Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance

Michelle Daniel, MD, MHPE, Joseph Rencic, MD, Steven J. Durning, MD, PhD, Eric Holmboe, MD, Sally A. Santen, MD, PhD, Valerie Lang, MD, MHPE, Temple Ratcliffe, MD, David Gordon, MD, Brian Heist, MD, MSc, Stuart Lubarsky, MD, MHPE, Carlos A. Estrada, MD, MS, Tiffany Ballard, MD, Anthony R. Artino Jr, PhD, Ana Sergio Da Silva, PhD, Timothy Cleary, PhD, Jennifer Stojan, MD, MHPE, and Larry D. Gruppen, PhD

Abstract

Purpose
An evidence-based approach to assessment is critical for ensuring the development of clinical reasoning (CR) competence. The wide array of CR assessment methods creates challenges for selecting assessments fit for the purpose; thus, a synthesis of the current evidence is needed to guide practice. A scoping review was performed to explore the existing menu of CR assessments.

Method
Multiple databases were searched from their inception to 2016 following PRISMA guidelines. Articles of all study design types were included if they studied a CR assessment method. The articles were sorted by assessment methods and reviewed by pairs of authors. Extracted data were used to construct descriptive appendixes, summarizing each method, including common stimuli, response formats, scoring, typical uses, validity considerations, feasibility issues, advantages, and disadvantages.

Results
A total of 377 articles were included in the final synthesis. The articles broadly fell into three categories: non-workplace-based assessments (e.g., multiple-choice questions, extended matching questions, key feature examinations, script concordance tests); assessments in simulated clinical environments (objective structured clinical examinations and technology-enhanced simulation); and workplace-based assessments (e.g., direct observations, global assessments, oral case presentations, written notes). Validity considerations, feasibility issues, advantages, and disadvantages differed by method.

Conclusions
There are numerous assessment methods that align with different components of the complex construct of CR. Ensuring competency requires the development of programs of assessment that address all components of CR. Such programs are ideally constructed of complementary assessment methods to account for each method's validity and feasibility issues, advantages, and disadvantages.

Definitions of clinical reasoning vary widely.[1] For the purposes of this paper, clinical reasoning is defined as *a skill, process, or outcome* wherein clinicians observe, collect, and interpret data to diagnose and treat patients.[2,3] Clinical reasoning entails both conscious and unconscious cognitive operations interacting with contextual factors.[4,5] Contextual factors include, but are not limited to, the patient's unique circumstances and preferences and the characteristics of the practice environment. Multiple components of clinical reasoning can be identified[1]: information gathering, hypothesis generation, forming a problem representation, generating a differential diagnosis, selecting a leading or working diagnosis, providing a diagnostic justification, and developing a management or treatment plan.[6] A number of theories (e.g., script, dual process, and cognitive load theories) from diverse fields (e.g., cognitive psychology, sociology, education) inform research on clinical reasoning.[7,8] This definition of clinical reasoning and these multiple theories provide the foundation for the current work.

Effective clinical reasoning is central to clinical competence. The Accreditation Council for Graduate Medical Education,[9] the CanMEDS framework,[10] and the Tuning Project (Medicine) in Europe[11] all describe clinical reasoning as a *core competency*. Ensuring the development of clinical competence (including clinical reasoning) across the medical education continuum requires an evidence-based approach to assessment. There is currently a wide array of clinical reasoning assessments, and the literature on which these tools are based is widely dispersed, crossing different fields and multiple medical specialties, which presents a challenge for medical educators attempting to select and implement assessments aligned with their particular goals, needs, and resources. These assessments are often designed for use in different contexts (e.g., workplace- and non-workplace-based environments).[12] The sheer number and diversity of clinical reasoning assessment methods create challenges for selecting assessments fit for the purpose, so a synthesis of the current evidence is needed to advance assessment practices for this core competency.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Michelle Daniel, University of Michigan Medical School, 6123 Taubman Health Sciences Library, 1135 Catherine St., SPC 5726, Ann Arbor, MI 48109; telephone: (734) 763-6770; e-mail: micdan@med.umich.edu; Twitter: @Emergdoc1975.

Acad Med. 2019;94:902–912.
First published online January 29, 2019
doi: 10.1097/ACM.0000000000002618

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A631, http://links.lww.com/ACADMED/A632, http://links.lww.com/ACADMED/A633, and http://links.lww.com/ACADMED/A634.

Our aim was to create a practical compendium of assessment methods to serve as a reference for medical educators. Given the richness and complexity of the clinical reasoning assessment literature, we chose to perform a scoping review to explore the following questions: What clinical reasoning assessment methods are available? What are the defining features of these assessment methods, and how are they typically used? What are the validity considerations (content, response process, internal structure, relationships to other variables, and consequences or outcomes on clinical practice performance) for each method? What are the feasibility issues, advantages, and disadvantages of each method? How might the relative strengths and weaknesses of each method be used to construct a clinical reasoning assessment program?

Method

Review methodology

We adopted a constructivist research paradigm in conducting this review. We chose a scoping methodology because our questions were exploratory and because preliminary searches had revealed a complex and heterogeneous body of literature.[13] We wanted to describe the broad field of clinical reasoning assessment methods,[14] yet remain focused on practical applications to ensure relevance for medical educators. We report on the most commonly used methods, but we do not seek to be exhaustive. This review is presented in accordance with the STORIES (Structured Approach to the Reporting in Healthcare Education of Evidence Synthesis) statement.[15]

Search strategy

We followed established PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines[16] for our initial search and article selection process. An experienced research librarian helped design the search strategy (see Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/A631). Numerous synonyms for clinical reasoning were combined with a broad range of assessment terms, as well as well-known clinical reasoning assessment methods. We ran the search in Ovid MEDLINE, CINAHL, ERIC, PsycINFO, Scopus, Google Scholar, and the New York Academy of Medicine Grey Literature Report from each database's inception through February 29, 2016, the date of our search. Retrieved citations were uploaded in DistillerSR (Evidence Partners, Ottawa, Ontario, Canada), an online data management system for performing systematic reviews.

Screening and review of articles

We began with broad inclusion criteria for our initial exploration of the clinical reasoning assessment literature: (1) any health profession (e.g., medicine, nursing, dentistry, physical or occupational therapy) at any stage of training or practice; (2) all study design types; and (3) any article that explicitly studied a method (or tool) of clinical reasoning assessment (or synonymous terms—e.g., clinical, diagnostic, therapeutic, or prognostic decision making or problem solving; see Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/A631). Articles were excluded if they were not in English, if decision making was applied only to a specific clinical problem (e.g., a case of atrial fibrillation) instead of the larger cognitive processes of clinical reasoning, or if the article was an essay or commentary that did not constitute research. Review articles were excluded from data extraction but were used to identify additional articles via snowballing. Prior to the final synthesis, we decided to focus on medical student, resident, or physician studies and de-emphasized the other health professions to both reduce the total number of articles for review and ensure that the focus was on clinical reasoning (and not on related but distinct constructs in the other health professions, such as critical thinking).[17]

Different combinations of authors (M.D., J.R., S.J.D., E.H., S.A.S., V.L., T.R., D.G., B.H., S.L., C.A.E., T.B., A.R.A., A.S.D.S., T.C., J.S., L.D.G.) reviewed the articles in multiple stages. Potentially relevant titles and abstracts were screened by pairs of authors. Full-text articles were then assessed by different pairs of authors for eligibility based on the inclusion and exclusion criteria. Prior to the assessment of full-text articles for eligibility, we sorted them by assessment methods based on our preliminary analyses of the abstracts and the collective expertise of our team. We were mindful that older methods may be more frequently represented in published articles (e.g., multiple-choice questions [MCQs]), that common educational practices may not necessarily be written about often (e.g., oral case presentations [OCPs]), and that feasibility may affect implementation and use (e.g., functional magnetic resonance imaging). Each assessment method was assigned to a pair of authors who further reviewed and synthesized those articles. Disagreements at any stage were resolved through discussion to reach consensus, with involvement of a third author if needed. Interrater agreement was assessed using Cohen kappa statistic at the data extraction level.

