Canadian Medical Education Journal
Revue canadienne de l'éducation médicale

Making judgments based on reported observations of trainee performance: A scoping review in Health Professions Education

L'évaluation de la performance de stagiaires basée sur des observations rapportées dans les programmes professionnalisants en santé : une étude de portée

Patricia Blanchette, Marie-Eve Poitras, Audrey-Ann Lefebvre and Christina St-Onge

Volume 15, Number 4, 2024

URI: https://id.erudit.org/iderudit/1113596ar
DOI: https://doi.org/10.36834/cmej.75522

See table of contents

Publisher(s)

Canadian Medical Education Journal

ISSN

1923-1202 (digital)

Explore this journal

Cite this document

Blanchette, P., Poitras, M.-E., Lefebvre, A.-A. & St-Onge, C. (2024). Making judgments based on reported observations of trainee performance: A scoping review in Health Professions Education. *Canadian Medical Education Journal / Revue canadienne de l'éducation médicale*, *15*(4), 63–75. https://doi.org/10.36834/cmej.75522

Article abstract

Background: Educators now use reported observations when assessing trainees' performance. Unfortunately, they have little information about how to design and implement assessments based on reported observations.

Objective: The purpose of this scoping review was to map the literature on the use of reported observations in judging health professions education (HPE) trainees' performances.

Methods: Arksey and O'Malley's (2005) method was used with four databases (sources: ERIC, CINAHL, MEDLINE, PsycINFO). Eligibility criteria for articles were: documents in English or French, including primary data, and initial or professional training; (2) training in an HPE program; (3) workplace-based assessment; and (4) assessment based on reported observations. The inclusion/exclusion, and data extraction steps were performed (agreement rate > 90%). We developed a data extraction grid to chart the data. Descriptive analyses were used to summarize quantitative data, and the authors conducted thematic analysis for qualitative data.

Results: Based on 36 papers and 13 consultations, the team identified six steps characterizing trainee performance assessment based on reported observations in HPE: (1) making first contact, (2) observing and documenting the trainee performance, (3) collecting and completing assessment data, (4) aggregating assessment data, (5) inferring the level of competence, and (6) documenting and communicating the decision to the stakeholders.

Discussion: The design and implementation of assessment based on reported observations is a first step towards a quality implementation by guiding educators and administrators responsible for graduating competent professionals. Future research might focus on understanding the context beyond assessor cognition to ensure the quality of meta-assessors' decisions.

Canadian Medical Education Journal

Reviews, Theoretical Papers, and Meta-Analyses

Making judgments based on reported observations of trainee performance: a scoping review in Health Professions Education

L'évaluation de la performance de stagiaires basée sur des observations rapportées dans les programmes professionnalisants en santé : une étude de portée

Patricia Blanchette,[1] Marie-Eve Poitras,[1] Audrey-Ann Lefebvre,[1] Christina St-Onge[1]

[1]University of Sherbrooke, Quebec, Canada

Correspondence to: Christina St-Onge; email: christina.st-onge@usherbrooke.ca

Published ahead of issue: Apr 22, 2024; published: Aug 30, 2024; CMEJ 2024 Available at https://doi.org/10.36834/cmej.75522

Abstract

Background: Educators now use reported observations when assessing trainees' performance. Unfortunately, they have little information about how to design and implement assessments based on reported observations.

Objective: The purpose of this scoping review was to map the literature on the use of reported observations in judging health professions education (HPE) trainees' performances.

Methods: Arksey and O'Malley's (2005) method was used with four databases (sources: ERIC, CINAHL, MEDLINE, PsycINFO). Eligibility criteria for articles were: documents in English or French, including primary data, and initial or professional training; (2) training in an HPE program; (3) workplace-based assessment; and (4) assessment based on reported observations. The inclusion/exclusion, and data extraction steps were performed (agreement rate > 90%). We developed a data extraction grid to chart the data. Descriptive analyses were used to summarize quantitative data, and the authors conducted thematic analysis for qualitative data.

Results: Based on 36 papers and 13 consultations, the team identified six steps characterizing trainee performance assessment based on reported observations in HPE: (1) making first contact, (2) observing and documenting the trainee performance, (3) collecting and completing assessment data, (4) aggregating assessment data, (5) inferring the level of competence, and (6) documenting and communicating the decision to the stakeholders.

Discussion: The design and implementation of assessment based on reported observations is a first step towards a quality implementation by guiding educators and administrators responsible for graduating competent professionals. Future research might focus on understanding the context beyond assessor cognition to ensure the quality of meta-assessors' decisions.

Résumé

Contexte : Les éducateurs utilisent désormais les observations rapportées pour évaluer la performance de leurs stagiaires. Malheureusement, ils disposent de peu d'informations sur la manière de concevoir et de mettre en œuvre des évaluations basées sur les observations rapportées. Objectif : L'objectif de cette étude de la portée des écrits était de recenser la littérature sur l'utilisation des observations rapportées lors d'évaluation de la performance de stagiaires dans les programmes professionnalisants en santé (PPS).

Méthodes : La méthode d'Arksey et O'Malley (2005) a été utilisée et quatre bases de données ont été interrogées (ERIC, CINAHL, MEDLINE, PsycINFO). Les critères d'admissibilité des articles étaient les suivants : (1) écrit en anglais ou en français ; (2) comprend des données primaires ; (3) traite de la formation initiale ou professionnelle ; (4) se situe en formation d'un PPS ; (5) traite de l'évaluation en stage ; et (6) traite de l'évaluation basée sur des observations rapportées. Les étapes d'inclusion/exclusion et d'extraction des données ont été réalisées à deux personnes (taux d'accord > 90%). Nous avons extrait les données avec une grille d'extraction des données préétablie et itérative. Des analyses quantitatives ont été menées pour résumer les données numériques et une analyse thématique pour résumer les données qualitatives.

Résultats : Sur la base de 36 articles et de 13 consultations, nous avons identifié six étapes caractérisant l'évaluation de la performances de stagiaires basée sur des observations rapportées dans les PPS : (1) établir un premier contact, (2) observer et documenter la performance du stagiaire, (3) recueillir et compléter les données d'évaluation, (4) agréger les données d'évaluation, (5) déduire le niveau de compétence, et (6) documenter et communiquer la décision aux parties prenantes.

Discussion : La conception et la mise en œuvre de l'évaluation sur la base d'observations rapportées constituent un premier pas vers la mise en œuvre d'une évaluation de qualité en guidant les éducateurs et les administrateurs responsables de la formation de professionnels compétents. Les recherches futures pourraient se concentrer sur la compréhension du contexte au-delà de la cognition de l'évaluateur afin de garantir la qualité des décisions prises par les métaévaluateurs.

