Validity of entrustment scales within anesthesiology residency training

Validité des échelles de confiance dans le cadre de la résidence en anesthésiologie

Daniel G. Dubois, MD · Alexander J. Lingley, MSc · Julie Ghatalia, Med · Meghan M. McConnell, PhD

Received: 9 October 2019 / Revised: 9 July 2020 / Accepted: 12 July 2020 / Published online: 20 October 2020
© Canadian Anesthesiologists' Society 2020

Abstract

Introduction Competency-based medical education requires robust assessment in authentic clinical environments. Using work-based assessments, entrustment scales have emerged as a means of describing a trainee's ability to perform competently. Nevertheless, psychometric properties of entrustment-based assessment are relatively unknown, particularly in anesthesiology. This study assessed the generalizability and extrapolation evidence for entrustment scales within a program of assessment during anesthesiology training.

Methods Entrustment scores were collected during the first seven blocks of training for three resident cohorts. Entrustment scores were assessed during daily evaluations using a Clinical Case Assessment Tool (CCAT) within the preoperative, intraoperative, and postoperative setting. The reliability of the entrustment scale was estimated using generalizability theory. Spearman's correlations measured the relationship between median entrustment scores and percentiles scores on the Anesthesia Knowledge Test

(AKT)-1 and AKT-6, mean Objective Structured Clinical Examination (OSCE) scores, and rankings of performance by the Clinical Competence Committee (CCC).

Results Analyses were derived from 2,309 CCATs from 35 residents. The reliability or generalizability (G) coefficient of the entrustment scale was 0.73 (95% confidence interval [CI], 0.70 to 0.76), and the internal consistency was 0.86 (95% CI, 0.84 to 0.88). Intraoperative entrustment scores significantly correlated with the AKT-6 ($\rho = 0.51$, $P = 0.01$), mean OSCE ($\rho = 0.45$, $P = 0.04$), and CCC performance rankings ($\rho = 0.52$, $P = 0.006$).

Conclusion As part of an assessment program, entrustment scales used early during anesthesiology training showed evidence of validity. Intraoperative entrustment scores had good reliability and showed acceptable internal consistency. Interpreting entrustment scores in this setting may constitute a valuable adjunct complementing traditional summative evaluations.

Résumé

Introduction La formation médicale fondée sur les compétences nécessite une évaluation rigoureuse dans des environnements cliniques authentiques. Se fondant sur des évaluations basées sur le travail, les échelles de confiance sont apparues comme une méthode pour décrire la capacité d'un résident à performer de façon compétente. Toutefois, les propriétés psychométriques de l'évaluation basée sur la confiance sont relativement peu connues, particulièrement en anesthésiologie. Cette étude a évalué les données de généralisabilité et d'extrapolation des échelles de confiance dans le cadre d'un programme d'évaluation pendant la formation en anesthésiologie.

Méthode Les notes sur les échelles de confiance ont été colligées pendant les sept premiers blocs de formation de trois cohortes de résidents. Les notes sur les échelles de

D. G. Dubois, MD (✉) · J. Ghatalia, Med
Department of Anesthesiology and Pain Medicine, University of
Ottawa, 1053 Carling Ave, Room B311, Ottawa, ON K1Y 4E9,
Canada
e-mail: dadubois@toh.ca

A. J. Lingley, MSc
Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

M. M. McConnell, PhD
Department of Anesthesiology and Pain Medicine, University of
Ottawa, 1053 Carling Ave, Room B311, Ottawa, ON K1Y 4E9,
Canada

Department of Innovation in Medical Innovation, University of
Ottawa, Ottawa, ON, Canada

confiance ont été évaluées pendant les évaluations quotidiennes à l'aide d'un Outil d'évaluation des compétences cliniques (CCAT – Clinical Case Assessment Tool) dans les cadres préopératoire, peropératoire et postopératoire. La fiabilité de l'échelle de confiance a été estimée à l'aide de la théorie de la généralisabilité. Les corrélations de Spearman ont mesuré la relation entre des notes moyennes sur l'échelle de confiance et les scores de percentile aux examens de connaissances en anesthésiologie AKT-1 et AKT-6, les scores moyens à l'Examen clinique objectif structuré (ECOS), et les classements de performance par le Comité de compétences cliniques (CCC).

Résultats Les analyses ont été dérivées à partir de 2309 résultats CCAT de 35 résidents. Le coefficient de fiabilité ou de généralisabilité (G) de l'échelle de confiance était de 0,73 (intervalle de confiance [IC] 95 %, 0,70 à 0,76), et la cohérence interne était de 0,86 (IC 95 %, 0,84 à 0,88). Les scores de confiance peropératoires étaient corrélés de manière significative aux scores sur l'AKT-6 ($\rho = 0,51$, $P = 0,01$), aux scores moyens à l'ECOS ($\rho = 0,45$, $P = 0,04$) et aux classements de performance du CCC ($\rho = 0,52$, $P = 0,006$).

