RBEM
Revista Brasileira de Educação Médica

abem associação brasileira de educação médica

REVIEW ARTICLE

DOI: https://doi.org/10.1590/1981-5271v49.3-2024-0089.ING

Evaluation of clinical reasoning in physicians and Medical students: an integrative review

Avaliação do raciocínio clínico de médicos e estudantes de Medicina: uma revisão integrativa

Maurício Prätzel Ellwanger[1] ellwangermp@gmail.com
Fernando Tureck[1] fernandotureck@gmail.com

ABSTRACT

Introduction: The development of clinical reasoning is crucial for medical students and practitioners, as it enables quality medical practice and rational case investigation. This process combines intuitive and analytical reasoning, adapting to the complexity of cases. During medical training, students progress from formulating diagnostic hypotheses to developing "scripts" of diseases. With the technological advancements in medicine, understanding how clinical reasoning adapts is essential to ensuring high-quality healthcare.

Objective: This study aims to analyze clinical reasoning assessment practices, identifying gaps and areas for improvement in evaluation methods to promote more effective curricula.

Method: This study reviews the primary methods of assessing clinical reasoning in doctors and medical students. Using an integrative review methodology, articles and essays published in the last twenty years, in any language, that discussed or compared assessment tools were selected.

Result: Of a total of 6,150 initially identified studies, after removing duplicates and evaluating titles and abstracts, 244 were selected. These articles underwent full reading, and a total of 122 articles were included in the review. The most commonly used methods are multiple-choice questions (MCQs), script concordance testing (SCT), and objective structured clinical examinations (OSCE), with over ten different assessment formats found and discussed.

Conclusion: Assessment instruments, besides providing objective grades for students, shape behavior and are thus fundamental for the increasingly better training of doctors. Numerous methods are currently being researched, while others, such as the script concordance test, are being widely implemented. This study contributes to the effort of compiling all existing literature data and discovering which methods, in aggregate, are considered the most reliable and valid.

Keywords: Clinical Reasoning; Education, Medical; Anthropometry; Clinical Decision-Making; Review.

RESUMO

Introdução: *O desenvolvimento do raciocínio clínico é crucial para estudantes e médicos, pois permite uma prática médica de qualidade e investigação racional de casos. Esse processo combina raciocínio intuitivo e analítico, adaptando-se à complexidade dos casos. Durante a formação médica, os alunos passam por fases de elaboração de hipóteses diagnósticas até o desenvolvimento de scripts das doenças. Com o avanço tecnológico na medicina, compreender como o raciocínio clínico se adapta é essencial para garantir cuidados de saúde de qualidade.*

Objetivo: *Este estudo visa analisar práticas de avaliação do raciocínio clínico, identificando lacunas e áreas de melhoria nos métodos avaliativos para promover currículos mais eficazes.*

Método: *Este estudo revisa as principais formas de avaliação do raciocínio clínico em médicos e estudantes de Medicina. Utilizando uma metodologia de revisão integrativa, foram selecionados artigos e ensaios publicados nos últimos 20 anos, em qualquer idioma, que abordavam comparações ou discussões sobre instrumentos de avaliação.*

Resultado: *Do total de 6.150 trabalhos inicialmente identificados, após a eliminação de duplicatas e a avaliação dos títulos e resumos, selecionaram-se 244 artigos que foram submetidos à leitura completa. Na revisão, mantiveram-se 122 artigos. Os métodos mais utilizados são questões de múltipla escolha (QME), teste de concordância de scripts (TCS) e exame clínico objetivo estruturado (OSCE), e mais de dez formatos de avaliação foram encontrados e discutidos.*

Conclusão: *O instrumento de avaliação, além de poder providenciar notas objetivas para os alunos, molda o comportamento e, portanto, é fundamental para a formação cada vez melhor de médicos. Inúmeros métodos têm sido pesquisados no momento, e outros, como o TCS, estão sendo implementados em larga escala. Este estudo contribui para o esforço de compilar todos os dados existentes da literatura e descobrir quais métodos, no agregado, são considerados mais confiáveis e válidos.*

Palavras-chave: *Raciocínio Clínico; Educação Médica; Avaliação; Tomada de Decisão Clínica; Revisão.*

[1] Universidade do Contestado, Mafra, Santa Catarina, Brazil

Chief Editor: Rosiane Viana Zuza Diniz.
Associate Editor: Mauricio Peixoto.

Received on 04/16/24; Accepted on 04/04/25.

Evaluated by double blind review process.

INTRODUCTION

Clinical reasoning is a complex concept, defined as the process by which physicians integrate theory and practice to formulate diagnostic hypotheses and plan interventions. This process involves data collection, hypothesis review, and continuous adaptation until confidence in the decision is achieved. Clinical reasoning is influenced by context, experience, and institutional factors, making decision-making complex and uncertain.[1,2]

Evaluating clinical reasoning in medical education is challenging. Structured methods are essential to train physicians capable of making informed decisions. Daniel et al. (2019), in their scoping review, explored a wide variety of evaluation strategies applied to physicians and medical students, highlighting everything from simulations to practical performance evaluations.[138] This study aims to complement this approach, focusing on the rigorous selection of studies with high methodological control, aiming to expand the analysis of how these tools impact clinical training.

Errors in clinical reasoning can lead to misdiagnosis, inappropriate treatments, and high costs to the healthcare system.[3] Statistics show that the rate of diagnostic error in medical emergencies, for example, can reach 10-15%, and that many of these errors are of cognitive origin, involving biases such as premature closure (stopping looking for new hypotheses once the diagnosis is made) and the search for confirmation.[4] Clinical reasoning can be understood from double processing models, with System 1 (intuitive and non-analytical reasoning) being fast and pattern-based, and System 2 (analytical reasoning) being slower, more deliberate, and logical.[3] Intuitive clinical reasoning can be advantageous in situations of high familiarity with the clinical case, but it is also susceptible to cognitive biases, such as availability and representativeness.[4] Analytical reasoning, on the other hand, although more exhaustive, can be limited by the capacity of working memory, generating errors due to the excess of information.[3] Both types of reasoning are present in clinical practice, and educational strategies that promote the combination of the two processes have shown moderate benefits in error reduction.[3]

Within this context, clinical reasoning emerges as fundamental to effectively deal with diagnostic errors. Recognizing its crucial role, internationally recognized research and science institutions have advocated for improvements in both clinical reasoning teaching and assessment, recognizing its importance in medical education. However, despite this recognition, deficiencies in the understanding of clinical reasoning concepts persist among many medical students and residents, highlighting possible gaps in both undergraduate and graduate medical studies[5].

Miller's pyramid describes four levels of clinical competence: 'Knows', 'Knows How', 'Shows How' and 'Does', highlighting the progression from theoretical knowledge to practice. Evaluations must follow this evolution, testing everything from factual knowledge to application in real scenarios. Moreover, recent reviews, such as the one by Cruess et al.[6], proposed the inclusion of a fifth level, 'Is', which reflects the physician's professional identity, highlighting that being a physician involves not only 'doing', but also incorporating the attitudes, values and behaviors that define medical practice. This addition to the pyramid highlights the importance of the formation of professional identity, a process that goes beyond simple technical performance, encompassing the internalization of the ethical and moral values of the profession. The evolution of clinical reasoning skills in medical education highlights the importance of balancing formal training and practical experience, recognizing the value of experiential knowledge alongside traditional curricula.[7–9]

It is important to emphasize that one of the main objectives of an assessment in this teaching-learning context is to promote stimuli to develop certain behaviors in students so that with these habits being formed, they are good doctors, and this process begins both in the classroom at the beginning of an undergraduate course and in the process of continuing education that occurs after graduation until retirement. According to Bacchi et al.,[10] students' behavior changes depending on the evaluation format they will take. In their study, the students reported they prefer to study through books and websites for multiple-choice question tests, but value face-to-face teaching from local consultants to study for the Script Concordance Test (SCT), that is, changes in the evaluation format can change habits, and consequently, can result in a significant change in the ways of learning and teaching medicine.

In the current context, in which medicine is undergoing significant transformations due to technological advances, especially with the increase in the use of artificial intelligence tools in the health area[11], assessing how clinical reasoning is evolving and adapting to these changes is extremely important to ensure the quality of medical training and health care.

The objectives of this study are to review the literature regarding the practices adopted for the evaluation of the development of physicians' and students' clinical reasoning, as well as to identify gaps in the evaluation formats and possible areas of improvement, contributing to the improvement of curricula and teaching and evaluation methodology, aiming to promote a more effective development of clinical reasoning among future physicians, promote more efficient clinical practices and ensure quality health care.

METHOD

This is an integrative literature review carried out with the objective of answering the following guiding question:

- What are the main ways of assessing the development of clinical reasoning in physicians and medical students?

The methodology used in this integrative review was adapted from the methodology proposed by *Mendes et al*.[12] and adopted by several other authors. This methodology consists of six steps: 1. creation of the research question; 2. definition of inclusion and exclusion criteria; 3. List of information to be extracted; 4. Evaluation of the included studies; 5. Interpretation of the results; and 6. presentation of the review.

A search for scientific publications was carried out in the PubMed, Scopus, Embase, Web of Science, Lilacs, and SciELO Citation Index databases, as shown in Chart 1. Subsequently, duplicate studies were eliminated. Then, the studies were evaluated through their titles and abstracts using the Rayyan[13] electronic platform to verify whether they met the inclusion and exclusion criteria.

Articles from the last twenty years were selected, without language filters, which compared or discussed the evaluation of clinical reasoning. The databases consulted, shown in Chart 1, predominantly include articles with titles and abstracts in English and Portuguese. In addition, the keywords used were formulated in English and Portuguese, which may have limited the retrieval of articles published in other languages.

Articles outside the guiding question, editorials, comments, other health professions, and studies on teaching without evaluation methods were excluded.

