative Medicine, North Shore Hospital, 124 Shakespeare Road, Takapuna, Auckland 0620, New Zealand
e-mail: nav.sidhu@waitematadhb.govt.nz

E. Clissold, MBBS
Institute for Innovation and Improvement, Waitemata District Health Board, Auckland, New Zealand

The educational environment was rated by trainees at $74.6 \pm 15.6\%$ with excellent internal consistency (Cronbach's $\alpha = 0.975$).

Conclusion The MATE survey tool generated valid and reliable scores when measuring the educational environment in the operating theatre. Further research is required to investigate possible differences between the training countries and age of junior doctors and the associated underlying factors. Other researchers are invited to administer the survey and share results within a central database.

Résumé

Objectif Nous avons cherché à élaborer une mesure contemporaine pour l'enseignement et l'apprentissage de l'anesthésie en salles d'opération qui pourrait être appliquée dans différents cadres de formation : la mesure pour l'environnement éducatif en salle d'anesthésie ou MATE (Measure for the Anaesthesia Theatre Educational Environment).

Méthodes Une revue systématique des publications et une approche de Delphes modifiée ont servi à identifier les éléments de validation du contenu. Des analyses de fiabilité et de facteurs exploratoires ont été menées après une enquête pilote auprès de stagiaires pour montrer la fiabilité du montage avec la suppression d'éléments redondants. Les domaines d'items ont été identifiés via une évaluation globale de l'exactitude des structures de facteurs et leurs relations avec des montages en situation réelle.

Résultats La recherche bibliographique a permis de créer une liste initiale de 73 éléments. Une approche de Delphes modifiée avec 24 experts a identifié 44 éléments pertinents. L'enquête pilote a généré 390 réponses. L'analyse de fiabilité, l'analyse des facteurs exploratoires et

l'évaluation globale ont permis d'affiner la mesure à 33 éléments. Quatre domaines ont été identifiés en fonction de la structure des facteurs : préparation et pratique de l'enseignement, évaluation et rétroaction, procédures et responsabilité, et environnement global. Les stagiaires ont attribué à l'environnement éducatif une cote de $74,6 \pm 15,6\%$ avec une excellente homogénéité interne (coefficient α de Cronbach = 0,975).

Conclusion *L'outil d'enquête MATE a généré des scores valides et fiables pour la mesure de l'environnement éducatif en salle d'opération. Des recherches supplémentaires sont nécessaires pour étudier les différences possibles entre les pays de formations, l'âge des jeunes médecins et les facteurs sous-jacents associés. D'autres chercheurs sont invités à administrer l'enquête et à en partager les résultats dans une base de données centrale.*

The term “educational environment” relates to how learners perceive their teaching and learning in the clinical setting. It is defined by the American Medical Association as “a social system that includes the learner (including the external relationships and other factors affecting the learner), the individuals with whom the learner interacts, the setting(s) and purpose(s) of the interaction, and the formal and informal rules/policies/norms governing the interaction”.¹ A suitable educational environment is critical for effective knowledge transfer, skills progression, and development in the affective domain. This is especially crucial in the clinically important, time-critical environment of the operating theatre. Routine direct observation of teaching encounters is resource-intensive and not feasible. Educational environment measures (EEMs) are survey instruments administered to learners that have been developed for a variety of clinical settings. Educational environment measures with high reliability and validity may be used as a surrogate for direct evaluation of teaching and learning, enabling continual professional development and quality improvement at departmental and regional levels. A systematic review by Soemantri *et al.* lists 31 published EEMs in health professions education, nine of which were designed for the postgraduate medical setting.²

There is no contemporary EEM for anesthesia teaching encounters in the operating theatre. Previously developed clinical anesthesia EEMs conducted validation studies in a specific area or region or were not focused on clinical teaching encounters. Holt and Roff published the Anaesthetic Trainee Theatre Educational Environment Measure with input from focus groups of anesthesiology

trainees, educational supervisors, and regional program directors, piloted on 218 trainees in one training region of the United Kingdom (UK).³ Smith and Castanelli developed an EEM with emphasis on the general learning environment rather than in-theatre teaching, piloted on 263 trainees in New Zealand (NZ) and Australia.⁴ They recently performed a second pilot on 172 trainees from one Australian training region, altering the Likert scale from the first pilot, to determine the minimum number of respondents required to maintain reliability of the measure.⁵

The objectives of our study were to:

1. Develop a clinical anesthesia EEM for teaching encounters in the operating theatre that was contemporary, based on most recent evidence, and generalizable to different training programs.
2. Interpret the pilot study results to guide future research

The Measure for the Anaesthesia Theatre Educational Environment (MATE) utilizes a series of methodologies to show validity, piloting the measure in different training jurisdictions. We chose to focus specifically on the operating theatre as the vast majority of teaching and learning in clinical anesthesia occurs here, providing insight into current practice and a focal point for potential quality improvement. Results from the pilot survey will be used to identify areas of future research.