Conclusion Dans le cadre d'un programme d'évaluation, la validité des échelles de confiance utilisées en début de la formation en anesthésiologie a été éprouvée. Les scores de confiance peropératoires ont démontré une bonne fiabilité et une cohérence interne acceptable. L'interprétation des scores de confiance dans ce cadre pourrait constituer un ajout précieux qui compléterait les évaluations sommatives.

Keywords entrustment · work-based assessment · competency-based medical education · performance evaluation · anesthesiology

The last two decades have seen competency-based medical education (CBME) frameworks increasingly integrated into physician training. Competency-based medical education aims to facilitate medical learners' acquisition of "competencies" or indices of professional aptitude through outcomes-based assessment. This learning model differs markedly from time-based training programs of preceding generations and denotes especial priority to assessing medical learners in authentic clinical environments.¹

The necessity for direct observation and assessment of trainees' performance has prompted the development of workplace-based assessments (WBAs) as a main modality for teaching and evaluation. Workplace-based assessments enable direct feedback on trainees' bedside performance

and appear in postgraduate medical education in various forms, including the Mini-Clinical Evaluation Exercise (mini-CEX), objective structured assessment of technical skill, direct observations of practical skills, and multisource feedback, among others.^{2,3} Since there is no single comprehensive assessment tool, experts suggest that decisions regarding trainees' progression be based on an assessment program using aggregates of multiple measures including qualitative and quantitative methods.^{4–6} Implementing multiple WBAs allows educators to comprehensively document the progression of knowledge, skills, and performance in medical learners.⁷

Workplace-based assessments in CBME have increasingly focused on assessing entrustment. Entrustment refers to the perception of supervising faculty members regarding the degree of independence with which trainees can competently perform clinical tasks. Entrustment scales constitute a convenient tool for measuring entrustment in clinical settings and are described across several fields including surgical specialties,^{8,9} anesthesiology,^{10,11} internal medicine,^{12–14} and pediatrics.¹⁵ An advantage of entrustment-based evaluation is that it links competencies to actual practice and facilitates stepwise incorporation of learners in progressively more advanced components of clinical tasks without jeopardizing patient safety.^{16,17}

In the era of CBME, it is imperative that modes of assessment provide reliable and valid measures of learner performance. Entrustment allocation is particularly important in anesthesia care where there is a demand for appropriate and timely clinical decisions in a setting where even minor events can become critical for the most vulnerable of patients. Despite increasing support for the use of entrustment constructs to document the progression of medical trainees' knowledge and skills, there is little research describing validity evidence for entrustment scales in anesthesiology, particularly on how they relate to other assessment methods commonly employed in the medical education setting. Our study aims to assess whether an entrustment-based construct in anesthesiology training is sufficiently valid and reliable for continued incorporation within an assessment program. To help support score interpretations and uses of our overlying construct, we will include in our discussion an interpretation argument for entrustment using Kane's validity framework.¹⁸

Methods

This retrospective cohort study was approved by the Ottawa Health Science Network Research Ethics Board

(protocol 20170972-OH1) with waived participant consent as analyses were conducted with a de-identified data set.

Setting and data source

The Department of Anesthesiology and Pain Medicine at the University of Ottawa introduced a novel competency-based training curriculum in July 2015. The CBME curriculum provides tailored learning experiences in a time-variable postgraduate training program by blending opportunities for faculty-curated self-directed learning, apprenticeship with content experts, and simulation-based learning. Entrustment scales are one of several assessment modalities integrated throughout the CBME curriculum and are completed using an online system called the Clinical Case Assessment Tool (CCAT; see Appendix). Residents are expected to complete one CCAT per shift, with a minimum of 12 CCATs to be submitted during each four-week block. The CCAT comprises entrustment scales allocated to the three phases of perioperative care: preoperative assessment, intraoperative performance, and postoperative plan. For each item, the faculty evaluates a trainee's performance using a five-point global rating scale with behavioural anchors related to entrustment as defined in the Ottawa surgical competency operating room evaluation (O-SCORE).¹⁹ The faculty assesses the trainee's performance during a specific clinical case, rather than provide a global assessment of overall performance for the day. The case is chosen by both the resident and the staff at the start of the clinical day, so as to minimize a biased selection of cases for assessment.

Within an assessment program, entrustment scales are complemented by several other quantitative assessments of performance incorporated in the CBME curriculum including the Anesthesia Knowledge Test (AKT), Objective Structured Clinical Examination (OSCE), and ranking of resident performance by the Clinical Competence Committee (CCC). The AKT (Interhospital Study Group for Anesthesia Education, <http://aktexam.org/index.php>) is a standardized multiple-choice exam that is administered at several stages of training to document trainee progression in anesthesia core knowledge. Learners' percentile AKT scores can be compared with those of all residents within North America who have written that specific exam. The present study included results from the AKT written in the first (AKT-1) and sixth (AKT-6) month of training. The OSCE takes place six months into the CBME curriculum and consists of ten stations designed to assess technical and non-technical skills. The OSCE stations are mapped to stage-specific milestones, as well as the National Curriculum for Canadian Anesthesiology Residency Training.²⁰ Resident performance was ranked annually at the end of the foundations stage of training and

was derived from consensus-based ordering of the enrolled resident cohort based on CCC member impressions of overall performance. This process involved distribution of an anonymous survey to members of the CCC ($n = 10$) requesting their individual rank order list of residents within the cohort. Subsequently, we amalgamated all responses and presented the tabulated data for group review at a confidential CCC meeting. We tested agreement by inviting members to share any discordant views from the rank order list and revisions were made until a verbal consensus on the overall order was reached.