RESULTS AND DISCUSSION

The search identified 6,150 studies. After eliminating duplicates (2862 articles) and evaluating titles and abstracts, 244 were selected for full reading. Figure 1 shows the process used.[14]

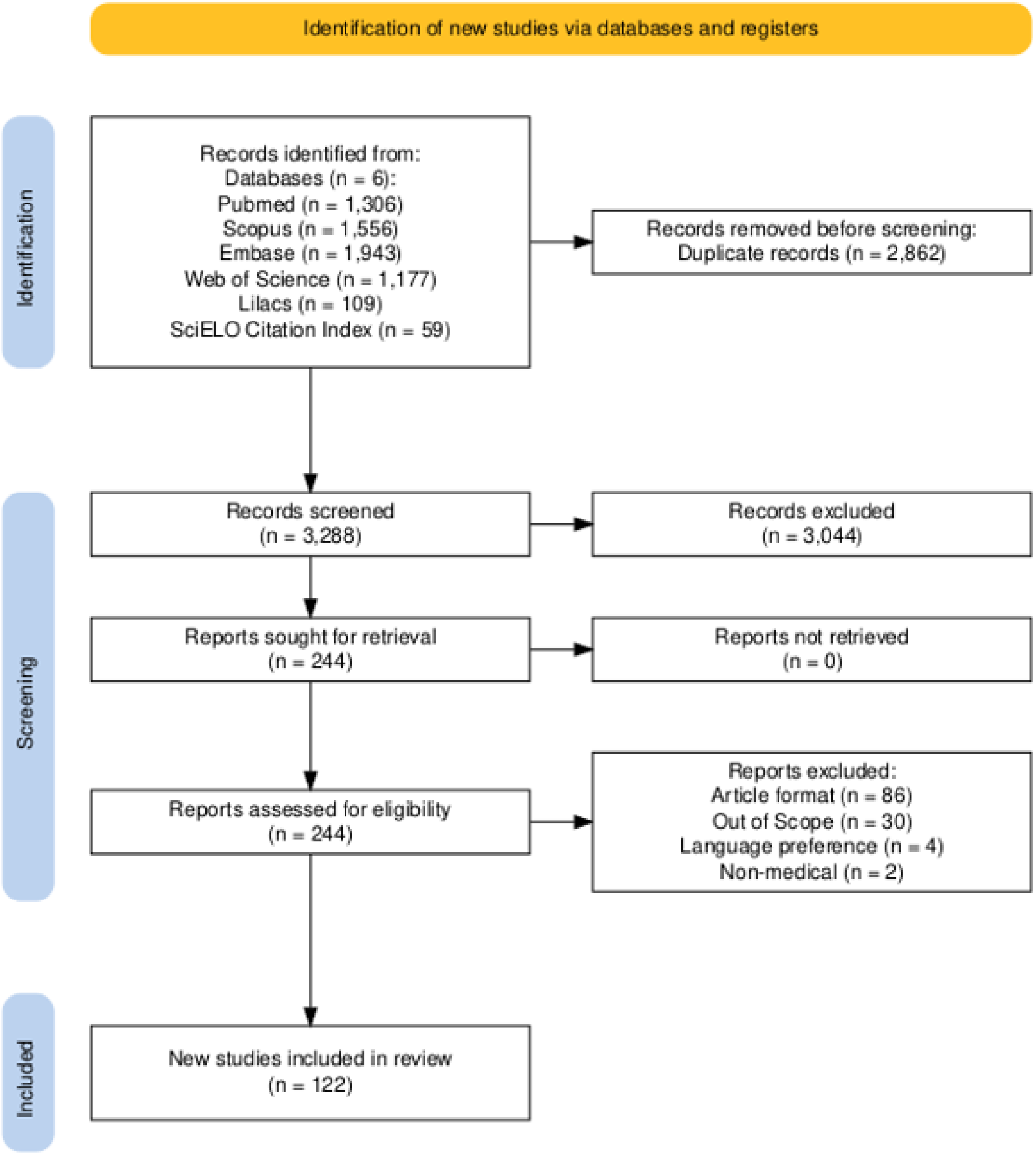
Among the articles selected for the review, it was observed that the year 2022 had the highest number of selected articles (n = 14) and that the articles that comprised the period 2014-2023 represented 68.85% of the total number of articles, compared to 31.15% in the years 2003-2013. This growing trend of publications over the years suggests an increasing interest and a greater reflection on the methods of evaluating clinical reasoning in the contemporary scientific scenario. The increase in the number of studies since 2014 reflects the greater appreciation of the evaluation of clinical reasoning.

Chart 1. Search keys.

Database	Search key	Field
Pubmed	((“clinical reasoning”) and (“measurement” or “evaluation” or “assessment” or “examination” or “assessing” or “exam” or “test” or “tests” or “testing” or “judgment” or “appraisal” or “analysis” or “performance”) and (“medical education” or “resident” or “residents” or “medical student” or “medical students” or “physician” or “physicians” or “medical doctor” or “medical doctors” or “medical school” or “medical schools” or “internship” or “intern” or “interns” or “residency” or “clerkship” or “attending” or “attendings”))	All Fields
Scopus		Article title, Abstract, Keywords
Embase		All fields
Web of Science		All fields
Lilacs	(((“clinical reasoning”) and (“measurement” or “evaluation” or “assessment” or “examination” or “assessing” or “exam” or “test” or “tests” or “testing” or “judgment” or “appraisal” or “analysis” or “performance”) and (“medical education” or “resident” or “residents” or “medical student” or “medical students” or “physician” or “physicians” or “medical doctor” or “medical doctors” or “medical school” or “medical schools” or “internship” or “intern” or “interns” or “residency” or “clerkship” or “attending” or “attendings”)) OR ((“raciocínio clínico”) and (“avaliação” or “teste” or “performance” or “desempenho” or “julgamento” or “análise” or “acurácia” or “prova” or “exame” or “verificação”) and (“estudante de medicina” or “estudantes de medicina” or “internato” or “interno” or “internos” or “residência” or “residente” or “residentes” or “médico” or “médicos” or “educação médica” or “ensino médico” or “escola médica” or “escolas médicas”)))	Title, Abstract, Subject
SciELO Citation Index	(((“clinical reasoning”) and (“measurement” or “evaluation” or “assessment” or “examination” or “assessing” or “exam” or “test” or “tests” or “testing” or “judgment” or “appraisal” or “analysis” or “performance”) and (“medical education” or “resident” or “residents” or “medical student” or “medical students” or “physician” or “physicians” or “medical doctor” or “medical doctors” or “medical school” or “medical schools” or “internship” or “intern” or “interns” or “residency” or “clerkship” or “attending” or “attendings”)) OR ((“raciocínio clínico”) and (“avaliação” or “teste” or “performance” or “desempenho” or “julgamento” or “análise” or “acurácia” or “prova” or “exame” or “verificação”) and (“estudante de medicina” or “estudantes de medicina” or “internato” or “interno” or “internos” or “residência” or “residente” or “residentes” or “médico” or “médicos” or “educação médica” or “ensino médico” or “escola médica” or “escolas médicas”)))	All fields

Source: Prepared by the authors.

Figure 1. Results (Constructed with the aid of the Haddaway et al.[137] tool and translated into Portuguese).



Source: Elaborated by the authors based on the tool by Haddaway et al.[137].

Validity and reliability are essential in the evaluation of instruments. Reliability measures consistency and validity depends on it. Cronbach's alpha is widely used for this measurement[136].

The evaluation of educational instruments requires reliability and validation to ensure consistent and accurate results. In the context of this study, Cronbach's alpha coefficient was used, which is widely used to measure the internal consistency of instruments in the area of medical education, ensuring that the instrument applied reliably reflects the clinical reasoning skills of medical students and professionals.[132]

The most common methods were SCT, OSCE, and MCQ. Digital assessments were also identified, but most of the methods were mentioned a few times. Below we will describe the different methods and for a better understanding of the reader, here is a list of abbreviations:

ART-R - Reconstructed tool for evaluating reasoning; ASCLIRE - Clinical Reasoning Assessment (electronic test); CBCRT - Computer-based clinical reasoning training system; CBA - Multimedia Case-Based Assessment; CCS - Computer-based case simulations; CDI - Clinical data interpretation; CIP - Integrative puzzles; CIVA - Clinical evaluation with image and video; CRANAPL - Clinical reasoning in admission grades - evaluation and planning; CRI-HT-S - Clinical reasoning indicators - anamnesis scale; CRC-DP - Clinical reasoning cases (digital platform); CPX - Clinical performance exam; crSBA - Context-rich questions with single best answer; CRT - Clinical reasoning task; CSE – Clinical Skills Exam; DTI - Diagnostic

Thinking Inventory; DPR - Diagnostic pattern recognition; EMQ - Extended matching questions; GCE - Graduation Competency Exam; HFS - High fidelity simulation; IDEA - Interpretive summary, differential diagnosis, explanation of reasoning, and alternatives; IGT - Information Gathering Tests; KF - Key features; LC - Long case evaluation; MEQ - Modified essay question; NAr - Resident admission grades; NPE - Standardized post-patient meeting notes; ORT - Observation rating form; OSCE - Objective Structured clinical examination; PD-R - Descriptive test with template through rubric; PCE - Meetings with real patients; PERT - Post-meeting evaluation tool; SP - Standardized patients; VP - Virtual Patients; MCQ - Multiple choice question; QRC - Short answer questions; REACT - Rapid evaluation assessment for clinical reasoning tool; SAVE - Short answer vignette exam; SCBD - Standardized cases-based discussion; SCOE - Structured clinical oral examination; SF - Scenario formation; SG - Simulation game; SRL - Self-regulated learning; SSAR - Summary assessment rubric; TCSe - Evolving Script Concordance Test; SCT - Script Concordance Test.

Script Concordance Test (SCT)

Script Concordance Tests consist of brief clinical scenarios that involve uncertainty, replicating the clinical reasoning process. Participants are presented with a series of questions that explore how new information influences their decisions. The participants' responses are then compared with those of a panel of experts, allowing the variability in the professionals' responses in different clinical contexts to be assessed[32,33,53–55,61,66,68,73,82,87,89,105,112,132].

Chart 2. Main results.

