JOURNAL OF CME
2024, VOL. 13, 2360137
https://doi.org/10.1080/28338073.2024.2360137

EUROPEAN CME Forum
Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Assessment of Entrustable Professional Activities Among Dutch Endocrine Supervisors

Joanne M. de Laat [a,b], Anouk N.A. van der Horst-Schrivers[c], Natasha M. Appelman-Dijkstra[d], Peter H. Bisschop[e], Koen M.A. Dreijerink[f], Madeleine L. Drent[f], Melanie M. van de Klauw[g], Wendela L. de Ranitz[h], Aline M.E. Stades[h], Nike M.M.L. Stikkelbroeck[a], Henri J.L.M. Timmers[a] and Olle ten Cate[b]

[a]Department of Internal Medicine, Division of Endocrinology, Radboud University Medical Center, Nijmegen, The Netherlands; [b]Utrecht Center for Research and Development of Health Professions Education, University Medical Center Utrecht, Utrecht, The Netherlands; [c]Proteion, Nursing Home Organization, Haelen, The Netherlands; [d]Department of Internal Medicine-Endocrinology, Leiden University Medical Center, Leiden, The Netherlands; [e]Department of Endocrinology and Metabolism, Amsterdam UMC, Location Academic Medical Center, Amsterdam, The Netherlands; [f]Department of Internal Medicine, Amsterdam UMC, Location VU University Medical Center, Amsterdam, The Netherlands; [g]Department of Endocrinology, University Medical Center Groningen, Groningen, The Netherlands; [h]Department of Endocrinology, University Medical Center Utrecht, Utrecht, The Netherlands

ABSTRACT

Entrustable Professional Activities (EPAs) are an important tool to support individualisation of medical training in a competency-based setting and are increasingly implemented in the clinical speciality training for endocrinologist. This study aims to assess interrater agreement and factors that potentially impact EPA scores. Five known factors that affect entrustment decisions in health profesions training (capability, integrity, reliability, humility, agency) were used in this study. A case-vignette study using standardised written cases. Case vignettes ($n = 6$) on the topics thyroid disease, pituitary disease, adrenal disease, calcium and bone disorders, diabetes mellitus, and gonadal disorders were written by two endocrinologists and a medical education expert and assessed by endocrinologists experienced in the supervision of residents in training. Primary outcome is the inter-rater agreement of entrustment decisions for endocrine EPAs among raters. Secondary outcomes included the dichotomous interrater agreement (entrusted vs. non-entrusted), and an exploration of factors that impact decision-making. The study protocol was registered and approved by the Ethical Review Board of the Netherlands Association for Medical Education (NVMO-ERB # 2020.2.5). Nine endocrinologists from six different academic regions participated. Overall, the Fleiss Kappa measure of agreement for the EPA level was 0.11 (95% CI: 0.03–0.22) and for the entrustment decision 0.24 (95% CI 0.11–0.37). Of the five features that impacted the entrustment decision, capability was ranked as the most important by a majority of raters (56%–67%) in every case. There is a considerable discrepancy between the EPA levels assigned by different raters. These findings emphasise the need to base entrustment decisions on multiple observations, made by a team of supervisors and enriched with factors other than direct medical competence.

ARTICLE HISTORY
Received 4 February 2024
Revised 14 May 2024
Accepted 19 May 2024

KEYWORDS
Entrustable professional activities; entrustment decisions; decision-making factors; capability assessment; interrater agreement; endocrinology; clinical speciality training

Introduction

As entrustable professional activities (EPAs, units of practice that can be entrusted to sufficiently competent trainees) have become an increasingly prevalent focus of workplace-base assessment [1], and insight is growing into factors that determine entrustment decisions [2], there is a need to understand the sources of variation between clinical supervisors when they evaluate learners with the purpose to make entrustment decisions for EPAs, i.e. to entrust trainees with specific tasks. Most studies have been conducted in residency training and undergraduate medical education, fewer in fellowship training. We chose endocrinology fellowship as a focus. Particularly, as endocrinology trainees increasingly pursue subspecialisation and rotate through specific expertise centres, understanding the variability in EPA scores becomes even more crucial.

