NBME®

WEBINAR

POWER OF COMMUNITY: ADDRESSING THE CHALLENGES OF CLINICAL REASONING ASSESSMENT

NOVEMBER 18, 2024

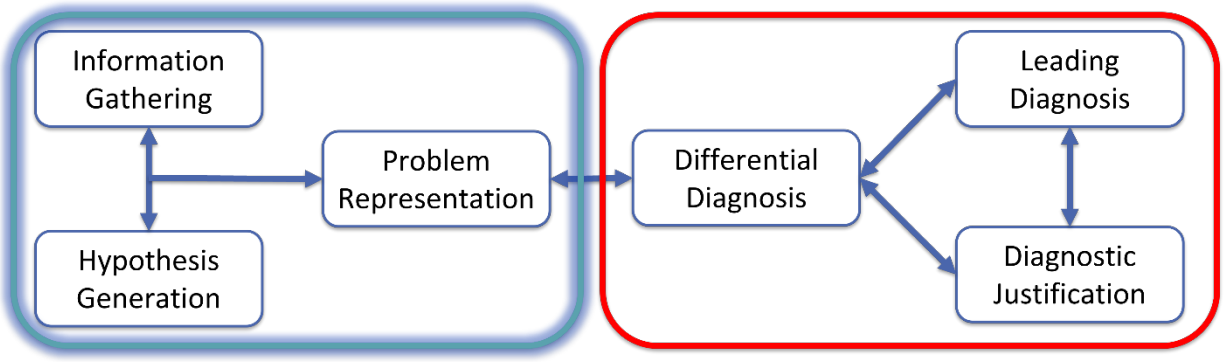
FEATURING

- **CHRIS FEDDOCK, MD**
  VP, COMPETENCY-BASED ASSESSMENT, NBME
- **SU SOMAY, EdD**
  LEAD MEASUREMENT SCIENTIST, NBME
- **THAI ONG, PhD**
  SENIOR PSYCHOMETRICIAN, NBME
- **ANALIA CASTIGLIONI, MD**
  UNIVERSITY OF CENTRAL FLORIDA COLLEGE OF MEDICINE
- **DAVID GORDON, MD**
  DUKE UNIVERSITY SCHOOL OF MEDICINE
- **MATTHEW KELLEHER, MD, MEd**
  UNIVERSITY OF CINCINNATI COLLEGE OF MEDICINE

INTRODUCTION

Given the association with diagnostic accuracy and medical errors, clinical reasoning (CR) skills are essential to the effective practice of medicine. Further, effective learner feedback is vital to support skills development and continued growth. However, the assessment of CR faces persistent challenges due to its complex nature. For example, CR has been conceptualized to include both diagnostic (collecting and analyzing clinical data to arrive at a diagnosis) and management reasoning (implementing an appropriate treatment plan).

Another barrier to the assessment of CR skills is the lack of a unified definition and framework.[1,2] Additionally, currently used assessments draw inferences from outcomes, such as questions that learners asked, diagnoses made and justifications provided for those diagnoses. Outcome-oriented tools appropriately assess if a learner reached correct or plausible conclusions, however, learners do not receive insight about how they reached a conclusion or where flaws may exist in their process of reasoning. Without these insights, it is not possible to provide learners with detailed feedback to improve their reasoning process, which is an essential component of an assessment for learning.



Acad Med. 2019;94:902–912.

Creating Community

Recognizing these challenges and understanding that effective partnerships are necessary to produce better learning outcomes,[3] NBME launched the Objective Structured Clinical Examination (OSCE) for Clinical Reasoning Creative Community (CC) initiative in January 2022. The CC initiative aims to leverage the expertise of NBME staff, medical education faculty and learners to support learner growth through the development of innovative assessments for learning. NBME received applications from 99 medical schools, representing 60% of all U.S. Liaison Committee on Medical Education-accredited institutions. From these, individuals from 10 schools were selected (see Table 1) to ensure contributions from diverse clinical learning environments and geographic regions.