Introduction

In many health professions education (HPE) reported observations inform programs, decisions about trainees' progression. These reported observations can be based on direct observation (e.g., observation of a trainee performing a clinical exam), or indirect observation (e.g., others providing information about trainees' performance, inferences made during a case-based discussion).[1] Faculty members, peers, patients, or other health professionals can contribute to the reported observations.[2]

Accreditation bodies have pushed for implementing competency-based education programs in HPE. Trainee assessment should reflect their performances during clinical education. In HPE, the trainee can be a student, clerk, or resident. A recent trend is for educators and programs to include reported observations to assess trainees' performance.[1] In this context, some supervisors observe and subsequently report their observations to assessors, which can be individuals or groups. These individuals or groups use the reported observations to make judgments about trainees' progression.[3] Unfortunately, there is little empirical evidence to guide the design and implementation of assessments based on reported observations. Explicit evidence-informed guiding principles for the design of assessments based on reported observations might contribute to higher quality implementations, contributing to the graduation of competent healthcare professionals.

One way to manage assessments based on reported observations is by creating committees to overview the assessment data analysis, such as competence committees (CC).[4] CC use reported observations comprising independent judgments to decide in a group process.[5–7] One challenge of CC using reported observations is to ensure fairness in their decision to limit legal repercussions,[8] and programs have to standardize CC processes.[9,10] HPE programs that use reported observations to determine trainee progression include, but are not limited to, medicine, nursing, and social work.[11]

Administrators and educators have implemented some strategies to manage multiple data from assessments based on reported observations. However, to our knowledge, we still know little about how these mechanisms and processes are designed and implemented to yield quality decisions. This scoping review is a first step towards structuring the use of reported observations in assessment decision processes and increasing the validity of assessment-data interpretation.[12–14] Thus, we summarize the scientific literature on assessment based on reported observations of trainee HPE performance. We conducted a scoping review, as opposed to a systematic review, to provide a comprehensive overview of what is known about the use and description of reported observations in the assessment of trainees' performance.

Methods

We conducted a scoping review mainly informed by Arksey & O'Malley's[15] work, with adjustments made based on recent methodological recommendations of Peters et al.,[16] Levac et al.,[17] and Trico et al.[18] More specifically, we have updated Arksey & O'Malley's work in identifying relevant studies,[16] selecting studies,[16-17] distributing data,[16-17] and collating results.[16,18] In addition, we were informed by Levac et al.[17] for seeking stakeholders' perspectives. This study received ethical approval from CER UQAC (Ref. No. 2019-1908/Poitras) and CER UdeS (Ref. No. 2019-204, 602.639.01).

Step 1: Identifying the research question

Our main research question was: What is known about the use and description of assessment based on reported observations of trainees' performance in health professions education (HPE)? To "identify, map, report, or discuss the characteristics" of knowledge,[16 (p2121)] our sub-questions were: (1) What are the assessment strategies used for reported observations in HPE?; and (2) How are assessments of trainee performance based on reported observations in HPE designed and implemented?

Step 2: Identifying relevant studies

The first author (PB) worked with two academic librarians to develop a search strategy[16] (see Appendix A). Team members' collective knowledge of performance assessment informed this strategy. We revised the keyword combinations and Boolean operators of this strategy iteratively to fully cover the depth and breadth of the literature about the use of reported observations for the assessment of trainees' performance in HPE. We conducted the final search on May 7, 2020, in ERIC (Education Resources Information Center); CINAHL (Cumulative Index to Nursing and Allied Health Literature); MEDLINE, and PsycINFO. Appendix 1 gives the keywords and Boolean operators used. The inclusion criteria were articles in French or English discussing (1) initial or professional training; (2) training in an HPE program; (3) workplace-based assessment; and (4) assessment based on reported observations. We included studies using

quantitative, qualitative, or mixed methods. We excluded knowledge synthesis, editorials, tables of contents, conference abstracts, and commentary papers. This was to include only primary data and avoid data redundancy.

Step 3: Study selection

Two team members (PB & A-AL) screened both titles and abstracts of references to determine their inclusion or exclusion.[16-17] They continued this process until they reached 90% agreement in their decision to include or exclude an abstract.[19] When these team members disagreed, they discussed all disagreements and consulted another team member (CS-O or/and M-EP). Once PB and A-AL achieved a 90% agreement for inclusion/exclusion (i.e., after reviewing 25% of the articles identified in the search), they divided the remaining references and screened them separately. During data charting, the identified snowball references and subsequently screened them for inclusion/exclusion.

Step 4: Data charting

First, PB and A-AL developed an extraction grid, which was subsequently reviewed and discussed with all team members informed by their experience in assessment (CS-O and M-EP). To ensure its clarity and standardized use, PB and A-AL tested the extraction grid on one article. They iteratively revised the initial extraction grid to include numerical variables (e.g., year of publication, country, trainee level, program, performance assessment tool or strategy, etc.) and qualitative data (e.g., design and implementation, facilitators of- or barriers to- the design and implementation, quality for the assessment of trainee performance based on reported observations, paper conclusions, and paper strengths and limitations).[15,17] PB and A-AL coded all papers independently in Dedoose (Dedoose, Manhattan Beach, CA, USA) and met to discuss extractions for all included papers at every 5 to 10 extracted articles.[16-17]

Step 5: Collating and reporting results

We summarized quantifiable data using frequencies when describing our archive (pool of manuscripts included).[16,18] These variables included: trainee discipline and study level, assessment strategies, and stakeholders. We used Dedoose and Microsoft® Excel for these analyses.

We summarized the qualitative data using thematic analysis.[16,17,19] We recognized the dual implication in the analysis process: deductive (theory-driven) and inductive (researcher subjectivity).[20] We followed Braun & Clarke's[21] five-step reflexive approach. First, the principal investigator (PB) deductively coded with an initial coding tree. She reread the data extracts to understand them. Second, she inductively refined the coding tree and applied it to the included manuscripts. PB used memos to document questions and code ideas to be validated with team members (CS-O & M-EP). Third, PB selected the codes concerning trainee performance assessment based on reported observations. Fourth, PB identified codes that could answer the research question and could form patterns in the data. She reviewed all the codes identified, grouped them to answer the research question, and examined all the extracts in each of the codes (internal consistency). Fifth, the team (PB, CS-O, and M-EP) defined, refined and organized the resulting themes.

Step 6: Consultation

We interviewed 13 nursing trainee assessors from two Québec (Canada) universities to see if their assessment of trainees' performance based on reported observations aligned with our findings.[17 (p4)] We interviewed nursing trainee assessors because they oversee the assessment of nursing trainees. There is a long tradition in these programs to make judgements about trainees' performances and progression based on reported observations. We believed them to be a rich source of information and complementary to the literature. To prevent biasing participant responses, we did not present the initial scoping findings. We obtained ethic approval from the institutional ethic committees.

