Canadian Medical Education Journal

Major Contributions

The critical role of direct observation in entrustment decisions

Le rôle fondamental de l'observation directe dans la décision de confier une responsabilité professionnelle

Matthew Sibbald,[1] Muqtasid Mansoor,[2] Michael Tsang,[2] Sarah Blissett,[3] Geoffrey Norman[1]

[1]McMaster Faculty of Health Sciences Education Research, Innovation and Program (MERIT), McMaster University, Ontario, Canada; [2]McMaster University, Ontario, Canada; [3]Centre for Education Research and Innovation, Schulich School of Medicine, Western University, Ontario, Canada

Correspondence to: Matthew Sibbald, Hamilton General Hospital, 5th floor McMaster Wing 237 Barton St, Hamilton Ontario L8L 2X2; phone: 905-526-7616; fax: 905-527-4463; email: sibbald@mcmaster.ca

Published ahead of issue: June 15, 2021; CMEJ 2021 Available at http://www.cmej.ca

https://doi.org/10.36834/cmej.72040

Abstract

Background: Entrustment decisions may be retrospective (based on past experiences with a trainee) or real-time (based on direct observation). We investigated judgments of entrustment based on assessor prior knowledge of candidates and based on systematic direct observation, conducted in an objective structured clinical exam (OSCE).

Methods: Sixteen faculty examiners provided 287 retrospective and real-time entrustment ratings of 16 cardiology trainees during OSCE stations in 2019 and 2020. Reliability and validity of these ratings were assessed by comparing correlations across stations as a measure of reliability, differences across postgraduate years as an index of construct validity, correlation to standardized in-training exam (ITE) as a measure of criterion validity, and reclassification of entrustment as a measure of consequential validity.

Results: Both retrospective and real-time assessments were highly reliable (all intra-class correlations >0.86). Both increased with a year of postgraduate training. Real-time entrustment ratings were significantly correlated with standardized ITE scores; retrospective ratings were not. Real-time ratings explained 37% (2019) and 46% (2020) of variance in examination scores vs. 21% (2019) and 7% (2020) for retrospective ratings. Direct observation resulted in a different level of entrustment compared with retrospective ratings in 44% of cases (p = <0.001).

Conclusions: Ratings based on direct observation made unique contributions to entrustment decisions.

Résumé

Contexte : La décision de confier une activité peut être rétrospective (basée sur les expériences antérieures avec un apprenant) ou en temps réel (basée sur l'observation directe). Nous avons étudié les évaluations de niveaux de confiance fondées sur des interactions antérieures des candidats par les évaluateurs et celles fondées sur l'observation directe systématique, dans le cadre d'un examen clinique objectif structuré (ECOS).

Méthodes : Seize évaluateurs du corps professoral ont fourni 287 évaluations rétrospectives et en temps réel du niveau de confiance faites lors des stations d'ECOS en 2019 et 2020 concernant 16 stagiaires en cardiologie. La fiabilité et la validité de ces évaluations ont été analysées en comparant les corrélations entre les stations comme mesure de la fiabilité, les différences entre les années d'études postdoctorales comme indice de la validité de construit, la corrélation avec l'examen normalisé en cours de formation (ITE) comme mesure de la validité de critère, et le reclassement des évaluations de la confiance comme mesure de la validité corrélative.

Résultats : Les évaluations rétrospectives et en temps réel étaient toutes les deux très fiables (toutes les corrélations intra-classes >0,86). Les deux augmentaient avec le niveau de formation postdoctorale. Les évaluations de la confiance en temps réel étaient significativement corrélées aux scores de l'examen normalisé en cours de formation; les évaluations rétrospectives ne l'étaient pas. Les évaluations en temps réel expliquaient 37 % (2019) et 46 % (2020) de la variance des notes d'examen, contre 21 % (2019) et 7 % (2020) pour les évaluations rétrospectives. L'observation directe a permis de reclasser 44 % des évaluations rétrospectives de la confiance (p=<0,001 dans les deux cas).

Conclusion : Les évaluations basées sur l'observation directe contribuent de façon importante à la décision de confier une activité.

Introduction

While the theoretical underpinning of competency based medical education (CBME) emphasizes the role of direct observation on entrustment decisions,[1,2] direct observation is not mandatory.[3–5] Direct observation is second nature in procedurally oriented specialties.[6] However, specialties focused on medical decision-making are less amenable to observation-based entrustment, prompting calls to nuance the universal application of a direct observation approach.[7,8] Understanding the extent to which direct observation informs entrustment decisions in cognitive tasks would advance the science of CBME, however this topic remains underexplored in the published literature.