ASSESSMENT DESIGN AND DEVELOPMENT

The CC team includes lead faculty from 10 medical schools (see Table 1) and key NBME staff members (see Table 2), bringing together a wide range of expertise in clinical reasoning, teaching, assessment and measurement science. The CC began its work by selecting a theoretical framework for clinical reasoning to guide the assessment design and development approach.[4] The CC team focused on diagnostic reasoning (DR) components within this framework, the process of identifying and understanding a patient's health problem through the collection and analysis of clinical data, the process of formulating differential diagnoses, and ultimately the process of arriving at a definitive diagnosis. Conceptualized as an unfolding process, the CC further prioritized two subdomains of DR: hypothesis-driven information gathering (HDIG) and problem representation (PR) processes. These two subdomains represent the foundational aspects of DR: generating and refining hypotheses based on gathered information and creating an accurate representation of the patient's problem.

Table 1

Schools and lead faculty members participating in the OSCE for Clinical Reasoning Creative Community

Duke University School of Medicine	David Gordon, MD
Howard University College of Medicine	Sharon Dowell, MBBS, MS Vishal Poddar, MD
Kaiser Permanente Bernard J. Tyson School of Medicine	Candace Pau, MD
Morehouse School of Medicine	Khadeja Johnson, MD
Southern Illinois University School of Medicine	Debra Klamen, MD, MHPE
University of Central Florida College of Medicine	Analia Castiglioni, MD
University of Cincinnati College of Medicine	Matthew Kelleher, MD, MEd
University of Connecticut School of Medicine	Laurie Caines, MD
University of New England College of Osteopathic Medicine	Kristen Mitchell, DO
University of New Mexico School of Medicine	Jan Veesart, MD

Table 2

NBME staff core team members participating in the OSCE for Clinical Reasoning Creative Community

Christopher Feddock, MD, MBA	Vice President, Educational Strategy
Marni Grambau	Director, Assessment Alliance
Scott Mandel	Senior Test Development Analyst
John Moore, PhD	Director, Assessment Data Initiatives
Ann Nolan	Process Expert
Thai Ong, PhD	Senior Psychometrician

Jeannette Sanger	Manager, Performance-Based Assessment
Su Somay, EdD	Lead Measurement Scientist

A New Approach

The CC used evidence-centered design (ECD), which is a principled approach to assessment design and development.[5] A systematic approach such as ECD is particularly necessary when developing assessments for complex competencies (e.g., problem solving, CR, collaboration) and when using multidimensional performance tasks (e.g., simulations or game-based assessments). Following this methodology, the CC undertook four key steps:

1. Specifying the purpose of the assessment and what the assessment results are intended to indicate about learners
2. Prioritizing the evidence that is necessary for evaluating the learner in the identified areas
3. Designing the most appropriate and fit-for-purpose tasks to yield that evidence
4. Identifying the data that should be collected from each of those tasks

To our knowledge, this is the first time a principled approach such as ECD has been implemented to design and develop assessments for evaluating CR skills.

This group focused on developing an OSCE assessment during the clerkship and post-clerkship years to support learners' DR development. Through the concurrent development of assessment tools and cases, the CC designed a process-oriented assessment of DR that provides process-related feedback. This parallel development approach guarantees that the behaviors and skills that are critical for evaluating learners' diagnostic reasoning were elicited by the new cases and accurately reflected in each rubric criterion.[6]

Hypothesis-Driven Information Gathering (HDIG)

The CC identified critical elements that impact overall diagnostic success and noted that some important dimensions of performance were missing from commonly used assessment instruments.[7-10] For example, the timing and sequence of questions asked to a patient represents an important dimension of performance. With the goal of providing process-related feedback, the CC developed the HDIG rubric to evaluate a learner's line of questioning during the patient encounter. The HDIG rubric characterizes learner performance using three criteria:

1. Characterization of the Chief Concern (CCC): the ability to elicit the essential components of the presenting symptom; adequate characterization of the chief complaint provides a foundation from which to pursue hypothesis-directed inquiry
2. Curiosity: the ability to identify relevant information available prior to the patient encounter (i.e., diagnostic clues) and explore the diagnostic clues in more depth by asking follow-up questions that change the likelihood or prioritization of possible hypotheses
3. Agility: the ability to promptly recognize critical information (i.e., a pivot point) elicited during the patient encounter, then explore that in more depth by asking follow-up questions that change the likelihood or prioritization of possible hypotheses