Methods

We obtained approval from the Awhina Research and Knowledge Centre (Protocol RM13266). We used a literature search to generate the initial item list (see Appendix I; available as Electronic Supplementary Material [ESM]), categorized into four provisional domains that were identified *a priori*. A modified Delphi approach was used to develop consensus on item inclusion to ensure content validity. The Delphi approach is an established objective method of obtaining expert consensus, allowing a large number of experts to contribute anonymously in a non-adversarial manner in a series of phases, with successive feedback of collective opinion and opportunity for correction.⁶ The Delphi methods are explicitly described in Appendix I (available as ESM).⁷

Pilot of MATE draft

We administered the pilot to junior doctors in seven countries. We defined junior doctors as medical practitioners working in clinical anesthesia that required some form of supervision. These included interns,

residents, house officers, medical officers, registrars, or fellows, not limited to those in vocational training programs. We anticipated the majority of participants to be vocational trainees, and the term “trainee” was utilized in department correspondence, with the above-emphasized definition. Residency program directors/coordinators in 16/17 Canadian, all 134 United States (US), and all three Singaporean anesthesiology residencies were sent an email with a request to forward the survey link to trainees in their departments. The head of school or trainee representative for 23/24 UK Schools of Anaesthesia and two large Hong Kong Special Administrative Region (HK SAR) teaching hospitals was sent a similar email. The Australian and New Zealand College of Anaesthetists (ANZCA) Clinical Trials Network facilitated email invitations to be sent to a random sample of 1,000 NZ and Australian trainees, and trainees in the Auckland region of NZ were individually emailed. United States military institutions, one Canadian institution, one UK School of Anaesthesia, and all but two Hong Kong departments were not approached because contact details could not be sourced.

Respondents were asked: “Please rate the following statements as they apply to your perception of teaching in the operating theatres of this department (applies to any site where anesthesia is delivered, including endoscopy or interventional suites)”. A seven-point rating scale was applied to all items (0 = strongly disagree, 6 = strongly agree). Participants were required to have been working in their department for a minimum of eight weeks to ensure adequate exposure to their clinical environment. The survey was administered using an online survey tool (Survey Monkey) and collected anonymously, with relevant demographic information. Stratification of results by demographic parameters was performed to guide possible future research.

Reliability and exploratory factor analyses (EFA)

Reliability and EFA of a pilot survey enabled refinement of the item list and demonstration of construct validity.^{8,9} Exploratory factor analysis is used to identify a set of latent constructs underlying a group of observed variables, as measured through items or questions.^{8,10} A series of mathematical iterations (factor rotations) creates linear combinations to explain the data, with each iteration revealing new information that allows the researcher to examine the relationships between items and factors.⁸ Redundant items may be removed if they load poorly onto factors or if they cross-load without strong primary loadings. The structure is refined until an efficient, mathematically sound, and theoretically grounded solution is reached.⁸ As this method of factor analysis is inherently designed to be exploratory,⁹ a global assessment

is required to show real-world constructs for item-factor relationships.

Statistical analysis

Basic analysis for the Delphi phases was performed using Microsoft Excel (Microsoft, Redmond, WA, USA). Data from the pilot survey were analyzed with IBM SPSS 24 (IBM, Armonk, NY, USA). $P \leq 0.05$ was considered statistically significant. We applied the Kolmogorov-Smirnov test to the pilot survey to determine if data were normally distributed. Cronbach’s α coefficients were generated to appraise internal consistency reliability, both prior to and after EFA.

Suitability for EFA was determined using the Kaiser-Meyer-Olkin Measure of Sampling Adequacy, Bartlett’s Test of Sphericity, and measures of communality. Factor extraction was performed using principal axis factoring (for non-parametric data). Eigenvalue analysis and the scree test were used to determine the number of factors retained—factors with eigenvalues of one or more and factors located above the inflection point on the scree plot.^{8,9} The eigenvalue describes the variance in the items explained by that factor.⁸ The scree test is a plot of the same eigenvalues on the y-axis and factor number on the x-axis, and can be open to subjective interpretation. Factor rotation was performed using an oblique method (promax), as we believed that the factors would be related to each other. During successive rotations, we removed items that failed to achieve a primary factor loading of at least 0.4 and items that exhibited cross-loadings of 0.3 or above without a strong primary factor loading (defined as ≥ 0.65). Rotations were performed until no items met the criteria for removal. Finally, a global assessment for accuracy of factor structure was performed to determine if the factors could be related to real-world constructs to determine the final MATE item structure.

Scores for each item were added to determine the overall MATE score, out of 198 (33 items $\times$ maximum score of 6). This was converted to a percentage score by dividing the total score by 198. For individual domains, total scores for items in each domain were divided by (number of items in that domain $\times$ 6). Respondents’ scores for the overall MATE were included only if they provided responses for all items, and for each domain only if they provided responses for all items within that domain. Demographic group comparisons were carried out using the Kruskal-Wallis test (for non-parametric data). If the Kruskal-Wallis test generated a P value ≤ 0.05 , Dunn’s non-parametric pairwise comparisons (two-way comparison between groups in each demographic category) with Bonferroni-adjusted significance were carried out.

Results

Literature review

The literature search yielded 6,820 results. Seventy-three papers were identified for further scrutiny after review of abstracts and 50 papers after bibliographic review, with seven found suitable for inclusion. These were two previously published anesthesia educational environment measures,^{3,4} three papers on characteristics of good teachers in anesthesia,¹¹⁻¹³ and two validated instruments for evaluation of anesthesiologists' supervision of trainees.^{14,15} Seventy-three discrete items were identified for the initial item list, grouped into four provisional domains.

Modified Delphi process

Thirty-five individuals were approached after being identified as potential "experts" for our panel, with 28 positive replies received. Four did not fulfill the inclusion criteria, either not having completed vocational training in anesthesiology or not possessing a formal qualification in medical education, resulting in a final figure of 24 experts. Response rates for phases 1-4 were 95.8%, 83.3%, 70.8%, and 66.7%, respectively. The demographic makeup of the panel and their response rates are listed in Table 1. Forty-four items achieved a mean score of ≥ 5 and a standard deviation (SD) of ≤ 1 , for inclusion in the draft measure to be piloted (Table 2).

