CHAPTER 5

Validity theory applied to entrustment as an approach to assessment

Claire Touchie, Olle ten Cate, Yoon Soo Park,
Benjamin Kinnear, David R. Taylor

Abstract

In adopting entrustment-based assessments, the construct has shifted from assessing learners' capability to provide competent care to their readiness for the responsibility for the welfare of patients and permission to perform clinical care with appropriate autonomy. Competence committees charged with making entrustment-based decisions must make decisions that are valid, fit for purpose, and interpreted appropriately. However, entrustment as a construct is complex and warrants a discussion regarding its relation to validity.

While many different validity questions may be asked in the context of entrustable professional activities (EPAs), this chapter focuses on what we believe is the most salient and novel feature of EPA-based programs, which is the introduction of entrustment decision-making as an approach to assessment of health professionals in training. Validity theory, with reference to the models of Messick and Kane, is discussed in the context of entrustment. This leads to reflections on how some assumptions regarding validity may need to be reconceptualized, how sources of evidence and validity arguments can support defensible decisions, and how threats to validity must be considered and minimized.

How to cite this book chapter:
Touchie C, ten Cate O, Park YS, Kinnear B, Taylor D. Validity theory applied to entrustment as an approach to assessment. In: ten Cate O, Burch VC, Chen HC, Chou FC, Hennus MP. (Eds). *Entrustable Professional Activities and Entrustment Decision-Making in Health Professions Education*, Chapter 5, pp. 51–63. [2024] London: Ubiquity Press. DOI: https://doi.org/10.5334/bdc.e

This chapter uses cross-references to other chapters of the same book. For those who read this chapter as a standalone publication: all cross-references can be found at: https://doi.org/10.5334/bdc

Authors

- Claire Touchie, MD, MHPE. University of Ottawa, Ottawa, Canada. Correspondence: ctouchie@toh.ca.
- Olle ten Cate, PhD. University Medical Center Utrecht, the Netherlands.
- Yoon Soo Park, PhD. University of Illinois College of Medicine, Chicago, Illinois, USA.
- Benjamin Kinnear, MD, MEd. Cincinnati Children's Hospital Medical Center/University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.
- David R. Taylor, MD, MHPE. Queen's University, Kingston, Canada.

Introduction

The emergence of EPAs and entrustment-based decisions in the context of competency-based education has led to questions of validity.[1,2] Many schools and programs have legitimate questions: Is the effort to change a program or improve assessment of trainees worth the investment? Will the change lead to better programs, better doctors, or safer patient care? As with other major shifts in educational approaches (e.g., problem-based learning), it is imperative that we provide validity evidence that these new approaches are fit for purpose. By this we mean the extent to which an educational and/or assessment approach fulfills its purpose or its function.[3]

Box 5.1 addresses some 'fit for purpose' validity questions relevant to EPAs. This chapter will focus on what we believe is the most salient and novel feature of EPA-based programs: the introduction of *entrustment decision-making* as an approach to assessment of health professionals in training. We will address how we think validity theory can be applied to this approach.[4]

Entrustment

Entrustment in health professions education involves confiding to a trainee the care of an individual or the execution of a task.[4] Entrustment happens when trainees are asked to look after a patient or perform tasks without direct supervision. Entrustment decisions can be made *in the moment*, when

Box 5.1: Examples of 'fit for purpose' validity questions around EPAs and entrustment.

Examples of 'fit for purpose' validity questions	Possible translations to operational questions	Examples of studies
How valid is this EPA?	Does this particular EPA reflect a relevant task? Can trainee readiness be measured?	Undergraduate medical education (UME) core EPA 5[5]
How valid is this EPA framework?	Does the framework of EPAs cover the breadth of activities in this profession? Is the framework workable in practice? Do these EPAs meet the expectations of employers or follow training?	EPAs in general surgery in the US,[6] pharmacy,[7] family medicine,[8] medical radiation technologists[9]
How valid is the entrustment-based discussion (EBD)?	Does the EBD increase a supervisor's insight into the readiness of the trainee for increased risks, compared to an alternative workplace-based assessment?	The procedure has been argued[10] but the validity question not investigated
How valid is the implementation of entrustment decision-making?	Do trainees qualified to be ready for distant supervision for an EPA actually receive the ensuing responsibility?	A survey-based study in dermatology addressed this[11]
How valid are entrustment-supervision (ES) scales to measure growth?	Do trainees with more experience require less supervision (or score better) on entrustment/supervision scales?	ES scales in anesthesia,[12,13] surgery,[14,15] pediatrics,[16,17,18] nursing,[19] internal medicine,[20,21] emergency medicine,[22] and UME[23] programs
How valid are ES scales compared to other measures?	Do scores on different scales to measure growth correlate with other scales?	ES scales in UME compared[24,25] ES scales versus milestone scales[26]
How do valid entrustment decisions come about?	Which trainee attributes account for the validity of entrustment decisions?	Supervisors'[27] or program directors'[28] opinions[29]