Collection of data

A program administrator (J.G.) external to the analysis extracted de-identified entrustment scores from the secured CCAT database for all enrolled residents during the first seven blocks of training after the CBME curriculum was introduced. The first seven blocks of training comprise three learning modules: "Transition to Discipline" (four weeks), "Boot Camp" (12 weeks), and "

Foundation of Discipline" (12 weeks). This temporal window was selected because three cohorts' (2015, 2016, 2017; $n = 35$ residents) entrustment data were available at the time of data extraction.

Median entrustment scores were calculated for each resident in three distinct perioperative phases over a predefined time period. In other words, for each resident, we determined the median score for each CCAT item (preoperative, intraoperative, and postoperative) across the three learning modules (Transition to Discipline, Boot Camp, and Foundations of Discipline). The data set also included trainees' percentile score on the AKT-1 and AKT-6, mean OSCE score, and overall rank score allocated by the CCC at the end of the foundations of discipline module.

Generalization

Generalization evidence refers to the relationship between individual assessment items and how these items relate to the overarching construct.²¹ Within medical education, validity evidence for generalization typically involves measures of reliability (e.g., reproducibility) of the scores, as well as item analyses such as item difficulty and item discrimination.²² In the present study, generalization of the entrustment scale (e.g., reliability) was assessed using generalizability (G) theory. Originally developed as a way to determine the amount of error involved in the measurement process, G-theory has been used to evaluate internal structure evidence in medical education contexts.²³ A G-coefficient of 0.80 is the generally accepted threshold for high stakes assessments, while lower coefficients of 0.70 are often accepted for

intermediate stakes assessments.^{24,25} Measures of internal consistency (i.e., Cronbach's alpha, or the average correlation between the three entrustment categories) were calculated by treating the item variance as a random facet. A strength of G-theory as a measure of reliability is its ability to identify which factors (e.g., resident, time in training, different CCAT items) contribute to the variability in entrustment scores. In our study, the object of measurement was labelled "r" for resident. A random effects model was used to quantify sources of variation using the following facets: training module (m) and CCAT item (i), and their interactions with one another and the resident. To account for the variance attributed to independent faculty raters, we identified a subset of raters who had rated the greatest number of residents across the study time period. Estimating the G-coefficient attributed to rater, we used faculty as our object of measurement and employed a random effects model to identify sources of variance associated with faculty members and CCAT items.

Extrapolation

As a further measure of validity evidence, extrapolation examines associations between assessment scores and other metrics or learner characteristics.²⁶ As the entrustment scale is intended to represent a resident's perceived independence, we predicted that scores would increase as residents progress through levels of CBME training. To evaluate this hypothesis, we collapsed data from the first seven blocks of training into their corresponding CBME modules: transitions to discipline module (block 1), Bootcamp Foundations module (blocks 2–4), and

Clinical Foundations module (blocks 5–7). This allowed us to calculate the median entrustment score for each CCAT item (preoperative, intraoperative, and postoperative) across the three learning modules (Transition to Discipline, Boot Camp, and Foundations of Discipline). To assess concurrent validity, we compared mean entrustment scores with performance on other known metrics.

Table 1 The frequency and percentage of scale rating selection, missing data points, and mean (standard deviation [SD]), median [interquartile range], and mode values for each Clinical Case Assessment Tool (CCAT) item obtained from 2,309 forms. 1—

CCAT item	Rating, n (%)					Missing Data	Mean (SD)	Median [IQR]
	1	2	3	4	5			
Preoperative	25 (1.1%)	245 (10.6%)	817 (35.4%)	761 (33.0%)	323 (14.0%)	138 (6.0%)	3.5 (0.9)	3 [3–4]
Intraoperative	84 (3.6%)	688 (29.8%)	1042 (45.1%)	380 (16.5%)	58 (2.5%)	57 (2.5%)	2.8 (0.8)	3 [2–3]
Postoperative	33 (1.4%)	279 (12.1%)	708 (30.7%)	588 (25.5%)	254 (11.0%)	447 (19.4%)	3.4 (0.9)	3 [3–4]

IQR = interquartile range

Statistical analysis

We conducted Friedman's analysis of variance for each CBME module, using the median entrustment scores as the dependent variable and entrustment score category (three levels: preoperative, intraoperative, and postoperative) as the independent variable. *Post hoc* analyses with Wilcoxon signed-rank test were conducted and Bonferroni corrections were applied to all multiple pairwise comparisons. Effect sizes were calculated using r statistic, as outlined by Field,²⁷ with small effect sizes $\approx r < 0.2$; medium effect sizes $\approx 0.3 < r < 0.5$; and large effect sizes $r > 0.5$. Spearman's correlations were used to measure the relationship between median entrustment scores for each CCAT item (preoperative, intraoperative, and postoperative), AKT-1 percentiles, AKT-6 percentiles, mean OSCE scores, and overall CCC rankings.