Pilot survey response

We received 390 responses. Twenty-six responses were excluded, 16 for not scoring any items and ten because of having worked in their department for under eight weeks,

leaving 364 responses available for analysis. We could not calculate the actual response rate as we were unable to determine what proportion of contacts forwarded the invitation email to their junior doctors and, for contacts that did so, how many junior doctors worked in those departments.

Exploratory factor analysis

The Kolmogorov-Smirnov test indicated that the pilot survey data were not normally distributed, and non-parametric statistical tests were henceforth applied. Detailed descriptions of the initial reliability analysis, preliminary analysis for suitability, factor extraction, and factor rotations are located in Appendices II and III (available as ESM). These EFA steps identified a further ten redundant items.

Global assessment for accuracy of factor structure showed that the provisional domains proposed in the draft MATE did not completely conform to the extracted factors, except for items in the provisional "Assessment and feedback" domain all loading to factor 1. Nevertheless, items in each factor could be related to real-world constructs, allowing for four distinct domains to be named and conferring construct validity to the MATE. The identified domains were "teaching preparation and practice" (factor 3), "assessment and feedback" (factor 1), "procedures and responsibility" (factor 4), and "overall atmosphere" (factor 2) (see table for rotation 3 in Appendix III; available as ESM). Minor adjustments were made to ensure consistency and avoid duplication under the new structure. The lowest-loading item under factor 2, "my clinical teachers provide appropriate support when I am performing a procedure for the first time", was moved to factor 4 ("procedures and responsibility"), and the item "The clinical teachers are easily accessible should

Table 1 Expert panel demographics and response rates

	Number (%)	Response rate
Medical education qualification		
Postgraduate/graduate certificate	12 (50.0)	89.6%
Postgraduate/graduate diploma	8 (33.3)	53.1%
Masters or doctorate	4 (16.7)	100%
Experience as specialist anesthetist		
< 5 yr	13 (54.2)	63.5%
5-10 yr	5 (20.8)	95.0%
> 10 yr	6 (25.0)	100%
Gender		
Female	12 (50.0)	85.4%
Male	12 (50.0)	72.9%

Table 2 Final rating of initial items in modified Delphi approach

Please rate the relevance of each statement to best-practice for anesthesia teaching encounters in the operating theatre (0 = not at all relevant, 6 = extremely relevant)	Mean	SD	Phase when consensus achieved
A. Preparation for teaching			
1. I am encouraged to visit patients preoperatively ³	5.5	0.7	2*
2. I discuss the anesthetic plan of cases with my clinical teacher ^{3,4,11,15}	5.9	0.4	2*
3. My clinical teachers seek to identify my current level of knowledge, if it is not already known to them ¹¹	5.6	0.5	2*
4. I have clear learning goals for theatre teaching sessions ^{3,4,14}	5.0	0.7	2*
5. I have freedom to set my own learning goals in the theatre setting ⁴	4.8	0.7	2
6. The learning goals formulated for a theatre session are relevant ¹⁴	5.4	0.6	3*
7. My clinical teachers engage with me when determining learning goals for the theatre session ¹⁴	5.2	0.8	3*
B. Teaching practice			
8. I am encouraged to actively participate with patient management ³	5.6	0.5	2*
9. Teaching occurs at appropriate times, not affecting vigilance ³	5.6	0.6	2*
10. I feel able to ask the questions I want to ^{3,4,14}	5.9	0.4	2*
11. My clinical teachers are accessible for advice ^{3,4,13}	5.4	0.5	2*
12. I receive supervision from clinical teachers that is appropriate for my level of training ^{3,4,12,13,15}	5.7	0.5	2*
13. Teaching is delivered in a clear manner ^{3,12}	5.1	0.7	2*
14. The teaching helps to develop my confidence ^{3,4}	4.9	0.6	2
15. My clinical teachers allocate adequate time to teach in the theatre setting ¹¹⁻¹³	4.4	1.0	4
16. My clinical teachers are patient when teaching ^{12,13}	4.9	0.7	3
17. References to the literature are used to support teaching ¹³	3.3	1.1	-
18. The teaching is appropriate for my level of training ¹³	5.4	0.7	2*
19. Much of what I am taught seems relevant to my career ^{3,4,13}	4.6	1.0	4
20. My clinical teachers demonstrate an active effort to teach in the operating theatre ^{13,14}	5.0	0.7	2*
21. My clinical teachers give priority to my learning goals when teaching in the operating theatre ¹⁴	4.2	1.0	4
22. My clinical teachers frequently refer to real clinical scenarios when teaching in the theatre setting ¹⁵	4.8	0.6	2
23. I have the opportunity to acquire the practical skills appropriate to my level of training ^{3,4,15}	5.0	0.8	3*
24. My clinical teachers provide appropriate support when I perform a procedure for the first time ¹³	5.8	0.5	2*
25. I receive theatre teaching in subspecialty areas targeted at my learning needs ^{3,4,13}	4.8	0.8	3
26. I have opportunities to learn about appropriate non-technical skills in the operating theatre ⁴	5.3	0.4	2*
27. I am able to achieve my learning goals in the operating theatre ^{4,14}	5.1	0.8	4*
28. My clinical teachers challenge me to be prepared for the unexpected ¹¹	4.9	1.0	4
29. My clinical teachers explain reasons for utilization of specific management strategies ¹³	5.2	0.6	2*
30. Discussions in the operating theatre are education-oriented ¹³	3.7	1.1	-
C. Assessment and feedback			
31. Assessment of my performance in the operating theatre occurs regularly ¹⁴	5.2	0.8	4*
32. My clinical teachers are fair in their assessment of my performance ^{3,4}	5.6	0.6	2*
33. My clinical teachers possess the necessary skills to assess my performance in the operating theatre ¹²	5.2	0.5	2*
34. My clinical teachers help to develop my competence ^{3,4}	5.3	0.6	2*
35. Feedback from clinical teachers is readily provided to me at all times ⁴	4.8	0.9	4
36. Feedback is provided on tasks that I perform under direct supervision ⁴	5.0	0.7	3*
37. I receive feedback that is appropriate for my level of training ⁴	5.3	0.6	2*
38. I receive feedback on specific performance issues ⁴	5.5	0.6	2*
39. Feedback is provided based on direct observation of my work ⁴	5.4	0.6	2*
40. Feedback is delivered soon after my work is observed ^{4,15}	5.4	0.6	3*
41. I receive honest feedback ^{4,11}	5.5	0.6	2*
42. I receive feedback that provides me with an opportunity to improve ^{4,13-15}	5.5	0.8	3*
43. Positive feedback is readily provided when indicated ¹⁴	5.1	0.6	2*
44. Corrective feedback is provided when indicated ¹⁴	5.6	0.5	2*
45. I have sufficient opportunities to reflect on my learning ^{4,11}	5.0	0.8	4*