CONTACT Joanne M. de Laat marieke.delaat@radboudumc.nl Department of Internal Medicine, Radboud University Medical Center, Division of Endocrinology, PO Box 9101, Nijmegen 6500 HB, The Netherlands

Supplemental data for this article can be accessed online at https://doi.org/10.1080/28338073.2024.2360137

The introduction of EPAs in postgraduate training [3] has shifted the focus of assessment from observed proficiency to entrustment decision-making [4], i.e. to the question: is the trainee ready to practice the EPA unsupervised? or, in other words, from a retrospective view to a prospective view [5]. Making such an entrustment decision involves more than the observation of skill; it includes an inference about the trainee's readiness to cope with unexpected situations, which may have not been observed or maybe even never have been encountered. In a review of pertinent literature, ten Cate & Chen summarised the trainee features that clinicians value when they need to trust a learner with critical activities in five categories: *Capability* (knowledge & skill; experience; adaptive expertise); *Integrity* (truthful, good intentions, patient-centred); *Reliability* (conscientious, predictable, accountable, responsible); *Humility* (observing limits, willing to ask help, receptive to feedback) and *Agency* (self-confident, proactive towards work, team, safety, development) [6]

While there are mixed reports about the reliability of entrustment decisions [7–10], the decision to award responsibilities in patient care should theoretically lead to more careful evaluations of trainees than a score or a proficiency scale that is known for low reliability [11]. At the same time, there may be legitimate variation among experienced clinical educators when they decide a learner is ready for autonomy [12]. Some may qualify this variation as an error due to subjective bias, but that notion can be questioned [13]. Expert judgement cannot always be captured in "objective" rating forms [14] and variation, to a certain extent, may be legitimate, as experience and expertise of raters can, and arguably should weigh in [15]. Yet, for the sake of fairness towards trainees, judgements must be defensible.

The aim of the current study was to investigate to what extent clinical supervisors agree on entrustment decisions in similar cases and how they weigh the general qualities of trainees (capability, integrity, reliability, humility, agency), using standardised written cases. We hypothesise that supervisors will show reasonable, but not perfect rating agreements.

Materials and Methods

Aims

The primary outcome of this study is the interrater reliability of entrustment determinations for endocrine EPAs among raters from different academic regions based on standardised written cases. We called the ratings "entrustment determinations" rather than entrustment decisions, as those would require live situations. Interrater variability was measured using a common five-point supervision scale as currently used for EPAs in endocrinology (1. Observe the EPA, 2. Practice the EPA under direct supervision, 3. Practice the EPA under indirect supervision, 4. Practice the EPA unsupervised, i.e. under clinical oversight, 5. Provide supervision for this EPA to juniors). Levels 4 and 5 on the supervision scale are equivalent to the EPA being entrusted, and lower levels imply that supervision is still required, i.e. the EPA is not entrusted. Thus, there is an educational and clinically meaningful boundary between levels 3 and 4. Therefore we also assessed interrater variability dichotomously as entrusted (level 4 or 5) vs. not entrusted (level 1 till 3).

We also assessed the interrater reliability of entrustment determinations for endocrine EPAs among raters within the same academic region. Other secondary outcomes comprise resident, supervisor, and contextual factors that impact the interrater reliability and the prevalence of discrepancies.

Design of the Survey

This is a case-vignette study using standardised written cases. Case vignettes ($n = 6$) on the topics thyroid disease, pituitary disease, adrenal disease, calcium and bone disorders, diabetes mellitus, and gonadal disorders were written by a team of two endocrinologists and a medical education expert. It is pivotal in vignette studies that the written scenarios are realistic and believable [16]. Vignettes were drafted by an endocrinologist with experience in supervision of resient training. Prior to the drafting of the vignettes, two authors acquainted themselves with a large number of completed EPA forms, taken from daily internal medicine practices (not restricted to endocrinology), to gain inspiration for vignettes that reflect real-life situations. Draft versions of these vignettes were send for feedback to the other endocrinologist and medical eduction expert for feedback and validation. Based on this feedback, content validity support was sought for three domains: clarity, relevance and importance. In an iterative process, the draft vignettes were improved until the three content validity domains were judged satisfactory by both readers. The vignettes describe fictive cases of residents in endocrinology to standardise the case being evaluated for all raters. Using fictive cases also ensures privacy and avoids impact on entrustment decisions of current residents. For each endocrine EPA ($n = 6$), one case was presented.