Results

Data were collected from 35 residents (three cohorts: 2015, 2016, and 2017) during the first seven blocks of training, resulting in 2,309 CCATs completed, with a mean (standard deviation) of 66 (10.1) CCAT forms submitted per resident. A total of 103 faculty members assessed at least one resident during the study time period, with a median [interquartile range (IQR)] of 67 [57–71] faculty members per resident (min = 9; max = 93).

Descriptive statistics for each of the three entrustment categories are presented in Table 1. These calculations only included data where scores were assigned by faculty raters and collapsed data within each module excluded cell counts associated with missing data. For preoperative and postoperative scores, the majority of ratings ranged from 3 to 5 (87.5% and 83.2%, respectively); however, for intraoperative scores, the ratings were less positively skewed, with the majority ranging from 2 to 4 (93.7% of scores). This pattern is further illustrated by the lower mean scores for intraoperative assessments (2.8) relative to preoperative (3.5) and postoperative (3.4) assessments.

"Staff had to do"; 2—"Staff had to talk resident through", 3—"Staff had to prompt resident from time to time"; 4—"Staff needed to be in the room just in case"; 5—"Staff did not need to be there"

The variance components in Table 2 were used to calculate a G-coefficient, revealing the reliability of the entrustment scales to be 0.73 (95% confidence interval [CI], 0.70 to 0.76). The G-coefficient associated with internal consistency of the entrustment scores was 0.86 (95% CI, 0.84 to 0.88). Residents (facet r) accounted for approximately 9% of the variance, which indicates that 9% of the variability in entrustment scores is attributable to the residents themselves. Nevertheless, most of the variance (approximately 36%) was attributed to entrustment scale items (facet i), suggesting that ratings varied across the three entrustment categories (e.g., preoperative, intraoperative, and postoperative scores). Time (facet m) accounted for approximately 18% of the variability in scores, which indicates that there is some variation across the seven blocks of training. Additionally, the interaction between resident and time (r^*m) accounted for 19% of the variability. This finding shows that for any given resident, there were some differences in their scores across the seven blocks of training. A total of approximately 16% of the variance was attributable to random error. The remaining facets and interactions accounted for less than 2% of the variability in entrustment score.

To account for the variance attributed to raters, we identified six faculty members who had assessed more than 70% of the trainees. No faculty member assessed all 35 trainees; the maximum number of trainees assessed was 29. From here, we cross referenced these faculty members with individual residents to identify a sub-sample of residents. This resulted in 19 residents who had been assessed at least once by each of these six faculty members. Across the six faculty members, the median [IQR] number of assessments was 2 [1–3] for an individual resident (min = 1; max = 3). Table 3 shows that most of the variance (44%) was

attributed to residual error. We again showed that residents accounted for 6% of the variance. The interaction between faculty member and resident (r^*f) accounted for 35% of the variability in scores, which suggests that scores associated with individual residents varied across faculty members. The items accounted for 21% of the variability, suggesting that ratings varied across the three entrustment categories (e.g., preoperative, intraoperative, and postoperative scores). The remaining facets and interactions accounted for less than 2% of the variability in entrustment scores. These variance components were used to calculate a G-coefficient, revealing the reliability of the entrustment scales with individual faculty raters to be 0.47 (95% CI, 0.17 to 0.69).

Table 4 displays median entrustment ratings across levels of CBME training. For all three CBME modules, median entrustment scores varied across the CCAT items (Transition to Discipline: $\chi^2(2) = 41.07$, adjusted $P < 0.001$; Bootcamp Foundations: $\chi^2(2) = 30.11$, adjusted $P < 0.001$; and Clinical Foundations: $\chi^2(2) = 37.00$, adjusted $P < 0.001$). Subsequent *post hoc* analyses using Wilcoxon signed-rank test were completed for each CBME module. During the Transition to Discipline module, median entrustment scores for intraoperative CCATs were less than both preoperative ($Z = 4.72$, adjusted $P < 0.001$, $r = 0.80$) and postoperative ($Z = 3.41$, adjusted $P = 0.04$, $r = 0.58$) CCAT scores. Median postoperative scores were also lower than preoperative scores during the Transition to Discipline module ($Z = 3.70$, adjusted $P < 0.001$, $r = 0.63$). During the Bootcamp Foundations module, median intraoperative CCAT scores were once again significantly lower than both preoperative ($Z = 4.18$, adjusted $P = 0.002$, $r = 0.71$) and postoperative scores ($Z = 3.83$, adjusted $P = 0.009$, $r = 0.65$). Nevertheless, there was no difference in

Table 2 Generalizability analysis for the Clinical Case Assessment Tool ($n = 2,309$ forms)