We recruited participants via an email sent by the program coordinator of each university. The principal investigator (PB) contacted individuals interested in participating in the study. The principal investigator (PB) conducted the interviews on a web platform. The interview guide, consistently used for all participants, aimed at understanding the role and responsibilities of the evaluator and how they conduct evaluations. An outside firm transcribed audio recordings. We summarized qualitative data using thematic analysis[21] during data collection. First, PB used the findings from the scoping as an initial coding tree. She reread her interviewer journal notes and transcript to further her understanding and appropriate the data. Second, PB kept an open mind to new codes during data collection and inductively refined the coding tree. She used memos to document questions and code ideas to be validated with other team members (CS-O & M-EP). Third, PB selected the codes concerning the design and implementation of trainee performance assessment based on reported observations. Fourth, she identified codes that

could inform the design and implementation and could form data patterns. She reviewed all the codes identified and grouped them to represent the strategy involved in trainee performance assessment based on reported observations. She also examined all the extracts in each of the codes (internal consistency). Fifth, team members (PB, CS-O, and M-EP) defined, refined, and organized the themes identified. Then, they merged the results from the consultations with the findings of the scoping review. MaxQDA was used to manage the qualitative data and Microsoft® Excel to merge the data phases. We reported our findings using the PRISMA Extension for Scoping Reviews (PRISMA-ScR)[18] and followed Levac et al.[17] recommendations.

Results

Figure 1 provides the flowchart diagram describing the processes for trainees' performance assessment based on reported observations in HPE determined through this scoping review.

Descriptive results for bibliometric data

We noted that more articles addressing assessment based on reported observations in HPE (*n* = 24) were published between 2015 and 2020 (representing 67% of our corpus, see Figure 2). Most articles were published in the United States (*n* = 23; 64%), Canada (*n* = 5; 14%), and the Netherlands (*n* = 3; 8%). One study was published in each of the following countries: England, Finland, Norway, and Singapore. The articles appeared in 19 periodicals, the most frequent being *Academic Medicine* (*n* = 6; 17%), followed by the *Journal of Graduate Medical Education* (*n* = 5; 14%), *Medical Education*, and *Medical Teacher* (*n* = 3; 8%). All were articles (*n* = 35; 97%) except for one conference report. Studies used varied methods including quantitative (*n* = 13; 36%), qualitative (*n* = 7; 19%) or mixed methods (*n* = 2; 6%). In fourteen articles the authors did not mention the method used (39%).

Descriptive results for quantitative data

Trainees' discipline and study level. The studies were about trainees in medicine (*n* = 32; 89%), and nursing (*n* = 4; 11%). Trainees were in their postgraduate medical education (*n* = 22: 61%); clerkship (*n* = 6; 17%), bachelor's degree or equivalent (*n* = 3; 8%), residency and clerkship (*n* = 1; 3%), or internship (*n* = 1; 3%). Three manuscripts did not specify trainee-level.

Workplace-based assessment strategies used. We identified 20 workplace-based assessment strategies used by observers and trainees to document their observations and assessors to make decisions during clinical education. We present the frequency and percentage of use in Table 1 for each assessment strategy.

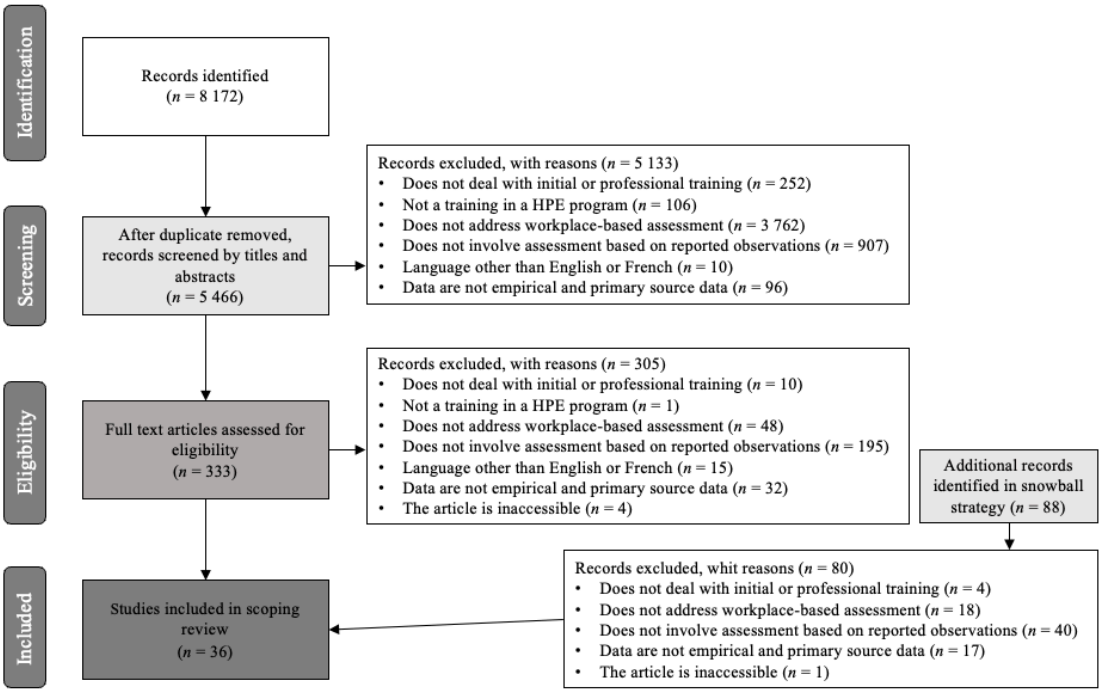

Figure 1. Flowchart diagram for 2020-21 scoping review of HPE trainee's performance assessment based on reported observations.

Table 1. Assessment strategies used in clinical education settings

Assessment Strategy	Number of times used	%
Uniform rating scale	10	28%
Milestones	10	28%
Multisource feedback (MSF)	7	19%
Rubrics	6	17%
Entrustable professional activities (EPAs)	5	14%
Narratives	5	14%
Mini-Clinical Evaluation Exercises (Mini-CEXs)	4	11%
Portfolios	2	6%
Numeric scale	2	6%
Assessment of Clinical Education (AssCE)	1	3%
Communication assessment tool	1	3%
Medical-record audit and feedback	1	3%
Anesthesiology and nontechnical skills	1	3%
Nontechnical surgical skills	1	3%
Operative performance rating system	1	3%
Set the phase, Elicit information, Give information, Understand the patient's perspective, End the encounter (SEGUE)	1	3%
Qualitative assessments	1	3%
Generic assessment	1	3%
Compliance formulary	1	3%
R-I-M-E terminology for narratives	1	3%

Stakeholders. We identified the terminology for the principal stakeholders and offer three generic terms and their respective definitions: (1) trainee, (2) observer, and (3) meta-assessors, which we use in our operationalization below (see Table 2). A trainee is an individual learning during clinical education. An observer is an individual who observes trainee performance. Given this central role in observation and documentation, consider that the observer supervises trainees. A meta-assessor is an individual or group of individuals responsible for assessing a trainee.