These EPAs had been defined for the curriculum developed by the Dutch society for Internal Medicine [17]. The vignettes and survey can be found in Appendix A (translated from Dutch).

Each vignette provided a fictitious portfolio including:

- Bio details of the fictitious resident (i.e. sex, age, part-time factor, and speciality training schedule)
- Number of cases seen and procedures performed related to the specified EPA
- Short resume of the resident's mini clinical evaluation exercise (Mini-CEX) results for patient consultations, case-based discussions, and multidisciplinary conferences
- Short resume of multi-source feedback on the resident's general performance
- Any scientific and educational activities of the resident

Raters filled out a survey for each vignette with open and closed-format questions via an electronic questionnaire, including:

- Baseline variables include educational position (program director or member of supervisory team), and years of experience with supervising residents.
- What entrustment level do you suggest to grant this fictitious resident? (Ordinal scale ranging from 1 to 5)
- Which trainee features contribute the most to entrustment determination? For this question five potential features (capability, integrity, reliability, humility, and agency) are sorted from most important to least important. This five feature-framework, known as A RICH, has previously been described by ten Cate & Chen [6].
- How confident are you with the entrustment level assigned? (Ordinal, 5-point scale, as previously described) [18,19]
- Do you have any further remarks relevant to the entrustment determination? (open question)

Distribution of the Survey

Potential raters for this study were endocrinologists, currently supervising endocrinology residents, for a period of at least twelve months. Raters were approached through the endocrine section of the Association for Internal Medicine, in which the program directors from each region take part.

All raters independently filled out a questionnaire describing entrustment determinations and the features that motivated their decision. From regions with two or more responding raters, both internal and external interrater agreement could be assessed. We hypothesise that ratings from within the same region are more similar, resulting in higher interrater agreement.

Informed consent was obtained from each participating rater. Upon providing informed consent, raters received the vignettes and survey digitally through secure online research software optimised for medical research (CastorEDC, Amsterdam, the Netherlands). The estimated time per case was 10–15 minutes. It was possible to save this questionnaire and continue it at a later moment. Raters received a follow-up email containing data about their progress and a reminder after 2 weeks to fill out the remaining cases, and 4 weeks after the original invitation. Personal contact with raters was established if the questionnaire has not been completed after 6 weeks.

Statistical Analysis

All data were anonymised prior to analysis. All regions and raters were coded. Several statistical test are available to measure agreement among raters. Fleiss' kappa is designed to evaluate categorical ratings with multiple raters, which aligns closely with the structured rating approach employed in our vignette assessments. Therefore we used this measure to evaluate inter-rater agreement.

Separate kappas were calculated for the interrater agreement within academic regions (internal), and across the regions (external). Secondary parameters, such as factors indicated to contribute to the entrustment determination and baseline variables are reported using descriptives such as frequencies. Missing data were excluded for each analysis.

The impact of rater characteristics on agreement was explored using descriptive statistics because the small numbers of data did not allow for formal testing of correlations.

Ethical Aspects

The study protocol was registered and approved by the Ethical Review Board of the Netherlands Association for Medical Education (NVMO-ERB # 2020.2.5).

Table 1. Results of EPA levels and entrustment decisions per case-vignette.