Facet	Variance component (VC)	VC%	Description
Resident (r)	0.032	9.3%	Variance due to differences between residents
Time (m)	0.060	17.5%	Variance due to different blocks in the CBME curriculum
Item (i)	0.123	36.0%	Variance attributable to the items on the tool
r^*m	0.065	19.0%	Variance attributable to the individual residents across the first seven blocks of training (i.e., did residents differ in how they were scored based on the block in which they were assessed?)
r^*i	0.002	0.6%	Variance attributable to the interaction between residents and items (i.e., did residents differ in how they were scored on particular items?)
m^*i	0.005	1.4%	Variance attributable to the interaction between blocks of training and items (i.e., did item means vary based on at what point in the curriculum the assessment took place?)
r^*m^*i	0.055	16.1%	Variance attributable to the residents, time, items, plus random error
G-coefficient: 0.73 (95% CI, 0.70 to 0.76)			

95% CI = 95% confidence interval, as described by Fan and Thompson.³⁴

CBME = competency-based medical education

Table 3 Generalizability analysis for variance in faculty

Facet	Variance component (VC)	VC%	Description
Resident (r)	0.033	6.0%	Variance due to differences between residents
Faculty (f)	0.003	0.5%	Variance due to different faculty raters
Item (i)	0.111	20.9%	Variance due to the items on the tool
r*f	0.134	35.3%	Variance due to different faculty members' assessments of different residents (i.e., did residents differ in how they were scored by different faculty members?)
r*i	0.011	2.0%	Variance attributable to the interaction between residents and items (i.e., did residents differ in how they were scored on particular items?)
f*i	0.006	1.2%	Variance attributable to the interaction between faculty members and items (i.e., did item means vary across different faculty members?)
r*f*i	0.234	44.0%	Variance attributable to the residents, faculty members, items, plus random error
G-coefficient: 0.47 (95% CI, 0.16 to 0.69)			

95% CI = 95% confidence interval, as described by Fan and Thompson.³⁴

Table 4 Median, 25th and 75th percentile, and mean rank values for pre-, intra-, and postoperative Clinical Case Assessment Tool (CCAT) scores across the three modules of the competency-based medical education (CBME) curriculum. Friedman analysis of variance, Wilcoxon signed-rank tests, and corresponding p-values compare mean rank of median CCAT scores across the three CBME modules

		CCAT item			Statistical analyses	
		Preoperative	Intraoperative	Postoperative	Friedman's analysis of variance	Wilcoxon signed-rank tests
Transition to Discipline	Median	3.0	2.0	3.0	$\chi^2 = 41.07, P < 0.001$	^a Z = 4.72, $P < 0.001$, r = 0.80
	25th percentile	3.0	2.0	2.0		^b Z = 3.41, $P = 0.04$, r = 0.58
	75th percentile	3.0	3.0	3.0		^c Z = 3.70, $P < 0.001$, r = 0.63
	Mean median rank	2.6	1.4	2.0		
	Sample size	313	329	281		
Bootcamp Foundations	Median	3.0	3.0	3.0	$\chi^2 = 30.11, P < 0.001$	^a Z = 4.18, $P = 0.002$, r = 0.71
	25th percentile	3.0	3.0	3.0		^b Z = 3.83, $P = 0.009$, r = 0.65
	75th percentile	4.0	3.0	3.5		^c Z = 1.70, $P = 1.0$, r = 0.29
	Mean median rank	2.4	1.5	2.1		
	Sample size	731	753	602		
Clinical Foundations	Median	4.0	3.0	4.0	$\chi^2 = 37.00, P < 0.001$	^a Z = 4.61, $P < 0.001$, r = 0.78
	25th percentile	3.0	3.0	3.0		^b Z = 4.36, $P < 0.001$, r = 0.74
	75th percentile	4.0	3.0	4.0		^c Z = 1.04, $P = 1.0$, r = 0.17
	Mean median rank	2.4	1.4	2.2		
	Sample size	1,127	1,170	979		

Effect sizes for Wilcoxon signed-rank tests were calculated using r statistic²⁷

Sample size refers to the number of data points used to calculate median and percentile values

^a Comparing medians for preoperative vs intraoperative scores

^b Comparing medians for postoperative vs intraoperative scores

^c Comparing medians for preoperative vs postoperative scores

Table 5 Spearman's rho correlations and percentage of variability explained (R^2) between Clinical Case Assessment Tool (CCAT) items, collapsed across blocks of training, and percentile scores for

Anesthesia Knowledge Test-1 (AKT-1), Anesthesia Knowledge Test-6 (AKT-6), mean Objective Structured Clinical Examination (OSCE) score, and Clinical Competency Committee (CCC) resident ranking

	CCAT item (rho (R^2))		
	Preoperative	Intraoperative	Postoperative
AKT-1 percentile	0.18 (3%)	0.24 (6%)	0.1 (1%)
AKT-6 percentile	0.29 (8%)	0.51* (26%)	0.39 (15%)
Mean OSCE score	0.29 (8%)	0.45* (20%)	0.35 (12%)
CCC rank	0.48* (23%)	0.55** (30%)	0.47* (22%)

*Bonferroni-adjusted $P < 0.05$; **Bonferroni-adjusted $P < 0.01$

median preoperative and postoperative scores ($Z = 1.70$, adjusted $P = 1.00$, $r = 0.29$). Lastly, during the Clinical Foundations module, median intraoperative entrustment scores were lower than preoperative ($Z = 4.61$, adjusted $P < 0.001$, $r = 0.78$) and postoperative ($Z = 4.36$, adjusted $P < 0.001$, $r = 0.74$) entrustment scores. Once again, there was no difference in preoperative and postoperative scores ($Z = 1.04$, adjusted $P = 1.0$, $r = 0.17$).