Table 2 continued

Please rate the relevance of each statement to best-practice for anesthesia teaching encounters in the operating theatre (0 = not at all relevant, 6 = extremely relevant)	Mean	SD	Phase when consensus achieved
D. Overall atmosphere			
46. I am aware of my duties and responsibilities in theatre ^{3,4}	5.2	0.6	2*
47. I have an appropriate level of clinical responsibility ^{3,4}	5.2	0.7	2*
48. I feel responsible and accountable for the care given to my patients ^{3,4}	5.1	1.1	-
49. I feel comfortable in theatre socially ^{3,4}	4.3	0.8	3
50. I have good collaboration with theatre staff ^{3,4}	4.8	1.0	4
51. I feel part of a team when working in the operating theatre ^{3,4}	5.1	0.7	3*
52. The people I work with in the operating theatre are friendly ³	4.4	0.7	2
53. The surgeons have no concerns about the noise of theatre teaching ³	2.8	1.3	-
54. There is no discrimination in this post ^{3,4}	5.1	1.1	-
55. A systematic clinical training program is implemented in this department ^{3,4,12}	4.9	0.8	3
56. The clinical training program allows me to get first-hand experience in a range of procedures ³	5.5	0.5	2*
57. There are good opportunities for trainees who fail to complete their training satisfactorily ³	4.2	1.3	-
58. There is an informative anesthesia trainee handbook ^{3,4}	4.3	0.8	3
59. I am given relief from duties to participate in formal educational programs ^{3,4}	5.0	1.2	-
60. The formal educational program is targeted to my learning needs ⁴	4.9	0.7	2
61. The clinical teachers are easily accessible should I require their help ^{3,14,15}	5.4	0.5	2*
62. I am aware to whom I should report, in a variety of circumstances ⁴	5.1	0.3	2*
63. My workload in this job is fair ^{3,4}	4.4	1.0	4
64. My time at work is utilized productively ⁴	4.7	0.7	2
65. My work is interesting with sufficient variety ⁴	4.1	1.1	-
66. I have access to up-to-date learning resources at work ⁴	4.5	0.8	3
67. Teaching and training are emphasized in this department ⁴	5.5	0.5	2*
68. The clinical teachers in this department are up-to-date with their medical knowledge ^{12,13,15}	4.9	0.6	2
69. My clinical teachers create a trusting and open learning climate ^{12,15}	5.4	0.7	2*
70. My clinical teachers are open to my suggestions regarding management of a patient ¹³⁻¹⁵	4.8	0.6	2
71. My clinical teachers promote an atmosphere of mutual respect ^{3,4,13-15}	5.5	0.7	2*
72. I have a good sense of rapport with my clinical teachers ^{3,4}	5.1	0.7	2*
73. I view the clinical teachers in this department as positive role models ¹⁴	5.1	0.5	2*

*Fulfilled criteria for inclusion in draft measure for pilot

I require their help” was removed as it was deemed to be very similar to the highest-loading item in that factor, “My clinical teachers are accessible for advice”. The completed factor analysis resulted in the final refined 33-item MATE survey tool (see Appendix IV; available as ESM).

Post hoc reliability

Overall internal consistency of the MATE was excellent (Cronbach's $\alpha = 0.975$). Internal consistency for the new domain labels was 0.945 for “teaching preparation and practice”, 0.964 for “assessment and feedback”, 0.833 for “procedures and responsibility”, and 0.936 for “overall atmosphere”. No improvement in reliability could be gained with the deletion of any item for any of the four domains.