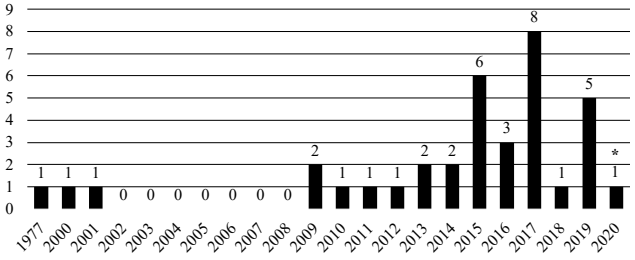

Figure 2. Article per publication year for 2020-21 scoping review of HPE trainee's performance assessment based on reported observations.

** Note: The search was conducted in May 2020, which can explain the lower number of articles in 2020.*

Design and implementation of trainee performance assessment based on reported observations. We identified six steps for trainee performance assessment based on reported observations: (1) making first contact (consultation data), (2) observing and documenting trainee performance (scoping and consultation data), (3) collecting and completing assessment data (scoping and consultation data), (4) aggregating assessment data (scoping and consultation data), (5) inferring the level of competence (scoping and consultation data), (6) documenting and communicating the decision to the stakeholders (scoping and consultation data). All except the first step were in both the scoping and consultation data. Because consultation is part of Arksey & O'Malley[15] methodology, we have combined the literature and participant data to present results that are more comprehensive.

Step 1: Making first contact. This first step allows all stakeholders (i.e., the meta-assessors, the observer, and the trainee) to discuss their respective roles and responsibilities, and their expectations about the workplace-based assessment. Considering the challenge in gathering sufficient quality data just in time, our participants believed this step promotes sound communication between these individuals. One participant highlighted how important it is for the meta-assessor and the observer to "be on the same page" (Part. No.13) regarding their views about the assessment.

Step 2: Observing and documenting trainee performance. During this step, an individual observes the trainee's performance (i.e., a peer),[22–24] a professional responsible for trainee supervision,[22–38] another professional,[22–24,27,33,39] and/or a patient.[22,23,33] Multiple observations[25,40] and observers[25,29,34,36,38–41] increase the reliability of the assessment.[25,29,32,34,36,38–41] Participants revealed assessments need to capture progression in performance. Establishing progression depends on the observations reported by both the observer and the trainee. Participants stated they encourage observers and trainees to observe and document the trainees' performance.

Participants shared that, in some situations, trainees had different assignments (e.g., ambulatory care, intensive care), limiting their opportunities to fully achieve their potential and limiting the opportunities to be observed.

Table 2. Terminology use for stakeholders involves in trainee's performance assessment based on reported observations.

	Medicine			Nursing		
	INDIVIDUALS					
Trainee	Term	Reference(s)	n	Term	Reference(s)	n
	Resident	(22-24, 26, 28, 32, 39, 41-43, 45, 46, 48, 49, 51, 53-55, 71)	19	Student	(33, 36, 38, 52)	4
	Trainee	(40, 41, 48, 50, 54, 55, 57)	7			
	Student	(25, 29, 30, 35, 37, 44)	6			
	Learner	(23, 24, 27)	3			
Observer	Faculty	(22, 24, 28, 31, 32, 45, 46, 51, 53)	9	Preceptor	(36, 38, 52)	3
	Preceptor	(26, 30)	2	Mentor	(33)	1
	House staff	(25, 48)	2			
	Faculty member	(35, 50)	2			
	Faculty attendee	(39)	1			
	Faculty rater	(50)	1			
	Faculty physician	(24)	1			
	Supervising physician	(46)	1			
	Clinical supervisor	(26)	1			
	Resident physician	(24)	1			
	Physician	(23)	1			
	Mentor	(29)	1			
	Medical educator	(40)	1			
	Supervising consultant	(41)	1			
	Master assessor	(24)	1			
	Supervisor	(31)	1			
Meta-assessor	Program director	(27, 44, 46, 50, 51, 54, 55, 71)	8	Lecturer	(52)	1
	Clerkship or assistant director	(25, 35, 37)	3	Teacher	(33)	1
	Medicine clerkship coordinator	(31)	1	Academic	(38)	1
	Faculty	(25, 44, 77)	3	Faculty	(36)	1
	Faculty attending member	(55)	1			
	Educational supervisor	(41)	1			
	Academic advisor	(28)	1			
	GROUPS OF INDIVIDUALS					
	Clinical competency committee	(22, 26, 27, 32, 39, 40, 42, 45, 46, 48, 53, 54, 71)	13			
	Competency committee	(28, 43, 49, 50, 57)	5			
	Evaluation grading committee	(30)	1			
	Assessment committee	(29)	1			
	Sub-committee	(42)	1			
	Entrustment committee	(24)	1			

The observers and trainees used different assessment forms to document their observations.[22–38,40,42–46] In addition, our participants specified they used notebooks, logbooks, and trainee self-assessment forms to document observations. The clinical environment in which the assessment occurred can facilitate (facilitator) or hinder (barrier), namely in terms of observation opportunities and length. In medicine, clinical education involves multiple observations and observers.[25] Ambulatory care, however, involves fewer observers who have less time to observe and document trainee performance.[25]

Both the observation and documentation of performance seem to be tainted by the observer's interpretation of the assessments.[33,38,40,42,47–50] For example, some observers find it difficult to understand abstract criteria on assessment forms[38,49] or misinterpret the scale used.[48,49] Our participants from Phase 6 of our scoping review highlighted similar concerns about the possible misinterpretation of "unclear" assessment-tool criteria (Part. No.03) and suggested reviewing assessment criteria with observers. A shared understanding and common vision of assessment, and support from the meta-assessors inform the observation and documentation of trainee performance.[28,38] Competent and available human resources might improve assessment feasibility.[26,32,36,43]

Step 3: Collecting and completing the assessment data. Collecting the assessment data occurs during,[26,28] at the midpoint, and/or at the end of the clinical education period.[29] The trainees,[26,28,29,34] the coordinator,[31,51] and/or the director[37] collect data using different information and communication technologies.[24,26,28,32] Participants reported trainees should provide assessment data to the meta-assessor halfway through and at the end of the clinical education period. Some studies suggest that

communication and information technology might facilitate assessment collection.[26] Electronic platforms might help in collecting assessment data generated by longitudinal assessment.[26] As they review, capture, synthesize, and present milestones, electronic systems might reduce the time competence committees spend on reviewing trainee performance.[40] Some participants in rural settings some posited that teleconferencing enhanced the communication between meta-assessor and trainee.[36] Participants identified Microsoft Teams and an on-line course management platform as facilitators in reporting collecting observations because they facilitated communication between stakeholders. Meta-assessors need enough reported observations to sketch a longitudinal picture of performance;[34,36,42,49] our participants corroborated this.

