Power of Community: Addressing the Challenges of Clinical Reasoning Assessment

On-screen: [Chris Feddock, MD. VP, Competency-Based Assessment, NBME.]

Chris: Hello, and thank you for joining us today. My name is Chris Feddock. I'm the Vice President of Educational Strategy at NBME. We've gathered a panel today of NBME staff and some of our faculty collaborators to discuss the challenges of assessing clinical reasoning.

More specifically, we'll be reviewing some of our collaborative work, that the NBME has undertaken with 10 medical schools over the last couple of years, with the goal of to provide feedback, improve the feedback given to learners on their diagnostic reasoning skills.

We're going to start with a video presentation that's a little under 30 minutes long. And after that, I'll introduce the panel and we'll move to some live question and answer session. While the video is playing, please feel free to use the zoom Q&A tool, to submit your questions, and we'll answer those questions as soon as the video is over. Thanks again and see you after the video.

On-screen: [NBME, Power of Community: Addressing the Challenges of Clinical Reasoning Assessment.

- Chris Feddock, MD. VP, Competency-Based Assessment, NBME.
- Analia Castiglioni, MD. University of Central Florida College of Medicine.
- Su Somay, EdD. Lead Measurement Scientist, NBME,
- David Gordon, MD. Duke University School of Medicine.
- Thai Ong, PhD. Senior Psychometrician, NBME.
- Matthew Kelleher, MD. MEd University of Cincinnati College of Medicine.]

Chris: Hello. Today we'll be summarizing the work of NBME First Creative Community. This was a collaboration between NBME staff and medical schools to really advance the assessment of clinical reasoning, specifically focusing on assessments for learning in the diagnostic reasoning space.

But let me begin with some introductions. My name is Chris Feddock. I'm the Vice President of Educational Strategy for the Growth and Innovation Division at NBME. Su?

Su: I'm Su Somay. I'm a lead measurement scientist in growth and innovation division as well.

Chris: Thai?

Thai: Thanks Chris. Hi, everyone. My name is Thai Ong. I'm a senior psychometrician and I work in the assessment operations division.

Chris: So ultimately, the purpose of our work is to improve the care that patients receive. That's why one of our main concerns were the number of medical errors that are currently occur in our country.

On-screen: ["We need to work to reduce medical errors. Diagnostic errors are one of the largest number of errors made, so that's why this Creative Community was so important." Analia Castiglioni, MD, University of Central Florida College of Medicine.]

Although surgical errors and medication errors are oftentimes recognized, we realized that a lot of diagnostic errors oftentimes go unrecognized until later.

And as we look at data that is available, every person in the United States is likely to experience a diagnostic error at some point in their lifetime. Given this background, we really wanted to develop some assessments to improve that diagnostic reasoning process.

On-screen: [Improving Diagnosis in Health Care.

National Academies of Sciences, Engineering, and Medicine; Institute of Medicine; Board on Health Care Services; Committee on Diagnostic Error in Health Care; Erin P. Balogh, Bryan T. Miller, and John R. Ball, Editors. DOI: https://doi.org/10.17226/21794.]

So, when we looked towards the National Academies of Science, Engineering and Medicine, they produced a report in which they recommended two directions to really reduce the number of diagnostic errors.

First was the promote teamwork amongst healthcare professionals. Second was to improve education and assessment, particularly towards those diagnostic skills. In recognition of this critical importance, enhancing clinical reasoning, education and assessment, and more specifically, diagnostic reasoning has remained a top priority for most medical educators.

However, progress is oftentimes challenging. Clinical reasoning and diagnostic reasoning, more specifically, are complex and multidimensional concepts that can be really hard to measure effectively.

Additionally, many of our existing tools overlook some of the critical dimensions of performance, particularly performance that can improve diagnostic outcomes. Those include aspects of taking a history, specifically the timing and sequence of how, how practitioners ask questions to their patients.

Recognizing these challenges, but also understanding that effective partnerships between learners, faculty members and systems of medicine are likely to be most effective in in creating change. The NBME launched its first Creative Community Initiative in January 2022. This initiative really was to combine the expertise and experience of multiple individuals to really address some of these challenges and assessment and medical education more broadly.