MATE scores

The mean (SD) % of the overall MATE score was 74.6 (15.6), with domain scores as follows: “teaching preparation and practice” [66.6 (19.2)], “Assessment and Feedback” [71.9 (19.0)], “procedures and responsibility” [85.5 (12.8)], and “overall atmosphere” [81.8 (16.2)]. Scores based on demographic background are listed in Table 3. A significant difference in MATE scores between groups was found in the country and age categories. With the former, *post hoc* testing using Dunn's non-parametric pairwise comparisons indicated that this was between Canada and Australia ($P = 0.013$) and Canada and NZ ($P = 0.036$). Significant differences between these two country pairs were observed in all MATE domains except “overall atmosphere”. For the age category, only the “assessment

Table 3 MATE scores based on demographic background

	<i>n</i>	Mean (%)	SD (%)	<i>P</i> value
Country*				0.003
Australia	125	72.0	14.8	
Canada	17	84.8	10.0	
Hong Kong SAR	19	76.0	12.8	
New Zealand	71	71.9	17.0	
United Kingdom	36	77.6	14.4	
United States	71	77.6	17.2	
Hospital†				0.197
Hospital A, NZ	25	71.0	19.6	
Hospital B, US	11	75.9	11.3	
Hospital C, NZ	13	74.0	13.8	
Hospital D, US	7	68.5	27.2	
Hospital E, NZ	11	72.3	13.9	
Hospital F, Canada	9	89.1	7.8	
Hospital G, Hong Kong SAR	18	75.6	13.1	
Hospital H, Australia	31	73.4	12.1	
Hospital I, US	10	79.6	12.7	
Gender				0.458
Female	165	74.1	15.6	
Male	175	75.1	15.9	
Training status in anesthesia				0.860
Vocational trainee	318	74.7	15.6	
Non-trainee	22	73.3	18.7	
Age				0.045
21-30 yr	150	76.5	14.7	
31 yr and over	190	73.1	16.4	
Clinical experience in anesthesia				0.094
Up to 12 months	74	77.3	15.1	
> 12 months	266	73.9	15.9	
Time in current department				0.947
8 weeks to 3 months	60	74.3	13.2	
3-6 months	53	75.1	15.8	
6-12 months	83	74.6	15.6	
> 12 months	144	74.6	16.9	

*One respondent failed to supply country/hospital information

†Only those with minimum seven valid respondents listed

NZ = New Zealand; SAR = Special Administrative Region; US = United States

and feedback” domain showed a significant difference, with junior doctors aged 30 yr and younger rating the mean (SD) domain higher than those over aged over 30 yr [74.8 (18.9) vs 69.7 (18.9); $P = 0.003$]. Less experienced junior doctors also rated this domain higher than their more experienced counterparts [75.7 (18.5) vs 70.8 (19.1); $P = 0.029$], although the overall MATE scores were not significantly different ($P = 0.094$).

Discussion

We have described the development of an instrument to measure the educational environment in the operating theatre for anesthesia, utilizing specific techniques at each stage of development to show different aspects of validity. A systematic literature review identified 73 items, reduced to 44 using a modified Delphi approach and further refined

to 33 items using EFA. The reliability and distribution of scores in this final instrument are described, with excellent reliability analysis and a successful pilot in different training programs and jurisdictions. The MATE shares only 11/33 items with a similar tool published 14 years ago,³ justifying the development of an updated measure.

Delphi approach

There is no strong evidence for the number of panel members or required response rates. For a homogeneous population (experts from the same discipline), 15-30 people is recommended.⁷ While it is generally accepted that higher response rates are better, at least 70% is recommended for each phase.⁶ Our 24 panel members achieved this in all but the final phase (66.7%). Combined with the systematic review of the literature for initial item generation, the Delphi approach confers content validity to the development of the MATE.

Three items were excluded at the Delphi stage because of a lack of consensus ($SD > 1.0$) despite achieving the target mean score. These were “I feel responsible and accountable for the care given to my patients”, “There is no discrimination in this post”, and “I am given relief from duties to participate in formal educational programs”. Free-text comments by expert panel members to justify outlying ratings alluded to the lack of direct relevance to in-theatre teaching. The issue of discrimination, along with sexual harassment and bullying, is an important one. We were comfortable with the removal of the aforementioned item as these issues were likely to be addressed by the retained item, “My clinical teachers promote an atmosphere of mutual respect”.

Interpretation of MATE survey tool findings

Thirty respondents (8.8%) submitted a MATE score of $< 50\%$. The measure developed by Holt and Roff showed 2.3% of respondents with an equivalent score of $< 50\%$,³ while a more recent measure encompassing the overall anesthesia clinical learning environment had 3.4% of respondents submitting a (corrected) score of $< 50\%$.⁴ At the opposite end, 194 MATE respondents (57.1%) rated their educational environment at $> 75\%$ compared with 37.6% and 41.1% in the two previous studies.^{3,4} This increase in both low and high ratings may be attributed to differences in the ratings scale, respondents, survey items, survey context, or other aspects of survey design. There is evidence that full labelling of the rating scale, as done in the two compared studies, results in respondents providing more central ratings and fewer ratings at the extreme ends of the scale.¹⁶ Almost half of all respondents in our study provided scores in the 50-80% range. The practice of full

labelling vs labelling only at the endpoints is a contentious one. An analysis of 13 surveys by Alwin and Krosnick indicated that fully labelled surveys were more reliable compared with endpoint-only labelling ($\alpha = 0.783$ vs 0.570),¹⁷ but this effect was not observed in our survey instrument ($\alpha = 0.975$). The use of descriptors such as “moderately agree” or “moderately disagree” with full labelling renders the variables as ordinal data, as one is unable to state with certainty that the intervals between the different anchors are equal. Respondents may also interpret differently what it means to “moderately” agree or disagree with a statement. Endpoint-only labelling results in a continuous rating scale that arguably conveys the idea of equal intervals between each point and are no less valid than fully labelled scales.¹⁸ Strictly speaking, fully labelled scales are called Likert scales, although the term is frequently used when referring to continuous rating scales.