A data extraction form (see Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/A632) was used to capture information on the characteristics of assessment methods, including the stimulus (e.g., written vignette, standardized patients [SPs], real patients); response format (e.g., selected response, constructed free text, performance); scoring (e.g., fixed answer, checklist, global rating scale); and common uses (e.g., low-, medium-, or high-stakes decisions). The form also captured information regarding a tool's feasibility and validity, as well as any themes (e.g., the influence of context) related to the method. For the purposes of this review, we viewed validity as a unified construct with multiple sources of evidence (e.g., content, response process).[18] Because this was a scoping review, the quality of articles was not formally assessed. Extraction proceeded until all articles for an assessment method had been fully reviewed or no new assessment insights were forthcoming.

Data synthesis

We used the extracted data to construct descriptive appendixes that summarize each assessment method, describing common stimuli, response formats, scoring, typical uses, validity considerations, feasibility issues, advantages, and disadvantages. Validity considerations are presented according to Messick's five domains as described in *Standards for Educational and Psychological Testing*.[19] These appendixes list some references to support the text, but they do not include the full list of the articles reviewed because, for some methods, there were over 60 articles. In some cases, we used additional seminal

references (outside of those included in the review) to support key points in these appendixes and in the Results below; these references were not included in the review because they did not meet the inclusion criteria.

Over the course of the review, it became apparent that certain assessment methods were better suited than others to measure different components of clinical reasoning (see above). Because we aimed to produce a practical guide for medical educators to select clinical reasoning assessment methods, we used our collective judgments to identify assessment methods more or less capable of measuring the different components of clinical reasoning. First, we agreed on working definitions for each of the different components (Table 1). Next, we sent a survey via Qualtrics (version from 2018, Qualtrics, Provo, Utah) to the full author group, asking them to rate each assessment method in terms of its ability to assess the different components (0 = not addressed, 1 = secondary or peripheral, 2 = primary focus, NA = cannot answer). We averaged the results and reported them on the following scale: 0.0–0.5 = poor, 0.6–1.0 = average, 1.1–1.5 = good, and 1.6–2.0 = very good.

Results

The initial database search and snowballing yielded 14,709 records. We removed 1,849 as duplicates, leaving 12,860 records to be screened by title and abstract. After this screening, 11,421 articles were excluded because they did not pertain to the assessment of clinical reasoning. The 1,439 remaining articles underwent full-text evaluation based on inclusion and exclusion criteria. At this stage, 901 articles were excluded from the analysis, with the main reason being that they did not explicitly study a clinical reasoning assessment method. In the end, 538 articles (from 1966 to 2016) were included in the review (see Figure 1 and Supplemental Digital Appendix 3 at http://links.lww.com/ACADMED/A633). Of these articles, 161 focused on other health professions. In the final synthesis, we focused exclusively on the 377 articles related to medical students, residents, and physicians. The interrater agreement calculated for the methods was high, ranging from 0.83 to 0.86.

The included articles encompassed a broad array of learners from preclinical medical students to clinical medical students, residents, and practicing physicians. The work in the articles came from many different countries; however, the majority came from the United States, Europe, and Canada. We clustered the articles into 20 different assessment methods (an experimental or novel category and 19 methods; see below). Some methods had a large number of articles (e.g., script concordance testing and technology-enhanced simulation each had over 60). Others had very small numbers of articles (e.g., clinical or comprehensive integrative puzzles [CIPs] and chart-stimulated recall [CSR] each had 3). Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/A634) shows the descriptive appendixes we constructed that summarize each assessment method, including common stimuli, response formats, scoring, typical uses, validity considerations, feasibility issues, advantages, and disadvantages, as well as supporting references.

Although the methods were quite heterogeneous, we identified three broad categories, along a continuum of authenticity: non-workplace-based assessments (non-WBAs), assessments in simulated clinical environments, and workplace-based assessments (WBAs). We recognize that these categories have exceptions and that some methods could realistically be placed in multiple categories (e.g., self-regulated learning microanalysis [SRL-M]). Assessments that were unique, novel, or exploratory were placed into an experimental or novel methods group. Although important methods may ultimately emerge from this body of work, it was not feasible to report on all of these methods in depth, and

Table 1

Working Definitions for the Different Components of Clinical Reasoning[a]

Component	Definition
Information gathering[72,73]	The process of acquiring the data needed to generate or refine hypotheses. This is usually an active process that includes taking a history, performing a physical, acquiring lab or radiographic data, reviewing the medical record, etc., but may be implicit (through observation) as well. The selection of information to gather is driven by knowledge representations of disease (i.e., scripts, schema).
Hypothesis generation[74,75]	An early nonanalytic or analytic process by which a physician tries to find diseases that can explain a patient's clinical findings. Hypothesis generation involves activation of knowledge representations of disease in an iterative process that feeds back on information gathering and vice versa (e.g., hypothesis generation leads to more information gathering, which leads to more hypothesis generation and/or refinement).
Problem representation[74,76]	A dynamic mental representation of all the relevant aspects of the case (including the patient's clinical findings, biopsychosocial dimensions, etc.) that can be communicated in a summary that includes semantic qualifiers and key findings.
Differential diagnosis[77,78]	A list of diagnostic hypotheses that represent the best summary categorizations of the problem representation (Note: Different specialties may have different priorities when it comes to ordering the differential; e.g., in EM, life-threatening diseases are often listed first, whereas in IM, the most likely diseases are usually listed first). As the strength of confidence and evidence for these representations change, a leading diagnosis emerges.
Leading or working diagnosis[79]	A diagnosis for which a physician's probability of a given disease has crossed his or her threshold to pursue additional testing or to initiate treatment, even if the diagnosis is not definitive.
Diagnostic justification[77,80]	The attempt to use the evidence (key clinical findings) from information gathering to choose one or more diagnoses as most likely and to defend that choice, comparing and contrasting other possible diagnoses. Justification often involves communication (orally or in writing) when socially required and may not be part of the a priori clinical reasoning process.
Management and treatment[79,81]	The actions that follow the clinical reasoning process, including prognostication, management, treatment, prevention strategies, and palliation of symptoms (including improvement of quality of life) and justification for such actions.

Abbreviations: EM indicates emergency medicine; IM, internal medicine.
[a]Used in a 2016 scoping review of clinical reasoning assessment methods.