The potential value of direct observation in entrustment decisions is made apparent through a contrast of two studies in emergency medicine. One used standardized assessment of observable tasks[9] showing a strong gradient in ratings over each postgraduate training year, whereas a second collected ratings after each shift[10] showing minimal gradient within each postgraduate year. The contrast between these two studies in their ability to identify growth provides some evidence of the importance of real-time direct observation in entrustment decision making.

Despite the emphasis on and the potential value of direct observation in CBME contexts, direct observation is underutilized in cognitive specialities. In busy, competency-based residency programs, faculty will frequently provide retrospective entrustment ratings of uncomplicated delegated acts that they did not directly observe.[7,10] Instead, the assessment derives from some kind of implicit mental averaging of the supervisor's observation of resident performance on the specific task over time. While such an approach, averaging over multiple observations, might be considered more reliable and valid than a single standardized observation, it depends on the supervisor's ability to recall and summarize. Such summative judgments are vulnerable to biases such as "primacy" and "recency" effects.[11] A central question for supervisors in cognitive specialities is 'how often does direct observation of these delegated acts lead to similar entrustment ratings to those provided by the supervisor from informal contact with the resident?"

In this study, we compared judgments of entrustment on specific stations of an objective structured clinical exam (OSCE) focused on cognitive tasks under three conditions: 1) expected level of performance, where the assessor was asked to rate typical performance for a resident at a given level with this station; 2) retrospective, where the assessor estimated how the resident would perform based on their prior observations with the resident; and 3) real-time, based on direct observation.

Methods

We conducted a prospective study comparing retrospective and real-time ratings to expected level of performance ratings in two sequential years of a residency program OSCE. The approach used the standard psychometric criteria of reliability and validity.[12] Reliability was determined across all stations in the OSCE. Construct validity was assessed by examining differences with years of training. Criterion validity was assessed by comparison with a standard in-training written examination. Consequential validity was assessed by examining change in entrustment decisions resulting from observed assessment.

Setting

Postgraduate cardiology trainees in postgraduate years (PGY) four through six participated in a formative 4-hour OSCE at a single center. The OSCE was blueprinted from the objectives of training for cardiology residency programs (Table 1). In February 2019, 10 residents participated in 12 stations with 12 different faculty examiners. In February 2020, 13 residents participated in 13 stations with 13 different faculty examiners. Seven residents and nine faculty examiners participated in both 2019 and 2020 examinations.

Table 1. OSCE Stations in 2019 and 2020

Station	2019 OSCE	2020 OSCE
1	Acute coronary disease	Acute coronary disease
2	Chronic coronary disease	Chronic coronary disease
3	Valvular heart disease	Valvular heart disease
4	Cardiac physical exam	Congenital heart disease: follow up visit of repaired tetralogy
5	Hypertension	Heart failure and cardiomyopathies
6	Pulmonary vascular disease	Hypertension related to aortic coarctation
7	Pericardial disease	Pulmonary vascular disease
8	Vascular medicine	Pericardial disease
9	Acute cardiac care	Vascular medicine
10	Electrophysiology	Acute cardiac care
11	Pregnancy in patients with cardiovascular disease	Electrophysiology
12	Congenital heart disease	Pregnancy in patients with cardiovascular disease
13	N / A	Cardiac physical exam on high fidelity simulator

Each station was constructed to mimic a clinical encounter, with the examiner playing the role of the patient. Residents were required to take a history, interpret physical exam data, interpret investigations (e.g. bloodwork, electrocardiogram, chest radiograph, echocardiography, angiograms etc.) and integrate these data into a management plan communicated to the patient. No procedural skills were tested. One faculty member was assigned to each station, based on content expertise. All faculty members knew all trainees for an average of 1.7 years and had worked with them in at least one clinical context in the last 10 months.

Entrustment ratings

Before the OSCE, faculty members were asked to review the station to which they were assigned, and decide the level of supervision they would provide for each resident: (1) based on the time the resident spent in the training program alone, i.e. postgraduate year (expected level of performance) and (2) based on prior experience with the resident (retrospective). During the OSCE, faculty members provided the level of supervision they felt appropriate after observing the trainee complete the station (real-time). All three types of entrustment ratings used the same entrustment scale based on prior scoring systems:[4,13]

1. Not yet developed
2. Competent to manage with proactive or direct supervision (i.e. needs to talk through it)
3. Competent to manage with reactive or on demand supervision (i.e. needs prompting for some management components)
4. Competent to manage without supervision (i.e. can provide definitive short- and long-term management for all aspects of the problem without prompting)
5. Ready to teach this (i.e. sophisticated understanding of the problem and its possible clinical variations and their impact on management)

Scores of four or higher are required for documentation of competence, whereas scores of three or lower imply some further development is required.