Based on our preliminary analysis, we propose that the following structure be used to evaluate scores for the MATE and its four domains: 0-50% = poor, 50.1-60% = below average, 60.1-70% = average, 70.1-80% = good, 80.1-90% = very good, and 90.1-100% = excellent. The use of a descriptive evaluation structure confers concrete meaning to the score generated by the measure, allows for ease of interpretation, and provides targets for quality improvement. Table 4 lists respondents' scores for the MATE and its constituent domains according to this evaluation structure. Interventions aimed at improving teaching and learning should focus on the “teaching preparation and practice” and “assessment and feedback” domains, as these obtained poor or below average evaluations from 33.5% and 23.9% of respondents, respectively.

Application of this evaluation structure requires an adequate sample size. A measure utilizing a four-point Likert scale recently showed adequate reliability with a minimum of eight respondents from a single department.⁵ For our study, we are unable to state a minimum sample size for individual departments. Conservatively, a sample size of < 10 may not allow for valid interpretation, and 10-20 should be interpreted with caution unless accompanied by a small variance. A standard deviation of 15% or less (0.9 on the 0-6 scale) may be sufficiently precise for a sample size of 10-20. Further research would be required to confirm the appropriateness of this evaluation structure and to determine minimum sample size for individual departments.

In subgroup analyses, a significant difference was observed between some countries. Nevertheless, we caution that firm conclusions cannot be drawn based on this evidence alone as sample sizes are insufficient to be representative of any single country and responses are biased towards participating departments, but it is an area

Table 4 MATE and domain scores according to evaluation structure

	MATE (<i>n</i> = 340)	Teaching preparation and practice (<i>n</i> = 358)	Assessment and feedback (<i>n</i> = 347)	Procedures and responsibility (<i>n</i> = 340)	Overall atmosphere (<i>n</i> = 340)
Excellent (90.1-100%)	58 (17.1%)	36 (10.1%)	62 (17.9%)	133 (39.1%)	118 (34.7%)
Very good (80.1-90%)	85 (25.0%)	47 (13.1%)	74 (21.3%)	93 (27.4%)	103 (30.3%)
Good (70.1-80%)	88 (25.9%)	96 (26.8%)	76 (21.9%)	65 (19.1%)	58 (17.1%)
Average (60.1-70%)	52 (15.3%)	59 (16.5%)	52 (15.0%)	28 (8.2%)	26 (7.6%)
Below average (50.1-60%)	27 (7.9%)	47 (13.1%)	35 (10.1%)	15 (4.4%)	13 (3.8%)
Poor (0-50%)	30 (8.8%)	73 (20.4%)	48 (13.8%)	6 (1.8%)	22 (6.5%)

MATE = Anaesthesia Theatre Educational Environment

that merits further investigation. Possible reasons may include differences in teaching culture, vocational training programs, institutional support, trainee expectations, educational resources, or clinical workload. Younger and less experienced trainees rated their experience of “assessment and feedback” significantly higher than their older and more experienced counterparts. One reason for this may be differences (real or perceived) in the quality and volume of feedback delivered, presumably higher in the younger and less experienced group because of their being at a stage of training that requires closer supervision and active teaching. Older and more advanced trainees may also be better equipped to critically rate a department because of their experience.

Exploratory factor analysis

There is no agreement on minimum sample size requirements for EFA, with figures ranging from 100-300.⁸ Others use the subject-to-variable (STV) ratio to base minimum sample size recommendations, with minimum ratios ranging from 5-10.^{8,9} A more contemporary view is that the required sample size is dependent on the strength of the item-factor relationship.⁸⁻¹⁰ For example, if all factors have at least four strong-loading items, the sample size may be irrelevant. Our study defined a strong loading as 0.65 or above, with other authors quoting as low as 0.5⁹ or as high as 0.7.^{8,10} If there are ten to 12 items with moderate loadings (0.4-0.6), a sample size of at least 150 is required.⁸ Factors that have few items and have moderate-to-low loadings require a minimum sample size of 300.⁸ EFA produces unreliable and non-valid results if performed with an inadequate sample size.⁹ With 364 valid responses, our data began with an STV ratio of 8.5, increasing to 11.0 after removal of redundant items. The final factor loading matrix (Appendix II; available as ESM) showed very strong item loading for all but one factor (factor four),

which loaded two items strongly (0.751 and 806) and two items moderately (0.492 and 0.543).

Conversion to percentage score

Deriving a percentage score from a rating scale is a common method for presenting EEM scores.^{4,5,19-24} The wider and consistent margins inherent in a 0-100 scale facilitate comparison between different measures or the same measure applied at a different time or place. A requirement for converting from a rating scale to a percentage score is factoring a zero point into the conversion if the lowest value in the original scale is not zero. For example, directly converting a 1-5 rating scale to a percentage score is erroneous because the lowest possible mean or median score is 1/5 or 20%, with a resultant 20-100% score. A 1-5 rating scale should therefore be recalculated as a 0-4 scale prior to conversion to generate a 0-100% score. Failure to adjust for this results in inflated percentage scores that are accentuated at the lower end of the scale, as shown in some studies.^{4,5,19} This inflation effect is worsened with narrower rating scales and produces inaccurate comparisons with other EEM results. One may also argue that fully labelled scales do not lend themselves to percentage conversion as one cannot be confident of equal intervals between rating points.