The Spearman's correlation evaluating concurrent validity is displayed in Table 5. Anesthesia Knowledge Test-1 scores did not correlate with any of the three CCAT items (all adjusted $P > 0.99$). In contrast, AKT-6 percentile scores (Spearman's rho = 0.51, adjusted $P = 0.01$) and mean OSCE scores (Spearman's rho = 0.45, $P_{\text{bonferroni}} = 0.04$) were positively associated with intraoperative CCAT scores. With regards to CCC rankings of residents' performance, there were significant correlations across all three items (all adjusted $P < 0.03$), indicating that residents with higher entrustment scores were ranked higher by the CCC.

Discussion

In the era of CBME, it is imperative that modes of assessment provide reliable and valid measures of learner performance. Our study aimed to assess whether an entrustment-based construct in anesthesiology training was sufficiently valid and reliable for continued incorporation within a program of assessment. In the present study, generalizability analysis revealed good reliability of entrustment scores, as indicated by a G-coefficient of 0.73. This value surpasses the threshold deemed appropriate for evaluation in intermediate stakes assessments^{25,28} but not for summative decisions. The G-coefficient associated with our results is greater than that calculated for similar entrustment-based assessment in the Department of Internal Medicine (G-coefficient = 0.61)¹³ but less than that in the Department of Surgery (G-coefficient = 0.88)⁸ at the same institution, although our

results are derived from a substantially greater number of submitted forms. Our analysis also revealed that the entrustment category (preoperative, intraoperative, or postoperative plan) contributed the most variability to scores in this study. We postulate two potential explanations for why intraoperative entrustment scores tended to be consistently lower than those for pre- and postoperative settings across the first seven blocks of training (Table 4): 1) preoperative scores may be initially elevated as they assess clinical skills that incoming residents have already acquired during undergraduate training (e.g., history taking and preoperative assessment); 2) direct observation by faculty members is anticipated to be more frequent during the intraoperative phase thus permitting higher accuracy of entrustment allocation.

Given that the entrustment scale derives from direct observation in the workplace, one would assume that it should yield information of high authenticity. The entrustment anchors employed in our study setting were derived from a previously validated tool developed for the assessment of surgical trainees.¹⁹ Although the original O-SCORE entrustment scale has not previously been assessed in anesthesiology training, Weller *et al.* described that the use of similar entrustment anchors significantly improved the reliability of anesthesiology WBA.¹⁰ Furthermore, subsequent studies have validated these entrustment scales in non-surgical fields such as internal medicine, suggesting that the construct is not specific to surgical specialties.^{12–15} To explore our assumption of authenticity, we evaluated the longitudinal progression of entrustment scores during the first seven blocks of training and via comparison with previously established assessment modalities. That entrustment scores significantly improved through the first three modules of training suggests they assessed an appropriate metric, namely a level of entrustment for clinical aptitude. Previous studies have similarly shown that entrustment scores are sensitive to residents' progression through training.^{8,12–15} In addition, entrustment scores for intra- and postoperative items

significantly correlated with the AKT and OSCE score at six months but not the AKT-1. Together, these results suggest entrustment scores provide a concurrently valid means of assessing resident performance during the intraoperative and postoperative setting.

A discussion of validation is relatively meaningless without an understanding of the implications for the assessment. In our program, the entrustment scale serves a formative purpose and is used to encourage narrative feedback with the goal of subsequent performance improvement. The reason we have selected this approach relates to the competency framework within our program, which is modelled after the original description by ten Cate¹⁶ and geared towards entrustment over a continuum that culminates in competence at the conclusion of training. Given that our study only included the first seven months of training and our reliability G-coefficient was < 0.8 , it is premature to posit that the entrustment scale using the CCAT alone provides a basis to make any high stakes decisions like promotion between stages of training. Rather, our study provides evidence that the entrustment scale is valid for CCC deliberations, though it should be carefully interpreted within the greater context of an assessment program.

The following limitations of this study are acknowledged. The findings are based on the implementation of an entrustment scale using the CCAT, which was collected at a single centre. This may limit translation of our generalizability to other programs who use different WBA methods or competency frameworks. When looking at the CCAT specifically, there are two main differences that may have affected our results. Firstly, the tool includes entrustment items for different phases of care (e.g., pre-, intra-, postoperative), which may have led to a form of measurement bias. Within the 2,309 CCAT forms submitted, 642 (28%) had data missing from at least one category. The majority of missing data were postoperative entrustment scores ($n = 447$), followed by preoperative scores ($n = 138$), with relatively few missing data points for intraoperative scores ($n = 57$). We speculate that this pattern of data omission may reflect the availability of the supervising anesthesiologist during each phase of care. While the evaluating anesthesiologist is often present intraoperatively, they may be absent during residents' preoperative assessment and postoperative care. Faculty participation during these times is important for quality of care given that entrustment scores for pre- and postoperative categories do not suggest residents are yet capable of independent practice in early training. Secondly, while residents are allocated by our schedulers to work with a variety of faculty members, the CCAT is a completely resident-initiated assessment tool. The fact that the residents alone can initiate this assessment may

constitute selection bias as learners can strategize to omit evaluations on cases which they perceive would lead to lower entrustment allocation. This exemplified behaviour may have led to a misrepresentation of our data given that an average of only 9.4 CCATs were submitted per resident over a four-week block; this accounts for half of the daily assessments that should occur over this period.