a trainee is asked to take over the care for a patient (which is ad hoc entrustment). The implicit assessment (i.e., observation + judgment ± feedback) of a trainee's readiness at the point of care (POC) is intended to direct learning and progression and to provide feedback to enhance growth as an emerging professional. These POC assessments are meant to be low in stakes, to be formative in purpose, and, on their own, not to be used to make promotion or credentialling decisions. However, such frontline assessments can be documented and integrated with other data points from different approaches to make a holistic, higher-stakes, summative decision about a trainee's capacity and permission to engage in patient care under less supervision. These summative entrustment decisions bring inherent consequences for both trainees and patients. Ensuring the validity of these entrustment decisions is a key step in incorporating them into an assessment strategy.[30]

Validity and entrustment

Validity in education refers to 'the degree to which evidence and theory support the interpretations and uses of scores of an assessment or test.'[31] The proposed interpretation of an assessment includes specifying the construct that is intended to be measured. In adopting entrustment, the construct has shifted from assessing trainees' *capability* to provide competent care to their readiness to be entrusted with the *responsibility* for the welfare of patients when performing an EPA with less or no supervision. Entrustment is a much more complex construct than capability; it requires additional consideration of other trainee qualities (e.g., conscientiousness, integrity, humility) as well as trainee-independent factors (e.g., patient acuity and complexity, and supervisor propensity to trust trainees); see also Chapter 4.[29,32] While entrustment is more meaningful for the purpose of making decisions to award clinical responsibility and autonomy, its complexity poses challenges from a construct validity perspective.

In addition, in gathering validity evidence, there is often reference to the objectivity of assessments. The search for objectivity (or measurement precision) in workplace-based assessment (WBA) has been pervasive; the lack of objectivity has often been framed as a lack of validity evidence for the use of competency-based frameworks in assessment, including that of EPAs.[2,33,34] However, the perceived necessity of objectivity in WBA has been challenged.[35,36,37] ten Cate and Regehr propose the concept of 'shared subjectivity,' where there is convergence of socially constructed perspectives rather than a focus on objectivity.[38] Constructing assessment approaches in health care often relies on consensus in the choice of test items, in standard setting, in the use of assessment tools, and similarly in judgments about trainee proficiency. Acknowledging that (a) expert judgment is indispensable and (b) experts differ in their unique and subjective judgments, subjectivity and its contribution to the variability of measurement should not be qualified as unwanted error.

On the contrary, using various perspectives to arrive at a coherent 'rich picture' through consensus rather than assuming a 'single truth' implies accepting, or even embracing, subjectivity or what could be called 'relevant variance.'[35,39,40,41] Nonetheless, in order to support the purpose of assessment, validity evidence must be gathered to support or refute the interpretation of whether an educator considers a trainee trustworthy for a clinical task and caring for patients. Readiness of trainees for unsupervised practice after training is a concern voiced in the literature and the importance of the validity of decisions to grant permission to act without supervision cannot be stressed enough.[42,43]

Understanding validity in the context of entrustment decision-making: Messick's and Kane's frameworks

Two dominant validity frameworks have been applied in health professions education[30,44]: Messick's sources of validity evidence and Kane's argument-based approach.[45,46] In Messick's approach, multiple sources of evidence are gathered to support the interpretations and uses of assessment data. These include evidence based on (a) test content (what construct is being assessed?); (b) response processes

(how do assessors or respondents operationalize the assessment?); (c) internal structure (are the tools or items together coherently measuring the intended construct?); (d) relations to other variables (does any other triangulating information support [or not] the interpretation?); and (e) consequences of testing (is there evidence that the intended and unintended impacts of assessment decisions are acceptable?). Table 5.1 translates Messick's sources of validity evidence when using EPAs as WBAs.

Table 5.1: Questions to guide the acquisition of Messick's five sources of validity evidence.