	2	3	4	5	Entrusted
Case 1	0 (0%)	5 (56%)	4 (44%)	0 (0%)	4 (44%)
Case 2	0 (0%)	4 (44%)	4 (44%)	1 (11%)	5 (56%)
Case 3	0 (0%)	1 (11%)	6 (67%)	2 (22%)	8 (89%)
Case 4	1 (11%)	8 (89%)	0 (0%)	0 (0)%	0 (0%)
Case 5	0 (0%)	2 (22%)	6 (67%)	1 (11%)	7 (78%)
Case 6	0 (0%)	5 (56%)	4 (44%)	0 (0%)	4 (44%)

Results

The case-vignette study was completed by nine endocrinologists from six different academic regions, out of twelve endocrinologists from eight regions that were invited. The response rate at both individual and regional levels was 75%. From three regions, the results of two raters were available, allowing analysis of variety between supervisors in the same region. There was minimal missing data, on only one question (rank potential features that could impact the entrustment decision). Data were complete for all raters on all other questions.

The EPA supervision level scores and entrustment decisions per case vignette are listed in Table 1. The most common EPA supervision levels recommended were 3 and 4. Occasionally level 5 and level 2 were assigned. For one case (#4) all raters agreed that the level of entrustment (level 4 or 5) had not been reached. In all other cases, there was variation regarding the entrustment determination, with a decision of entrustment being assigned by 44 (89% of raters per case). Overall, Fleiss Kappa for the EPA level was 0.11 (95% CI: 0.03–0.22) and for the entrustment determination (levels 4 and 5 combined) 0.24 (95% CI 0.11–0.37).

In general, raters were quite confident with the scores they assigned based on the case descriptions. In one case all raters scored a 4 or higher on the 5-point Likert scale measuring confidence in the EPA level assigned, in 2 cases 89% scored 4 or higher, and in three cases 67% scored a 4 or higher. When raters were less confident in a particular case a score of 3 (neutral) was used; scores associated with lack of confidence [1] and [2] on the Likert scale were never used. Confidence was not associated with agreement on the EPA level or entrustment determination (Figure 1).

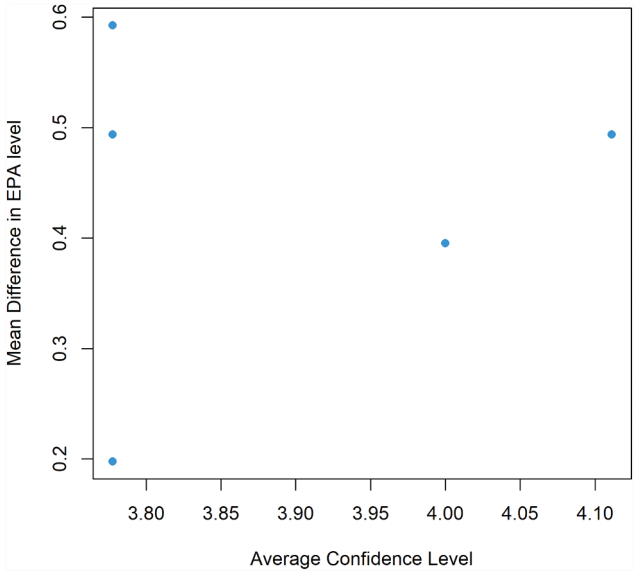

Figure 1. Relationship between the average rater confidence level about a trainee and the mean difference in recommended supervision levels between raters.

In all raters, EPA scores varied among cases. Although numbers were small, gazing at plots, we found no trend in rater characteristics such as years of experience, and educational position.

Of the five factors that impacted the entrustment decision, the capability was ranked as the most important by a majority of raters (56%–67%) in every case (Table 2). All other factors (integrity, reliability, humility and agency) scored roughly equally throughout the different cases and were typically selected as the most important factor by one rater for each case.