To complete the assessment data collection process, the meta-assessor organizes a discussion with the trainee and/or a group of individuals. In medicine, the program director (meta-assessor) collects the assessment data during a formal assessment session with the trainee[25,34,35,37,39,44,51] or by a group of individuals that might include the observer and/or trainee.[22,24,28,29,32,42] In nursing, the meta-assessor completes the assessment data during a formal assessment session with the observer and trainee.[33,36,38,52] To complete assessment data, participants use direct observations of trainee performance during formal assessment sessions (e.g., clinical judgment in a case-based discussion), and reported observations (e.g., reported observations by a chief nurse when visiting clinical settings). Some participants share more data in formal assessment sessions with only the trainee and/or observer. They felt that this might preserve the observer–trainee relationship. Others prefer formal assessment-group sessions to observe and manage relationship conflicts. Participants recognized that communication skills are essential in supporting trainees and in collecting sufficient quality-assessment data during the short time dedicated to formal assessment sessions.

Step 4: Aggregating assessment data. In medical programs, an individual or a group analyzes the reported observations before and/or during a formal assessment session.[34,39,44,51] Formal assessment-group sessions,[25,35,37] competence committees,[22,24,26,28–30,32,40,42,45,47,49,53,54] subcommittees[42] and/or entrustment committees[24,26,28] are groups of individuals that aggregate assessment data. Sometimes, the observer and trainee take part in that process. The aggregated assessment data comprises multiple independent judgments.[24–26,28,30,32,34,40,42,45] In nursing, only one individual revises observations during a formal assessment session composed of a meta-assessor, observer, and trainee.[28,33,36,38] To facilitate aggregation, our participants reviewed the assessment data before the formal assessment session.

Meta-assessors aggregate observations composed of different data types (e.g., numerical assessments, narrative comments) and data sources (e.g., observer and trainee). Dealing with different data types can be difficult.[28,34,42,53–55] For example, aggregating formative-assessment data to make high-stakes summative decisions is challenging;[28,40] the participants confirmed this. Aggregating data from different sources is also a challenge.[42] Participants consider reported observations from observers as more credible than information provided by trainees. High-quality assessment practices might facilitate the aggregation of data. One example would be an assessment with clear language and criteria[42,49] or rubrics[38,42,49] informed by a competency framework.[26,32,43,45,49] Narrative expressions like “solid” and “good” are examples of vague language that leads to reading between the lines and misinterpretation that are a “kind of faint praise.”[55(p299)] Participants view high-quality assessment data as being a "concrete, specific, and detailed" example of observed performances (Part. No.11) or "facts" that illustrate the target competence (Part. No.12). They also consider that consistency between the different data sources relates to assessment quality.

Step 5: Inferring the level of competence. Inferring is the action of transforming some data into affirmations.[12] This step can take different forms but comprises establishing trainee development, progression, or mastery regarding given performance criteria. We refer to it as inferring a trainee's level of competence. Meta-assessors and participants recognize the necessity to monitor trainees' progression. They use different strategies to do so, such as comparing actual trainee performance to a target performance or to a trainee's past performance.[42] Participants infer a trainee's level of competence in an individual decision-making process and acknowledged the input of observer judgment in this cognitive process.

Several authors reported that a group decision-making process might facilitate inferring the level of competence[22,24–26,28,34,40,43,45] including the observer in the decision-making process might facilitate the assessment process as observers provide insights and precise written assessment data by orally reporting their

observations.[22,25,26,28] Committee members resolving disagreements can facilitate the decision-making group process.[45] The quality of the judgment in assessing a trainee based on reported observations increases when stakeholders take action to limit potential bias and make the process transparent. Including a meta-assessor in the assessment process[24] and using assessments with milestones (clear assessment rubrics) might mitigate potential biases related to the observer interpretation.[48,49]

Step 6: Documenting and communicating the decision. Meta-assessors document their judgement of trainees' competence level into a numerical scale.[28,39,45,47–49,53,54,56,57] Those scales can be in the form of milestones (e.g., entrustable professional activities (EPAs),[24,26,28] compliance forms,[51] etc.). In this context, milestones refer to « educational statements that illustrate how a physician's competence is expected to progress over the course of his/her career from novice to mastery."[58] Meta-assessors use different means to communicate their assessment decisions[22,24,28,34,40] to the trainee and program administrators.[22,28,34,40] Some meta-assessors report using personal portable devices (such as cell phones); while others use computers or portable devices provided by their workplace.[24,26,28,36,40] Participants document their decisions and transform them by using the same trainee's self-assessment form. They communicate their decisions to the trainee during the final formal assessment session and/or via an electronic platform.

Discussion

In this scoping review, we summarized the scientific literature on assessment that uses reported observations of trainee HPE performance. Our analysis of the data suggests assessments based on reported observations have become more prevalent in the fields of medicine and nursing in recent years. Given that our sample of articles only pertains to the fields of medicine and nursing, we will restrict our generalizations solely to these two professions, as opposed to the broader domain of HPE. We also identified several workplace-based assessment strategies used by meta-assessors. Using studies included in our review and consultation, we could make explicit six steps characterizing the design and implementation of trainee performance assessment based on reported observations in medical education (ME) and in nursing education (NE). The last phase of our work (consultation), emphasized that among all the issues related to assessment based on reported observations of trainee performance in ME and NE, negative consequences of graduating noncompetent professionals are a major issue. The results lead us to the following observations. First, meta-assessors should ensure high-quality decisions and graduating competent future physicians and nurses. Second, participants use schemas to assess trainees' performances based on reported observations. Third, there are several challenges when using reported observations to assess trainees' performances. Given the widespread implementation of competency-based medical education and CC structures, our results are timely.

We found that meta-assessors should ensure high-quality decisions and graduating competent physicians and nurses. Our participants shared their concerns about the negative consequences of graduating noncompetent professionals, and their motivation to gather supporting evidence of validity for their assessment data interpretation, similar to what has been documented in the literature.[59] Even if the validation of assessment data interpretation is mainly an institutional responsibility,[59,60] we observed our participants were cognizant of its importance in the decision process. We observed validation practices in participants' inferences, such as scoring, generalizations, and decision-making that are reminiscent of Kane's approach to validity and validation.[12] Participants used different inferences to support their interpretations of assessment data as suggested in Kane, for example.[12] Our findings suggest that participants might have had implicit concerns for the validity of assessment data interpretation without naming those concerns explicitly. The next steps could be to study meta-assessors' validation practices and their impact on the quality of the decisions made about the trainees' performances based on reported observations.