Sources	Questions to ask	
	Individual raters	**Competency committees**
Content	Was the right activity observed or discussed and assessed? Was it a sound representation of the intended EPA?	Were all aspects of the EPA sufficiently represented in the various observations and discussions?
Response process	Did the assessor understand what to observe and how to complete the rating tool? What was taken into account when making the entrustment decision?	Have all committee members been trained to use the data? Did they review and understand the available information about the trainee? Has the assessor thought about the perspective they bring while assessing the trainee?
Internal structure	Is the entrustment decision supported by the information provided on the rating tool? Does the rating tool provide sufficient information to provide meaningful feedback on readiness for entrustment?	Were multiple different observers involved in assessing the trainee? How did the judgments converge?
Relations to other variables	Are there sources of evidence that support (or contradict) the available information?	How do the outcomes of EPA entrustment decisions compare to other assessments the trainee completed?
Consequences	Did the observer follow up after their recommendation? Are there unintended consequences of the decision? What is the impact of trainees identified as entrustable versus trainees that still need additional training and remediation?	Does the committee keep track of decisions and trainee action to justify their decisions? Were the decisions fair? Is there evidence of bias or equity concerns (e.g., gender, race)?

Kane uses an argument-based approach to validation whereby evidence is prioritized and used to support or refute a chain of inferences connecting the moment of assessment to the resulting decision or use from the assessment or, in this case, an entrustment decision. Evidence is collected to support multiple different types of inferences: scoring, generalization, extrapolation, and implication (Figure 5.1).

Inferences in validity arguments are claims drawn from available information.[47] The information is there to support the entrustment decisions. In Kane's model, each sequential inference requires additional information in support of those claims, thus making the argument in support of the decision.

Both validity frameworks can support each other with the information gathered through Messick's sources of evidence supporting Kane's inferences, as depicted in Table 5.2.

Most traditional assessments in education use supporting evidence for scoring and generalization inferences, leading to judgments about knowledge and skill, and to decisions about student progress, often as passing or failing tests and receiving grades. Summative entrustment decisions bring deliberate operational decisions that affect patient care, and are thus related to *consequences* in Messick's model and Kane's *implication inference*. In the latter model, extrapolation to determine the readiness for entrustment and autonomy is the step made in many programs using EPAs: a decision that reflects trust in the trainee, or an 'entrustment determination.' However, that is not the ultimate step. The proof of the pudding is the actual summative entrustment, reflecting the willingness to schedule a trainee for lesser supervision or unsupervised clinical service. This

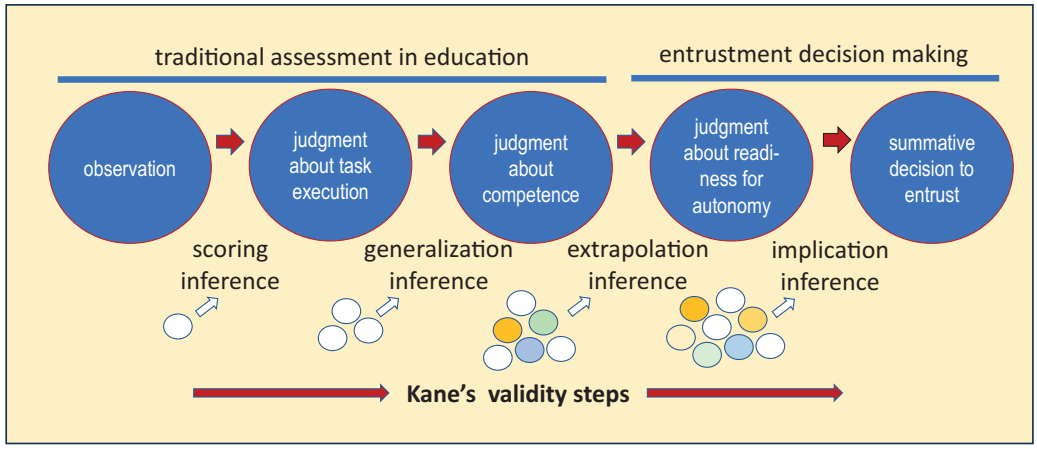

Figure 5.1: Entrustment decisions using Kane's validity argument.

step has implications for the trainee in the philosophy of EPA-based education, emphasizing assessment directing progressive autonomy and entrustment decisions. It also has implications for the patient, because of the direct relationship to patient care responsibilities. In reality, at least to date, rules and regulations often restrict true entrustment, which may explain why implication inferences and consequences evidence are not yet commonly reported.[11,48,49,50]

Table 5.2: Blending Messick's and Kane's validity models.