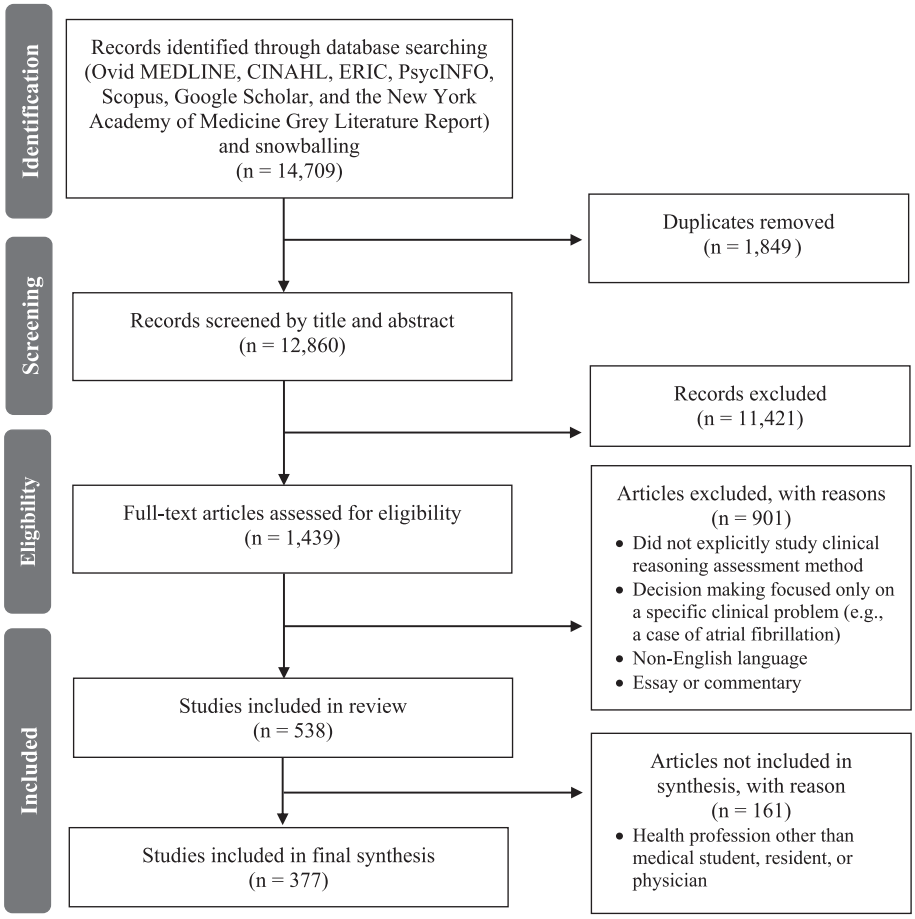

Figure 1 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram for a 2016 scoping review of clinical reasoning assessment methods.

they are only addressed in Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/A634).

Non-WBAs

We identified 10 methods that largely focused on "classroom" assessments or non-WBAs.

1. *MCQs* consist of a clinical vignette followed by up to five potential answers or alternatives and may be structured as to require a single best answer, a combination of alternatives, true or false for each alternative, or matching.[20]
2. *Extended matching questions (EMQs)* resemble MCQs in their use of a clinical vignette with a single best answer selected from a list of alternatives, but they contain longer lists of potential answers (more than five) that are applied to multiple questions.[21,22]
3. *Short- or long-answer (essay) questions* describe a method wherein a clinical vignette is followed by one or more questions answered using constructed free-text responses that range in length from a few words to several sentences.[23,24]
4. *Modified essay questions (MEQs)* are a method wherein serial information is provided about a case chronologically.[25,26] After each item, learners must document a decision in a constructed free-text (essay) format before they can view subsequent items.
5. *Patient management problems (PMPs)* consist of context-rich clinical scenarios, where specific resources are available for diagnosis and management.[27,28] The learner must select among multiple alternatives for action, and the results of those actions are then provided (e.g., electrocardiogram [ECG] findings) as they continue working through the case.
6. *Key feature examinations (KFEs)* contain clinical vignettes followed by two to three questions focused on the critical steps in clinical decision making.[29,30] Key features are case specific (e.g., a thunderclap headache is a key feature in the diagnosis of subarachnoid hemorrhage).
7. *Script concordance tests (SCTs)* comprise short clinical scenarios associated with uncertainty that are designed to represent the way new information is processed during clinical reasoning.[31,32] Learners must answer a series of questions (e.g., if you were thinking X and then you found Y, this answer would become more likely, less likely, or no change). Responses are compared with those acquired from a reference panel of "experts," accounting for the variability of clinicians' responses in different clinical situations.
8. *CIPs* take the form of a grid, often analogized to an extended matching crossword puzzle.[33,34] A number of findings are placed in columns (e.g., history, physical, ECG, labs, pathophysiology, pharmacology), and related diagnoses are placed in rows (e.g., myocardial infarction, pulmonary embolism, aortic dissection). The learner is asked to compare and contrast items within a column as well as across the rows (selecting the best "match" for the finding), building basic illness scripts for each diagnosis.
9. *Concept maps* are a schematic assessment method wherein learners represent their knowledge of a domain, as well as the organization of that knowledge, by creating a graphical illustration.[35,36] Maps may be free-form or hierarchical, outlining both concepts and the relationships between the concepts.
10. *Oral examinations* are verbal assessments conducted by one or more faculty member in either an unscripted or semiscripted fashion to assess clinical reasoning and decision-making abilities, as well as professional values.[37,38]

The majority of non-WBAs use written clinical vignettes or scenarios as the stimuli, though images, videos, and other formats may be used to supplement or complement the written testing materials. Only one non-WBA method uses a verbal stimulus (oral examinations).

The response formats are predominately written, though there is variability in type (e.g., selected answers, constructed free text). Scoring processes vary. Aggregated, fixed-answer responses are common (e.g., MCQs, EMQs, PMPs, KFEs). Scoring can be weighted (i.e., certain items count more than others) or unweighted (i.e., all items count equally) and compensatory (i.e., can get some percentage wrong and still pass) or noncompensatory (i.e., a score of 100% is required to pass). Itemized and global rating scales are used for short- or long-answer constructed free-text responses and MEQs, and they can be norm- or criterion-referenced. CIP grids and concept maps have more complex scoring systems. SCT responses are compared for fit to a "gold standard" (i.e., the expert panel's responses), and the examinee receives partial to full credit for each item depending on the proportion of the expert panel that chose that response. Several non-WBA methods are used for medium- to high-stakes examinations (e.g., MCQs and KFEs are commonly used for summative end-of-course assessments and medical licensing examinations). Other methods (e.g., CIPs, concept maps) are less well explored and are currently most suitable for formative assessments or research.

Validity considerations, feasibility issues, advantages, and disadvantages are highly specific to each method. Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/A634) details these differences, but a few themes for non-WBAs warrant mention here. MCQs, EMQs, and KFEs are the most frequently used non-WBAs, and they have the advantage of broad sampling that helps minimize context specificity. They offer the best chance of high internal consistency and thus have the greatest utility for high-stakes assessments. Content validity evidence for these methods can be strong because of expert consensus and blueprinting. These methods also offer the advantage of content control and consistency; there is a "right" answer to each problem, a feature not always possible in WBAs, which allows a measurement of accuracy. Further, all non-WBA methods allow students to be assessed across a standardized set of problems, something that is not possible in the workplace. The greatest validity challenge for non-WBA methods is in response process evidence. Selecting a correct answer from a number of possibilities, developing a graphic representation of knowledge organization, or even selecting information from a predefined list are not generally representative of authentic clinical reasoning activities in practice. Many of these methods emphasize part-task, rather than whole-task assessment (i.e., they measure fewer components of clinical reasoning than WBA methods; see Chart 1). The defensibility of relying heavily on non-WBAs to determine clinical reasoning competence is questionable because part-task assessments cannot ensure successful transfer of skills into clinical practice. Several of these methods have extensive evidence of their relationship to other variables, as well as internal structure evidence, but others lack these forms of validity evidence. Consequences or outcomes on clinical practice performance are significant because non-WBAs are often used to make summative pass or fail judgments as well as licensing, certification, and credentialing decisions. Formative assessment for learning can occur when non-WBAs are used as progress tests and for the effect they have on the development of clinical reasoning (e.g., using concept maps to help develop cognitive networks).