We also found that participants use schemas to assess trainees' performances based on reported observations. Similarly to previous work on assessor cognition,[61] we documented the use of schemas -by participants- that are based on their conceptions of competency, assessment strategies, and context specificity. Quite like in the literature on assessor cognition,[61–65] we documented participants make inferences about trainees by using various exemplars. There are however differences between our findings and those in the assessor-cognition literature. For example, our understanding of the design and implementation seems to be more linear than what has been documented previously,[61,62] for which, assessors navigate more fluidly between components or elements of the assessment process. Unfortunately, meta-assessors do

not seem sufficiently supported when assessing a performance. Given that the observer-assessor position can be a source of subjectivity and bias,[64–66] our research highlights the potential benefits of using a meta-assessor, or group of meta-assessors, which can provide greater objectivity in interpreting assessment data and mitigate potential assessor bias. Future work could focus on how individual decision processes, compared to group decision-making processes, impact the assessment of trainee performance.

Finally, we found that there are several challenges when using reported observations to assess trainees' performances. As researchers have noted before,[67] one challenge is the deficiencies in the assessment data that are reported (e.g., incomplete, uncertain, and difficult to interpret). Further research could be conducted to explore the psychometric properties of these data. Meanwhile, building on this specific challenge and the observer usefulness in CC,[4,68] we suggest CC should include a person who observed a performance. The recommendation of including a person who observed a performance could allow for more nuance when interpreting the assessment data. Challenges associated with trainees' assessments might be overcome with a better understanding of the subjective nature of performance assessment rather than by producing more data with strong psychometric properties.[67] Our review did not identify challenges of direct observations, such as lack of time[69,70] or struggling to create observation opportunities.[71] Time challenges in collecting and completing assessment data (Step 3) were documented. This absence of direct observation challenges might be explained by including the meta-assessor in the assessment that (trainee's performance assessment based on reported observations) considers assessment beyond the observer and the trainee[69] and might encompass the observer availability and clinical workload.[72,73] Our work clearly identifies that there is a need to describe interventions to encompass the lack in reported assessment data. Further work is required to assess the feasibility of this six-step approach before moving to widespread implementation.

One limitation of our scoping review is that our corpus consisted only of documents from medicine and nursing. Since the programs in medicine and nursing tailor their assessments based on reported observations differently, we can presume that other HPE programs could also be different and that might impact the design and implementation phases. Also, because of the limited number of manuscripts included in our study, we could not analyze the data based on specific assessment purposes. Publication bias may also be a limitation, as some journals may selectively publish positive results while neglecting negative findings. We haven't underscored the validity assessment of the approach employed, we noticed that has a limitation. Next steps should include the verification of the assessment approach's validity. Three studies do not indicate the level of the trainees, and this might be considered as a weakness of our review. Therefore, the challenges reported for clinical education may not be specific to the particular assessment purposes, which could vary. The limited corpus of literature might reflect a taboo regarding the use of reported observation considering it as subpar. However, with the increased reliance on assessment based on reported observations, more researchers may undertake and publish work that tackles its challenges and potential.

Conclusion

We observed that validation of assessment data interpretation are a part of participants' experiential knowledge and are probably driven by social imperatives.[59] As such, assessors might need better support to ensure the quality of their decisions. Gaining a full understanding of the variability of meta-assessor judgments means exploring and understanding the context beyond meta-assessor cognition (e.g., observer and trainee). To facilitate future research, our work provides overall terminology to name the various steps and stakeholders involved in trainees' performance assessment based on reported observations. Future research might build on these findings to identify contextual factors influencing the use of trainee performance assessment based on reported observations.

Conflicts of Interest: The authors have no conflicts of interest.
Funding: The research project received funding from Social Sciences and Humanities Research Council.
Disclosure: We used the service of a professional linguistic reviewer.
Acknowledgments: The authors wish to thank Josée Toulouse and Annie Plourde for their contribution in developing the search strategy. In addition, a special thanks to Linda Bergeron and Kathleen Ouellet for their constructive feedback on the article.
Edited by: Doug Archibald (section editor); Marco Zaccagnini (senior section editor); Marcel D'Eon (editor-in-chief).

References

1. Gofton W, Dudek N, Barton G, Bhanji F. Workplace-based assessment implementation guide: formative tips for medical teaching practice. *R Coll Physicians Surg Can*. 2017;1-12.
2. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach*. 2010;32(8):676-82. https://doi.org/10.3109/0142159X.2010.500704
3. 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):1055-68. https://doi.org/10.1111/medu.12546
4. 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):1110-5. https://doi.org/10.1080/0142159X.2018.1474191
5. Schuwirth LWT, van der Vleuten CPM. A history of assessment in medical education. *Adv Health Sci Educ*. 2020;25(5):1045-56. https://doi.org/10.1007/s10459-020-10003-0
6. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-14. https://doi.org/10.3109/0142159X.2012.652239
7. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-8. https://doi.org/10.3109/0142159X.2013.789134
8. Colbert CY, French JC, Herring ME, Dannefer EF. Fairness: the hidden challenge for competency-based postgraduate medical education programs. *Perspect Med Educ*. 2017;6:347-55. https://doi.org/10.1007/s40037-017-0359-8
9. Colbert CY, Dannefer EF, French JC. Clinical competency committees and assessment: changing the conversation in graduate medical education. *J Grad Med Educ*. 2015;7(2):162-5. https://doi.org/10.4300/JGME-D-14-00448.1
10. French JC, Dannefer EF, Colbert CY. A systematic approach toward building a fully operational clinical competency committee. *J Surg Educ*. 2014;71(6):e22-7. https://doi.org/10.1016/j.jsurg.2014.04.005
11. Larocque S, Luhanga FL. Exploring the issue of failure to fail in a nursing program. *Int J Nurs Educ Scholarsh*. 2013;10(1):115-22. https://doi.org/10.1515/ijnes-2012-0037
12. Kane M. The argument-based approach to validation. *Sch Psychol Rev*. 2013;42(4):448-57. https://doi.org/10.1080/02796015.2013.12087465
13. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1-73. https://doi.org/10.1111/jedm.12000
14. Zumbo BD, Chan EKH. Validity and validation in social, behavioral, and health sciences. Cham: Springer; 2014. https://doi.org/10.1007/978-3-319-07794-9
15. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. 2005;8(1):19-32. https://doi.org/10.1080/1364557032000119616
16. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid Synth*. 2020;18(10):2119-26. https://doi.org/10.11124/JBIES-20-00167
17. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci*. 2010;5:69. https://doi.org/10.1186/1748-5908-5-69
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-73. https://doi.org/10.7326/M18-0850
19. Thomas A, Lubarsky S, Durning SJ, Young ME. Knowledge syntheses in medical education: demystifying scoping reviews. *Acad Med*. 2017;92(2):161-6. https://doi.org/10.1097/ACM.0000000000001452
20. Braun V, Clarke V. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Couns Psychother Res*. 2021;21(1):37-47. https://doi.org/10.1002/capr.12360
21. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77-101. https://doi.org/10.1191/1478088706qp063oa
22. Donato AA, Alweis R, Wenderoth S. Design of a clinical competency committee to maximize formative feedback. *J Community Hosp Intern Med Perspect JCHIMP*. 2016;6(6):33533. https://doi.org/10.3402/jchimp.v6.33533
23. Moonen-van Loon JMW, Overeem K, Govaerts MJB, Verhoeven BH, van der Vleuten CPM, Driessen EW. The reliability of multisource feedback in competency-based assessment programs: the effects of multiple occasions and assessor groups. *Acad Med*. 2015;90(8):1093-9. https://doi.org/10.1097/ACM.0000000000000763
24. Keeley MG, Gusic ME, Morgan HK, Aagaard EM, Santen SA. Moving toward summative competency assessment to individualize the postclerkship phase. *Acad Med*. 2019;94(12):1858-64. https://doi.org/10.1097/ACM.0000000000002830
25. Hemmer PA, Hawkins R, Jackson JL, Pangaro LN. Assessing how well three evaluation methods detect deficiencies in medical students' professionalism in two settings of an internal medicine clerkship. *Acad Med*. 2000;75(2):167-73. https://doi.org/10.1097/00001888-200002000-00016
26. Murray KE, Lane JL, Carraccio C, et al. Crossing the gap: using competency-based assessment to determine whether learners are ready for the undergraduate-to-graduate transition. *Acad Med*. 2019;94(3):338-45. https://doi.org/10.1097/ACM.0000000000002535
27. Hicks PJ, Margolis MJ, Carraccio CL, et al. A novel workplace-based assessment for competency-based decisions and learner feedback. *Med Teach*. 2018;40(11):1143-50. https://doi.org/10.1080/0142159X.2018.1461204
28. Rich JV, Fostaty Young S, Donnelly C, et al. Competency-based education calls for programmatic assessment: but what does this look like in practice? *J Eval Clin Pract*. 2019;26(4):95. https://doi.org/10.1111/jep.13328
29. Driessen EW, van Tartwijk J, Govaerts M, Teunissen P, van der Vleuten CP. The use of programmatic assessment in the clinical workplace: a Maastricht case report. *Med Teach*. 2012;34(3):226-31. https://doi.org/10.3109/0142159X.2012.652242
30. Cianciolo AT, Hingle S, Hudali T, Beason AM. Evaluating clerkship competency without exams. *Clin Teach*. 2020;17(6):624-8. https://doi.org/10.1111/tct.13114
31. Lass SL, Kornreich HK, Hoffmann KI, Friedman DB. Consistency in ratings of clinical performance of the same students throughout medical school and internship. Annual Conference on Research in Medical Education. Conference on Research in Medical Education. 1977;16:147-52. PMID: 606069
32. Perry M, Linn A, Munzer BW, et al. Programmatic assessment in emergency medicine: implementation of best practices. *J Grad*