		Kane's validity inferences			
		Scoring	**Generalization**	**Extrapolation**	**Implication**
Messick's sources of validity evidence	Content	•	•	•	
	Response process	•			
	Internal structure		•	•	
	Relation to other variables			•	•
	Consequences				•

Defensible summative entrustment decisions

Optimizing entrustment decisions relies on sampling and gathering the right information. This may sound simple but it is not. Entrustment-based WBAs occur in authentic clinical environments. The clinical workplace is a complex, adaptive environment with many variables that cannot be controlled or standardized for the purpose of trainee assessment. Assessments thus vary across multiple facets—the time when administered, the patient or case of interest, the assessor in charge, and more. At the very least, it is much less prescribed and controlled than either written or simulated assessments. Thus the validity of entrustment within the context of WBAs warrants further exploration.

Sampling

With most assessments using written or simulated settings, trainees are assessed on a standardized set of items or scenarios representing a sampling of the universe of possible items. This sample

is representative, one of convenience, and similar for a group of individuals being assessed in standardized conditions. Items are scored in a prescribed fashion and the data are analyzed and scrutinized for reliability using tried and true approaches. Quantitative methods are available to describe the level of validity evidence at every step of either or both of the Messick and Kane approaches.

In general, EPA-based data collection has been accepted to be a convenience sample of trainee performance. EPA assessments are initiated opportunistically within the daily clinical workflow. We assume the convenience sample is representative of the trainee's larger body of work. Recent research is calling this into question.[51,52,53] First, the sampling approach used can appear to be purposive rather than for convenience, generating assessment data that intentionally select certain observations and are, therefore, not producing a representative dataset. Second, the purposes for initiating the assessment of a particular encounter are highly variable depending on the motivation of the person making the initiation decision, potentially leading to bias or underrepresentation.

To avoid bias and underrepresentation in sampling, it is important for programs to have a clear blueprint for the sampling expected. It may be helpful to also gather information on context such as the complexity of the patient or case to better understand the sample upon which an entrustment decision is being made. Finally, bringing different assessments together (e.g., EPA observations, case-based discussions, multisource feedback, product evaluation) in a trainee portfolio can then be used for summative entrustment decisions.

Mitigating threats to validity

Threats to validity occur when the assessment measures something other than what is intended. Two different categories that threaten validity are (a) construct underrepresentation and (b) construct-irrelevant variance. Construct underrepresentation (CU) occurs when the assessment does not fully represent the construct intended. For example, if the construct is the care of an adult population and the trainee has only been assessed with male patients, then there is the underrepresentation (or, in this case, no representation) of female patients. Construct-irrelevant variance (CIV) is a systematic error whereby the assessment scores are affected by variables that are extraneous to the assessment's intended purpose.[31,54] CU and CIV can affect the validity argument put forth for decision-making. If significant enough, these can negatively impact decisions and refute the argument. Not attending to these can impact patient and trainee safety. Table 5.3 provides examples of threats to validity and measures to mitigate them.

Reconceptualizing reliability

Bringing together the different assessments from a trainee's portfolio is necessary to make holistic decisions. Based on this data, competence committees (CCs) consider whether the trainee is ready to act with less supervision. In order for decisions to be robust and reproducible, clear specifications about which assessments will be included and how the data will be interpreted and used should be clearly defined.[55,56]

Establishing reliability for trainee assessments requires demonstrating the reproducibility of ratings across multiple assessment occasions.[57] The greater the extent to which assessment ratings are dependent on factors external to the trainee, the more challenging it is to establish this reproducibility. Entrustment intentionally incorporates factors outside of the control of trainees, such as an authentic clinical setting and varying patient acuity/complexity, into the rating construct itself. In addition, the concept of reproducibility is problematic as individual observations are usually followed by feedback to improve performance next time, changing the conditions for reproduction. Thus, when looking at entrustment of a trainee for an EPA over time, we are

Table 5.3: Threats to validity and examples, related to Messick's sources of validity evidence.

Sources of validity evidence	Threats to validity	Measures to consider
Sampling/content	• Observed cases have been relatively simple • Too many favorable observation moments chosen by trainees	• Include 'case complexity' scores in observation ratings • Include unannounced observations
	• Trainees lack critical experiences in patient care	• Including logs of patient encounters in portfolio to evaluate experience • Carefully designing schedules and rotational experiences • Including entrustment-based discussions (with what-if probes)
Response process	• Benefit-of-the-doubt ratings given	• Faculty development and frame-of-reference training • Forcing raters to think prospectively (will you trust your next patient with this trainee?)
	• Trainee adjusts behavior, aware of observer present	• Weighing longitudinal (MSF) information more heavily
	• CC members have not absorbed relevant trainee data	• Require preparation for CC meetings • Present aggregated trainee data in highly digestible (visual) way
Internal structure	• Contradictory data at the CC table • Insufficient variety of data available	• Discuss trainee only when sufficient data available • Explore sources of contradictions
Relationship with other variables	• Variable personal experiences of CC members with individual trainees • Circumstantial information reflecting presumptive trust diverges from observational data	• Evaluate trainee data against general framework (e.g., A RICH) • Analyze and understand external source of data
Consequences	• Incidents reported about the trainee after the summative entrustment decision	• Evaluate trainees after summative decisions • Analyze incidents to disentangle competence from unusual case complexity