Author – Year	Country	Evaluation tool	Key Points
Aubart et al. - 2020[20]	France	SCT	Cronbach's alpha between 0.08 and 0.59.
Bhardwaj et al. - 2022[27]	United States	SCT	Weak correlation between SCT and various tests of medical competence.
Blunk et al. - 2019[29]	United States	SCT	SCT for psychiatry improves students' clinical reasoning.
Boulouffe et al. - 2014[30]	Belgium	SCT	SCT scores depend on the panel used for reference.
Carrière et al. - 2009[32]	Canada	SCT	Residents approved the representation of cases in the SCT .
Charlin et al. - 2006[33]	Canada	SCT	The SCT demonstrated efficacy in the evaluation of clinical reasoning.
Compère et al. - 2016[37]	France	SCT	Residents' results were lower than experts' results.
Cooke, Lemay and Beran - 2017[38]	United States	TCSe	TCSe is considered engaging and close to real cases.
Couraud et al. - 2015[39]	France	SCT	SCT are an appropriate assessment for continuing education.
Ducos et al. - 2015[43]	France	SCT	SCT is reliable for assessing clinical reasoning in residents.
Duggan and Charlin - 2012[45]	Australia	SCT	Reasons for SCT questions being inadequate are not obvious.
Enser et al. - 2017[47]	France	SCT	SCT measured that noise affects residents' clinical reasoning.
Gagnon et al. - 2005[53]	France	SCT	20-member panel for important examinations; more than 10 for acceptable exams.
Gagnon et al. - 2008[54]	Canada	SCT	Tests that are considered reliable usually include few questions but cover a larger number of clinical cases (20-30), while less reliable tests tend to contain fewer cases (4-5 questions) for each set of 25-30 clinical cases.
Gagnon et al. - 2011[55]	Canada	SCT	Results suggest robustness of the methodology, requiring confirmation in other contexts.
Gómez et al. - 2021[58]	Spain	SCT	SCT is effective in measuring clinical reasoning.
Gómez, Sequeros and Martínez - 2022[59]	Spain	SCT	Supervised SCT can improve pediatric care at home.
Goos et al. - 2016[60]	Germany	SCT	SCT shows discrepancy between specialists and students in visceral surgery.
Goulet et al. - 2010[61]	United States	SCT	The Cronbach's alpha coefficient for SCT was 0.90. Indicating a high reliability test.
Hamui et al. - 2018[64]	Argentina	SCT	First viable experience of SCT on a national scale (Argentina) in pediatrics.

Continue...

DOI: https://doi.org/10.1590/1981-5271v49.3-2024-0089.ING

Chart 2. Continuation.

Author – Year	Country	Evaluation tool	Key Points
Humbert et al. - 2011[66]	United States	SCT	SCT differentiates students from different years and specialists.
Humbert and Miech - 2014[67]	United States	SCT	Students have improved at dealing with uncertainties in clinical scenarios, as measured by SCT .
Humbert, Besinger and Miech - 2011[68]	United States	SCT	SCT shows to be promising in the evaluation of clinical reasoning.
Iravani et al. - 2016[71]	Iran	SCT	The SCT is a reliable tool for assessing clinical reasoning.
Kania et al. - 2011[73]	France	SCT	Participants found the test realistic, interesting, relevant, and intuitive.
Kaur, Singla and Mahajan - 2020[74]	India	SCT	Medical students showed high satisfaction with this method.
Kazour et al. - 2016[75]	Lebanon	SCT	Students improved their SCT scores by 11% after training.
Kün-Darbois et al. - 2022[79]	France	SCT	The global opinion of medical students and faculty about SCT was globally negative.
Lambert et al. - 2009[82]	Canada	SCT	SCT is useful for discriminating between participants according to their level of experience.
Lineberry et al. - 2019[83]	Argentina	SCT	It explains two hypothetical situations with the potential for misinterpretation of the SCT responses.
Talvard, Olives and Mas - 2014[86]	France	SCT	Although SCT is a useful method, it can be difficult to develop.
Marie et al. - 2005[87]	France	SCT	SCT allows you to differentiate groups of candidates based on their level of competence.
Mathieu et al. - 2013[88]	France	SCT	Limitations: Cost and time of development and recruitment.
Meterissian et al. - 2007[89]	Canada	SCT	Useful test for training and evaluation in general surgery.
Mzoughi et al. - 2018[91]	Tunisia	SCT	Average scores in the SCT exceeded the reference test. Correlation absent.
Nomura et al. - 2021[93]	Japan	SCT	Created and validated SCT in Japanese to assess clinical education in Japan.
Nouh et al. - 2012[94]	Canada	SCT	The test was able to differentiate junior residents (R1-R2) from senior residents (R3-R5).
Omega et al. - 2022[95]	Indonesia	SCT	SCT was able to discriminate between groups of different clinical experiences.
Petrucci et al. - 2013[98]	Canada	SCT	Article presents contradictory results.
Peyrony et al. - 2020[99]	France	SCT	SCT scores varied statistically but differences were weak.
Piovezan et al. - 2012[100]	Brazil	SCT	SCT is feasible and easy to develop, administer, and correct.
Power, Lemay and Cooke - 2016[102]	Canada	SCT	The study showed that SCT can be improved with a written Think-Aloud protocol.
Ruiz et al. - 2010[105]	United States	SCT	SCT can be useful as formative feedback for trainees and physicians.
Sibert et al. - 2006[112]	France	SCT	SCT online construct validity is not straightforward.
Steinberg et al. - 2020[116]	United States	SCT	SCT has limited utility in assessing clinical reasoning among residents.
Subra et al. - 2017[117]	France	SCT	Experts' scores were higher than students' scores on all assessments (p <0.001).
Van den Broek et al. - 2012[122]	Holland	SCT	Including differential diagnosis resulted in lower reliability, contrary to expectations.
Wan, Tor and Hudson - 2019[125]	Australia	SCT	Candidates' scores increased with increasing level of clinical experience.
Wan, Tor and Hudson - 2020[126]	Australia	SCT	Test responses reflect thoughts, validating the SCT process.
Wilson, Pike and Humbert - 2015[128]	United States	SCT	Ambiguity in the SCT dimensionality impairs the interpretation of the scores.

Continue...

Chart 2. Continuation.

Author – Year	Country	Evaluation tool	Key Points
Wilson, Pike and Humbert - 2014[129]	United States	SCT	Scoring on a 5-point scale is more reliable than on a 3-point scale.
Yassin et al. - 2021[130]	Iraq	SCT	The junior residents' score on the SCT did not differ much from that of the senior residents.
Zavaleta-Hernández et al. - 2011[132]	Mexico	SCT	Residents showed significant improvement in performance with clinical experience.

Source: Prepared by the authors.

Among the tools for assessing clinical reasoning, the most frequently used in the studies that are part of the review was the script concordance test (SCT). The SCT has been tested in several areas of medicine, such as psychiatry[29], otorhinolaryngology[71,73], anesthesiology[36,43,47,95], surgery[60,89], pediatrics[32,59], among others. Several studies[37,60,67,75,92,94,110,117,125] report that SCT was able to successfully differentiate students according to their different stages of education, such as between a 5th year student and a first-year one, or between a resident and a student, and recognize the skills of a specialist in relation to a resident, for example. This is an important characteristic, since it reflects the expectation that students and physicians with a higher degree of experience should perform better than those who have had a shorter training time. Some authors have also reported in their results that the SCT can be an assessment with high degree of reliability measured by Cronbach's alpha[17,29,30,33,36,37,47,53,59–61,71,73,75,82,84,87–89,91,93,94,98,100,105,106,112,129].

Nomura et al. concluded that SCT works in languages other than English, such as Japanese[93]. In addition to these findings, several authors suggest that SCT is a valid instrument to assess clinical reasoning[33,36,39,53–55,58,74,82,84,87,100,105].

Although a wide range of studies indicate success when using SCT, other studies point to divergent results. Some studies show that the SCT was unable to distinguish the different levels of education of physicians[130]. Many reported that it is too expensive and/or difficult to construct the questions for the SCT [45,86,88,98,98,122], and that the test reliability measured by Cronbach's alpha is low[20,27,39,43,63,64,67,92,95,99,116,122,130].

Some feedbacks on the use by both students and some teachers were not positive[64,76,79]. This fact can be explained by the novelty of the SCT environment and the unusual nature of the medical reasoning required[79]. Kelly et al.[76] show in their study that students preferred MCQ to SCT; the students mentioned familiarity and perceived objectivity with the MCQ format.

Other feedbacks were that they can be useful more for teaching than for evaluation[63], that the construct validity is limited or ambiguous, compromising the meaning and interpretation of the SCT scores[112,116,129].

Another point to be mentioned is that there was a low correlation between the SCT and other evaluation formats, which means that a good grade in the SCT does not imply that one will have a good grade if another style of evaluation is applied. Amini et al.[17], found a weak correlation between the SCT scores and the multiple choice questions (MCQ). Humbert et al.[68] found a weak correlation between the SCT and Step 2 CK scores. Bhardwaj et al.[27] found a weak correlation between the SCT scores and the USMLE Step 1 and Step 2 tests, among others.

Other findings include those of Mzoughi et al.[91] that did not find a correlation between the scores of the SCT and the reference cardiology test. Ducos et al.[43] explain in their study that the SCT measures different domains of other tests. Groves et al.[63] showed that the SCT and CRP also measure different aspects of clinical reasoning. This is extremely relevant, because, in theory, all assessment instruments measure clinical reasoning, but the literature reveals that there are divergences in what could be considered clinical reasoning and, as clinical reasoning is an especially broad term, a more precise definition is still lacking. During our review of 122 articles, we observed that some authors tried to define and characterize the construct in its parts, from the anamnesis, physical examination, diagnosis, and so on[10,21,23,44,48,52,75,85,104,113,114,123]. Others assume that it is only the diagnosis, and for most it is loosely classified within the large umbrella that this term can encompass.

Several other studies on SCT have been carried out over the years, ranging from tests on the Likert scale with 3 to 5 points, with the 5-point scale being preferred.[129] It was investigated whether the students selected the answers on the SCT (Script Concordance Test) that reflected their actual clinical reasoning and, in general, it was found that they did[121]. Based on that, it was suggested to incorporate a think-aloud technique[97] to improve the SCT assessment. Additional difficulties were analyzed, ranging from divergences in the interpretation of the question and answer[83], to the vulnerability to the use of extreme values[110] on *the* Likert scale, to the possibility of answering the SCT questions without reading the problem vignette[90], the importance of disclosing the score of the panel and the standard deviation to have a reference of scores[30], and even an evolution of the SCT[38].