Standardized testing

Each October, all residents completed an international six hour standardized in-training examination (ITE) constructed by the American College of Cardiology. The examination was separate in time from the OSCE. The ITE contained approximately 150 items blueprinted from the objectives of training for cardiology residency programs. Resident scores are reported as percentiles.

Analysis

Psychometric analyses were conducted separately for 2019 and 2020, and the two analyses were treated as replications. While some residents were in both cohorts this was not accounted for in the analysis.

Descriptive statistics: We calculated means and standard deviations separately for exam year (2019, 2020), postgraduate year (PGY) (4,5,6) and scoring method (expected, retrospective, real time). Histograms were constructed for average OSCE score and each type of entrustment by PGY, for 2019 and 2020.

Reliability

Test reliability for each rating type was computed across all 12 (2019) or 13 stations (2020). As each OSCE station had different content and raters, reliability estimates incorporate both variances related to content and raters. We performed a repeated measures ANOVA on individual station scores then calculated the G coefficient for the mean score across stations. In this analysis, reliability was assessed in comparison with other residents *at the same level* as Resident scores were nested in the PGY variable.

Validity: We considered three pieces of evidence that entrustment ratings reflect trainees' abilities to practice safely and independently:[12] the ability of the rating type to distinguish among residents at each level of training, comparison to an external standard as a form of criterion validity, and recategorization of decisions through direct observation as a form of consequential validity.

Relation to PGY

The previous analysis of variance, by rating type and PGY level, was also used to test for differences among means for both 2019 and 2020 OSCEs. Since every assessor was aware of the level of each resident, this was a weak test of validity.

Relation to In-Training Exam (ITE)

As an objective standard of performance, ITE multiple choice test can be criticized for not comprehensively assessing important domains such as communication skills. Despite this limitation, multiple choice testing has been shown superior to OSCEs in predicting malpractice,[14] peer review problems 10 years after graduation.[15] and 30 day mortality in the coronary care unit.[16]

We computed simple correlations between average OSCE score and ITE score for 2020 and 2019. Postgraduate year was ignored. Following this analysis, the additional variance accounted for by retrospective and real-time ratings was calculated by first computing R^2 for each Pearson correlation coefficient then taking the difference between this and the expected rating $R^{2.}$

Recategorization of entrustment decisions through direct observation

We examined consequential validity by examining how frequently real-time judgements resulted in recategorization of retrospective entrustment decisions, both using a 5-point ordinal scale typical of most entrustment measurements[4,13] and recategorization around the binary threshold typically used for summative decision making. Net recategorization by observation was calculated separately for both real time and retrospective ratings by creating tables of real time and retrospective ratings as columns and observed ratings as rows and using chi square testing to determine the significance of recategorization. Net recategorization was defined as 1 – the percentage of equivalently categorized trainees by retrospective ratings compared to observed real-time ratings of entrustment.

Critical value for significance

In the setting of multiple statistical tests, we applied a Bonferroni correction to maintain a type 1 error rate of 0.05, which resulted in the statistical threshold of $p < 0.0025$ being considered significant.

Ethical approval was obtained by the Hamilton Integrated Ethics Review Board protocol #7567.

Results

Descriptive statistics

Expected entrustment ratings (i.e., based on training time), retrospective entrustment ratings (i.e., based on previous experience with the trainee) and real-time entrustment ratings (based on observed performance in the OSCE station) all varied by trainee level (Figure 1). Interestingly, real-time entrustment ratings had trainees from each level in each category whereas expected and retrospective ratings did not (Figure 1).

Reliability

Reliability coefficients for both retrospective and real-time ratings were large (R2>0.855), as shown in Table 2.