looking for growth or improvement, not consistency of performance. So, for a trainee performing an individual EPA over time, using growth rate reliability and growth curve reliability might be more appropriate. Park et al. used these types of reliability calculations to estimate longitudinal consistency in milestone ratings.[58] This provided reliable longitudinal data to track individual progress in a manner that would likely be appropriate for EPAs as well.

When evaluating the reliability of a portfolio for summative decisions such as transitions to practice, other forms of reliability calculations can be considered. To improve the reliability of the decisions based on multiple assessments, composite reliability of multiple data sources or assessment systems can be considered.[59] A qualitative approach can also be used. Trustworthiness of the data and triangulation with corroboration of data across assessments can be used when viewing the entire body of data for a trainee.[60] In addition, the reliability of CC decisions could be explored through decision consistency and investigating the extent to which different CCs would make similar decisions using the same trainee data. Studies such as this have yet to be done. Regardless

of the approach used, the collective judgment made by experts in a CC with multiple data points should lead to decisions that are valid.[61,62]

Constructing an argument for defensible decisions

Variability of cases, contexts, and raters is inherent to WBAs. Sampling can be purposive but is always limited and assessments require a shared subjectivity or argued 'intersubjective judgment,' rather than proof of absolute objectivity. Decisions must be made with inherently incomplete data, and a prediction that this trainee will absolutely not make mistakes after summative entrustment is impossible. In a critical review, Kinnear et al. suggest that argumentation theory can help frame validity arguments over whether one chooses Messick's framework, Kane's, or both.[47] Those building the arguments and intended audiences need to develop a shared understanding of the validity argumentation process and its standards. Arguments should be tailored to the needs of, and clearly understood by, the audiences, be they trainees, teachers, programs, or credentialling agencies. Strength and cogency of argumentation should determine interpretations and inferences to arrive at best possible decisions.

Various examples are offered in the literature on how to construct a validity argument for decision-making. Touchie et al. discuss validity in the setting of summative decision-making using both Messick's and Kane's approaches.[30] Rotthoff et al. posit that assessments are not necessarily analytic or holistic but rather may be on a continuum.[63] Kinnear et al., in two different studies, offer a validity map also using both Messick and Kane to support decision-making in residency training and use theory to support time-variable training and decisions about readiness for practice.[64,65] Consistent across these examples is the reliance on established experts to review diverse sources of data, draw conclusions, and make summative decisions. Reliability evidence in this context argues that a separate set of experts would likely come to similar judgments on the adequacy of the data and decisions made.

Conclusions

Entrustment decision-making has implications for trainees and for patient care. Entrustment as a construct is complex and poses challenges when gathering validity evidence. It has validity implications that differ from other assessment formats. Using the validity frameworks of Messick and Kane, we can apply theory to gather the evidence necessary for the defensibility of decision-making. These provide a platform to reconceptualize assumptions underlying sampling, reliability, and decision-making and to understand how to mitigate threats to validity.

Competing interests

The authors declare that they have no competing interests.

References

1. Guralnick S, Yedowitz-Freeman J. Core entrustable professional activities for entry into residency: curricular gap or unrealistic expectations. *J Grad Med Educ*. 2017;9(5):593–594. DOI: https://doi.org/10.4300/JGME-D-17-00559.1
2. Krupat E. Critical thoughts about the core entrustable professional activities in undergraduate medical education. *Acad Med*. 2018;93(3):371–376. DOI: https://doi.org/10.1097/ACM.0000000000001865