The results of this review indicate that methods such as the Script Concordance Test (SCT) and the Objective Structured Clinical Examination (OSCE) are widely used in the evaluation of clinical reasoning in physicians and medical students. Daniel et al. (2019) also identified and classified these methods in their scoping review, highlighting the importance of different types of assessment, such as practical assessments and simulations.[138] The main difference between the studies is in the methodological approach: while Daniel et al. mapped a wide range of methods from multiple sources, including gray literature, our study uses an integrative review with more restrictive selection criteria, prioritizing studies with greater methodological control.

Objective Structured Clinical Examination (OSCE)

OSCEs consist of multiple stations where examinees perform different clinical tasks, using standardized patients, observer's assessments, written notes, and other methods for a comprehensive assessment. This was the second most common type of evidence found in our literature review. According to the authors, OSCE is not a test performed exclusively to measure clinical reasoning, as it involves anamnesis, physical examination, communication, clinical reasoning, procedures, and professionalism skills[114]. Volkan et al.[123] in their studies, were able to differentiate two major factors that fit the data well. The first factor was related to Physical Examination and Anamnesis, labeled as information collection, while the second factor was related to Differential Diagnosis/Clinical Reasoning and Patient Interaction, labeled as reasoning and dissemination of information. Sim et al.[114] highlighted that the students who were analyzed had their highest grades during the procedural stage of the OSCE and their lowest grades in the clinical reasoning topic, indicating that clinical reasoning, in addition to being comprehensive, has a teaching format that is not so direct.

Numerous researchers consider the use of this instrument to be positive[23,24,46,57,85,123], and many have obtained good results[48,50] in reliability measured by Cronbach's alpha. Volkan et al.[118] performed the examination with 9 separate stations (obstetrics/gynecology, pediatrics, breast examination, pathology, psychiatry, asthma, radiology, cardiology, and neurology categories), obtaining good reliability except in the obstetrics/gynecology and pathology stations.

Berger et al.[24] carried out a study that states that lay evaluators, in possession of a rubric, can provide results similar to medical evaluators. This finding is very interesting because the financial amount spent on a doctor is expensive, and thus, could reduce costs when giving grades. Gilkes et al.[57] also used a rubric to evaluate and support the usefulness of a simple and standardized feedback tool for teaching and evaluating medical students.

Chart 3. Main results.

Author – Year	Country	Evaluation tool	Key Points
Behrens et al. - 2018[23]	Chile	OSCE	Viable collaborative exam for Chilean medical students.
Berger et al. - 2012[24]	United States	OSCE	Laypeople can evaluate clinical reasoning with appropriate training.
Durning et al. - 2012[46]	United States	OSCE	OSCE is a valid format for evaluating second-year medical students.
Fleiszer et al. - 2017[50]	Canada	OSCE	Clinical reasoning improves with clinical practice; evaluation is feasible.
Gallagher et al. - 2020[56]	United States	OSCE	OSCE written notes associated with performance in clinical rotation.
Gilkes, Kealley & Frayne - 2022[57]	Australia	OSCE	Standardized feedback is useful for teaching and evaluating students.
Lukas et al. - 2014[85]	United States	OSCE	OSCE participants obtained higher scores subsequently.
Park et al. - 2015[96]	South Korea	OSCE	OSCE scores may not reflect a medical student's clinical reasoning ability.
Sim et al. - 2014[114]	Malaysia	OSCE	OSCE measures anamnesis, physical examination, communication, clinical reasoning, procedural and professionalism skills.
Volkan et al. - 2004[123]	United States	OSCE	A model with two correlated factors fitted the data well.

Source: Prepared by the authors

Not all studies were positive for the OSCE. In the perception of Park et al.[96] OSCE scores may not reflect a medical student's clinical reasoning ability. It further adds that efforts should be made to improve the assessment of medical students' clinical reasoning skills using OSCE. Fleiszer et al.[50] report that there are still limitations in the use of rubrics for evaluation, as they noticed that clinical reasoning, as measured by some of its "procedural" characteristics, improves throughout the year of the clinical internship. Rubrics can be created to objectively assess "procedural" issues, but not for "semantic" issues, i.e., aspects such as the effectiveness of clinical reasoning, decision-making, and the performance of medical procedures would be easier to measure using a rubric than comprehension, interpretation, and accuracy in communication.

Multiple Choice Questions (MCQ)

A multiple-choice question in medicine may present a clinical scenario, followed by a question with several answer options, usually with five, and only one is correct.

Multiple-choice questions (MCQ) are currently the most common method[76] for evaluation in medical schools. It is perceived that studies on multiple-choice questions have a more in-depth character since it is already so widespread, or of a comparative nature, as they try to compare different methods with the gold standard to find some correlation.

Silberman et al.[113] in their study prepared a tool that can be a formative assessment of psychiatric reasoning, as a complement to the various methodologies already available in this difficult area of education. Surry et al.[119] evaluated the students' cognitive dispositions when taking a multiple-choice test. Coderre et al.[35] obtained results that support the idea that well-designed multiple-choice questions can, in fact, test higher-order clinical reasoning. Furthermore, they point out that, when testing clinical reasoning, the wording of the question or the content remains more important than the number of alternatives. One of the studies[134] obtained a questionable degree of reliability measured by Cronbach's alpha, indicating that even a gold standard test, depending on the context in which it was developed or applied, may not be such a reliable measurement instrument.

Digital, online and simulation assessments

In recent years, several forms of evaluation have been introduced, whether based on multimedia and video, computers, or online. However, it was observed that the studies are in smaller quantity and may constitute material for studies in the coming years. This was useful in cases such as the global Covid-19 pandemic, in which authors such as Duffy, Tully, and Stanton[44] adapted the "long case" version to a 100% digital version, eliminating the physical exam. However, they concluded that ultimately, the use of simulated patients alone to develop clinical skills is no substitute for encounters with real patients. Similar results were seen in the study by Fink et al.[49] which concludes that perceived authenticity and diagnostic accuracy was higher for standardized patients (SP) than for virtual patients (VP), although the cognitive load when performing the activity was equivalent.

The use of a digital platform has numerous advantages such as the automation of numerous services. Waechter et al.[124] studied a digital platform with clinical reasoning cases, in which, with only two instructors, they were able to provide detailed feedback in 376 cases, demonstrating a viable and sustainable workload. Schaye et al.[109] developed and validated a high-performance Machine Learning model that classifies the quality of clinical reasoning in the documentation of residents' admission scores in the clinical environment, that is, they included sophisticated algorithms for evaluation that greatly reduce the manpower of the evaluation process. Kotwal et al.[78] conducted a similar study, studying the Clinical reasoning in admission grades - evaluation and planning; (CRANAPL) tool and found that given the tool's easy use and versatility, hospitalist divisions in academic and non-academic settings can use the CRANAPL tool to evaluate and provide feedback on the documentation of hospitalists' clinical reasoning. All of these studies respond to ways in which educators and clinicians can provide more detailed and consistent feedback in a feasible and sustainable manner.

Zuo et al.[135] used a computer-based clinical reasoning training system (CBCRT) and found that test scores increased with years of training, i.e., there are indications that the

Chart 4. Main results.

Author – Year	Country	Evaluation tool	Key Points
Silberman et al. - 2020[113]	United States	MCQ	Tool can be a useful formative assessment of psychiatric reasoning.
Surry et al. - 2018[119]	United States	MCQ	The article evaluates the cognitive disposition of students when performing an MCQ.

Source: Prepared by the authors.

Chart 5. Main results.

Author – Year	Country	Evaluation tool	Key Points
Dong et al. - 2020[42]	China	CCS	Resident CCS scores decreased with years of training.
Jerant and Azari - 2004[72]	United States	CCS	CCS software scores (D&X Clinician) seem to have no criterion validity.
Kotwal et al. - 2019[78]	United States	CRANAPL	Hospitalists may use CRANAPL to evaluate clinical documentation.
Lai et al. - 2022[81]	Taiwan	SP	Getting the "most likely diagnosis" correlates with different skills.
Pottier et al. - 2016[101]	France	SP	Global evaluation is preferred for speed and effectiveness.
Zuo et al. - 2021[135]	China	CBCRT	Scores increased with training.

Source: Prepared by the authors.

Chart 6. Main results.

Author – Year	Country	Evaluation tool	Key Points
Kunina-Habenicht et al. - 2015[80]	Germany	ASCLIRE	It is argued that ASCLIRE measures a third construct.
Schauber et al. - 2021[107]	Germany	ASCLIRE	Diversity of response formats is essential in the evaluation.

Source: Prepared by the authors

software is evaluating coherently, being able to differentiate students in their different stages of training.

Dong et al.[42] used computer-based case simulations (CCS) with the DxR Clinician software and suggested training actions, as the residents' scores were lower than the international scores. This implication is extremely relevant, since with standardized software systems it is possible to analyze and evaluate different institutions worldwide and create systems to ensure universal quality medical education. Fida and Kassab, [48], also used the same software and reported that it has good internal consistency reliability and that it measures the student's clinical competence in a construct related to those measured by the OSCEs. However, it seems that CCS scores measure a different construct than that measured through encounter-with-real-life patient examinations, i.e., CCS is reliable and measures clinical competence but may not fully capture the skills assessed in encounters with real patients. Another, older study by Jerant and Azari[72] reports that the automated diagnostic reasoning scores generated by a widely available CCS software (D&X Clinician) appear to have no criterion validity. This implies that, despite the great potential of the tool, more studies are needed to ensure that the use of these new study methods is reliable for use on a global scale.

A new approach was found in the study of Fonteneau et al.[51] in which it uses gamification concepts to build a Simulation Game (SG) and compares it to a high-fidelity simulation (HFS) and multiple choice questions (MCQ). The authors concluded that the used SG better reflected the clinical competence of students on HFS than an MCQ in the same clinical case of pediatric asthma exacerbation. Moreover, they add that the students appreciated the game as an evaluation method.