Validity

Each type of entrustment rating increased with PGY in each OSCE year as shown in Table 2 (all $p < 0.001$). The relation between scores derived from the three methods and the ITE are shown in Table 3. Only the real-time ratings were significantly correlated with the ITE. These ratings accounted for substantially more variance in the examination than retrospective ratings. Real-time observation resulted in both increased and decreased entrustment compared with expected and retrospective judgments (Figure 2). Observation reclassified 38% of expected entrustment ratings (X^2= 73, df = 16, $p < 0.00001$) and 44% of retrospective entrustment ratings (X^2 = 102, df = 16, $p < 0.00001$). When entrustment ratings were reclassified in a binary system (with scores equal to or greater than four considered competent), observation reclassified 33% of expected entrustment ratings (X^2= 31.5, df = 1, $p = 0.0001$) and 29% of retrospective ratings (X^2= 49.1, df=1, $p = 0.0001$).

Table 2. Entrustment ratings and test reliability in 2019 and 2020 objective structured clinical exam (OSCE)

Rating type	2019 OSCE				2020 OSCE				Mean rating for all trainees		Rating reliability (intra-class correlation)	
	PGY4	PGY5	PGY6	p across PGY	PGY4	PGY5	PGY6	P across PGY	2019	2020	2019	2020
Expected for level	2.50 ± 0.66	3.42 ± 0.65	4.33 ± 0.63	<0.001	2.00 ± 0.40	3.08 ± 0.48	4.08 ± 0.48	<0.001	3.42 ± 0.96	3.05 ± 0.96	0.982	0.998
Retrospective	3.00 ± 0.70	3.71 ± 0.77	4.22 ± 0.68	<0.001	2.52 ± 0.90	3.31 ± 0.77	3.57 ± 0.81	<0.001	3.66 ± 0.86	3.13 ± 0.94	0.924	0.932
Real-time	3.33 ± 0.86	3.44 ± 1.01	4.06 ± 0.67	<0.001	2.62 ± 0.88	3.38 ± 0.89	3.62 ± 0.97	<0.001	3.59 ± 0.92	3.16 ± 1.01	0.855	0.887

Ratings are out of 5. A score of 4 or more represents the ability to perform the station independently.

Table 3. Correlations with in-training exam scores and variance explained by expected for level, retrospective and real time entrustment ratings

Entrustment rating type	2019			2020		
	Correlation	% variance	Additional % variance**	Correlation	% variance	Additional % Variance**
Expected for level	0.253	0.064	---	-0.023	0.0#	---
Retrospective	0.525	0.275	0.212	0.258	0.066	0.066
Real time	.658 *	0.432	0.368	.678 *	0.459	0.459

Discussion

The data presented in this study indicate that ratings derived from real-time observation in a standardized setting contributed unique information to assessment of individual residents. Direct observation resulted in net reclassification of entrustment ratings often; one in three ratings was reclassified across the threshold of entrustment typically used for summative decision making. Expected ratings explained only from 0 to 6% of the variance in ITE scores; retrospective judgments explained an additional 7-21% of examination performance. However, real-time ratings explained an additional 37 - 46% of the variance. Only real-time entrustment ratings correlated significantly with standardized testing.

These findings provide validity evidence for standardized direct observation in entrustment ratings, even in predominantly cognitive tasks. While this study involved a small number of trainees in a single discipline, multiple tasks were assessed in a rigorous format, with replication of findings across two years. While this does not guarantee generalization to other disciplines or contexts, it is consistent with the inferences drawn between studies to suggest greater discriminatory ability of ratings based on direct observation.[9,10]

Practically, more reliance on direct observation could substantially impact opportunities and supervision provided to trainees. Currently, supervisors often rely on informal observation to form entrustment judgments and allow trainees to engage in activities in the workplace, sometimes without direct supervision. Faculty in this study were asked to make a similar judgment by considering a specific situation and assigning a level of supervision based on their prior impressions of the resident, then directly observed the resident. Interestingly, over 40% of the time faculty changed their minds after observing the trainee. Based on this finding, relying on retrospective judgments of entrustment will result in a substantive percentage of trainees being given more independence (and some less independence) in the workplace than if the degree of supervision were determined by direct observation. Relying only on retrospective judgements for entrustment potentially denies some junior trainees an opportunity for independent learning and places some senior trainees in situations of inadequate supervision.

The high frequency of reclassification of entrustment, particularly the reclassification of senior trainees to lower levels of entrustment, strengthens the validity argument for the use of direct observation entrustment ratings in a competency-based residency framework. This also calls into question the risky practice of presumptively entrusting residents, then documenting entrustment when no complications occurred from a delegated act. This is indirect evidence at best.

Further, the substantive reclassification of trainees based on direct observation highlights potential validity risks in assigning entrustment ratings based on integrating prior experiences, as is sometimes done in residency in-training assessments. These ratings will frequently differ from ratings based on observation, as documented in this study.