Assessments in simulated clinical environments

Two methods were identified that occur in simulated clinical environments.

1. *Objective structured clinical examinations (OSCEs)* are performance-based evaluations of students' clinical skills including, but not exclusively focused on, clinical reasoning.[39,40] OSCEs comprise multiple stations where examinees execute different clinical tasks, incorporating SPs, observer ratings, written notes, and other methods, to provide a comprehensive assessment.
2. *Technology-enhanced simulation* describes a variety of assessment methods wherein learners physically interact with a tool or device that mimics clinical care.[41,42] These can encompass a range of instruments from static high-fidelity mannequins to virtual reality patient avatars that can change in response to learner input.

Assessments in simulated clinical environments typically use SPs, high-fidelity mannequins, or virtual patient avatars as stimuli. The response format for OSCEs and technology-enhanced simulations is usually task performance or constructed verbal or written responses. Scoring is often via itemized checklists that may be dichotomous (i.e., done or not done) or behaviorally anchored. Global rating scales are also common. OSCEs are used for both formative and high-stakes summative assessments (e.g., the United States Medical Licensing Examination Step 2 Clinical Skills and the Medical Council of Canada's Qualifying Examination Part 2), whereas technology-enhanced simulations are mainly used for formative assessments.

Validity considerations, feasibility issues, advantages, and disadvantages are detailed in Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/A634), but a few themes warrant highlighting. In terms of content validity, these methods can be blueprinted, and their alignment with clinical practice is reasonable (higher than most non-WBAs, yet less authentic than true WBAs). Highly organized, standardized, reproducible stations require attention to SP and rater training. There is greater ability to control contextual factors in these standardized environments than in assessments that occur during actual clinical practice. Blueprinting for these assessments must attend to content specificity and distinguish what essential features are required to pass (with clear anchors for global rating scales). Performance correlations with other assessment measures (i.e., non-WBAs and/or WBAs) are only low to moderate, which is acceptable for formative assessments but is less than desirable for high-stakes summative decisions. Assessments in simulated environments are valued for their ability to measure multiple clinical reasoning components (Chart 1), but a major practical problem is that they are resource-intensive to both develop and administer.

WBAs

Seven methods were identified that focus on assessments in authentic clinical environments or WBAs.

1. *Direct observation*, also known as performance or clinical observation, describes the presence of an observer

Chart 1
Strength of Assessment Methods for Measuring the Different Components of Clinical Reasoning[a]

Assessment method: Definition	Clinical reasoning component						
	IG	HG	PR	DD	LD	DJ	MT
Non-workplace-based assessments							
Clinical or comprehensive integrative puzzles: An extended matching crossword puzzle designed to assess a learner's ability to relate clinical vignettes to specific diagnoses and diagnostic or therapeutic interventions.	0.4	0.3	0.6	1.1	1.9	0.4	1.3
Concept maps: A schematic method for learners to organize and represent their knowledge and knowledge structures through a graphical illustration of the complex processes and relationships between concepts within a subject domain.	0.4	0.4	1.2	1.0	0.4	0.8	0.9
Extended matching questions: A written exam format consisting of a lead-in question (clinical vignette) followed by multiple answer options in a list where more answer options are given than in multiple-choice questions (i.e., > 5).	0.2	0.3	0.2	0.8	1.7	0.3	1.3
Key feature examinations: Problems typically consist of a clinical vignette followed by 2–3 questions that assess the critical elements ("key features") or challenging decisions that clinicians must make.	0.9	0.5	0.4	1.5	1.4	0.6	1.4
Multiple-choice questions: A clinical vignette is followed by up to 5 alternatives. Questions may take the following formats: single best alternative, matching, true or false, and combinations of alternatives.	0.9	0.3	0.0	0.6	1.9	0.0	1.8
Modified essay questions: A method wherein serial information about a clinical case is presented chronologically. After each item, the learner must document a decision. The student cannot preview subsequent items until a decision is made.	1.3	1.2	1.0	1.6	1.7	1.3	1.7
Oral examinations: A verbal examination conducted by one or more faculty members through unscripted or semiscripted questions that assess clinical reasoning and decision-making abilities, as well as professional values.	1.3	1.3	1.1	1.8	1.8	1.9	1.9
Patient management problems: A clinical scenario is presented in real-life settings with specific resources available for diagnosis or management. The learner chooses among multiple alternatives. The results of actions (e.g., labs, images) are provided.	1.6	1.0	0.3	1.4	1.9	0.6	1.7
Script concordance tests: Clinical scenarios with uncertainty are followed by a series of questions (e.g., if you are thinking X and you find Y, the answer becomes more likely, less likely, or no change). Responses are compared with those of experts.	0.4	0.8	0.6	0.8	1.3	0.9	1.1
Short- or long-answer (essay) questions: A clinical vignette is followed by one or more questions. Learners provide free-text responses that range in length from a few words to several sentences.	0.8	1.2	1.2	1.8	1.7	1.8	1.7
Assessments in simulated clinical environments							
Objective structured clinical examinations: Performance-based evaluations comprising multiple stations where examinees execute different clinical tasks, incorporating standardized patients, observer ratings, written notes, etc.	2.0	1.3	1.3	1.8	1.7	1.3	1.7
Technology-enhanced simulation: An educational tool or device with which the learner physically interacts to mimic an aspect of clinical care. Tools range from high-fidelity mannequins to dynamic virtual reality patients.	1.6	0.6	0.5	1.1	1.7	0.6	1.9
Workplace-based assessments							
Chart-stimulated recall: A hybrid assessment format that combines review of a written note from an actual patient encounter and an oral examination to probe the learner's underlying thought processes, with feedback to improve decision making.	1.1	1.2	1.4	2.0	1.9	1.9	2.0
Direct observation: A method that involves an instructor watching a learner in the workplace environment. Assessment tools for this include the mini-clinical evaluation exercise (mini-CEX).	1.9	1.1	1.5	1.7	1.7	1.5	1.6
Global assessment: Individual judgment or preceptor gestalt of learner clinical reasoning performance, often expressed on clinical rating forms (e.g., end-of-shift, end-of-clerkship).	1.3	1.0	1.3	1.6	1.6	1.4	1.6
Oral case presentation: A structured verbal report of a clinical case. The learner makes deliberate choices about what to include, what not to include, the order in which data are presented, and the structure and content of the assessment and plan.	1.1	1.1	1.8	2.0	2.0	2.0	1.9
Self-regulated learning microanalysis: A structured interview protocol designed to gather in-the-moment, task-level information on a learner's thoughts, actions, and feelings as they approach, perform, and reflect on a clinical activity.	1.4	1.6	1.6	1.7	1.7	1.6	1.7
Think aloud: A method in which participants are given a task and asked to voice their thoughts in an unfiltered form while completing or immediately after completing the task.	1.4	1.8	1.7	1.8	1.7	1.9	1.6
Written notes: A structured written report about a patient case. Postencounter notes are one specific format with expectations for expressing clinical reasoning in the form of a summary statement, problem list, prioritized differential diagnosis, etc.	1.2	0.6	1.4	1.9	1.7	1.6	2.0

Abbreviations: IG indicates information gathering; HG, hypothesis generation; PR, problem representation; DD, differential diagnosis; LD, leading diagnosis; DJ, diagnostic justification; MT, management and treatment.