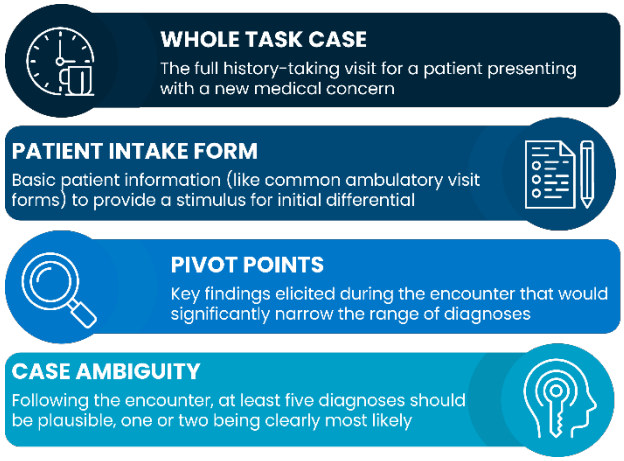
Together, these criteria allow evaluation of the breadth, depth, quality and responsiveness of learners' information gathering during the patient encounter—providing an authentic process measure of reasoning.

Problem Representation (PR)

Learners' PR skills were evaluated by their pre- and post-encounter responses, wherein they were asked to create a PR and a differential diagnosis based on the available patient information. Specifically, learners were asked to list their differential diagnosis, indicate which diagnosis they consider "most likely" and explain the critical elements of the patient information for determining their differential (i.e., the key features).

Case Development

The case development process was modeled after the Association of Standardized Patient Educators Standards of Best Practice.[11] As the assessment constructs and rubrics evolved, the group centered on a few design specifications to enhance the effectiveness of the cases.



PILOT STUDY

After several rounds of proof-of-concept tests, the CC implemented final revisions to the assessment design, stimuli and rubrics for a pilot study to collect initial validity evidence. During a two-week testing period, 76 post-clerkship medical students completed four cases (see Table 3) using a virtual OSCE format. The pilot produced a total of 304 unique student-patient encounter videos, and each encounter consisted of three different assessment elements: a pre-encounter task, the patient encounter, and a post-encounter task.

Table 3

Descriptions of cases used in the pilot study

Case	Patient Demographics	Chief Concern
1	33 y/o cisgender woman	"I can't catch my breath."
2	40 y/o cisgender man	"I can't stop throwing up."
3	46 y/o cisgender woman	"My legs have been feeling really weak."
4	65 y/o cisgender man	"I haven't been able to sleep through the night."

Evaluative Approach

Learner performance in each patient encounter was evaluated by two independent faculty raters using the HDIG rubric. Each HDIG rubric criterion was rated on a four-point scale, with 1 reflecting no demonstration of the specified behavior, 2 indicating partial demonstration, 3 signaling expected demonstration for a learner ready for supervised practice and 4 showing full demonstration. The rubric provided detailed descriptions of each criterion and the associated performance levels. Before faculty members began rating, they completed a frame-of-reference training using examples showcasing the relationship between student performance and the rubric scale (see Table 4 for the HDIG scores for an example learner).

One of the key objectives of the CC team was to ensure timely feedback for learners, which required automated scoring as part of the solution. To achieve this, faculty members were trained to annotate, rather than rate, pre- and post-encounter activity responses. Annotation in this context involved manually tagging elements of learner responses that aligned with the key essentials and the expected differential diagnoses for each case. These annotations were subsequently used to build automated scoring models to streamline the feedback process.

Table 4

HDIG ratings for an example learner

Learner	Case	CCC	Curiosity	Agility
A	1	4	2	3
A	2	4	3	2
A	3	3	2	2
A	4	4	1	3

KEY FINDINGS

The NBME core team evaluated four research questions (RQ) following the pilot study.

RQ1: How accurate were Standardized Patients (SPs) in portraying DR cases as developed specifically for the OSCE for CR assessment? What factors were associated with SP case portrayal accuracy?