3. Dijkstra J, Galbraith R, Hodges BD, et al. Expert validation of fit-for-purpose guidelines for redesigning programmes of assessment. *BMC Med Educ*. 2012;12:20. DOI: https://doi.org/10.1186/1472-6920-12-20
4. ten Cate O, Hart D, Ankel F, et al. Entrustment decision-making in clinical training. *Acad Med*. 2016:91(2):191–198. DOI: https://doi.org/10.1097/ACM.0000000000001044
5. Aulet TH, Moore JS, Callas PW, Nicholas C, Hulme M. (En)trust me: validating an assessment rubric for documenting clinical encounters during a surgery clerkship clinical skills exam. *Am J Surg*. 2020;219(2):258–262. DOI: https://doi.org/10.1016/j.amjsurg.2018.12.055
6. Brasel KJ, Lindeman B, Jones A, et al. Implementation of entrustable professional activities in general surgery: results of a national pilot study. *Ann Surg*. 2023;278(4):578–586. DOI: https://doi.org/10.1097/SLA.0000000000005991
7. McDowell L, Hamrick J, Fetterman J, Brooks K. Preceptors' perceptions of an entrustable professional activities-based community introductory pharmacy practice experience curriculum. *Curr Pharm Teach Learn*. 2024;16(2):109–118. DOI: https://doi.org/10.1016/j.cptl.2023.12.026
8. Newton WP, Magill M, Barr W, Hoekzema G, Karuppiah S, Stutzman K. Implementing competency based ABFM board eligibility. *J Am Board Fam Med*. 2023;36(4):703–707. DOI: https://doi.org/10.3122/jabfm.2023.230201R0
9. Tu CY, Huang KM, Cheng CH, Lin WJ, Liu CH, Yang CW. Development, implementation, and evaluation of entrustable professional activities (EPAs) for medical radiation technologists in Taiwan: a nationwide experience. *BMC Med Educ*. 2024;24:95. DOI: https://doi.org/10.1186/s12909-024-05088-9
10. ten Cate O, Hoff RG. From case-based to entrustment-based discussions. *Clin Teach*. 2017;14(6):385–389. DOI: https://doi.org/10.1111/tct.12710
11. Sigurdsson V, ten Cate O. Do summative entrustment decisions actually lead to entrustment? *Clin Teach*. October 10, 2023:e13668. DOI: https://doi.org/10.1111/tct.13668
12. Dubois DG, Lingley AJ, Ghatalia J, McConnell MM. Validity of entrustment scales within anesthesiology residency training. *Can J Anaesth*. 2021;68(1):53–63. DOI: https://doi.org/10.1007/s12630-020-01823-0
13. Weller JM, Castanelli DJ, Chen Y, Jolly B. Making robust assessments of specialist trainees' workplace performance. *Br J Anaesth*. 2017;118(2):207–214. DOI: https://doi.org/10.1093/bja/aew412
14. Sandhu G, Nikolian VC, Magas CP, et al. Optrust: validity of a tool assessing intraoperative entrustment behaviors. *Ann Surg*. 2018;267(4):670–676. DOI: https://doi.org/10.1097/SLA.0000000000002235
15. Liebert CA, Melcer EF, Keehl O, et al. Validity Evidence for ENTRUST as an assessment of surgical decision-making for the inguinal hernia entrustable professional activity (EPA). *J Surg Educ*. 2022;79(6):e202-e212. DOI: https://doi.org/10.1016/j.jsurg.2022.07.008
16. Li S, Qi X, Li H, Zhou W, Jiang Z, Qi J. Exploration of validity evidence for core residency entrustable professional activities in Chinese pediatric residency. *Front Med (Lausanne)*. 2023;10:1301356. DOI: https://doi.org/10.3389/fmed.2023.1301356
17. Pitts S, Schwartz A, Carraccio CL, et al. Fellow entrustment for the common pediatric subspecialty entrustable professional activities across subspecialties. *Acad Pediatr*. 2022;22(6):881–886. DOI: https://doi.org/10.1016/j.acap.2021.12.019
18. 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: https://doi.org/10.1097/ACM.0000000000001820
19. Lau ST, Ang E, Shorey S, Lau Y. Entrustable professional activity assessment tool for clinical procedures: a psychometric study. *J Clin Nurs*. 2021;30(19–20):2822–2831. DOI: https://doi.org/10.1111/jocn.15788