[a]From a 2016 scoping review of clinical reasoning assessment methods. The strength of each assessment method is indicated by shading: black indicates poor (0.0–0.5); dark gray, average (0.6–1.0); light gray, good (1.1–1.5); white, very good (1.6–2.0).

(typically a faculty member) who collects data about learners in authentic clinical contexts.[43] A variety of assessment tools have been used for direct observation[43] (e.g., the mini-clinical evaluation exercise [mini-CEX]),[44] though they are not all explicitly designed to assess clinical reasoning.

2. *Global assessments* are common components of faculty evaluation forms.[45,46] They capture individual judgments or preceptor gestalt about clinical reasoning performance based on direct or indirect observations.
3. *OCPs* are structured verbal reports of clinical cases.[47,48] Evidence of a learner's diagnostic and therapeutic reasoning is assessed as the learner makes deliberate choices about what to include or exclude, data organization, and the structure and content of the assessment and plan. Raters can probe learners for understanding and additional information.
4. *Written notes* are another means of communicating clinical information about a case in a structured way—in this case, via a written report.[49] They may be assessed by using one of a variety of tools (e.g., postencounter notes,[50] the IDEA [interpretive summary, differential diagnosis, explanation of reasoning, and alternatives] assessment tool[51]). Similar to OCPs, clinical reasoning may be assessed from multiple features of a note, particularly the summary statement (an encapsulation of the case containing key features and semantic qualifiers), problem list, prioritization of the differential diagnosis, justification, and management plan.
5. *CSR* is a hybrid format consisting of clinical documentation review from an actual clinical encounter, an oral examination where an evaluator probes underlying thought processes, and feedback that may include action plans to improve future diagnostic decision making.[52,53]
6. *Think aloud (TA)* is a technique where learners are given a discrete task and asked to voice the unfiltered thoughts they have or had while performing the work.[54,55] TAs are typically administered while completing the task (simultaneous) but may also be performed immediately following task completion (delayed).
7. *SRL-M* describes a structured interview protocol designed to gather in-the-moment, task-level information about learners' thoughts, actions, and feelings as they approach, perform, and reflect on a clinical activity that has a beginning, middle, and end.[56,57] Combined with features of the TA, it can assess metacognition.

WBA methods rely on real patients as stimuli. Response formats for these methods include clinical performance with patients (direct observation, global assessment) or constructed verbal or written free text (OCPs, written notes, CSR, TA, SRL-M). Scoring mechanisms vary widely and include itemized or global rating scales of various types (norm referenced, criterion referenced, entrustment scales, supervision scales), as well as checklists, etc. WBAs are most commonly used for formative assessment during clinical clerkships and residency. When they are used to make summative decisions, multiple observations or global assessments are typically aggregated. The workhorses of WBAs are direct observation (e.g., mini-CEX), which is typically used for formative assessments; and global assessments, which are typically used for end-of-rotation summative assessments during clinical clerkships and residency rotations. Oral presentations and written notes may influence a faculty rater's final global assessment but are infrequently used for high-stakes assessments. TA and SRL-M are typically more involved in research contexts but have been used for the remediation of struggling learners.[58,59]

The details of validity considerations, feasibility issues, advantages, and disadvantages of WBA methods are summarized in Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/A634), but we will highlight a few themes here. A great strength of WBAs is their ability to measure multiple components of clinical reasoning (Chart 1). Because these methods are embedded in authentic clinical environments, there is reasonable content and response process validity evidence. The nonsystematic nature of clinical practice, however, can present challenges with regard to content coverage and over- or underrepresentation of certain clinical problems. Internal structure evidence (e.g., item analysis data, score scale reliability, standard errors of measurement) is problematic in that many of these methods require an observer (faculty member) to quantify their observation of a complex behavior into a small number of assessment outcomes. Biases and inconsistencies are inherent in this judgment process.[60–62] A key strategy to reduce these threats to validity is to ensure an adequate number of observations across a diverse set of clinical problems by multiple raters over time. The defensibility of using WBAs for summative pass/fail and remediation decisions is questionable without this because, from a generalizability theory perspective, 12 to 14 mini-CEXs are needed to reach acceptable reliability for judgments. Challenges to implementing WBAs include time, faculty development, accountability, and recognition for faculty who engage in these assessments, as clinical environments often value productivity over the supervision and evaluation of trainees.

Discussion

This review summarizes the currently available menu of clinical reasoning assessment methods and highlights validity considerations, feasibility issues, advantages, and disadvantages for each. Chart 1 and Supplemental Digital Appendix 4 (at http://links.lww.com/ACADMED/A634) in particular can help inform the construction of *programs of assessment.*[63] Educators can select from a number of different but complementary clinical reasoning assessment methods, each with different validity considerations. Practical guidance based on our findings is given in List 1.

The value of the existing menu of clinical reasoning assessment methods can perhaps best be understood through the lens of competency-based education. If medical educators want to ensure that learners are competent in clinical reasoning, they must provide robust assessment of *all* components of clinical reasoning[12] (see Table 1). Further, they must also arrange for adequate sampling. This can only be accomplished by employing multiple assessment methods.[63]

A close look at Chart 1 demonstrates that many forms of non-WBAs in common use (MCQs, EMQs, KFEs, SCTs) are only

List 1

Practical Guidance for Clinical Reasoning Assessment From a 2016 Scoping Review of Clinical Reasoning Assessment Methods

- Multiple assessment methods (i.e., non-WBAs, assessments in simulated clinical environments, and WBAs) should be used as part of a clinical reasoning assessment program.
- Many individual assessment methods can obtain adequate reliability for high-stakes assessment (≥ 0.8) with an adequate number of items or cases, broad sampling, and sufficient testing time.
- To ensure competence, a large number of assessments are needed, administered longitudinally, that cover a variety of clinical problems in diverse settings to accommodate content and context specificity.
- Methods should be chosen based on coverage of the different components of clinical reasoning, validity, feasibility, defensibility, and fit for the purpose of the assessment.
- Whole- and part-task assessment methods (i.e., those that cover all versus a few components of clinical reasoning) used together can ensure measurement of the whole construct *and* adequate sampling.
- Non-WBAs (e.g., MCQs, EMQs, KFEs) have the advantage of broad sampling, blueprinting, control, and consistency. They can also assess accuracy.
- MCQs and KFEs have the best validity evidence regarding content, internal structure, and consequences or outcomes on clinical practice performance; however, they have significant issues with cueing when it comes to response process.
- Non-WBAs measure a more limited number of components of clinical reasoning compared with simulations and WBAs, which tend to measure more of the whole task.
- WBAs are embedded in actual clinical practice, lending authenticity to content and response process validity; however, content coverage is not systematic.
- The defensibility of using WBAs for summative decisions is questionable because, from a generalizability theory perspective, a large number of measurements are needed to reach acceptable reliability for judgments. Ensuring evaluation by multiple raters over time is also essential for WBAs.
- Whole-task clinical reasoning assessments (i.e., those that cover the full range of tasks from information gathering to differential diagnosis to management and treatment) are essential for formative feedback and assessment for learning.
- Assessments in simulated clinical environments and WBAs are essential parts of any comprehensive assessment strategy because they ensure that learners are assessed on the whole task, though they are time- and resource-intensive to develop and administer.