Three raters who were experts in SP quality control evaluated the accuracy of SPs using a quality control rubric derived from scoring guidelines and SP training materials. The quality control rubric included criteria relevant to the specific cases, such as pivot point delivery and aspects of patient interaction. A random set of 101 encounters across the four cases were reviewed, which constituted 38% of the total eligible encounters from the pilot. There was a strong consensus among raters, with two or more raters agreeing in 92% to 97% of instances. Learners elicited responses from SPs with a high rate of correctness, ranging from 86% to 97%, depending on the case. The predominant type of error concerned "volunteering" of information, which was the most frequent in 3 out of 4 cases, accounting for 3% to 12% of responses. These results indicate SP portrayal accuracy was above the desirable standard across all four cases, which eliminated SP portrayal errors as a potential confounding factor to the interpretation of subsequent analyses.

RQ2: How did learners perform on the pre-encounter and post-encounter tasks?

Since the diagnostic value of key features for a given case can vary significantly, the CC team decided to categorize them into two distinct groups: Critical and Relevant features. Critical features are the key elements of each case that are crucial for making a correct diagnosis. In contrast, Relevant features are important elements that offer additional context and detail to enhance the overall understanding of the patient's condition, though they are less essential for identifying the correct diagnosis. Notable variability existed in the proportion of Critical and Relevant features acquired by learners both across different cases and between pre- and post-encounter tasks within each case (see Table 5). If the post-encounter critical key feature proportions are considered a proxy for case difficulty, Case 2 appeared to be the most challenging for learners, whereas Case 4 seemed to be the least challenging.

Table 5

Average proportion of features that learners noted in pre- and post-encounter activities, by Case

	Critical Features		Relevant Features	
Case	Pre-Encounter	Post-Encounter	Pre-Encounter	Post-Encounter
1	44%	59%	44%	25%
2	57%	50%	29%	40%
3	59%	63%	59%	16%
4	49%	71%	39%	29%

The percentage of learners who listed the correct diagnosis on their post-encounter task for each case is shown in Table 6. Across cases, there was variation in learner performance, especially when comparing the percentage of those who listed the correct diagnosis across cases. If the post-encounter differentials are considered a proxy for case difficulty, Case 1 appeared to be the most challenging for learners, whereas Case 4 appeared to be the least challenging.

Table 6

Percentage of learners who listed the correct diagnosis on their differential list

Case	% Differential List	% Differential List & Marked as Most Likely
1	16%	5%
2	61%	34%
3	57%	49%
4	89%	76%

These results indicate that learners' performance on the pre- and post-encounter tasks varied substantially by case. This has been a consistent finding with DR, given varying learner familiarity with the most likely diagnosis for the case alongside other factors that may be related to case difficulty (defined as the percentage of learners who listed the correct diagnosis in the post-encounter differential).

RQ3: What was the inter-rater reliability of the HDIG ratings?

Inter-rater reliability was evaluated using both consistency and agreement measures across the rater pairs. Inter-rater consistency assesses how well the raters aligned in their rank ordering of learners (e.g., Did Rater A rank learners in the same order as Rater B?), whereas inter-rater agreement measures how closely the raters matched in their absolute ratings of learners (e.g., Did Rater A give the same absolute rating to Learner A as Rater B did?). Evaluating both inter-rater consistency and agreement is crucial for a comprehensive understanding of rater performance. Inter-rater consistency and agreement values closer to 1 are desirable, as a value of 1 indicates perfect consistency and agreement between rater pairs.

The inter-rater consistency and agreement values for the HDIG rubric were similar, indicating comparable interpretations across the two sets of values. The mean inter-rater consistency and agreement values were ~0.40 across the three HDIG criteria, with raters showing the least agreement and consistency on the Agility criterion and the most agreement and consistency on the Curiosity criterion. The low to moderate inter-rater reliability, as anticipated in this initial pilot, suggests that improving the clarity of the scoring criteria or offering additional training for raters—especially for the Agility criterion—could enhance consistency in evaluations.

RQ4: What was the overall reliability of HDIG scores? How did changes in assessment design elements and different weighting schemes affect this reliability? Additionally, what was the relationship between composite scores and diagnostic accuracy?