Subgroups of Raters Within the Same Region

For 3 academic regions in the Netherlands, the results of two raters each per region were available. In one of these

Table 2. Most important factor for determining EPA level per case.

	capability	integrity	reliability	humility	agency
Case 1	5 (56%)	1 (11%)	2 (22%)	1 (11%)	0 (0%)
Case 2	5 (56%)	0 (0%)	1 (11%)	1 (11%)	2 (22%)
Case 3	6 (67%)	1 (11%)	0 (0%)	1 (11%)	1 (11%)
Case 4	5 (56%)	1 (11%)	1 (11%)	1 (11%)	0 (0%)
Case 5	6 (67%)	1 (11%)	1 (11%)	0 (0%)	1 (11%)
Case 6	6 (67%)	1 (11%)	1 (11%)	0 (0%)	1 (11%)

centres, the interrater agreement seemed substantially higher than in the study overall, with a Fleiss Kappa of 0.67 for both the EPA level and the entrustment decision. In the other two regions, there was no agreement between the raters from that region (Fleiss Kappa 0.0) for both the EPA level and the entrustment decision.

Qualitative Assessment of Remarks

Although raters expressed a high level of confidence in the EPA level assigned based on the case-vignettes multiple raters remarked in the free text field of the survey that for a definitive judgement they would incorporate the context of results from previous rotations and the residents' overall knowledge of endocrinology. In addition to the five factors that were ranked for their impact on the assigned EPA level, raters mentioned that other aspects of professional behaviour that were found in the case also impacted their judgement on EPA level. These aspects were described as preparedness for patient consultations, communication with nurses or other staff, and knowing what to do after a medical incident. These aspects of professional behaviour seem to relate in part to the factors capability, humility and agency respectively.

Discussion

We found a rather low agreement between raters on the entrustment-supervision levels assigned to six written cases of endocrinology residents. Raters particularly diverged between levels 3 and 4 (readiness to perform the EPA with indirect supervision versus unsupervised). Even within the same academic regions we found low agreement levels. Medical skill (capability) was ranked as most important for the entrustment level, although all other factors suggested (integrity, reliability, humility and agency) were of influence.

This is an early study in its kind measuring agreement in EPA levels in standardised cases and exploring the impact of the reported A RICH factors on EPA scores. The use of standardised case vignettes provides some unique and interesting insights into the interrater agreement and the factors that impact entrustment decisions as compared to real-life assessments [9,20]. Using standardised fictive cases ensures that each rater based his judgements on the same set of information, and provides the opportunity to consciously consider the factors on which scoring is based as well as what factors might be missing to provide a more confident scoring.

Obviously, these case-vignettes also come with some considerable limitations. To make such cases as realistic as possible we also provided a brief fictive personal background including familial situation and social stressors. We included a variety of realistic and believable issues on a variety of domains in our vignettes. Some specific issues might cause more doubt on the entrustment decision than others. The prevalence of such issues might be different than in real life and increase the number of cases with large variety in the entrustment decision. Further, vignettes miss out on some contextual aspects including social behaviour, and experiences with this resident in past rotations. Written cases seemed the most appropriate method to standardise the cases. Videos would be less useful, as they only capture observations on single moments. We called the ratings "entrustment determinations" rather than entrustment decisions, as those would require observations of live situations and the option to actually decide on entrustment with a health care task [21]. Despite these limitations, raters expressed high confidence in the EPA levels they assigned, demonstrating that the case description provided the most important details to come to a justified conclusion.

Workplace-based assessment (WBA) in postgraduate medical education has received wide attention since the 1990s [15]. The introduction of the mini clinical evaluation exercise or miniCEX [16,17] was an early example, followed by tools for direct observation of technical skills [18], case-based discussions [19], 360 degree evaluation [20] and many more [22]. The worldwide movement of competency-based medical education with its focus on standardised outcomes of training [21,22], largely since the turn of the century, has not only further stimulated WBA but reinforced the need for standards of validity [23]. With the introduction of EPAs in postgraduate training [3] the focus of WBA shifted from retrospective observed proficiency to prospective entrustment decision-making.

The findings contribute to a recent research agenda for EPA development on the meso and macro levels [18].