Sulaiman and Handy[118] performed a Clinical Evaluation with Image and Video (CIVA) and found a strong correlation between the new method (CIVA) and the OSCE ($r = 0.83$, $p < 0.001$). Kim et al. did a Multimedia Case-Based Assessment (CBA) and concluded that the CBA more strongly stimulates data analysis and synthesis than other evaluation instruments such as modified essay questions (MEQ) and clinical performance examination (CPX).

Kunina-Habenicht et al.[80] conducted a study using an electronic form called the Clinical Reasoning Assessment (ASCLIRE) and tested the construct validity. Correlations were calculated with variables that were external to the test under study, such as MCQ and OSCE results, and no strong correlations were found. In fact, ASCLIRE correlated to approximately the same extent with MCQ and OSCE results, just as these two correlate with each other. Given that most agree that MCQ and OSCE measure different aspects of clinical reasoning (declarative and procedural knowledge, respectively), it is stated that ASCLIRE measures a third construct. Schauber et al.[107] also used the same electronic form and concluded that there is not a generally "more valid" response format than another to assess clinical reasoning. Exclusive use of one format may impair the validity of the test. The diversity of response formats is essential in the conception of evaluations in medical education, as it is not the format alone, but the task as a whole, that influences the reasoning processes.

This statement by Schauber et al.[107] is extremely relevant, as they consider that evaluation is a method to condition behavior. The act of combining different formats has been tested several times by different authors in medical

Olympiads in Iran[16,90,106]. Amini et al.[16] built an assessment instrument that combined questions such as: Key Features (KF), Script Concordance Test (SCT), Clinical Reasoning Problems (CRP), Comprehensive Integrative Puzzles (CIP). It is noteworthy that the new test, which encompasses the different question formats, obtained excellent reliability values using Cronbach's alpha method. Sadeghi et al.[106] reproduced something similar to Amini et al.[16] and obtained similar results. Monajemi et al.[90] built an assessment instrument that combined questions such as: Key Features (KF), Script Concordance Test (SCT), Clinical Reasoning Problems (CRP), Comprehensive Integrative Puzzles (CIP), Information Gathering Tests (IGT), Scenario Formation (SF), some innovative ones. However, they found problems such as unfamiliarity with the tests, too much time to construct IGT-type questions, and that constructing SCT questions is not a trivial process, since the questions have to be very well formulated, as they often do not need to read the scenario to answer the questions.

Joint evaluation methods

The combination of different evaluation methods has been explored as a strategy to obtain a more comprehensive and reliable view of students' and physicians' clinical reasoning. This integrative approach allows you to take advantage of and compensate for the limitations of each method, providing a more complete picture of clinical skills. Chart 7 presents a summary of the main studies that applied joint methods, highlighting the reliability and key points of each combination.

As shown in Chart 7, the combination of methods improves the reliability of the assessment, identifies educational skills and gaps, and reinforces the importance of methodological diversity.

Chart 7. Main results.

Author – Year	Country	Evaluation tool	Key Points
Amini et al. - 2011[16]	Iran	KF; SCT; CRP; CIP and New	Combination of tests shows high reliability.
Amini et al. - 2017[17]	Iran	SCT and MCQ	A study showed a weak correlation between SCT and MCQ.
Bacchi et al. - 2019[10]	Australia	SCT and MCQ	Preference for books and websites for MCQ; face-to-face teaching for SCT.
Coderre et al. 2004[35]	Canada	MCQ and EMQ	MCQ test higher-order clinical reasoning.
Compère et al. - 2015[36]	France	SCT, MCQ and QRC	SCT indicates that educational intervention improves clinical reasoning in residents.
Covin et al. - 2020[40]	United States	CRT, PNS, SSAR and CDI	Different instruments measure different components of clinical reasoning.
Derakhshandeh et al. - 2018[41]	Iran	CRP and MCQ	CRP can be an alternative to measure clinical reasoning skills.
Fida and Kassab - 2015[48]	Bahrain	CCS, MCQ, QRC, OSCE and PCE	CCS scores have good reliability but differ from OSCEs.
Fink et al. - 2021[49]	Germany	PP and PV	PP is superior to PV in authenticity and diagnostic accuracy; equivalent cognitive load.
Fonteneau et al. - 2020[51]	France	HFS, SG and MCQ	SG reflects clinical competence better than MCQ
Fürstenberg et al. - 2020[52]	Germany	CRI-HT-S	CRI-HT-S is consistent for assessing clinical reasoning.
Groves et al. - 2003[62]	United States	DTI and CRP	DTI may limit diagnostic diversity
Groves et al. - 2013[63]	Australia	SCT and CRP	CRP and SCT complement each other; useful more for teaching than evaluation.
Haring et al. - 2020[65]	Holland	ORT and PERT	ORT and PERT: reliable, valid, weak correlation between them.
Huwendiek et al. - 2017[69]	Germany	crSBA, KF, OSCE	KF more realistic, reliable and efficient than crSBA.
Kelly, Durning and Denton - 2012[76]	United States	SCT and MCQ	SCT was preferred over MCQ by only a minority of students.
Kim et al. - 2019[77]	South Korea	CPX, CBA and MEQ	Distinct assessments encourage different aspects of participants' clinical reasoning.
Loh et al. - 2020[84]	Singapore	SCT and EMQ	Experience and training predict after-graduation reasoning ability.
Monajemi et al. - 2012[90]	Iran	KF, TCS, CRP, CIP, IGT and SF	Problems include familiarity, extensive format, and formulation of the test.

Continue...

Chart 7. Continuation.

Author – Year	Country	Evaluation tool	Key Points
Nazim et al. - 2019[92]	Pakistan	SCT and EMQ	SCT and EMQs discriminate apprentices by clinical experience in urology.
Sadeghi et al. - 2019[106]	Iran	KF; TCS; CRP; CIP; New	Combination of tests is recommended for high-risk examinations.
See, Tan and Lim - 2014[110]	Singapore	SCT and MCQ	SCT distinguishes competencies, but vulnerable to extreme values.
Williams et al. - 2011[127]	United States	DPR and CDI	Student performance increases with years of medical training.
Zia et al. - 2019[134]	Pakistan	MCQ and KF	Limited time affects test reliability.

Source: Prepared by the authors.

Other assessment tools

Other methods were found in a few references, suggesting that they may be methods that are not so common in medical education settings or that these methods represent promising areas for future studies. Key Features (KF)[69,133], for example, are effective for evaluating decisions in specific clinical situations, but their applicability may be limited in broader contexts where clinical reasoning is sought to be evaluated in general. Descriptive tests[115] and oral tests[103,111] offer opportunities for direct evaluation of critical thinking, although they face practical limitations, such as subjectivity and lack of standardization.

The Short Answer Vignette Exam (SAVE)[28] and the 'one-minute preceptor' method[15] provide immediate feedback but lack robustness in terms of objective measurement. Methods such as IDEA (Interpretive summary, differential diagnosis, explanation of reasoning, and alternatives)[21,108] and the Diagnostic Thinking Inventory (DTI)[25,62] are useful for capturing nuances of diagnostic reasoning, although their implementation requires specialized resources for detailed analysis of the results.

Tools such as REACT (Rapid Evaluation Assessment for Clinical Reasoning Tool)[97] and MATCH (Measuring Analytical Thinking in Clinical Health Care)[22] bring innovations by integrating reasoning and practical analysis, but their application still depends on additional validation in different cultural and educational contexts. Instruments such as the crSBAs (Context-rich questions with single best answer)[69] and the Extended Matching Questions (EMQ)[26] facilitate the standardized assessment of knowledge, while Comprehensive Integrative Puzzles (CIP)[31] and Standardized Patients (SP)[81,101] are tools that integrate practical and theoretical skills, although they require time and resources for their implementation.

These methods, although varied, still face challenges such as the complexity of implementation, the need for resources, and limitations in the assessment of more complex skills, highlighting the importance of future research that can validate and adapt these instruments to different contexts in medical education.

This study has limitations. Clinical reasoning is broad and complex, making it difficult to fully address it in a single review. Furthermore, as this is an integrative review, we did not perform a detailed methodological analysis of each article, which may impact the evidence validity. Despite this, the study contributes by organizing and synthesizing the models for evaluating clinical reasoning.

Chart 8. Main results.

Author – Year	Country	Evaluation tool	Key Points
Aagaard, Teherani and Irby[15]- 2004[10]	United States	OMP	Preceptors: OMP improves diagnosis, reliability, effectiveness, and efficiency.
Anderson et al. - 2008[18]	Australia	CRP	Additional validation is required for application.
Artino et al. - 2014[19]	United States	SRL	Students are aware, but few apply diagnostic reasoning strategies.
Baker et al. - 2015[21]	United States	IDEA	Valid and reliable tool for low-risk formative assessment.
Baño et al. - 2011[22]	Chile	MATCH	Score increased with experience.
Beullens, Struyf and Damme (2006)[25]	Belgium	DTI	Grades, assessed with DTI, increased after the clinical seminars.

Continue...

Chart 8. Continuation.

Author – Year	Country	Evaluation tool	Key Points
Beullens, Stryf e Damme - 2005[26]	Belgium	EMQ	EMQ distinguishes specialists from less specialized ones.
Blazek et al. - 2014[28]	United States	SAVE	SAVE is useful in the evaluation of medical students in psychiatry.
Capaldi et al. - 2015[31]	United States	CIP	Individual CIP did not correlate with multiple-choice tests.
Cheung et al. - 2022[34]	United States	NPE	Creating scoring rules and guidelines is resource-intensive.
Duffy, Tully and Stanton - 2023[44]	Ireland	LC without physical examination	Simulated patients is not a substitute for encounters with real patients.
Im et al. - 2016[70]	South Korea	CPX	CPX: High reliability, problems with validity.
Peterson et al. - 2022[97]	United States	REACT	Educational objectives compete with real clinical situations.
Rajeswaran et al. - 2022[103]	Canada	SCOE	Positive performance of clinicians in the SCOE, with no correlation with other evaluations.
Reinert et al. - 2014[104]	United States	CSE	CSE is valid to evaluate the performance of students in the surgery internship.
Schaye et al. - 2021[108]	United States	IDEA	Reliable and easy tool for feedback on clinical reasoning.
Schaye et al. - 2022[109]	United States	Nar	Machine Learning model classifies quality of clinical reasoning.
Shenoy et al. - 2021[111]	United States	SOE	SOE for formative feedback has educational advantages, as noted by fellows and faculty.
Simpkins et al. - 2019[115]	United States	PD-R	Rubric assisted students and teachers in feedback considered significant.
Sulaiman and Hamdy - 2013[118]	United Arab Emirates	CIVA	The strongest correlation was found between CIVA and OSCE (r = 0.83, p < 0.001).
Sutherland et al. - 2019[120]	Australia	SCBD	SCBD has acceptable internal consistency, measures related to construct.
Thammasitboon et al. - 2020[121]	United States	ART-R	ART-R demonstrated validity as an assessment tool with five domains.
Waechter et al. - 2022[124]	Canada	CRC-DP	Two instructors, with a viable workload, provided detailed feedback on 376 cases, contributing to the improvement of students' performance in clinical reasoning.
Yudkowsky et al. - 2019[131]	United States	GCE	Non-clinical evaluators score grades with reliability comparable to clinical ones.
Zegota et al. - 2022[133]	Germany	KF	Validation of the clinical reasoning test using response theory.