Although the inter-rater reliability values in RQ3 provide some information about rater agreement and consistency, they do not fully account for other factors that may influence overall HDIG score reliability, such as case difficulty. Therefore, it is important to account for both rater and case variations and their impact on overall HDIG score reliability. Additionally, exploring different weighting methods for creating a composite HDIG score and their impact on composite score reliability is necessary for future use of these scores. For example, is it more reliable to report HDIG scores as one composite score than three individual scores? If so, what weighting method should be used in calculating the HDIG composite scores? These questions were evaluated using multivariate generalizability theory.[13,14] There were three primary takeaways from these analyses:

1. Individual HDIG scores were moderately reliable.
2. Composite HDIG scores (weighted average of CCC, Curiosity and Agility scores) were more reliable than the individual HDIG scores. Composite HDIG scores based on theoretical weighting (subject matter experts weighted based on relative importance) resulted in the highest reliability.
3. HDIG score reliability is positively impacted by number of cases.

Finally, learners with higher HDIG scores achieved significantly greater diagnostic accuracy than their peers. Considering the favorable reliability of the composite HDIG scores, their predictive power for diagnostic outcomes was analyzed using generalized linear mixed modeling.[15] The effect of the HDIG composite score on diagnostic outcomes was statistically significant ($p = 0.03$) when learners' composite HDIG scores (fixed effect) were regressed on diagnostic accuracy while accounting for the case and student effects (random effects).

CONCLUSION

NBME's inaugural CC initiative made important theoretical and practical contributions to the assessment of DR. The CC initiative aimed to address the complex challenges associated with assessing DR in medical education. Key findings from the initiative and its pilot study are as follows:

- Students' cognitive processes during the pre- and post-encounter tasks, as well as during the patient encounters themselves, aligned with the intended processes the cases were designed to elicit.
- Learners varied in their ability to identify Critical features across different cases and between pre- and post-encounter tasks within a case.
- The percentage of learners who listed the correct diagnosis on their differential list varied significantly across cases, pointing to differences in case difficulty.
- Individual and composite HDIG score reliability was notably influenced most by number of cases while the rater effect was negligible; therefore, increasing the number of cases (as opposed to raters) may be a worthwhile approach to increasing individual and composite score reliabilities.
- HDIG composite scores with theoretical weights were more reliable than individual HDIG scores, suggesting composite scores should be included in future score interpretation and reporting.
- HDIG composite scores were significantly predictive of diagnostic outcomes, indicating learners with higher HDIG scores had significantly better diagnostic accuracy compared to their peers.

By focusing on formative *process-oriented* evaluation, this approach represents a significant shift toward a more nuanced and effective assessment of CR skills, addressing one of the critical needs in competency-based medical education and providing a model for future advancements in assessment of hard-to-measure constructs.

Throughout this two-year initiative, the CC members attended numerous regional and national meetings, continuously sharing their progress and findings. This consistent dissemination allowed the team to receive ongoing feedback from the medical education community, affirming the importance of the CC's targeted constructs and demonstrating that conceptualization and operationalization (i.e., HDIG criteria) of these constructs have broad applicability beyond the OSCE method.

The CC findings underscore the necessity for comprehensive and theory-driven strategies to support the development and enhancement of CR skills in medical students. Further, this work highlights the importance of standardizing and validating CR assessments to optimize the assessment and feedback quality. While the pilot study provided initial validity evidence, it also identified several potential next steps to improve the quality of the assessment. First, the current case development framework needs to be expanded to allow for a more comprehensive conceptualization and integration of the elements contributing to case difficulty and complexity. Further, given the considerable time investment required from the CC members, future steps should include exploration of leveraging large language models to facilitate case development. Second, the rating and annotation tasks place a significant burden on faculty. Developing a fully automated evaluation of the pre- and post-encounter and the patient encounter performances is necessary for scalability. Automation is also critical for the provision of prompt feedback. Last, continued piloting and validation efforts are essential to build a robust validity argument.