Yet, concerns about the validity of workplace-based assessments have been frequently voiced. Much of the assessment of learners in the workplace involves single trainees, single assessors, and unique contexts (e.g. dealing with a unique patient), leading to persistent psychometric problems, such as rater leniency bias or generosity error, halo effects, restriction of range, poor discrimination between trainees, lack of documentation of deficits, low intra-rater and inter-rater and cross-occasion consistency [23–25]

In our study, we also found a low inter-rater agreement, despite the extensive shared definition process for EPAs in our curriculum plan. These results are indicative of the subjectivity in the scoring of EPA. To improve the

reliability of entrustment ratings it is needed to approach EPA as a continuous process, requiring sufficient exposure of a supervisor to a learner, building a trust relation, with regular grading of clinical activities, and evaluation of finding within the group of supervisors [26,27]. Faculty development can also help improve the reliability of entrustment-supervision ratings [28]. For summative entrustment decisions that lead to formal qualifications, team decisions have been recommended [29,30] in which all relevant sources of information are taken into consideration. Ideally descrepancies in entrustment decisions between supervisors should be resolved through discussion after considering all relevant information. Literature on team decisions in EPAs has not described how to handle persisting descrepancies (i.e. by majority of vote, or large majority).

Understanding the factors that impact variance in EPA-level ratings is important to appreciate discrepancies. As expected, we found much emphasis on capability as the most important factor in assessing the EPA level. This finding is in agreement with a recent natural language processing study of comments made during EPA assessment [31]. The study showed that entrustment levels were associated with detailed feedback on specific steps for performing a clinical task, rather than non-specific comments [31]. These findings are also in accordance with the observation that in interventional medical specialities there is increasing interest in using procedural videos and images in combination with motion analysis for the assessment of competence [32,33].

Although medical competence that could directly impact patient outcomes will always be a key component in entrustment decisions, we found that in almost half of the ratings other A RICH factors were selected as the most important factor impacting EPA scores. These factors include agency, reliability, integrity, and humility. Moreover, all these A RICH factors were selected at least once as the most important contributing factor, validating the concept of these factors [6]. Adding such factors provides richness to the entrustment decisions, although these factors can be challenging to be expressed in words [15,34].

In conclusion, there is rather a substantial discrepancy between the EPA levels assigned by different raters even within the same academic region. These findings emphasise the need to base the entrustment decision on multiple observations, made by a team of supervisors and enriched with factors other than direct medical competence.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Authors' Contributions

Dr. J.M. de Laat Study design, Writing of Case-vignets, conduct of study, analysis of results, drafting of manuscript.

Dr. A.N.A. van der Horst-Schrivers is an endocrinologist with a special interest in clinical education at Maastricht University Medical Center, Maastricht, the Netherlands

Dr. N.M. Appelman-Dijkstra Study design, Interpretation of results, critically reviewing of manuscript.

Prof. Dr. P.H. Bisschop Study design, Interpretation of results, critically reviewing of manuscript.

Dr. K.M.A. Dreijerink Study design, Interpretation of results, critically reviewing of manuscript.

Prof. Dr. M.L. Drent Study design, Interpretation of results, critically reviewing of manuscript.

Dr. M.M. van de Klauw Study design, Interpretation of results, critically reviewing of manuscript.

Dr. W.L. De Ranitz Study design, Interpretation of results, critically reviewing of manuscript.

Dr. A.M.E. Stades Study design, Interpretation of results, critically reviewing of manuscript.

Dr. N.M.M.L. Stikkelbroeck Study design, Interpretation of results, critically reviewing of manuscript.

Prof. Dr. H.J.L.M. Timmers Study design, Interpretation of results, critically reviewing of manuscript.

Dr. O. ten Cate, PhD Study design, Reviewing of case vignets, supervision of analysis, Interpretation of results, critically reviewing of manuscript.

Availability of Data and Materials

The data supporting the findings of this article are accessible through the corresponding author upon reasonable request.

Ethics Approval and Consent to Participate

The study protocol was registered and approved by the Ethical Review Board of the Netherlands Association for Medical Education (NVMO-ERB # 2020.2.5). Written Informed consent was obtained from each participating rater.