There are several important limitations of the study. It involved a small number of participants taken from a single centre and discipline. However, it had a large number of assessors, spanned multiple content domains, had a reasonable variation of competency across years of training and was adequately powered to draw robust conclusions. Further, all critical conclusions were replicated across the two cohorts of trainees. The choice of criterion measure, the ITE, is proximate and empirically defensible, but does leave unanswered the relation between measures in the educational setting and longer-term outcomes.

In that regard, all entrustment ratings in this study were based on an OSCE setting without relevant patient outcomes. This has significant downsides. First, the retrospective decision of supervisors in the OSCE setting may be prone to recall bias compared with an assessment

done at the end of rotation. Second, the supervisors were acting as patients in the scenarios which requires trainees and supervisors to suspend their disbelief around the simulation of the scenario, potentially reducing the authenticity of the interaction. However, the advantage of this setup is that it frees supervisors from the urge to prompt or step in, simplifying the entrustment decision process. While this theoretically might lead to overestimating entrustment, a greater percentage of ratings obtained through direct observation were reclassified at a lower level than a higher level of entrustment.

While all faculty examiners worked with all trainees, the degree of clinical experiences likely varied. Whether or not retrospective entrustment ratings are more correlated when faculty supervise trainees for longer periods of time in the clinical environment is unknown.

Conclusions

In summary, direct observation adds to the validity of entrustment ratings. Even among senior residents performing cognitive tasks, direct observation affects faculty impressions.

Conflicts of Interest: The authors declare no conflicts of interest.
Funding: None.

References

1. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Med Teach*. 2010;32(8):638–45. https://doi.org/10.3109/0142159X.2010.501190
2. Ten Cate O, Hart D, Ankel F, et al. Entrustment decision making in clinical training. *Acad Med*. 2016;91(2):191–8. https://doi.org/10.1097/ACM.0000000000001044
3. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. *Acad Manag Rev*. 1995;20(3):709–34. https://doi.org/10.5465/amr.1995.9508080335
4. Holzhausen Y, Maaz A, Cianciolo AT, ten Cate O, Peters H. Applying occupational and organizational psychology theory to entrustment decision-making about trainees in health care: a conceptual model. *PME.* 2017;6(2):119–26. https://doi.org/10.1007/s40037-017-0336-2
5. 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(9):1408–17. https://doi.org/10.1097/ACM.0b013e3181eab0ec
6. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med*. 2007;82(6):542–7. https://doi.org/10.1097/ACM.0b013e31805559c7
7. Hatala R, Ginsburg S, Hauer KE, Gingerich A. Entrustment ratings in internal medicine training: capturing meaningful supervision decisions or just another rating? *Gen Intern Med*. 2019;34(5):740–3. https://doi.org/10.1007/s11606-019-04878-y
8. Landreville JM, Cheung WJ, Hamelin A, Frank JR. Entrustment Checkpoint: Clinical Supervisors' Perceptions of the Emergency Department Oral Case Presentation. *TLM*. 2019;31(3):250–7. https://doi.org/10.1080/10401334.2018.1551139
9. Chan T, Sherbino J, Collaborators M. The McMaster Modular Assessment Program (McMAP): a theoretically grounded work-based assessment system for an emergency medicine residency program. *Acad Med*. 2015;90(7):900–5. https://doi.org/10.1097/ACM.0000000000000707
10. Chan TM, Sherbino J, Mercuri M. Nuance and noise: lessons learned from longitudinal aggregated assessment data. *JGME*. 2017;9(6):724–9. https://doi.org/10.4300/JGME-D-17-00086.1
11. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *TLM*. 2003;15(4):270–92. https://doi.org/10.1207/S15328015TLM1504_11
12. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–7. https://doi.org/10.1046/j.1365-2923.2003.01594.x
13. 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 Oct;87(10):1401–7. https://doi.org/10.1097/ACM.0b013e3182677805
14. Tamblyn R, Abrahamowicz M, Dauphinee D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA*. 2007 Sep 5;298(9):993. https://doi.org/10.1001/jama.298.9.993
15. Tamblyn R. Association between licensure examination scores and practice in primary care. *JAMA*. 2002 Dec 18;288(23):3019. https://doi.org/10.1001/jama.288.23.3019
16. Norcini JJ, Lipner RS, Kimball HR. Certifying examination performance and patient outcomes following acute myocardial infarction. *Med Educ*. 2002 Sep;36(9):853–9. https://doi.org/10.1046/j.1365-2923.2002.01293.x