Another limitation is that we were unable to control for case exposure and faculty raters in the present study. Throughout the duration of the study, faculty members assessed different residents across cases of varying complexity, meaning that faculty ratings were confounded with residents/patient cases. For this reason, it was not possible to incorporate these measures as further subgrouping precluded sufficient statistical powering. Given that case complexity and variability in faculty assessors likely influence entrustment,²⁵ our inability to quantify the influence of these factors is a clear limitation of our study. In favour of our results, the fact that each resident was assessed by a wide range of faculty members (e.g., an average of 63 faculty members per resident), it is unlikely that residents were rated by a biased sample of raters based on the central limit theorem. To further disentangle the effect of rater on entrustment scores, we ran a generalizability analysis by identifying a subset of faculty members who had rated the greatest number of residents across the study time period. While faculty raters accounted for little variability in entrustment scores (0.5%), substantial variation (35%) was associated with the interaction between faculty member and resident, suggesting that scores associated with individual residents varied across faculty members. This interaction may be because faculty members have their own individual rater biases or that faculty members observed residents during cases of varying complexity and at different times in the CBME curriculum. Future studies are needed to examine inter-rater scoring of entrustment scales to better inform current practice habits and facilitate future faculty development initiatives in entrustment-based assessment.²⁹

Lastly, we would like to make special note regarding the incorporation of Kane's Validity Framework and the CCC ranking process. While the framework enabled a cohesive argument for evaluation of the construct, we felt that it relied too heavily on the judgement of the researchers, particularly in the implications section given that no observable metrics for intended use were identified *a priori*. Similar to previous work describing the use of this framework in simulation,³⁰ future studies looking to evaluate the validity of entrustment in CBME may benefit from beginning with clearly identified assumptions and outcome measures for each inference with prioritization afforded to the weakest link (e.g., implications).^{31,32} While the apparent "wisdom of the

crowds” should provide better data than any individual or single assessment, the idea of CCC ranking made many of our members feel uncomfortable and question the utility of such a process. Future research on CCC deliberations for summative assessment should acknowledge the potential biases proposed by Dickey³³ and seek to evaluate the unintended consequences related to each individual assessment practice.

In summary, identifying WBAs that can identify genuine progression towards competency has emerged as a major theme in contemporary medical education. Measuring objective independence in the workplace helps provide clear outcomes for all stakeholders and represents a crucial component of trainee performance assessment. Our study found that perioperative entrustment scores, particularly for intraoperative phases of care, provide a reliable and valid means of assessing clinical performance during the early stages of anesthesiology residency training. Future studies are necessary to discern if these results translate to a wider breadth of competencies such as those proposed by the Royal College of Physicians and Surgeons of Canada and whether entrustment scores can be used to provide criteria or normative data for summative decision-making by the CCC. Broaching these challenges

will require robust validity evidence and it would be desirable to consider a longitudinal national approach to collectively answer these questions.

Author contributions *Daniel Dubois* contributed to study conception and design; acquisition and interpretation of data; and drafting the article. *Alexander Lingley* contributed to drafting the article. *Julie Ghatalia* contributed to acquisition of data and drafting the article. *Meghan McConnell* contributed to study design, data analysis and interpretation, and drafting the article.

Disclosures None.

Funding statement None.

Editorial responsibility This submission was handled by Dr. Steven Backman, Associate Editor, *Canadian Journal of Anesthesia*.

Appendix: Sample Clinical Case Assessment Tool (CCAT) administered at the University of Ottawa Faculty of Medicine Department of Anesthesiology and Pain Medicine

uOttawa

Department of Anesthesiology Clinical Case Assessment Tool

Resident:

Staff:

Date:

Case:

The purpose of this assessment is to evaluate the resident's ability to complete the case safely and independently. With that in mind, please use the scale below to evaluate each stage of this case, irrespective of the resident's level of training.