Abbreviations: WBAs indicates workplace-based assessments; MCQs, multiple-choice questions; EMQs, extended matching questions; KFEs, key feature examinations.

poor to average at assessing information gathering, hypothesis generation, and problem representation. Their strengths lie more in assessing differential diagnosis, leading diagnosis, and management and treatment. Assessments in simulated clinical environments and WBAs are better at assessing information gathering, with direct observation and OSCEs being the strongest in this domain. SRL-M and TA strategies are effective tools for measuring hypothesis generation and problem representation because they force learners to articulate these otherwise hidden steps in the reasoning process.[64] By carefully combining strategies that are strong at assessing the different components of clinical reasoning (e.g., MCQs + SRL-M + OSCEs), educators can begin to ensure assessment of all components of the larger competency.

Of course, clinical reasoning competence as a "whole" is more than the sum of its "parts."[65] When constructing an assessment program, it is necessary, but not sufficient, to ensure assessment of all components of clinical reasoning. Whole-task assessments (i.e., those that cover the full range of clinical reasoning) are needed to ensure that learners can transfer skills into clinical practice,[66] while part-task assessments are needed to achieve broad sampling. Combinations of whole- and part-task assessments (e.g., direct observations, OSCEs, and global assessments combined with MCQs, KFEs, and EMQs) can form a foundation for a program of assessment.

Educators must also consider the validity, feasibility, and defensibility of assessments when choosing among methods. Looking at Chart 1, one might conclude that if assessors predominately used WBAs, they would obtain robust coverage of all components of clinical reasoning in authentic clinical environments and easily be able to deem a learner competent. Although WBAs are critically important and deserve greater emphasis in current competency-based educational programs,[67,68] the limitations of an exclusively WBA approach to assessing clinical reasoning rest in the problem of content and context specificity because feasibility and cost (with regard to faculty time and money) often limit the number and variety of cases that can be sampled. Seen in this light, non-WBAs (e.g., MCQs, EMQs, KFEs) add important value to a program of clinical reasoning assessment by ensuring broad sampling, while lessening issues of context specificity and providing opportunities for blueprinting, control, consistency, and accuracy. Thus, for validity and feasibility reasons, it is critical to have a balance of non-WBAs, assessments in simulated clinical environments, and WBAs in any assessment program.

Creating such a balance can be challenging depending on the educational context. For example, undergraduate medical education programs often use a combination of MCQs, OSCEs, global assessments, oral presentations, and written notes to assess reasoning. These programs may wish to improve the use of certain methods, such as direct observation, while also incorporating novel methods, such as TA or SRL-M to get at components of clinical reasoning that may be currently underassessed. In graduate medical education, the bulk of learning and assessment happens in the clinical environment, augmented occasionally by technology-enhanced simulation and in-training examinations, which are largely comprised of MCQs. Incorporating a wider range of assessment methods, improving on assessment methods currently in use, and training raters on tools in busy clinical settings will be daunting. As WBAs improve, it may be possible that these more holistic assessments can predominate, and non-WBAs can be used largely for situations of uncertainty and remediation; however, much research is still needed to make this transition effectively.

Ultimately, institutions must ensure that their programs of assessment offer complete coverage of the components of clinical reasoning (Table 1 and Chart 1).

Programs will need to use both whole- and part-task methods as well as provide a balanced representation of methods with regard to various threats to validity (see Supplemental Digital Appendix 4 at http://links.lww.com/ACADMED/A634). Programmatic assessment for clinical reasoning is still a nascent concept at many institutions, yet this is where this review suggests the field needs to move in the future. Institutions need to conduct frequent assessments of clinical reasoning, gathering information longitudinally from multiple sources, using multiple methods, across various contexts or settings. This is challenging in the real world because of time and the necessity of faculty development, yet it is critical for the defensibility of an assessment program when making high-stakes summative decisions or competency determinations. It is also critical to ensure patient safety.[69] Whether our current assessment practices strike the right balance of non-WBAs, assessments in simulated clinical environments, and WBAs is debatable but beyond the scope of this review to fully address.

Although our discussion has largely focused on determining clinical reasoning competency and assessment *of* learning, it is also important to consider assessment *for* learning. While many of the same principles apply, assessment for learning is more formative and may employ methods that have a different range of validity evidence because of their high value for learning and teaching the clinical reasoning process (i.e., the method is fit for the purpose). For example, CIPs and concept maps have great utility *for* learning in that they help students develop illness scripts and form connections, facilitating the development of coding and retrieval networks, which are thought to be the basis of diagnostic expertise.[70,71] Whole-task clinical reasoning assessments, such as direct observations and technology-enhanced simulations, are essential means of obtaining formative feedback, even if they are not well suited for making summative judgments without large numbers of observations.

Our review had several limitations. The currency of the review was impacted by the time required to analyze all the references uncovered in our search. Thus, some new developments may not be included. However, our comprehensive search process makes it unlikely that we missed assessment methods that have significant usage or evidence. During the scoping process, we made decisions relatively late in the process not to include articles from other health professions, largely for pragmatic reasons. When constructing the appendixes, we had to make judgments concerning the advantages, disadvantages, and feasibility of different methods, which were not always explicitly addressed in the articles included in the review.

Although we have made some suggestions on how to combine various types of assessment methods, we need future studies that rigorously evaluate such assessment programs as opposed to only evaluating the validity of the individual tools. Defining the prevalence of use of assessment methods and gaps in current practice was beyond the scope of this review, but it is an area ripe for future research.

The importance of clinical reasoning as a physician competency mandates rigor and innovation in the assessment of it. This review demonstrates that there has been considerable innovation in clinical reasoning assessment methods, but there remains much work to be done. We hope this collated resource will help educators become more aware of the existing menu of clinical reasoning assessment methods and how to choose among them. We emphasize the need for ongoing and rigorous gathering of validity evidence to guide improvements in each of these methods. Future research is also needed to determine how to best combine various methods into valid programs of clinical reasoning assessment to allow medical schools, residency programs, and licensing boards to confidently determine the competence of their learners.

Acknowledgments: The authors would like to thank their librarians, Nancy Allee, Donna Berryman, Elizabeth Richardson, and Whitney Townsend, for their expertise and support. They would also like to thank the innumerable other contributors to this multiyear project and the reviewers for their constructive suggestions for improving this article.

Funding/Support: None reported.

Other disclosures: E. Holmboe works for the Accreditation Council for Graduate Medical Education and receives royalties for a textbook on assessment from Elsevier.

Ethical approval: Reported as not applicable.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of their universities, the U.S. Navy, the Department of Defense, or the United States Government.

Previous presentations: A portion of this work was accepted for presentation as a workshop entitled Clinical Reasoning Assessment Tools: So Many Methods How to Choose at the Association of American Medical Colleges 2018 Annual Meeting; Austin, Texas; November 6, 2018.

M. Daniel is assistant dean for curriculum and associate professor of emergency medicine and learning health sciences, University of Michigan Medical School, Ann Arbor, Michigan; ORCID: http://orcid.org/0000-0001-8961-7119.

J. Rencic is associate program director of the internal medicine residency program and associate professor of medicine, Tufts University School of Medicine, Boston, Massachusetts; ORCID: http://orcid.org/0000-0002-2598-3299.

S.J. Durning is director of graduate programs in health professions education and professor of medicine and pathology, Uniformed Services University of the Health Sciences, Bethesda, Maryland.

E. Holmboe is senior vice president of milestone development and evaluation, Accreditation Council for Graduate Medical Education, and adjunct professor of medicine, Northwestern Feinberg School of Medicine, Chicago, Illinois; ORCID: http://orcid.org/0000-0003-0108-6021.

S.A. Santen is senior associate dean and professor of emergency medicine, Virginia Commonwealth University, Richmond, Virginia; ORCID: http://orcid.org/0000-0002-8327-8002.