Med Educ. 2018;10(1):84-90. https://doi.org/10.4300/JGME-D-17-00094.1
33. Helminen K, Tossavainen K, Turunen H. Assessing clinical practice of student nurses: views of teachers, mentors and students. *Nurse Educ Today*. 2014;34(8):1161-6. https://doi.org/10.1016/j.nedt.2014.04.007
34. Duitsman ME, Fluit CRMG, van der Goot WE, et al. Judging residents' performance: a qualitative study using grounded theory. *BMC Med Educ*. 2019;19(1):13. https://doi.org/10.1186/s12909-018-1446-1
35. Hemmer PA, Dadekian GA, Terndrup C, et al. Regular formal evaluation sessions are effective as frame-of-reference training for faculty evaluators of clerkship medical students. *J Gen Intern Med*. 2015;30(9):1313-8. https://doi.org/10.1007/s11606-015-3294-6
36. Yonge O, Myrick F, Ferguson LM. Preceptored students in rural settings want feedback. *Int J Nurs Educ Scholarsh*. 2011;8(1). https://doi.org/10.2202/1548-923X.2047
37. Battistone M, Pendleton B, Milne C, et al. Global descriptive evaluations are more responsive than global numeric ratings in detecting students' progress during the inpatient portion of an internal medicine clerkship. *Acad Med*. 2001;76(10):S105-7. https://doi.org/10.1097/00001888-200110001-00035
38. Wu XV, Enskär K, Pua LH, Heng DGN, Wang W. Clinical nurse leaders' and academics' perspectives in clinical assessment of final-year nursing students: a qualitative study. *Nurs Health Sci*. 2017;19(3):287-93. https://doi.org/10.1111/nhs.12342
39. Borman KR, Augustine R, Leibrandt T, Pezzi CM, Kukora JS. Initial performance of a modified milestones global evaluation tool for semiannual evaluation of residents by faculty. *J Surg Educ*. 2013;70(6):739-49. https://doi.org/10.1016/j.jsurg.2013.08.004
40. Hauer KE, Chesluk B, Iobst W, et al. Reviewing residents' competence: a qualitative study of the role of clinical competency committees in performance assessment. *Acad Med*. 2015;90(8):1084-92. https://doi.org/10.1097/acm.0000000000000736
41. Goodyear HM, Lakshminarayana I, Wall D, Bindal T. A multisource feedback tool to assess ward round leadership skills of senior paediatric trainees: (2) Testing reliability and practicability. *Postgrad Med J*. 2015;91(1075):268-73. https://doi.org/10.1136/postgradmedj-2015-133308
42. Ekpenyong A, Baker E, Harris I, et al. How do clinical competency committees use different sources of data to assess residents' performance on the internal medicine milestones? A mixed methods pilot study. *Med Teach*. 2017;39(10):1074-83. https://doi.org/10.1080/0142159X.2017.1353070
43. Swing SR, Clyman SG, Holmboe ES, Williams RG. Advancing resident assessment in graduate medical education. *J Grad Med Educ*. 2009;1(2):278-86. https://doi.org/10.4300/JGME-D-09-00010.1
44. Berger JS, Pan E, Thomas J. A randomized, controlled crossover study to discern the value of 360-degree versus traditional, faculty-only evaluation for performance improvement of anesthesiology residents. *J Educ Perioper Med JEPM*. 2009;11(2):E053. https://doi.org/10.46374/volxi-issue2-berger
45. Nabors C, Forman L, Peterson SJ, et al. Milestones: a rapid assessment method for the Clinical Competency Committee. *Arch Med Sci*. 2017;13(1):201-9. https://doi.org/10.5114/aoms.2016.64045
46. Chan TM, Sherbino J, Mercuri M. Nuance and noise: lessons learned from longitudinal aggregated assessment data. *J Grad Med Educ*. 2017;9(6):724-9. https://doi.org/10.4300/JGME-D-17-00086.1
47. Hauer KE, Clauser J, Lipner RS, et al. The internal medicine reporting milestones: cross-sectional description of initial implementation in U.S. residency programs. *Ann Intern Med*. 2016;165(5):356-62. https://doi.org/10.7326/M15-2411
48. Friedman KA, Balwan S, Cacace F, Katona K, Sunday S, Chaudhry S. Impact on house staff evaluation scores when changing from a Dreyfus- to a Milestone-based evaluation model: one internal medicine residency program's findings. *Med Educ Online*. 2014;19(1). https://doi.org/10.3402/meo.v19.25185
49. Aagaard E, Kane GC, Conforti L, et al. Early feedback on the use of the internal medicine reporting milestones in assessment of resident performance. *J Grad Med Educ*. 2013;5(3):433-8. https://doi.org/10.4300/JGME-D-13-00001.1
50. Chan TM, Sebok-Syer SS, Sampson C, Monteiro S. The quality of assessment of learning (Qual) score: validity evidence for a scoring system aimed at rating short, workplace-based comments on trainee performance. *Teach Learn Med*. 2020;32(3):319-29. https://doi.org/10.1080/10401334.2019.1708365
51. Ogunyemi D, Eno M, Rad S, Fong A, Alexander C, Azziz R. Evaluating professionalism, practice-based learning and improvement, and systems-based practice: utilization of a compliance form and correlation with conflict styles. *J Grad Med Educ*. 2010;2(3):423-9. https://doi.org/10.4300/JGME-D-10-00048.1
52. Engström M, Löfmark A, Vae KJU, Mårtensson G. Nursing students' perceptions of using the Clinical Education Assessment tool AssCE and their overall perceptions of the clinical learning environment - A cross-sectional correlational study. *Nurse Educ Today*. 2017;51:63-7. https://doi.org/10.1016/j.nedt.2017.01.009
53. Park YS, Zar FA, Norcini JJ, Tekian A. Competency evaluations in the next accreditation system: contributing to guidelines and implications. *Teach Learn Med*. 2016;28(2):135-45. https://doi.org/10.1080/10401334.2016.1146607
54. Watson RS, Borgert AJ, O Heron CT, et al. A multicenter prospective comparison of the accreditation council for graduate medical education milestones: clinical competency committee vs. resident self-assessment. *J Surg Educ*. 2017;74(6):e8-14. https://doi.org/10.1016/j.jsurg.2017.06.009
55. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49(3):296-306. https://doi.org/10.1111/medu.12637
56. Douglass KA, Jacquet GA, Hayward AS, et al. Development of a global health milestones tool for learners in emergency medicine: a pilot project. *AEM Educ Train*. 2017;1(4):269-79. https://doi.org/10.1002/aet2.10046
57. Bartlett KW, Whicker SA, Bookman J, et al. Milestone-based assessments are superior to likert-type assessments in illustrating trainee progression. *J Grad Med Educ*. 2015;7(1):75-80. https://doi.org/10.4300/JGME-D-14-00389.1
58. The Royal College of Physicians and Surgeons of Canada. CanMEDS Milestones. Ottawa (ON): The Royal College of Physicians and Surgeon of Canada; 2015. Available from: https://canmeds.royalcollege.ca/en/milestones
59. Marceau M, Gallagher F, Young M, St-Onge C. Validity as a social imperative for assessment in health professions education: a concept analysis. *Med Educ*. 2018;52(6):641-53. https://doi.org/10.1111/medu.13574
60. American Educational Research Association., American Psychological Association., National Council on Measurement in Education., Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and