Scale

- 1 – "Staff had to do"** (e.g., resident required complete hands-on guidance by staff, resident did not do case, resident was not given the opportunity to do case)
2 – "Staff had to talk resident through" (e.g., resident able to perform tasks but needed constant direction)
3 – "Staff had to prompt resident from time to time" (e.g., resident demonstrated some independence but required intermittent direction)
4 – "Staff needed to be in the room just in case" (e.g., resident was independent but unaware of risks and required supervision for safe practice)
5 – "Staff did not need to be there" (e.g., resident completely independent, understood risks and performed safely, ready for practice)

All 6 questions are mandatoryQs 1-3 to be completed by staff:*

	1	2	3	4	5
1. Pre-operative / History-taking	☐	☐	☐	☐	☐
2. Intra-operative / Physical examination	☐	☐	☐	☐	☐
3. Post-operative plan / Management plan	☐	☐	☐	☐	☐

Qs 4-6 to be completed by resident and staff:

4. What did the resident do well?

- ☐ Medical Expert
- ☐ Communicator
- ☐ Collaborator
- ☐ Manager
- ☐ Health Advocate
- ☐ Scholar
- ☐ Professional

5. What should the resident do differently next time?

- ☐ Medical Expert
- ☐ Communicator
- ☐ Collaborator
- ☐ Manager
- ☐ Health Advocate
- ☐ Scholar
- ☐ Professional

6. What are the next steps in the resident's learning plan?

7. General Comments:☐ Discussed with resident

Staff Signature:

Resident Signature:

Put completed forms in the locked box at your site. Comments and questions should be directed to the resident site coordinator.

References

1. Snell LS, Frank JR. Competencies, the tea bag model, and the end of time. *Med Teach* 2010; 32: 629-30.
2. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach* 2007; 29: 642-7.
3. Frank JR, Snell LS, ten Cate O, et al. Competency-based medical education: theory to practice. *Med Teach* 2010; 32: 638-45.
4. 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.
5. Wilkinson JR, Crossley JG, Wragg A, et al. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008; 42: 364-73.
6. van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach* 2012; 34: 205-14.
7. van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005; 39: 309-17.
8. Rekman J, Hamstra SJ, Dudek N, et al. A new instrument for assessing resident competence in surgical clinic: The Ottawa Clinic Assessment Tool. *J Surg Educ* 2016; 73: 575-82.
9. 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: e90-6.
10. 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.
11. Weller JM, Castanelli DJ, Chen Y, Jolly B. Making robust assessments of specialist trainees' workplace performance. *Br J Anaesth* 2017; 118: 207-14.
12. Warm EJ, Mathis BR, Held JD, et al. Entrustment and mapping of observable practice activities for resident assessment. *J Gen Intern Med* 2014; 29: 1177-82.
13. Halman S, Rekman J, Wood T, et al. Avoid reinventing the wheel: implementation of the Ottawa Clinic Assessment Tool (OCAT) in internal medicine. *BMC Med Educ* 2018; DOI: <https://doi.org/10.1186/s12909-018-1327-7>.
14. Warm EJ, Held JD, Hellmann M, et al. Entrusting observable practice activities and milestones over the 36 months of an internal medicine residency. *Acad Med* 2016; 91: 1398-405.
15. Mink RB, Schwartz A, Herman BE, et al. Validity of level of supervision scales for assessing pediatric fellows on the common pediatric subspecialty entrustable professional activities. *Acad Med* 2018; 93: 283-91.
16. ten Cate O. Nuts and bolts of entrustable professional activities. *J Grad Med Educ* 2013; 5: 157-8.
17. ten Cate O. Entrustment as assessment: recognizing the ability, the right, and the duty to act. *J Grad Med Educ* 2016; 8: 261-2.
18. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015; 49: 560-75.
19. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med* 2012; 87: 1401-7.
20. Association of Canadian University Departments of Anesthesia. National Curriculum for Canadian Anesthesiology Residency. Ottawa, Ontario; 2015. Available from URL: <http://www.royalcollege.ca/rcsite/documents/ibd/anesthesiology-national-curriculum-e.pdf>. Accessed July 2020.
21. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract* 2014; 19: 223-50.
22. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003; 37: 830-7.
23. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006; 119(166): e7-16.
24. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002; 36: 972-8.
25. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004; 38: 1006-12.
26. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul* 2016; DOI: <https://doi.org/10.1186/s41077-016-0033-y>.
27. Field A. Discovering Statistics Using IBM SPSS Statistics, 5th Edition. Sage Publications; 2017.
28. Sterkenburg A, Barach P, Kalkman C, Gielen M, ten Cate O. When do supervising physicians decide to entrust residents with unsupervised tasks? *Acad Med* 2010; 85: 1408-17.
29. Holmboe ES, Ward DS, Reznick RK, et al. Faculty development in assessment: the missing link in competency-based medical education. *Acad Med* 2011; 86: 460-7.
30. Tavares W, Brydges R, Myre P, et al. Applying Kane's validity framework to a simulation based assessment of clinical competence. *Adv Health Sci Educ Theory Pract* 2018; 23: 323-38.
31. Govaerts MJ, van der Vleuten CP, Holmboe ES. Managing tensions in assessment: moving beyond either-or thinking. *Med Educ* 2019; 53: 64-75.
32. Hawkins RE, Welcher CM, Holmboe ES, et al. Implementation of competency-based medical education: are we addressing the concerns and challenges? *Med Educ* 2015; 49: 1086-102.
33. Dickey CC, Thomas C, Feroze U, Nakshabandi F, Cannon B. Cognitive demands and bias: challenges facing clinical competency committees. *J Grad Med Educ* 2017; 9: 162-4.
34. Fan X, Thompson B. Confidence intervals about score reliability coefficients, please: an EPM guidelines editorial. *Educ Psychol Meas* 2001; 61: 517-31.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.