V. Lang is associate professor of medicine, University of Rochester School of Medicine and Dentistry, Rochester, New York; ORCID: http://orcid.org/0000-0002-2157-7613.

T. Ratcliffe is associate professor of medicine, University of Texas Long School of Medicine at San Antonio, San Antonio, Texas.

D. Gordon is medical undergraduate education director, associate residency program director of emergency medicine, and associate professor of surgery, Duke University School of Medicine, Durham, North Carolina.

B. Heist is clerkship codirector and assistant professor of medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania.

S. Lubarsky is assistant professor of neurology, McGill University, and faculty of medicine and core member, McGill Center for Medical Education, Montreal, Quebec, Canada; ORCID: http://orcid.org/0000-0001-5692-1771.

C.A. Estrada is staff physician, Birmingham Veterans Affairs Medical Center, and director, Division of General Internal Medicine, and professor of medicine, University of Alabama, Birmingham, Alabama; ORCID: https://orcid.org/0000-0001-6262-7421.

T. Ballard is plastic surgeon, Ann Arbor Plastic Surgery, Ann Arbor, Michigan.

A.R. Artino Jr is deputy director for graduate programs in health professions education and professor of medicine, preventive medicine, and biometrics pathology, Uniformed Services University of the Health Sciences, Bethesda, Maryland; ORCID: http://orcid.org/0000-0003-2661-7853.

A. Sergio Da Silva is senior lecturer in medical education and director of the masters in medical education program, Swansea University Medical School, Swansea, United Kingdom; ORCID: http://orcid.org/0000-0001-7262-0215.

T. Cleary is chair, Applied Psychology Department, CUNY Graduate School and University Center, New York, New York, and associate professor of applied and professional psychology, Rutgers University, New Brunswick, New Jersey.

J. Stojan is associate professor of internal medicine and pediatrics, University of Michigan Medical School, Ann Arbor, Michigan.

L.D. Gruppen is director of the master of health professions education program and professor of learning health sciences, University of Michigan Medical School, Ann Arbor, Michigan; ORCID: http://orcid.org/0000-0002-2107-0126.

References

1 Young M, Thomas A, Lubarsky S, et al. Drawing boundaries: The difficulty in defining clinical reasoning. Acad Med. 2018;93:990–995.
2 Eva KW, Hatala RM, Leblanc VR, Brooks LR. Teaching from the clinical reasoning literature: Combined reasoning strategies help novice diagnosticians overcome misleading information. Med Educ. 2007;41:1152–1158.
3 Gruppen LD. Clinical reasoning: Defining it, teaching it, assessing it, studying it. West J Emerg Med. 2017;18:4–7.
4 Durning SJ, Artino AR. Situativity theory: A perspective on how participants and the environment can interact: AMEE guide no. 52. Med Teach. 2011;33:188–199.
5 Kahneman D. Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux; 2011.
6 Gruppen LD, Frohna AZ. Clinical reasoning. In: Norman GR, van der Vleuten CPM, Newble DI, eds. International Handbook of Research in Medical Education. Vol 1. Dordrecht, The Netherlands: Kluwer Academic; 2002:205–230.
7 Young ME, Dory V, Lubarsky S, Thomas A. How different theories of clinical reasoning influence teaching and assessment. Acad Med. 2018;93:1415.
8 Ratcliffe TA, Durning SJ. Theoretical concepts to consider in providing clinical reasoning instruction. In: Trowbridge RL, Rencic JA, Durning SJ, eds. Teaching Clinical Reasoning. Philadelphia, PA: American College of Physicians; 2015:13–30.
9 Accreditation Council for Graduate Medical Education. ACGME common program requirements. http://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf. Revised July 1, 2017. Accessed January 4, 2019.
10 Royal College of Physicians and Surgeons of Canada. CanMEDS: Better standards, better physicians, better care. http://www.royalcollege.ca/rcsite/canmeds/canmeds-framework-e. Accessed January 4, 2019.
11 The Tuning Project (Medicine). Learning outcomes/competencies for undergraduate medical education in Europe. http://www.unideusto.org/tuningeu/images/stories/Summary_of_outcomes_TN/Learning_Outcomes_Competences_for_Undergraduate_Medical_Education_in_Europe.pdf. Accessed January 4, 2019.
12 Ilgen JS, Humbert AJ, Kuhn G, et al. Assessing diagnostic reasoning: A consensus statement summarizing theory, practice, and future needs. Acad Emerg Med. 2012;19:1454–1461.
13 Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc. 2015;13:141–146.
14 Colquhoun HL, Levac D, O'Brien KK, et al. Scoping reviews: Time for clarity in definition, methods, and reporting. J Clin Epidemiol. 2014;67:1291–1294.
15 Kirkpatrick D. Evaluation of training. In: Craig RL, Bittel LR, eds. Training and Development Handbook. New York, NY: McGraw-Hill; 1967:87–112.
16 Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6:e1000097.
17 Adams MH, Whitlow JF, Stover LM, Johnson KW. Critical thinking as an educational outcome: An evaluation of current tools of measurement. Nurse Educ. 1996;21:23–32.
18 Messick S. Validity. In: Linn RL, ed. Educational Measurement. 3rd ed. New York, NY: Macmillan; 1989:13–103.
19 American Psychological Association; National Council on Measurement in Education; Joint Committee on Standards for Educational and Psychological Testing. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.
20 Norcini JJ, Swanson DB, Grosso LJ, Shea JA, Webster GD. A comparison of knowledge, synthesis, and clinical judgment. Multiple-choice questions in the assessment of physician competence. Eval Health Prof. 1984;7:485–499.
21 Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Acad Med. 1994;69(10 suppl):S1–S3.
22 Beullens J, Struyf E, Van Damme B. Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. Med Educ. 2006;40:1173–1179.
23 Day SC, Norcini JJ, Diserens D, et al. The validity of an essay test of clinical judgment. Acad Med. 1990;65(9 suppl):S39–S40.
24 de Graaff E, Post GJ, Drop MJ. Validation of a new measure of clinical problem-solving. Med Educ. 1987;21:213–218.
25 Feletti GI. Reliability and validity studies on modified essay questions. J Med Educ. 1980;55:933–941.
26 Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper. BMC Med Educ. 2007;7:49.
27 McCarthy WH, Gonnella JS. The simulated patient management problem: A technique for evaluating and teaching clinical competence. Br J Med Educ. 1967;1:348–352.
28 Newble DI, Hoare J, Baxter A. Patient management problems. Issues of validity. Med Educ. 1982;16:137–142.
29 Page G, Bordage G. The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. Acad Med. 1995;70:104–110.
30 Hrynchak P, Takahashi SG, Nayer M. Key-feature questions for assessment of clinical reasoning: A literature review. Med Educ. 2014;48:870–883.
31 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47:1175–1183.
32 Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. Eval Health Prof. 2004;27:304–319.
33 Ber R. The CIP (comprehensive integrative puzzle) assessment method. Med Teach. 2003;25:171–176.
34 Capaldi VF, Durning SJ, Pangaro LN, Ber R. The clinical integrative puzzle for teaching and assessing clinical reasoning: Preliminary feasibility, reliability, and validity evidence. Mil Med. 2015;180(4 suppl):54–60.
35 Pottier P, Hardouin JB, Hodges BD, et al. Exploring how students think: A new method combining think-aloud and concept mapping protocols. Med Educ. 2010;44:926–935.
36 Daley BJ, Torre DM. Concept maps in medical education: An analytical literature review. Med Educ. 2010;44:440–448.
37 Anastakis DJ, Cohen R, Reznick RK. The structured oral examination as a method for assessing surgical residents. Am J Surg. 1991;162:67–70.
38 Wass V, Wakeford R, Neighbour R, Van der Vleuten C; Royal College of General Practitioners. Achieving acceptable reliability in oral examinations: An analysis of the Royal College of General Practitioners membership examination's oral component. Med Educ. 2003;37:126–131.
39 Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: An historical and theoretical perspective. Med Teach. 2013;35:e1437–e1446.
40 Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part II: Organisation & administration. Med Teach. 2013;35:e1447–e1463.
41 Ilgen JS, Sherbino J, Cook DA. Technology-enhanced simulation in emergency medicine: A systematic review and meta-analysis. Acad Emerg Med. 2013;20:117–127.
42 Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. Acad Med. 2013;88:872–883.
43 Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. JAMA. 2009;302:1316–1326.
44 Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: A method for assessing clinical skills. Ann Intern Med. 2003;138:476–481.