REFERENCES

1. Norman G. (2005). Research in Clinical Reasoning: Past history and Current Trends. *Medical Education*, 39(4), 418–427. https://doi.org/10.1111/j.1365-2929.2005.02127.x
2. Connor, D. M., Durning, S. J., & Rencic, J. J. (2020). Clinical Reasoning as a Core Competency. *Academic Medicine: Journal of the Association of American Medical Colleges*, 95(8), 1166–1171. https://doi.org/10.1097/ACM.0000000000003027
3. Englander, R., Holmboe, E., Batalden, P., Caron, R. M., Durham, C. F., Foster, T., Ogrinc, G., Ercan-Fang, N., & Batalden, M. (2020). Coproducing Health Professions Education: A Prerequisite to Coproducing Health Care Services? *Academic Medicine: Journal of the Association of American Medical Colleges*, 95(7), 1006–1013. https://doi.org/10.1097/ACM.0000000000003137
4. Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C. A., Ballard, T., Artino, A. R., Jr, Sergio Da Silva, A., Cleary, T., Stojan, J., & Gruppen, L. D. (2019). Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic Medicine: Journal of the Association of American Medical Colleges*, 94(6), 902–912. https://doi.org/10.1097/ACM.0000000000002618
5. Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A Brief Introduction to Evidence-Centered Design. *ETS Research Report Series*, 2003(1), i-29.
6. Bowen J. L. (2006). Educational Strategies to Promote Clinical Diagnostic Reasoning. *The New England Journal of Medicine*, 355(21), 2217–2225. https://doi.org/10.1056/NEJMra054782
7. Hasnain, M., Bordage, G., Connell, K.J., Sinacore, J.M. (2001) History-Taking Behaviors Associated With Diagnostic Competence of Clerks: An Exploratory Study. *Academic Medicine*, 76(10 Suppl), S14-7. https://doi.org/10.1097/00001888-200110001-00006
8. Nendaz, M.R., Gut, A.M., Perrier, A., Louis-Simonet, M., Blondon-Choa, K., Herrmann, F.R., Junod, A.F., Vu, N.V. (2006) Brief Report: Beyond Clinical Experience: Features of Data Collection and Interpretation That Contribute to Diagnostic Accuracy. *Journal of General Internal Medicine*, 21(12), 302-5. https://doi.org/10.1111/j.1525-1497.2006.00587.x
9. LaRochelle, J., Durning, S. J., Boulet, J. R., van der Vleuten, C., van Merrienboer, J., & Donkers, J. (2016). Beyond Standard Checklist Assessment: Question Sequence May Impact Student Performance. *Perspectives on Medical Education*, 5(2), 95–102. https://doi.org/10.1007/s40037-016-0265-5
10. Haring, C.M., Cools, B.M., van Gurp, P.J.M., van der Meer, J.W.M., Postma, C.T. (2017) Observable Phenomena That Reveal Medical Students' Clinical Reasoning Ability During Expert Assessment of Their History Taking: A Qualitative Study. *BMC Medical Education*, 17(1), 147. https://doi.org/10.1186/s12909-017-0983-3
11. Lewis, K. L., Bohnert, C. A., Gammon, W. L., Hölzer, H., Lyman, L., Smith, C., Thompson, T. M., Wallace, A., & Gliva-McConvey, G. (2017). The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). *Advances in Simulation* (London, England), 2, 10. https://doi.org/10.1186/s41077-017-0043-4
12. Page, G., Bordage, G., & Allen, T. (1995). Developing Key-Feature Problems and Examinations to Assess Clinical Decision-Making Skills. *Academic Medicine: Journal of the Association of American Medical Colleges*, *70*(3), 194–201. https://doi.org/10.1097/00001888-199503000-00009
13. Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer-Verlag.
14. Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of Generalizability: A Liberalization of Reliability Theory. *British Journal of Statistical Psychology*, 16(2), 137-163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

15. McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2001). *Generalized, Linear, and Mixed Models* (Vol. 325). New York: John Wiley & Sons.

If you have any questions about content, please email cfeddock@nbme.org.

ABOUT NBME

NBME offers a versatile selection of high-quality assessments and educational services for students, professionals, educators and institutions dedicated to the evolving needs of medical education and health care. To serve these communities, we collaborate with a comprehensive array of professionals, including test developers, academic researchers, scoring experts, practicing physicians, medical educators, state medical board members and public representatives. Learn more at nbme.org.