Utility

Potential applications of the MATE in the context of anesthesiology training are numerous. In other settings, EEMs have been used to evaluate interventions designed to improve teaching and learning,²⁵ monitor the impact of curricular change,²⁶⁻²⁸ longitudinal changes over time and between cohorts,²³ and differences in training locations.²⁴ In a recent review of a generic postgraduate training EEM, 8/9 studies reported significant differences in overall EEM scores for rural vs urban training locations.²⁹

Residency program directors or education coordinators in individual departments may use the MATE as an educational key performance index to address areas of concern as they are identified. There is evidence of correlation between positive EEM scores and improved academic performance. A study of 206 general medical residents in 21 training hospitals showed a positive correlation between EEM scores and performance in the in-training examination.³⁰ A study of dental undergraduates showed correlation between scores in the perception of learning domain and higher grades, while low scores in three domains were associated with failing grades.³¹ One study with medical undergraduates showed no differences in overall scores but correlation between high scores in selected domains and superior academic performance.³² Nevertheless, a survey of nursing undergraduates in one institution showed no correlation between EEM scores and academic performance.³³ A large study of 1,350 medical students from 22 medical schools showed a positive correlation between overall EEM scores and levels of resilience.³⁴

Bodies responsible for accreditation of vocational training could identify outliers among institutions, learning from well-performing departments and providing assistance or remediation measures for poorly performing ones. While face-to-face interviews with trainees during site visits provide invaluable information for training accreditation decisions, the MATE allows for a more feasible and objective assessment of the educational environment. The measure allows input from all trainees, an impractical task with individual face-to-face interviews. Accreditation bodies may identify potential problems earlier and enable targeted enquiry during site visits. Training institutions, regions, or countries that report uniformly average or less-than average scores may seek to investigate why this difference exists.

Limitations

The primary limitation of this study is the inability to determine an accurate response rate due to the method in which the survey was distributed. Certain subgroup comparisons do not allow for firm conclusions because of a lack of a representative sample. Nevertheless, this approach allowed us to obtain a much larger sample size than previously published anesthesia EEMs.^{3,4} It also allowed for sampling of a heterogeneous population with the implication that the results are likely to be generalizable to different training regions and systems. Future work comparing differences between training regions or countries should be designed to ensure representative sampling. Our reliability analysis could have been supplemented with a multivariate

generalizability analysis to identify and assess the effects of various possible sources of error.

Future research

We invite educational supervisors to utilize the MATE on an ongoing basis. We offer to share (at no cost) a customized electronic survey and subsequent results analysis to any department that wishes to administer the measure. The complete measure is included in Appendix IV (available as ESM). As responses are generated, we aim to add these to a central database with the consent of participating institutions, maintaining anonymity of departments and individual respondents. This will enable participating institutions to compare their results with mean and median scores in their region or country and to track temporal changes or effects of interventions. Using this database, future studies may focus on differences between training regions/countries, changes over time, confirmation of the proposed evaluation structure through qualitative analysis, and determination of minimum sample size for individual departments. Further factor analyses on a new population should be performed to reconfirm the underlying structure. Multivariate generalizability analysis on future samples would identify and assess possible sources of error and determine a minimal sample size for individual departments.

Conclusion

The MATE is potentially a valid and reliable tool to measure the educational environment in the operating theatre, specific to anaesthesia. It can be used by individual institutions or vocational training bodies as a key performance index in education or to evaluate effects of interventions in teaching and learning. Further research is required to investigate differences in training countries and possible underlying factors. The authors will maintain a database of responses, preserving the anonymity of respondents and their institutions. Educational supervisors and researchers are invited to administer the measure and collaborate with the authors to enable further investigation in this area.

Acknowledgements We thank the following individuals for their input: Tom Burrows, Damien Castanelli, Nina Civil, Marlin De Silva, Kirsty Forrest, Alistair Kan, Laura Kwan, David Law, Emelyn Lee, Helen Lindsay, Neil Macdonald, Nola Ng, Lindy Roberts, Ross Scott-Weekly, Natalie Smith, Ben Snow, Melanie Speer, Timothy Starkie, Ghassan Talab, Michael Tan, Kersi Taraporewalla, Jennifer Weller, Eva Wilson, and Caroline Zhou. We also thank the Australian and New Zealand College of Anaesthetists (ANZCA) Clinical Trials Network for facilitating survey distribution to New Zealand and

Australian trainees and all educational supervisors and residents internationally who engaged in the study.

Declaration of interests No external funding and no competing interests declared

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

Author contributions Navdeep Sidhu contributed substantially to all aspects of this manuscript, including the conception and design, acquisition, analysis and interpretation of data, and drafting the article. Eleri Clissold contributed substantially to the conception and design of the manuscript, analysis of data, and drafting the article.