20. Colbert-Getz JM, Lappe K, Gerstenberger J, Milne CK, Raaum S. Capturing growth curves of medical students' clinical skills performance. *Clin Teach*. 2023;20(6):e13623. DOI: https://doi.org/10.1111/tct.13623
21. 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(10):1398–1405. DOI: https://doi.org/10.1097/ACM.0000000000001292
22. Dewhirst S, Wood TJ, Cheung WJ, Frank JR. Assessing the utility of a novel entrustment-supervision assessment tool. *Med Educ*. 2023;57(10):949–957. DOI: https://doi.org/10.1111/medu.15156
23. Violato C, Cullen MJ, Englander R, et al. Validity evidence for assessing entrustable professional activities during undergraduate medical education. *Acad Med*. 2021;96(7S):S70-S75. DOI: https://doi.org/10.1097/ACM.0000000000004090
24. Ryan MS, Khan AR, Park YS, et al. Workplace-based entrustment scales for the core EPAs: a multisite comparison of validity evidence for two proposed instruments using structured vignettes and trained raters. *Acad Med*. 2022;97(4):544–551. DOI: https://doi.org/10.1097/ACM.0000000000004222
25. Ryan MS, Gielissen KA, Shin D, et al. How well do workplace-based assessments support summative entrustment decisions? A multi-institutional generalisability study. *Med Educ*. January 2, 2024. DOI: https://doi.org/10.1111/medu.15291
26. Brazelle M, Zmijewski P, McLeod C, Corey B, Porterfield JR, Lindeman B. Concurrent validity evidence for entrustable professional activities in general surgery residents. *J Am Coll Surg*. 2022;234(5):938–946. DOI: https://doi.org/10.1097/XCS.0000000000000168
27. Kennedy TJT, Regehr G, Baker GR, Lingard L. Point-of-care assessment of medical trainee competence for independent clinical work. *Acad Med*. 2008;83(10 Suppl):S89-S92. DOI: https://doi.org/10.1097/ACM.0b013e318183c8b7
28. Yoon MH, Kurzweil DM, Durning SJ, et al. It's a matter of trust: exploring the basis of program directors' decisions about whether to trust a resident to care for a loved one. *Adv Health Sci Educ Theory Pract*. 2020;25(3):691–709. DOI: https://doi.org/10.1007/s10459-019-09953-x
29. ten Cate O, Chen HC. The ingredients of a rich entrustment decision. *Med Teach*. 2020;42(12):1413–1420. DOI: https://doi.org/10.1080/0142159X.2020.1817348
30. Touchie C, Kinnear B, Schumacher D, et al. On the validity of summative entrustment decisions. *Med Teach*. 2021;43(7):780–787. DOI: https://doi.org/10.1080/0142159X.2021.1925642
31. American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA). *Standards for Educational and Psychological Testing*. American Educational Research Association; 2014.
32. Hauer K, 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*. 2014;19(3):435–456. DOI: https://doi.org/10.1007/s10459-013-9474-4
33. Lurie S. History and practice of competency-based assessment. *Med Educ*. 2012;46:49–57. DOI: https://doi.org/10.1111/j.1365-2923.2011.04142.x
34. Norman G, Norcini J, Bordage G. Competency-based education: milestones or millstones? *J Grad Med Educ*. 2014;6(1):1–16. DOI: https://doi.org/10.4300/JGME-D-13-00445.1
35. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564–568. DOI: https://doi.org/10.3109/0142159X.2013.789134
36. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068. DOI: https://doi.org/10.1111/medu.12546