45 Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011;45:560–569.
46 Shea JA, Norcini JJ, Kimball HR. Relationships of ratings of clinical competence and ABIM scores to certification status. Acad Med. 1993;68(10 suppl):S22–S24.
47 Lewin LO, Beraho L, Dolan S, Millstein L, Bowman D. Interrater reliability of an oral case presentation rating tool in a pediatric clerkship. Teach Learn Med. 2013;25:31–38.
48 Bordage G, Connell KJ, Chang RW, Gecht MR, Sinacore JM. Assessing the semantic content of clinical case presentations: Studies of reliability and concurrent validity. Acad Med. 1997;72(10 suppl 1):S37–S39.
49 Durning SJ, Artino A, Boulet J, et al. The feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. Med Teach. 2012;34:30–37.
50 Park YS, Lineberry M, Hyderi A, Bordage G, Riddle J, Yudkowsky R. Validity evidence for a patient note scoring rubric based on the new patient note format of the United States Medical Licensing Examination. Acad Med. 2013;88:1552–1557.
51 Baker EA, Ledford CH, Fogg L, Way DP, Park YS. The IDEA assessment tool: Assessing the reporting, diagnostic reasoning, and decision-making skills demonstrated in medical students' hospital admission notes. Teach Learn Med. 2015;27:163–173.
52 Goulet F, Jacques A, Gagnon R, Racette P, Sieber W. Assessment of family physicians' performance using patient charts: Interrater reliability and concordance with chart-stimulated recall interview. Eval Health Prof. 2007;30:376–392.
53 Schipper S, Ross S. Structured teaching and assessment: A new chart-stimulated recall worksheet for family medicine residents. Can Fam Physician. 2010;56:958–959, e352–e354.
54 Chatterjee S, Ng J, Kwan K, Matsumoto ED. Assessing the surgical decision making abilities of novice and proficient urologists. J Urol. 2009;181:2251–2256.
55 Sibbald M, de Bruin AB. Feasibility of self-reflection as a tool to balance clinical reasoning strategies. Adv Health Sci Educ Theory Pract. 2012;17:419–429.
56 Cleary TJ, Callan GL, Zimmerman BJ. Assessing self-regulation as a cyclical, context-specific phenomenon: Overview and analysis of SRL microanalysis protocols. Educ Res Int. 2012;428639:1–19.
57 Artino AR Jr, Cleary TJ, Dong T, Hemmer PA, Durning SJ. Exploring clinical reasoning in novices: A self-regulated learning microanalytic assessment approach. Med Educ. 2014;48:280–291.
58 Durning SJ, Cleary TJ, Sandars J, Hemmer P, Kokotailo P, Artino AR. Perspective: Viewing "strugglers" through a different lens: How a self-regulated learning perspective can help medical educators with assessment and remediation. Acad Med. 2011;86:488–495.
59 Andrews MA, Kelly WF, DeZee KJ. Why does this learner perform poorly on tests? Using self-regulated learning theory to diagnose the problem and implement solutions. Acad Med. 2018;93:612–615.
60 Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. Acad Med. 2010;85(10 suppl):S25–S28.
61 Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: Rethinking the etiology of rater errors. Acad Med. 2011;86(10 suppl):S1–S7.
62 Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the "black box" differently: Assessor cognition from three research perspectives. Med Educ. 2014;48:1055–1068.
63 van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012;34:205–214.
64 Cleary TJ, Durning SJ, Artino AR Jr. Microanalytic assessment of self-regulated learning during clinical reasoning tasks: Recent developments and next steps. Acad Med. 2016;91:1516–1521.
65 Schuwirth L, van der Vleuten C, Durning SJ. What programmatic assessment in medical education can learn from healthcare. Perspect Med Educ. 2017;6:211–215.
66 Vandewaetere M, Manhaeve D, Aertgeerts B, Clarebout G, Van Merriënboer JJ, Roex A. 4C/ID in medical education: How to design an educational program based on whole-task learning: AMEE guide no. 93. Med Teach. 2015;37:4–20.
67 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;90:1084–1092.
68 Gruppen LD, Ten Cate O, Lingard LA, Teunissen PW, Kogan JR. Enhanced requirements for assessment in a competency-based, time-variable medical education system. Acad Med. 2018;93(3 suppl):S17–S21.
69 Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf. 2013;22(2 suppl):ii21–ii27.
70 Gomes AP, Dias-Coelho UC, Cavalheiro PDO, Siqueira-Batista R. The role of concept maps in the medical education. Rev Br Educ Med. 2011;35:275–282.
71 Groothoff JW, Frenkel J, Tytgat GA, Vreede WB, Bosman DK, ten Cate OT. Growth of analytical thinking skills over time as measured with the MATCH test. Med Educ. 2008;42:1037–1043.

References cited in Table 1 only

72 Gruppen LD, Wolf FM, Billi JE. Information gathering and integration as sources of error in diagnostic decision making. Med Decis Making. 1991;11:233–239.
73 Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: A narrative review and a proposal. Med Educ. 2015;49:961–973.
74 Krupat E, Wormwood J, Schwartzstein RM, Richards JB. Avoiding premature closure and reaching diagnostic accuracy: Some key predictive factors. Med Educ. 2017;51:1127–1137.
75 Pelaccia T, Tardif J, Triby E, et al. Insights into emergency physicians' minds in the seconds before and into a patient encounter. Intern Emerg Med. 2015;10:865–873.
76 Cutrer WB, Sullivan WM, Fleming AE. Educational strategies for improving clinical reasoning. Curr Probl Pediatr Adolesc Health Care. 2013;43:248–257.
77 Monteiro SD, Sherbino JD, Ilgen JS, et al. Disrupting diagnostic reasoning: Do interruptions, instructions, and experience affect the diagnostic accuracy and response time of residents and emergency physicians? Acad Med. 2015;90:511–517.
78 Graber ML, Tompkins D, Holland JJ. Resources medical students use to derive a differential diagnosis. Med Teach. 2009;31:522–527.
79 Stojan JN, Daniel M, Morgan HK, Whitman L, Gruppen LD. A randomized cohort study of diagnostic and therapeutic thresholds in medical student clinical reasoning. Acad Med. 2017;92(11 suppl):S43–S47.
80 Williams RG, Klamen DL, Markwell SJ, Cianciolo AT, Colliver JA, Verhulst SJ. Variations in senior medical student diagnostic justification ability. Acad Med. 2014;89:790–798.
81 Goldszmidt M, Minda JP, Bordage G. Developing a unified list of physicians' reasoning tasks during clinical encounters. Acad Med. 2013;88:390–397.