References

1. American Medical Association. Report of the Council on Medical Education 7-A-09. Transforming the medical education learning environment. Available from URL: [https://www.ama-assn.org/sites/default/files/media-browser/public/about-ama/councils/Council Reports/council-on-medical-education/a09-cme-transforming-medical-education-learning-environment.pdf](https://www.ama-assn.org/sites/default/files/media-browser/public/about-ama/councils/Council%20Reports/council-on-medical-education/a09-cme-transforming-medical-education-learning-environment.pdf) (accessed May 2018).
2. Soemantri D, Herrera C, Riquelme A. Measuring the educational environment in health professions studies: a systematic review. *Med Teach* 2010; 32: 947-52.
3. Holt MC, Roff S. Development and validation of the Anaesthetic Theatre Educational Environment Measure (ATEEM). *Med Teach* 2004; 26: 553-8.
4. Smith NA, Castanelli DJ. Measuring the clinical learning environment in anaesthesia. *Anaesth Intensive Care* 2015; 43: 199-203.
5. Castanelli DJ, Smith NA. Measuring the anaesthesia clinical learning environment at the department level is feasible and reliable. *Br J Anaesth* 2017; 118: 733-9.
6. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs* 2000; 32: 1008-15.
7. Clayton MJ. Delphi: a technique to harness expert opinion for critical decision-making tasks in education. *Educ Psychol* 1997; 17: 373-86.
8. Beavers AS, Lounsbury JW, Richards JK, Huck SW, Skolits GJ, Esquivel SL. Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation* 2013; 18: 1-13. Available from URL: <http://pareonline.net/pdf/v18n6.pdf> (accessed May 2018).
9. Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation* 2005; 10: 1-9. Available from URL: <http://pareonline.net/getvn.asp?v=10&n=7> (accessed May 2018).
10. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods* 1999; 4: 272-99.
11. Cleave-Hogg D, Benedict C. Characteristics of good anaesthesia teachers. *Can J Anaesth* 1997; 44: 587-91.
12. Ortwein H, Blaum WE, Spies CD. Anesthesiology residents' perspective about good teaching - a qualitative needs assessment. *Ger Med Sci* 2014; 12: Doc05.
13. Haydar B, Charnin J, Voepel-Lewis T, Baker K. Resident characterization of better-than- and worse-than-average clinical teaching. *Anesthesiology* 2014; 120: 120-8.
14. Lombarts KM, Bucx MJ, Arah OA. Development of a system for the evaluation of the teaching qualities of anesthesiology faculty. *Anesthesiology* 2009; 111: 709-16.
15. de Oliveira Filho GR, Dal Mago AJ, Garcia JM, Goldschmidt R. An instrument designed for faculty supervision evaluation by anesthesia residents and its psychometric properties. *Anesth Analg* 2008; 107: 1316-22.
16. Moors G, Kieruj ND, Vermunt JK. The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociol Methodol* 2014; 44: 369-99.
17. Alwin DF, Krosnick JA. The reliability of survey attitude measurement: the influence of question and respondent attributes. *Sociol Methods Res* 1991; 20: 139-81.
18. Saris WE, Gallhofer IN. Design, Evaluation, and Analysis of Questionnaires for Survey Research. Hoboken, NJ: John Wiley & Sons; 2007.
19. Kanashiro J, McAleer S, Roff S. Assessing the educational environment in the operating room - a measure of resident perception at one Canadian institution. *Surgery* 2006; 139: 150-8.
20. Mahoney A, Crowe PJ, Harris P. Exploring Australasian surgical trainees' satisfaction with operating theatre learning using the 'surgical theatre educational environment measure'. *ANZ J Surg* 2010; 80: 884-9.
21. Yin T, Child S. The Auckland Surgical Theatre Educational Environment Measure: does attending surgery benefit house officers? *N Z Med J* 2015; 128: 94-8.
22. Binsaleh S, Babaeer A, Rabah D, Madbouly K. Evaluation of urology residents' perception of surgical theater educational environment. *J Surg Educ* 2014; 72: 73-9.
23. Palmgren PJ, Sundberg T, Laksov KB. Reassessing the educational environment among undergraduate students in a chiropractic training institution: a study over time. *J Chiropr Educ* 2015; 29: 110-26.
24. Wong PN, John DN, Deslandes RE, Hughes ML. Same syllabus, different country - using DREEM to compare the educational environments at two Pharmacy schools. *Pharm Educ* 2015; 15: 87-92.
25. Leung Y, Salfinger S, Mercer A. The positive impact of structured teaching in the operating room. *Aust NZ J Obstet Gynaecol* 2015; 55: 601-5.
26. Finn Y, Avalos G, Dunne F. Positive changes in the medical educational environment following introduction of a new systems-based curriculum: DREEM or reality? *Curricular change and the environment*. *Ir J Med Sci* 2014; 183: 253-8.
27. Shankar PR, Bharti R, Ramireddy R, Balasubramaniam R, Nuguri V. Students' perception of the learning environment at Xavier University School of Medicine, Aruba: a follow-up study. *J Educ Eval Health Prof* 2014; 11: 9.
28. Qin Y, Wang Y, Floden RE. The effect of problem-based learning on improvement of the medical educational environment: a systematic review and meta-analysis. *Med Princ Pract* 2016; 25: 525-32.
29. Chan CY, Sum MY, Lim WS, Chew NW, Samarasekera DD, Sim K. Adoption and correlates of Postgraduate Hospital Educational Environment Measure (PHEEM) in the evaluation of learning environments - a systematic review. *Med Teach* 2016; 38: 1248-55.
30. Shimizu T, Tsugawa Y, Tanoue Y, et al. The hospital educational environment and performance of residents in the General Medicine In Training Examination: a multicenter study in Japan. *Int J Gen Med* 2013; 6: 637-40.

31. *Al-Ansari AA, El Tantawi MM.* Predicting academic performance of dental students using perception of educational environment. *J Dent Educ* 2015; 79: 337-44.
32. *Sarwar S, Tarique S.* Perception of educational environment: does it impact academic performance of medical students? *J Pak Med Assoc* 2016; 66: 1210-4.
33. *Payne LK, Glaspie T.* Associations between baccalaureate nursing students' perceptions of educational environment and HESITM scores and GPA. *Nurse Educ Today* 2014; 34: e64-8.
34. *Tempski P, Santos IS, Mayer FB, et al.* Relationship among medical student resilience, educational environment and quality of life. *PLoS One* 2015; 10: e0131535.