psychological testing. Washington (DC): American Educational Research Association; 2014.
61. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. *Med Educ*. 2016;50(5):511-22. https://doi.org/10.1111/medu.12973
62. St-Onge C, Chamberland M, Lévesque A, Varpio L. Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Adv Health Sci Educ*. 2016;21:627-42. https://doi.org/10.1007/s10459-015-9656-3
63. Govaerts M, van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164-74. https://doi.org/10.1111/medu.12289
64. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45(10):1048-60. https://doi.org/10.1111/j.1365-2923.2011.04025.x
65. Yeates P, Cardell J, Byrne G, Eva KW. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. *Med Educ*. 2015;49(9):909-19. https://doi.org/10.1111/medu.12777
66. Lee V, Brain K, Martin J. From opening the 'black box' to looking behind the curtain: cognition and context in assessor-based judgements. *Adv Health Sci Educ*. 2019;24:85-102. https://doi.org/10.1007/s10459-018-9851-0
67. Pack R, Lingard L, Watling CJ, Chahine S, Cristancho SM. Some assembly required: tracing the interpretative work of Clinical Competency Committees. *Med Educ*. 2019;53(7):723-34. https://doi.org/10.1111/medu.13884
68. Hauer KE, Cate O ten, Boscardin CK, et al. Ensuring resident competence: a narrative review of the literature on group decision making to inform the work of Clinical Competency Committees. *J Grad Med Educ*. 2016;8(2):156-64. https://doi.org/10.4300/JGME-D-15-00144.1
69. Cheung WJ, Patey AM, Frank JR, Mackay M, Boet S. Barriers and enablers to direct observation of trainees' clinical performance: a qualitative study using the theoretical domains framework. *Acad Med*. 2019;94(1):101-14. https://doi.org/10.1097/ACM.0000000000002396
70. Watling C, LaDonna KA, Lingard L, Voyer S, Hatala R. 'Sometimes the work just needs to be done': socio-cultural influences on direct observation in medical training. *Med Educ*. 2016;50(10):1054-64. https://doi.org/10.1111/medu.13062
71. St-Onge C. Enjeux et défis de l'évaluation longitudinale: quelques pistes de réflexion préalables à son implantation. *Pédagogie Médicale*. 2018;19(3):137-42. https://doi.org/10.1051/pmed/2019022
72. Madan R, Conn D, Dubo E, Voore P, Wiesenfeld L. The enablers and barriers to the use of direct observation of trainee clinical skills by supervising faculty in a psychiatry residency program. *Can J Psychiatry*. 2012;57(4):269-72. https://doi.org/10.1177/070674371205700411
73. Kogan JR, Conforti LN, Yamazaki K, Iobst W, Holmboe ES. Commitment to change and challenges to implementing changes after workplace-based assessment rater training. *Acad Med*. 2017;92(3):394-402. https://doi.org/10.1097/ACM.0000000000001319

Appendix A. The Search strategy used in 2020-21 for the scoping review on the assessment of trainees' performance based on reported observations in HPE.

TI (C1 AND (C2 OR C3 OR C4)) AND **AB** (C1 AND C2 AND C3) AND C4	
C1	assess* OR evaluat* OR rat* OR feedback* OR tool* OR judg*
C2	authenti* OR perform* OR competenc* OR skill* OR abilit* OR attitud* OR aptitude
C3	learn* OR student* OR trainee* OR supervisee* OR workplace* OR "work-bas*" OR "work place*"
C4	(medical OR medicine OR nurs* OR "physical therap*" OR "occupational therap*" OR dentist* OR pharmac* OR "health sciences" OR "health prof*" OR "physiotherap*) N0 education*

Abbreviation: TI = title; C1 = first concept; C2 = second concept; C3 = third concept; C4 = fourth concept; AB = Abstract.