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-Onge1

1University of Sherbrooke, Quebec, Canada

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

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

© 2024 Blanchette, Poitras, Lefebvre, St-Onge. Synergies Partners. This is an Open Journal Systems article distributed under the terms of the Creative Commons Attribution License. (https://creativecommons.org/licenses/by-nc-nd/4.0) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited.

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


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 métaveilleurs.
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). Faculty members, peers, patients, or other health professionals can contribute to the reported observations.

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. 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.

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). CC use reported observations comprising independent judgments to decide in a group process.

One challenge of CC using reported observations is to ensure fairness in their decision to limit legal repercussions, and programs have to standardize CC processes. HPE programs that use reported observations to determine trainee progression include, but are not limited to, medicine, nursing, and social work.

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. 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 work, with adjustments made based on recent methodological recommendations of Peters et al., Levac et al., and Trico et al. More specifically, we have updated Arksey & O’Malley’s work in identifying relevant studies, selecting studies, distributing data, and collating results. In addition, we were informed by Levac et al. 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, 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 (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. They continued this process until they reached 90% agreement in their decision to include or exclude an abstract. 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). 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.

Step 5: Collating and reporting results
We summarized quantifiable data using frequencies when describing our archive (pool of manuscripts included). 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. We recognized the dual implication in the analysis process: deductive (theory-driven) and inductive (researcher subjectivity). We followed Braun & Clarke's five-step reflexive approach. First, the principal investigator (PB) deductively coded with an initial coding tree. She rered 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. 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 progressions based on the 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 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) and followed Levac et al. 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.
number of articles in
of HPE trainee
Figure
individual or group of
observer supervises trainees. A meta
observes trainee performance. Given this central role in
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.

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

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.

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’Malley15 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 observations25,40
and observers25,29,34,36,38–41 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 trainee's 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 1. Assessment strategies used in clinical education settings

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

<table>
<thead>
<tr>
<th>Trainee</th>
<th>Medicine</th>
<th>Nursing</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDIVIDUALS</td>
<td>Term</td>
<td>Reference(s)</td>
</tr>
<tr>
<td>Resident</td>
<td>(22-24, 26, 28, 32, 39, 41-43, 45, 46, 48, 49, 51, 53-55, 71)</td>
<td>19</td>
</tr>
<tr>
<td>Trainee</td>
<td>(40, 41, 48, 50, 54, 55, 57)</td>
<td>7</td>
</tr>
<tr>
<td>Student</td>
<td>(25, 29, 30, 35, 37, 44)</td>
<td>6</td>
</tr>
<tr>
<td>Learner</td>
<td>(23, 24, 27)</td>
<td>3</td>
</tr>
<tr>
<td>Observer</td>
<td>Faculty</td>
<td>(22, 24, 28, 31, 32, 45, 46, 51, 53)</td>
</tr>
<tr>
<td></td>
<td>House staff</td>
<td>(25, 48)</td>
</tr>
<tr>
<td></td>
<td>Faculty member</td>
<td>(35, 50)</td>
</tr>
<tr>
<td></td>
<td>Faculty attendee</td>
<td>(39)</td>
</tr>
<tr>
<td></td>
<td>Faculty rater</td>
<td>(50)</td>
</tr>
<tr>
<td></td>
<td>Supervising physician</td>
<td>(46)</td>
</tr>
<tr>
<td></td>
<td>Clinical supervisor</td>
<td>(26)</td>
</tr>
<tr>
<td></td>
<td>Resident physician</td>
<td>(24)</td>
</tr>
<tr>
<td></td>
<td>Physician</td>
<td>(23)</td>
</tr>
<tr>
<td></td>
<td>Mentor</td>
<td>(29)</td>
</tr>
<tr>
<td></td>
<td>Medical educator</td>
<td>(40)</td>
</tr>
<tr>
<td></td>
<td>Supervising consultant</td>
<td>(41)</td>
</tr>
<tr>
<td></td>
<td>Master assessor</td>
<td>(24)</td>
</tr>
<tr>
<td></td>
<td>Supervisor</td>
<td>(31)</td>
</tr>
<tr>
<td></td>
<td>Program director</td>
<td>(27, 44, 46, 50, 51, 54, 55, 71)</td>
</tr>
<tr>
<td></td>
<td>Clerkship or assistant director</td>
<td>(25, 35, 37)</td>
</tr>
<tr>
<td></td>
<td>Medicine clerkship coordinator</td>
<td>(31)</td>
</tr>
<tr>
<td></td>
<td>Faculty</td>
<td>(25, 44, 77)</td>
</tr>
<tr>
<td></td>
<td>Faculty attending member</td>
<td>(55)</td>
</tr>
<tr>
<td></td>
<td>Educational supervisor</td>
<td>(41)</td>
</tr>
<tr>
<td></td>
<td>Academic advisor</td>
<td>(28)</td>
</tr>
<tr>
<td>GROUPS OF INDIVIDUALS</td>
<td>Clinical competency committee</td>
<td>(22, 26, 27, 32, 39, 40, 42, 45, 46, 48, 53, 54, 71)</td>
</tr>
<tr>
<td></td>
<td>Competency committee</td>
<td>(28, 43, 49, 50, 57)</td>
</tr>
<tr>
<td></td>
<td>Evaluation grading committee</td>
<td>(30)</td>
</tr>
<tr>
<td></td>
<td>Assessment committee</td>
<td>(29)</td>
</tr>
<tr>
<td></td>
<td>Sub-committee</td>
<td>(42)</td>
</tr>
<tr>
<td></td>
<td>Entrustment committee</td>
<td>(24)</td>
</tr>
</tbody>
</table>