Source: Prepared by the authors.

FINAL CONSIDERATIONS

The review highlighted that assessment methods such as the Script Concordance Test (SCT), the Objective Structured Clinical Examination (OSCE) and multiple choice questions (MCQ) play key roles in medical education, each contributing in a unique way to the development and evaluation of clinical reasoning. The SCT has been shown to be effective in measuring clinical reasoning ability in uncertainty scenarios, while the OSCE evaluates the application of theoretical knowledge in simulated clinical practice contexts. Artificial intelligence brings advances to medical evaluation, but it still lacks standardization and adaptation. We recommend that future studies investigate the integration of these technologies to improve the assessment of clinical reasoning, promoting more effective practices that are aligned with the current demands of medical education.

AUTHORS' CONTRIBUTION

Fernando Tureck contributed to the study conception and supervision; data collection, analysis and interpretation; critical review of the manuscript; final approval of the version to be published. Maurício Prätzel Ellwanger contributed to data collection, analysis and interpretation; writing of the manuscript; final approval of the version to be published.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

SOURCES OF FUNDING

This work was carried out with the support of the National Council for Scientific and Technological Development – CNPq – Brazil.

DATA AVAILABILITY

Research data is available in the body of the document.

REFERENCES

1. Norman GR, Van Der Vleuten CPM, Newble DI, Dolmans DHJM, Mann KV, Rothman A, et al., organizers. International handbook of research in medical education. Dordrecht: Springer Netherlands; 2002. v. 7 [acesso em 11 set 2024]. Disponível em: http://link.springer.com/10.1007/978-94-010-0462-6.
2. Gruppen L. Clinical reasoning: defining it, teaching it, assessing it, studying it. West J Emerg Med. 2017;18(1):4-7.
3. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med. 2017;92(1):23-30.
4. Norman GR, Eva KW. Diagnostic error and clinical reasoning: diagnostic error and reasoning. Med Educ. 2010;44(1):94-100.
5. Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Acad Med. 2020;95(8):1166-71.
6. Cruess RL, Cruess SR, Steinert Y. Amending Miller's pyramid to include professional identity formation. Acad Med. 2016;91(2):180-5.
7. Miller GE. The assessment of clinical skills/competence/performance. Acad Med J Assoc Am Med Coll. 1990;65(9 Suppl):S63-67.
8. Witheridge A, Ferns G, Scott-Smith W. Revisiting Miller's pyramid in medical education: the gap between traditional assessment and diagnostic reasoning. Int J Med Educ. 2019;10:191-2.
9. Lee Y-S. OSCE for the medical licensing examination in Korea. 2008;24(12).
10. Bacchi S, Tan Y, Chim I, Dabarno M, Lubarsky S, Duggan P. Script concordance test examinations: student perception and study approaches. 2019;20(2).
11. Haug CJ, Drazen JM. Artificial intelligence and Machine Learning in clinical medicine. N Engl J Med. 2023;388(13):1201-8.
12. Mendes KDS, Silveira RCDCP, Galvão CM. Revisão integrativa: método de pesquisa para a incorporação de evidências na saúde e na enfermagem. Texto Contexto Enferm. 2008;17(4):758-64.
13. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan – a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.
14. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;(160).
15. Aagaard E, Teherani A, Irby DM. Effectiveness of the one-minute preceptor model for diagnosing the patient and the learner: proof of concept. Acad Med. 2004;79(1):42-9.
16. Amini M, Moghadami M, Kojuri J, Abbasi H, Abadi AAD, Molaee NA, et al. An innovative method to assess clinical reasoning skills: clinical reasoning tests in the second national medical science Olympiad in Iran. BMC Res Notes. 2011;4(1):418.
17. Amini M, Shahabi A, Moghadami M, Shams M, Anooshirvani A, Rostamipour H, et al. Psychometric characteristics of script concordance test (SCT) and its correlation with routine multiple choice question (MCQ) in internal medicine department. Biomed Res. 2017;28(19).
18. Anderson K, Peterson R, Tonkin A, Cleary E. The assessment of student reasoning in the context of a clinically oriented PBL program. Med Teach. 2008;30(8):787-94.
19. Artino AR, Cleary TJ, Dong T, Hemmer PA, Durning SJ. Exploring clinical reasoning in novices: a self-regulated learning microanalytic assessment approach. Med Educ. 2014;48(3):280-91.
20. Cohen Aubart F, Papo T, Hertig A, Renaud MC, Steichen O, Amoura Z, et al. Are script concordance tests suitable for the assessment of undergraduate students? A multicenter comparative study. Rev Médecine Interne. 2021;42(4):243-50.
21. 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(2):163-73.
22. Baño G, Di Lalla S, Domínguez P, Noel Seoane M, Wainsztein R, Ossorio MF, et al. Evaluación de una prueba para valorar el razonamiento clínico en diferentes niveles de capacitación médica. Rev Médica Chile. 2011;139(4):455-61.
23. Behrens C, Morales V, Parra P, Hurtado A, Fernández R, Giaconi E, et al. Diseño e implementación de OSCE para evaluar competencias de egreso en estudiantes de medicina en un consorcio de universidades chilenas. Rev Médica Chile. 2018;146(10):1197-204.
24. Berger AJ, Gillespie CC, Tewksbury LR, Overstreet IM, Tsai MC, Kalet AL, et al. Assessment of medical student clinical reasoning by "lay" vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination. Am J Surg. 2012;203(1):81-6.
25. Beullens J, Struyf E, Van Damme B. Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. Med Educ. 2006;40(12):1173-9.
26. Beullens J, Struyf E, Van Damme B. Do extended matching multiple-choice questions measure clinical reasoning? Med Educ. 2005;39(4):410-7.
27. Bhardwaj P, Black EW, Fantone JC, Lopez M, Kelly M. Script concordance tests for formative clinical reasoning and problem-solving assessment in general pediatrics. MedEdPORTAL. 2022;11274.
28. Blazek M, Bess J, Hirshbein L, Chiang C, Sastry D, Ravindranath D. The Short-Answer Vignette Examination (SAVE): an assessment tool for the core psychiatry clerkship. Acad Psychiatry. 2014;38(5):615-8.
29. Blunk DI, Tonarelli S, Gardner C, Quest D, Petitt D, Leiner M. Evaluating medical students' clinical reasoning in psychiatry using clinical and basic science concepts presented in session-level integration sessions. Med Sci Educ. 2019;29(3):819-24.
30. Boulouffe C, Doucet B, Muschart X, Charlin B, Vanpee D. Assessing clinical reasoning using a script concordance test with electrocardiogram in an emergency medicine clerkship rotation. Emerg Med J. 2014;31(4):313-6.
31. 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(Suppl 4):54-60.
32. Carrière B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a script concordance test. Ann Emerg Med. 2009;53(5):647-52.
33. Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, et al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. Med Educ. 2006;40(9):848-54.
34. Cheung JJH, Park YS, Aufderheide K, Holden J, Yudkowsky R. Optimizing clinical reasoning assessments with analytic and holistic ratings: examining the validity, reliability, and cost of a simplified patient note scoring procedure. Acad Med. 2022;97(11S):S15-21.
35. Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. BMC Med Educ. 2004;4(1):23.
36. Compère V, Moriceau J, Gouin A, Guitard PG, Damm C, Provost D, et al. Residents in tutored practice exchange groups have better medical reasoning as measured by the script concordance test: a pilot study. Anaesth Crit Care Pain Med. 2015;34(1):17-21.
37. Compère V, Abily J, Moriceau J, Gouin A, Veber B, Dupont H, et al. Residents in tutored practice exchange groups have better medical reasoning as measured by script concordance test: a controlled, nonrandomized study. J Clin Anesth. 2016;32:236-41.
38. Cooke S, Lemay JF, Beran T. Evolutions in clinical reasoning assessment: the evolving script concordance test. Med Teach. 2017;39(8):828-35.
39. Couraud S, Girard P, Girard N, Souquet PJ, Coiffard B, Charlin B, et al. Évaluation des connaissances sur le dépistage du cancer par test de concordance de script. Rev Mal Respir. 2016;33(5):333-42.

40. Covin Y, Longo P, Wick N, Gavinski K, Wagner J. Empirical comparison of three assessment instruments of clinical reasoning capability in 230 medical students. BMC Med Educ. 2020;20(1):264.

41. Derakhshandeh Z, Amini M, Kojuri J, Dehbozorgian M. Psychometric characteristics of Clinical Reasoning Problems (CRPs) and its correlation with routine multiple choice question (MCQ) in Cardiology department. J Adv Med Educ Prof. 2018;6(1):37-42.

42. Dong L, Li W, Ma D, Lv C, Chen C. The evaluation and comparison on different types of resident doctors in training through DxR Clinician system. J Phys Conf Ser. 2020;1549(4):042076.

43. Ducos G, Lejus C, Sztark F, Nathan N, Fourcade O, Tack I, et al. The Script Concordance Test in anesthesiology: validation of a new tool for assessing clinical reasoning. Anaesth Crit Care Pain Med. 2015;34(1):11-5.