Keypoints

- Understanding the variability among raters in EPA scores and identifying its sources is crucial for evaluating the clinical competence of residents in managing endocrine diseases.
- There is considerable inter-rater variability in EPA levels assigned by different endocrinology supervisors. The discrepancies were potentially relevant to entrustment decision-making.
- High inter-rater variability was also observed among raters from the same academic region.
- Factors that might impact EPA have previously been summarised by the acronym A RICH: Agency,

Reliability, Integrity, Capability, Humility. All these factors were found to be of importance for the EPA level, although capability was the single most important factor.

- Entrustment decisions should be made by team decision, and based on multiple and enriched observations.

ORCID

Joanne M. de Laat http://orcid.org/0000-0003-3316-1540

References

[1] Chen HC, ten Cate O. Assessment through entrustable professional activities. In: Delany C, editor. Learning & teaching in clinical contexts: a practical guide. Chatswood: Elsevier Australia; 2018. p. 286–304.

[2] Hauer KE, Ten Cate O, Boscardin C, et al. Understanding trust as an essential element of trainee supervision and learning in the workplace. Adv Health Sci Educ Theory Pract. 2014;19(3):435–456. doi: 10.1007/s10459-013-9474-4

[3] ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med. 2007;82(6):542–547. doi: 10.1097/ACM.0b013e31805559c7

[4] Ten Cate O, Hart D, Ankel F, et al. Entrustment decision making in clinical training. Acad Med. 2016;91(2):191–198. doi: 10.1097/ACM.0000000000001044

[5] Ten Cate O, Carraccio C, Damodaran A, et al. Entrustment decision making: extending Miller's pyramid. Acad Med. 2021;96(2):199–204. doi: 10.1097/ACM.0000000000003800

[6] Ten Cate O, Chen HC. The ingredients of a rich entrustment decision. Med Teach. 2020;42(12):1413–1420. doi: 10.1080/0142159X.2020.1817348

[7] Eltayar AN, Aref SR, Khalifa HM, et al. Do entrustment scales make a difference in the inter-rater reliability of the workplace-based assessment? Med Educ Online. 2022;27(1):2053401. doi: 10.1080/10872981.2022.2053401

[8] Weller JM, Castanelli DJ, Chen Y, et al. Making robust assessments of specialist trainees' workplace performance. Br J Anaesth. 2017;118(2):207–214. doi: 10.1093/bja/aew412

[9] 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(2):283–291. doi: 10.1097/ACM.0000000000001820

[10] Kelleher M, Kinnear B, Sall D, et al. A reliability analysis of entrustment-derived workplace-based assessments. Acad Med. 2020;95(4):616–622. doi: 10.1097/ACM.0000000000002997

[11] Moonen-van Loon JM, Overeem K, Donkers HH, et al. Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. Adv Health Sci Educ Theory Pract. 2013;18(5):1087–1102. doi: 10.1007/s10459-013-9450-z

[12] Sterkenburg A, Barach P, Kalkman C, et al. When do supervising physicians decide to entrust residents with unsupervised tasks? Acad Med. 2010;85(9):1408–1417. doi: 10.1097/ACM.0b013e3181eab0ec

[13] ten Cate O. Bias or legitimate subjectivity in entrustment? 2022. Available from: https://icenetblog.royalcollege.ca/2022/05/12/bias-or-legitimate-subjectivity-in-entrustment/

[14] van Enk A, Ten Cate O. "Languaging" tacit judgment in formal postgraduate assessment: the documentation of ad hoc and summative entrustment decisions. Perspect Med Educ. 2020;9(6):373–378. doi: 10.1007/S40037-020-00616-X

[15] Ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. Acad Med. 2019;94(3):333–337. doi: 10.1097/ACM.0000000000002495

[16] Hughes RH, Huby M. The construction and interpretation of vignettes in social research. Soc Work & Soc Sci Rev—An Inte J Appl Res. 2004;11(1):36–51. doi: 10.1921/17466105.11.1.36

[17] [medicine] NIVDsoi. Landelijk opleidingsplan Interne geneeskunde 2019. 2019.