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 forms38,49 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 director17 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. As electronic platforms might help in collecting assessment data generated by longitudinal assessment. As they review, capture, synthesize, and present milestones, electronic systems might reduce the time competence committees spend on reviewing trainee performance. Some participants in rural settings some posited that teleconferencing enhanced the communication between meta-assessor and trainee. 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, 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 or by a group of individuals that might include the observer and/or trainee. In nursing, the meta-assessor completes the assessment data during a formal assessment session with the observer and trainee. 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. Formal assessment-group sessions, competence committees, subcommittees and/or entrustment committees 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. In nursing, only one individual revises observations during a formal assessment session composed of a meta-assessor, observer, and trainee. 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. For example, aggregating formative-assessment data to make high-stakes summative decisions is challenging. The participants confirmed this. Aggregating data from different sources is also a challenge. Participants consider reported observations from observers 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 or rubrics informed by a competency framework. 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." 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. 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. 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 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
Committee members resolving disagreements can facilitate the decision-making group process.\textsuperscript{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\textsuperscript{24} and using assessments with milestones (clear assessment rubrics) might mitigate potential biases related to the observer interpretation.\textsuperscript{48,49}

**Step 6: Documenting and communicating the decision.**

Meta-assessors document their judgement of trainees’ competence level into a numerical scale.\textsuperscript{22,25,26,28} Those scales can be in the form of milestones (e.g., entrustable professional activities [EPAs],\textsuperscript{14,26,29} compliance forms,\textsuperscript{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.”\textsuperscript{58} Meta-assessors use different means to communicate their assessment decisions\textsuperscript{22,24,28,34,40} to the trainee and program administrators.\textsuperscript{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.\textsuperscript{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.\textsuperscript{59} Even if the validation of assessment data interpretation is mainly an institutional responsibility,\textsuperscript{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.\textsuperscript{12} Participants used different inferences to support their interpretations of assessment data as suggested in Kane, for example.\textsuperscript{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,\textsuperscript{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,\textsuperscript{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,\textsuperscript{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, 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, 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, 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. Our review did not identify challenges of direct observations, such as lack of time or struggling to create observation opportunities. 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 and might encompass the observer availability and clinical workload. 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. 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


31. Lass SL, Kornreich HK, Hoffmann Kl, 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: 60609


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.

<table>
<thead>
<tr>
<th></th>
<th>C1 = assess* OR evaluat* OR rat* OR feedback* OR tool* OR judg*</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>authenti* OR perform* OR competenc* OR skill* OR abilit* OR attitud* OR aptitude</td>
</tr>
<tr>
<td>C3</td>
<td>learn* OR student* OR trainee* OR supervisee* OR workplace* OR &quot;work-bas**&quot; OR &quot;work place**&quot;</td>
</tr>
<tr>
<td>C4</td>
<td>(medical OR medicine OR nurs* OR &quot;physical therap***&quot; OR &quot;occupational therap***&quot; OR dentist* OR pharmac* OR &quot;health sciences&quot; OR &quot;health prof***&quot; OR &quot;physiotherap&quot;) N0 education*</td>
</tr>
</tbody>
</table>

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