44. Duffy B, Tully R, Stanton AV. An online case-based teaching and assessment program on clinical history-taking skills and reasoning using simulated patients in response to the Covid-19 pandemic. BMC Med Educ. 2023;23(1):4.

45. Duggan P, Charlin B. Summative assessment of 5thyear medical students' clinical reasoning by script concordance test: requirements and challenges. BMC Med Educ. 2012;12(1):29.

46. Durning SJ, Artino A, Boulet J, La Rochelle J, Van Der Vleuten C, Arze B, et al. The feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. Med Teach. 2012;34(1):30-7.

47. Enser M, Moriceau J, Abily J, Damm C, Occhiali E, Besnier E, et al. Background noise lowers the performance of anaesthesiology residents' clinical reasoning when measured by script concordance: a randomised crossover volunteer study. Eur J Anaesthesiol. 2017;34(7):464-70.

48. Fida M, Kassab SE. Do medical student's scores using different assessment instruments predict their scores in clinical reasoning using a computer-based simulation? Adv Med Educ Pract. 2015;135.

49. Fink MC, Reitmeier V, Stadler M, Siebeck M, Fischer F, Fischer MR. Assessment of diagnostic competences with standardized patients versus virtual patients: experimental study in the context of history taking. J Med Internet Res. 2021;23(3):e21196.

50. Fleiszer D, Hoover ML, Posel N, Razek T, Bergman S. Development and validation of a tool to evaluate the evolution of clinical reasoning in trauma using virtual patients. J Surg Educ. 2018;75(3):779-86.

51. Fonteneau T, Billion E, Abdoul C, Le S, Hadchouel A, Drummond D. Simulation game versus multiple choice questionnaire to assess the clinical competence of medical students: prospective sequential trial. J Med Internet Res. 2020;22(12):e23254.

52. Fürstenberg S, Helm T, Prediger S, Kadmon M, Berberat PO, Harendza S. Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. BMC Med Educ. 2020;20(1):368.

53. Gagnon R, Charlin B, Coletti M, Sauve E, Van Der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med Educ. 2005;39(3):284-91.

54. Gagnon R, Charlin B, Lambert C, Carrière B, Van Der Vleuten C. Script concordance testing: more cases or more questions? Adv Health Sci Educ. 2008;14(3):367-75.

55. Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? Adv Health Sci Educ. 2011;16(5):601-8.

56. Gallagher BD, Nematollahi S, Park H, Kurra S. Comparing students' clinical grades to scores on a standardized patient note-writing task. J Gen Intern Med. 2020;35(11):3243-7.

57. Gilkes L, Kealley N, Frayne J. Teaching and assessment of clinical diagnostic reasoning in medical students. Med Teach. 2022;44(6):650-6.

58. Iglesias Gómez C, González Sequeros O, Sarquella Brugada G, Padilla Del Rey ML, Salmerón Martínez D. Usefulness of SCT in detecting clinical reasoning deficits among pediatric professionals. Prog Pediatr Cardiol. 2021;61:101340.

59. Iglesias Gómez C, González Sequeros O, Salmerón Martínez D. Clinical reasoning evaluation using script concordance test in primary care residents. An Pediatría Engl Ed. 2022;97(2):87-94.

60. Goos M, Schubach F, Seifert G, Boeker M. Validation of undergraduate medical student script concordance test (SCT) scores on the clinical assessment of the acute abdomen. BMC Surg. 2016;16(1):57.

61. Goulet F, Jacques A, Gagnon R, Charlin B, Shabah A. Poorly performing physicians: does the script concordance test detect bad clinical reasoning? J Contin Educ Health Prof. 2010;30(3):161-6.

62. Groves M, O'Rourke P, Alexander H. The clinical reasoning characteristics of diagnostic experts. Med Teach. 2003;25(3):308-13.

63. Groves M, Dick ML, McColl G, Bilszta J. Analysing clinical reasoning characteristics using a combined methods approach. BMC Med Educ. 2013;13(1):144.

64. Hamui M, Ferreira JP, Torrents M, Torres F, Ibarra M, Ossorio MF, et al. Script Concordance Test: first nationwide experience in pediatrics. Arch Argent Pediatr. 2018;116(1):e151-5.

65. Haring CM, Klaarwater CCR, Bouwmans GA, Cools BM, Van Gurp PJM, Van Der Meer JWM, et al. Validity, reliability and feasibility of a new observation rating tool and a post encounter rating tool for the assessment of clinical reasoning skills of medical students during their internal medicine clerkship: a pilot study. BMC Med Educ. 2020;20(1):198.

66. Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA. Assessment of clinical reasoning: a script concordance test designed for pre-clinical medical students. Med Teach. 2011;33(6):472-7.

67. Humbert AJ, Miech EJ. Measuring gains in the clinical reasoning of medical students: longitudinal results from a school-wide Script Concordance Test. Acad Med. 2014;89(7):1046-50.

68. Humbert AJ, Besinger B, Miech EJ. Assessing Clinical reasoning skills in scenarios of uncertainty: convergent validity for a Script Concordance Test in an emergency medicine clerkship and residency. Acad Emerg Med. 2011;18(6):627-34.

69. Huwendiek S, Reichert F, Duncker C, De Leng BA, Van Der Vleuten CPM, Muijtjens AMM, et al. Electronic assessment of clinical reasoning in clerkships: a mixed-methods comparison of long-menu key-feature problems with context-rich single best answer questions. Med Teach. 2017;39(5):476-85.

70. Im S, Kim DK, Kong HH, Roh HR, Oh YR, Seo JH. Assessing clinical reasoning abilities of medical students using clinical performance examination. Korean J Med Educ. 2016;28(1):35-47.

71. Iravani K, Amini M, Doostkam A, Dehbozorgian M. The validity and reliability of script concordance test in otolaryngology residency training. 4(2).

72. Jerant AF, Azari R. Validity of scores generated by a web-based multimedia simulated patient case software: a pilot study: Acad Med. 2004;79(8):805-11.

73. Kania RE, Verillaud B, Tran H, Gagnon R, Kazitani D, Huy PTB, et al. Online Script Concordance Test for clinical reasoning assessment in otorhinolaryngology: the association between performance and clinical experience. Arch Otolaryngol Neck Surg. 2011;137(8):751-5.

74. Kaur M, Singla S, Mahajan R. Script Concordance Test in pharmacology: maiden experience from a medical school in India. J Adv Med Educ Prof. 2020;8(3):115-20.

75. Kazour F, Richa S, Zoghbi M, El-Hage W, Haddad FG. Using the Script Concordance Test to evaluate clinical reasoning skills in psychiatry. Acad Psychiatry. 2017;41(1):86-90.

76. Kelly W, Durning S, Denton G. Comparing a Script Concordance Examination to a multiple-choice examination on a core internal medicine clerkship. Teach Learn Med. 2012;24(3):187-93.

77. Kim S, Choi I, Yoon BY, Kwon MJ, Choi S-jin, Kim SH, et al. Medical students' thought process while solving problems in 3 different types of clinical assessments in Korea: clinical performance examination, multimedia case-based assessment, and modified essay question. J Educ Eval Health Prof. 2019;16:10.

78. Kotwal S, Klimpl D, Tackett S, Kauffman R, Wright S. Documentation of clinical reasoning in admission notes of hospitalists: validation of the CRANAPL assessment rubric. J Hosp Med. 2019;14(12):746-53.

79. Kün-Darbois JD, Annweiler C, Lerolle N, Lebdai S. Script concordance test acceptability and utility for assessing medical students' clinical reasoning: a user's survey and an institutional prospective evaluation of students' scores. BMC Med Educ. 2022;22(1):277.

80. Kunina-Habenicht O, Hautz WE, Knigge M, Spies C, Ahlers O. Assessing clinical reasoning (ASCLIRE): instrument development and validation. Adv Health Sci Educ. 2015;20(5):1205-24.

81. Lai JH, Cheng KH, Wu YJ, Lin CC. Assessing clinical reasoning ability in fourth-year medical students via an integrative group history-taking with an individual reasoning activity. BMC Med Educ. 2022;22(1):573.

82. Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. Radiat Oncol. 2009;4(1):7.

83. Lineberry M, Hornos E, Pleguezuelos E, Mella J, Brailovsky C, Bordage G. Experts' responses in script concordance tests: a response process validity investigation. Med Educ. 2019;53(7):710-22.

84. Loh KW, Rotgans JI, Tan K, Tan NCK. Predictors of clinical reasoning in neurological localisation: a study in internal medicine residents. Asia Pac Sch. 2020;5(3):54-61.

85. Lukas RV, Blood A, Park YS, Brorson JR. Assessment of neurological clinical management reasoning in medical students. J Clin Neurosci. 2014;21(6):919-22.

86. Talvard M, Olives JP, Mas E. Évaluation des étudiants en médecine lors de leur stage en gastro-entérologie pédiatrique par un test de concordance de script. Arch Pédiatrie. 2014;21(4):372-6.

87. Marie I, Sibert L, Roussel F, Hellot MF, Lechevallier J, Weber J. Le test de concordance de script: un nouvel outil d'évaluation du raisonnement et de la compétence clinique en médecine interne? Rev Médecine Interne. 2005;26(6):501-7.

88. Mathieu S, Couderc M, Glace B, Tournadre A, Malochet-Guinamand S, Pereira B, et al. Construction and utilization of a script concordance test as an assessment tool for dcem3 (5th year) medical students in rheumatology. BMC Med Educ. 2013;13(1):166.

89. Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? Am J Surg. 2007;193(2):248-51.

90. Monajemi A, Soltani A, Hadadgar A, Hadizadeh F, Adibi P, Akbari R, et al. A comprehensive test of clinical reasoning for medical students: an olympiad experience in Iran. J Educ Health Promot. 2012;1(1):10.

91. Mzoughi K, Zairi I, Kedous MA, Mhamdi SE, Dhiab MB, Mghaieth F. Le test de concordance de script comme outil d'évaluation sanctionnante en cardiologie script concordance test as a sanctionnal evaluation in cardiology.