[18] Ten Cate O, Balmer DF, Caretta-Weyer H, et al. Entrustable professional activities and entrustment decision making: a development and research agenda for the next decade. Acad Med. 2021;96(7S):S96–S104. doi: 10.1097/ACM.0000000000004106

[19] Ten Cate O, Schwartz A, Chen HC. Assessing trainees and making entrustment decisions: on the nature and use of entrustment-supervision scales. Acad Med. 2020;95(11):1662–1669. doi: 10.1097/ACM.0000000000003427

[20] Ryan MS, Richards A, Perera R, et al. Generalizability of the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE) scale to assess medical student performance on core EPAs in the workplace: findings from one institution. Acad Med. 2021;96(8):1197–1204. doi: 10.1097/ACM.0000000000003921

[21] Ten Cate O, Jarrett JB. Would I trust or will i trust? The gap between entrustment determinations and entrustment decisions for trainees in pharmacy and other health professions. Pharmacy (Basel). 2023;11(3):107. doi: 10.3390/pharmacy11030107

[22] Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 2009;302(12):1316–1326. doi: 10.1001/jama.2009.1365

[23] Govaerts MJ, van der Vleuten CP, Schuwirth LW, et al. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. Adv Health Sci Educ Theory Pract. 2007;12(2):239–260. doi: 10.1007/s10459-006-9043-1

[24] Albanese MA. Challenges in using rater judgements in medical education. J Eval Clin Pract. 2000;6(3):305–319. doi: 10.1046/j.1365-2753.2000.00253.x

[25] Massie J, Ali JM. Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. Adv Health Sci Educ Theory Pract. 2016;21(2):455–473. doi: 10.1007/s10459-015-9614-0

[26] Hirsh DA, Holmboe ES, ten Cate O. Time to trust: longitudinal integrated clerkships and entrustable professional activities. Acad Med. 2014;89(2):201–204. doi: 10.1097/ACM.0000000000000111

[27] Bonnie LHA, Visser MRM, Kramer AWM, et al. Insight in the development of the mutual trust relationship between trainers and trainees in a workplace-based postgraduate medical training programme: a focus group study among trainers and trainees of the Dutch general practice training programme. BMJ Open. 2020;10(4):e036593. doi: 10.1136/bmjopen-2019-036593

[28] Kogan JR, Dine CJ, Conforti LN, et al. Can rater training improve the quality and accuracy of workplace-based assessment narrative comments and entrustment ratings? a randomized controlled trial. Acad Med. 2022;98(2):237–247. doi: 10.1097/ACM.0000000000004819

[29] Touchie C, Kinnear B, Schumacher D, et al. On the validity of summative entrustment decisions. Med Teach. 2021;43(7):780–787. doi: 10.1080/0142159X.2021.1925642

[30] Ekpenyong A, Padmore JS, Hauer KE. The purpose, structure, and process of clinical competency committees: guidance for members and program directors. J Grad Med Educ. 2021;13(2 Suppl):45–50. doi: 10.4300/JGME-D-20-00841.1

[31] Gin BC, Ten Cate O, O'Sullivan PS, et al. Exploring how feedback reflects entrustment decisions using artificial intelligence. Med Educ. 2022;56(3):303–311. doi: 10.1111/medu.14696

[32] Mason JD, Ansell J, Warren N, et al. Is motion analysis a valid tool for assessing laparoscopic skill? Surg Endosc. 2013;27(5):1468–1477. doi: 10.1007/s00464-012-2631-7

[33] Sanchez-Margallo JA, Sanchez-Margallo FM, Oropesa I, et al. Objective assessment based on motion-related metrics and technical performance in laparoscopic suturing. Int J Comput Assist Radiol Surg. 2017;12(2):307–314. doi: 10.1007/s11548-016-1459-3

[34] Gingerich A, Sebok-Syer SS, Larstone R, et al. Seeing but not believing: Insights into the intractability of failure to fail. Med Educ. 2020;54(12):1148–1158. doi: 10.1111/medu.14271