37. Valentine N, Durning SJ, Shanahan EM, Schuwirth L. Fairness in assessment: identifying a complex adaptive system. *Perspect Med Educ*. 2023;12(1):315–326. DOI: https://doi.org/10.5334/pme.993
38. ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2019;94(3):333–337. DOI: https://doi.org/10.1097/ACM.0000000000002495
39. Crossley J. Validity and truth in assessment. *Med Educ*. 2013;47(12):1152–1154. DOI: https://doi.org/10.1111/medu.12317
40. Virk A, Joshi A, Mahajan R, Singh T. The power of subjectivity in competency-based assessment. *J Postgrad Med*. 2020;66(4):200–205. DOI: https://doi.org/10.4103/jpgm.JPGM_591_20
41. Tavares W, Kinnear B, Schumacher DJ, Forte M. 'Rater training' re-imagined for work-based assessment in medical education. *Adv Health Sci Educ*. 2023;28(5):1697–1709. DOI: https://doi.org/10.4103/jpgm.JPGM_591_20
42. Jonker G, Ochtman A, Marty AP, et al. Would you trust your loved ones to this trainee? Certification decisions in postgraduate anaesthesia training. *Brit J of Anaesth*. 2020;125(5):e408-e410. DOI: https://doi.org/10.1016/j.bja.2020.07.009
43. Fletcher KE, O'Connor AB, Kisielewski M, Willett LL. Why do residency program directors consider resigning? A mixed-methods analysis of a national program director survey. *Amer J Med*. 2020;133(6):761–767. DOI: https://doi.org/10.1016/j.amjmed.2020.02.016
44. Schuwirth L, van der Vleuten C. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38–48. DOI: https://doi.org/10.1097/ACM.0000000000002495
45. Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. 3rd ed. Macmillan; 1989:13–103.
46. Kane MT. Validating the interpretations and uses of test scores. *J Educ Measur*. 2013;50(1):1–73.
47. Kinnear B, Schumacher DJ, Driessen EW, Varpio L. How argumentation theory can inform assessment validity: a critical review. *Med Educ*. 2022;56:1064–1075. DOI: https://doi.org/10.1111/medu.14882
48. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Int Med*. 2005;20:1159–1164. DOI: https://doi.org/10.1111/j.1525-1497.2005.0258.x
49. 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*. 2014;19(2):233–250. DOI: https://doi.org/10.1007/s10459-013-9458-4
50. 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: https://doi.org/10.3390/pharmacy11030107
51. Teunissen PW, Stapel DA, van der Vleuten C, Scherpbier A, Boor K, Scheele F. Who wants feedback? An investigation of the variables influencing residents' feedback-seeking behavior in relation to night shifts. *Acad Med*. 2009;84(7):910–917. DOI: https://doi.org/10.1097/ACM.0b013e3181a858ad
52. Gaunt A, Patel A, Rusius V, Royle TJ, Markham DH, Pawlikowska T. 'Playing the game': how do surgical trainees seek feedback using workplace-based assessment? *Med Educ*. 2017;51(9):953–962. DOI: https://doi.org/10.1111/medu.13380
53. Gauthier S, Braund H, Dalgarno N, Taylor D. Assessment-seeking strategies: navigating the decision to initiate workplace-based assessment. *Teach Learn Med*. 2023, June 29:1–10. DOI: https://doi.org/10.1080/10401334.2023.2229803
54. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. 2004;38(3):327–333. DOI: https://doi.org/10.1046/j.1365-2923.2004.01777.x
55. Lineberry M, Park YS, Cook DA, Yudkowski R. Making the case for mastery learning assessments: key issues validation and justification. *Acad Med*. 2015;90(11):1445–1450. DOI: https://doi.org/10.1097/ACM.0000000000000860

56. Hu WCY, Dillon HCB, Wilkinson TJ. Educators as judges: applying judicial decision-making principles to high-stakes education assessment decision. *Teach Learn Med*. 2023;35(2):168–179. DOI: https://doi.org/10.1080/10401334.2022.2038176
57. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ*. 2004;38(9):1006–1012. DOI: https://doi.org/10.1111/j.1365-2929.2004.01932.x
58. Park YS, Hamstra SJ, Yamakazi K, Holmboe E. Longitudinal reliability of milestones-based learning trajectories in family medicine residents. *JAMA Network Open*. 2021;4(12):e2137179. DOI: https://doi.org/10.1001/jamanetworkopen.2021.3717
59. Park YS, Lineberry M, Hyderi A, Bordage G, Xing K, Yudkowski R. Differential weighting for sub-component measures of integrated clinical encounter scores based on the USMLE Step-2 CS examination: effects on composite score reliability and pass-fail decisions. *Acad Med*. 2016;91(10):S24-S30. DOI: https://doi.org/10.1097/ACM.0000000000001359
60. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49:560–575. DOI: https://doi.org/10.1111/medu.12678
61. Kinnear B, Warm EJ, Hauer KE. Twelve tips to maximize the value of a clinical competency committee in postgraduate medical education. *Med Teach*. 2018;40(11):1110–1115. DOI: https://doi.org/10.1080/0142159X.2018.1474191
62. Hauer KE, ten Cate O, Boscardin CK, et al. Ensuring resident competence: a narrative review of the literature on group decision-making to inform the work of clinical competence committees. *J Grad Med Educ*. 2016;8(2):156–164. DOI: https://doi.org/10.4300/JGME-D-15-00144.1
63. Rotthoff T, Kadmon M, Harendza S. It does not have to be either or! Assessing competence in medicine should be a continuum between an analytic and a holistic approach. *Adv Health Sci Educ*. 2021;26:1659–1673. DOI: https://doi.org/10.1007/s10459-021-10043-0
64. Kinnear B, Kelleher M, May B, et al. Constructing a validity map for a workplace-based assessment system: cross-walking Messick and Kane. *Acad Med*. 2021;96:S64-S69. DOI: https://doi.org/10.1097/ACM.0000000000004112
65. Kinnear B, Martini A, Varpio L, et al. How do validity experts conceptualise argumentation? It's a rhetorical question. *Med Educ*. 2024, January 18. DOI: https://doi.org/10.1111/medu.15311