92. Nazim S, Talati J, Pinjani S, Biyabani SR, Ather M, Norcini J. Assessing clinical reasoning skills using Script Concordance Test (SCT) and extended matching questions (EMQs): a pilot for urology trainees. J Adv Med Educ Prof. 2019;7(1). doi: https://doi.org/10.30476/jamp.2019.41038

93. Nomura O, Itoh T, Mori T, Ihara T, Tsuji S, Inoue N, et al. Creating clinical reasoning assessment tools in different languages: adaptation of the pediatric emergency medicine Script Concordance Test to Japanese. Front Med. 2021;8:765489.

94. Nouh T, Boutros M, Gagnon R, Reid S, Leslie K, Pace D, et al. The script concordance test as a measure of clinical reasoning: a national validation study. Am J Surg. 2012;203(4):530-4.

95. Omega A, Wijaya Ramlan AA, Soenarto RF, Heriwardito A, Sugiarto A. Assessing clinical reasoning in airway related cases among anesthesiology fellow residents using Script Concordance Test (SCT). Med Educ Online. 2022;27(1):2135421.

96. Park WB, Kang SH, Myung SJ, Lee YS. Does Objective Structured Clinical Examinations Score reflect the clinical reasoning ability of medical students? Am J Med Sci. 2015;350(1):64-7.

97. Peterson BD, Magee CD, Martindale JR, Dreicer JJ, Mutter MK, Young G, et al. REACT: Rapid Evaluation Assessment of Clinical Reasoning Tool. J Gen Intern Med. 2022;37(9):2224-9.

98. Petrucci AM, Nouh T, Boutros M, Gagnon R, Meterissian SH. Assessing clinical judgment using the Script Concordance Test: the importance of using specialty-specific experts to develop the scoring key. Am J Surg. 2013;205(2):137-40.

99. Peyrony O, Hutin A, Truchot J, Borie R, Calvet D, Albaladejo A, et al. Impact of panelists' experience on script concordance test scores of medical students. BMC Med Educ. 2020;20(1):313.

100. Piovezan RD, Custódio O, Cendoroglo MS, Batista NA, Lubarsky S, Charlin B. Assessment of undergraduate clinical reasoning in geriatric medicine: application of a Script Concordance Test. J Am Geriatr Soc. 2012;60(10):1946-50.

101. Pottier P, Castillo JM, Boet S, Le Pabic E, Hardouin JB. Validation d'une grille d'évaluation des compétences cliniques utilisée par des patients standardisés. Rev Médecine Interne. 2016;37(12):802-10.

102. Power A, Lemay JF, Cooke S. Justify your answer: the role of written think aloud in Script Concordance Testing. Teach Learn Med. 2017;29(1):59-67.

103. Rajeswaran V, Devine L, Lorens E, Robertson S, Huszti E, Panisko DM. Types of clinical reasoning in a summative clerkship oral examination. Med Teach. 2022;44(6):657-63.

104. Reinert A, Berlin A, Swan-Sein A, Nowygrod R, Fingeret A. Validity and reliability of a novel written examination to assess knowledge and clinical decision making skills of medical students on the surgery clerkship. Am J Surg. 2014;207(2):236-42.

105. Ruiz JG, Tunuguntla R, Charlin B, Ouslander JG, Symes SN, Gagnon R, et al. The Script Concordance Test as a measure of clinical reasoning skills in geriatric urinary incontinence. J Am Geriatr Soc. 2010;58(11):2178-84.

106. Sadeghi A, Ali Asgari A, Moulaei N, Mohammadkarimi V, Delavari S, Amini M, et al. Combination of different clinical reasoning tests in a national exam. J Adv Med Educ Prof. 2019;7(4):230-4.

107. Schauber SK, Hautz SC, Kämmer JE, Stroben F, Hautz WE. Do different response formats affect how test takers approach a clinical reasoning task? An experimental study on antecedents of diagnostic accuracy using a constructed response and a selected response format. Adv Health Sci Educ. 2021;26(4):1339-54.

108. Schaye V, Miller L, Kudlowitz D, Chun J, Burk-Rafel J, Cocks P, et al. Development of a clinical reasoning documentation assessment tool for resident and fellow admission notes: a shared mental model for feedback. J Gen Intern Med. 2021;37(3):507-12.

109. Schaye V, Guzman B, Burk-Rafel J, Marin M, Reinstein I, Kudlowitz D, et al. Development and validation of a Machine Learning Model for automated assessment of resident clinical reasoning documentation. J Gen Intern Med. 2022;37(9):2230-8.

110. See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. Med Educ. 2014;48(11):1069-77.

111. Shenoy RV, Newbern D, Cooke DW, Chia DJ, Panagiotakopoulos L, DiVall S, et al. The structured oral examination: a method to improve formative assessment of fellows in pediatric endocrinology. Acad Pediatr. 2022;22(7):1091-6.

112. Sibert L, Darmoni SJ, Dahamna B, Hellot MF, Weber J, Charlin B. On line clinical reasoning assessment with Script Concordance Test in urology: results of a French pilot study. BMC Med Educ. 2006;6(1):45.

113. Silberman EK, Ramesh S, Adler D, Sargent J, Moore T, Blanco M. Assessing residents' skills in psychiatric reasoning: the tufts test of formulation and treatment planning. Acad Psychiatry. 2020;44(6):701-8.

114. Sim JH, Aziz YFA, Mansor A, Vijayananthan A, Foong CC, Vadivelu J. Students' performance in the different clinical skills assessed in OSCE: what does it reveal? Med Educ Online. 2015;20(1):26185.

115. Min Simpkins AA, Koch B, Spear-Ellinwood K, St. John P. A developmental assessment of clinical reasoning in preclinical medical education. Med Educ Online. 2019;24(1):1591257.

116. Steinberg E, Cowan E, Lin M, Sielicki A, Warrington S. Assessment of emergency medicine residents' clinical reasoning: validation of a Script Concordance Test. West J Emerg Med . 2020;21(4) [acesso em 7 fev 2024]. Disponível em: https://escholarship.org/uc/item/92b825mf.

117. Subra J, Chicoulaa B, Stillmunkès A, Mesthé P, Oustric S, Rougé Bugat ME. Reliability and validity of the script concordance test for postgraduate students of general practice. Eur J Gen Pract. 2017;23(1):209-14.

118. Sulaiman ND, Hamdy H. Assessment of clinical competencies using clinical images and videos "CIVA". BMC Med Educ. 2013;13(1):78.

119. Surry LT, Torre D, Trowbridge RL, Durning SJ. A mixed-methods exploration of cognitive dispositions to respond and clinical reasoning errors with multiple choice questions. BMC Med Educ. 2018;18(1):277.

120. Sutherland RM, Reid KJ, Chiavaroli NG, Smallwood D, McColl GJ. assessing diagnostic reasoning using a standardized case-based discussion. J Med Educ Curric Dev. 2019;6:238212051984941.

121. Thammasitboon S, Sur M, Rencic JJ, Dhaliwal G, Kumar S, Sundaram S, et al. Psychometric validation of the reconstructed version of the assessment of reasoning tool. Med Teach. 2021;43(2):168-73.

122. Van Den Broek WES, Van Asperen MV, Ten Cate OThJ, Custers E, Valk GD. Effects of two different instructional formats on scores and reliability of a script concordance test. Perspect Med Educ. 2012;1(3):119-28.

123. Volkan K, Simon SR, Baker H, Todres ID. Psychometric structure of a comprehensive Objective Structured Clinical Examination: a factor analytic approach. Adv Health Sci Educ. 2004;9(2):83-92.

124. Waechter J, Allen J, Lee CH, Zwaan L. Development and pilot testing of a data-rich clinical reasoning training and assessment tool. Acad Med. 2022;97(10):1484-8.

125. Wan MSH, Tor E, Hudson JN. Construct validity of script concordance testing: progression of scores from novices to experienced clinicians. Int J Med Educ. 2019;10:174-9.

126. Wan MSH, Tor E, Hudson JN. Examining response process validity of script concordance testing: a think-aloud approach. Int J Med Educ. 2020;11:127-35.

127. Williams RG, Klamen DL, White CB, Petrusa E, Fincher RME, Whitfield CF, et al. Tracking development of clinical reasoning ability across five medical schools using a Progress Test. Acad Med. 2011;86(9):1148-54.

128. Wilson AB, Pike GR, Humbert AJ. Testing for multilevel dimensionality: a higher-order factor analysis of a Script Concordance Test. Med Sci Educ. 2015;25(4):439-46.

129. Wilson AB, Pike GR, Humbert AJ. Analyzing Script Concordance Test scoring methods and items by difficulty and type. Teach Learn Med. 2014;26(2):135-45.

130. Yassin BAG, Almothaffar A, Aladhadh M, Hussein MF, Ghafour ZA, Gorial FI. Studying clinical reasoning of internal medicine residents in medical city campus using Script Concordance Test.

131. Yudkowsky R, Hyderi A, Holden J, Kiser R, Stringham R, Gangopadhyaya A, et al. Can nonclinician raters be trained to assess clinical reasoning in postencounter patient notes? Acad Med. 2019;94(11S):S21-7.

132. Zavaleta-Hernández S, Cerón-Rodríguez M, Olivar-López V, Espinoza R, Rizzoli-Córdoba A. Validation of the Script Concordance Test as an instrument to assess clinical reasoning of residents in pediatric emergency medicine in Mexico. Bol Med Hosp Infant Mex. 2011;68.

133. Zegota S, Becker T, Hagmayer Y, Raupach T. Using item response theory to appraise key feature examinations for clinical reasoning. Med Teach. 2022;44(11):1253-9.

134. Zia S, Obaid M, Ahmed J, Qazi S. Key-feature questions for assessment of clinical reasoning: reliable and valid or not? RAWAL Med J. 2019;44(1):189-191.

135. Zuo T, Sun B, Guan X, Zheng B, Qu B. Evidence of construct validity of computer-based tests for clinical reasoning: instrument validation study. JMIR Serious Games. 2021;9(4):e17670.

136. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 27 2011;2:53-5.

137. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA 2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Syst Rev. 2022;18(2):e1230.

138. Daniel M, Rencic J, Durning SJ, Holmboe E, Heist B, Lubarsky S, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med. 2019;94(6):901-12.