NBME, we received a robust response from medical schools receiving nearly a hundred applications from which we selected 10 schools.

These are the lead faculty members from each of those 10 schools who were intimately involved in every stage of the process that we went through to develop an assessment for learning.

On-screen: [

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

Su, Thai and I are here representing a larger NBME team who worked very closely with those lead faculty members.

On-screen: [

- John Moore, PhD Director, Assessment Data Initiatives, NBME.
- Ann Nolan Process Expert, NBME.
- Marni Grambau Director, NBME Assessment Alliance, NBME.
- Jeannette Sanger Manager, Computer Based Case Simulations, NBME.
- Scott Mandel Senior Test Development Analyst, NBME.]

And together we really brought in a wide range of expertise, not only of diagnostic reasoning, but teaching assessment and measurement science.

Another challenge in the assessment of diagnostic reasoning is the lack of a unified framework. So we began our work by selecting a single framework to build from, and we chose the one published by Daniel and Colleagues because we felt it was the best reflection of current medical school curricular.

On-screen: [Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance.

Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, Ratcliffe T, Gordon D, Heist B, Lubarsky S, Estrada CA, Ballard T, Artino AR Jr, Sergio Da Silva A, Cleary T, Stojan J, Gruppen LD.

Acad Med. 2019 Jun;94(6):902-912. doi: 10.1097/ACM.0000000000002618. PMID: 30720527.]

Further current assessment tools primarily evaluate whether learners reached a correct or plausible outcome, focusing on perhaps differential diagnosis or leading diagnosis or diagnostic justification.

On-screen: [Diagram of diagnostic Reasoning Framework.

Need for process measures as most current assessments focus on outcomes!

From left to right six boxes are interconnected with double-sided arrows, organized into two halves.

Left side (earlier phase): Contains Information Gathering at the top left and Hypothesis Generation at the bottom left, linked together by a vertical double-headed arrow. A horizontal double-headed arrow extends from this connection to a middle box labeled Problem Representation. Beneath this section, the text reads: "Enhances deeper learning" and "Necessary for behavior change".

Right side: A horizontal double-headed arrow connects Problem Representation to a box labeled Differential Diagnosis. Two diverging arrows connect Differential Diagnosis to Leading Diagnosis (top right) and Diagnostic Justification (bottom right). A vertical double-headed arrow links Leading Diagnosis and Diagnostic Justification. Beneath this section, the text reads: "Does not generalize" and "Minimal impact on future behavior".]

Unfortunately, those types of measures are less generalizable to other contexts, and therefore the feedback that you're able to give on them is less likely to impact future behaviors. So we focused our work on the earlier phases of the clinical reasoning process. We conceptualized diagnostic reasoning as an unfolding process and noted the complex and interdependent relationship between hypothesis generation and information gathering.

So, we advocated for combining those two into something we called hypothesis driven information gathering. So, our work focused on essentially two subdomains: hypothesis driven information gathering and problem representation.

On-screen: [Information Gathering + Hypothesis Generation: Hypothesis Driven Information (HDIG). Problem Representation]

So, following a thorough and thoughtful process, the creative community decided to focus the assessment on learners in the clerkship and post clerkship phase, recognizing that those learners likely had the best ability to demonstrate the behaviors associated with those two subdomains.

Our work together was really guided by three principles. One was co-creation, really building on and using the unique talents that we had assembled. Second is that this was really research oriented. We were really focused on whether or not we could design something new, a new approach to assessing that could really drive assessment for learning. And then last, we really were focused on using principled assessment, design something with an evidence-based that would really support the design of our instruments.

So, with that, let me turn it over to Su to talk more about that assessment design.

Su: Thank you, Chris. I'm going to be talking about how we develop the assessment, but before we go into the details of our processes of assessment development, I wanted to have a word on validity.

On-screen: [A word on validity. Validity is the body of evidence supporting the accuracy of inferences drawn from assessment outcomes.

Purpose: Validity is defined in relation to the specific purpose for which the assessment is used.

The question is...(Crossed out: "Is this assessment valid?") (Not crossed out: "Are the inferences I'm making based on assessment outcomes valid for my purpose?)]

Validity is the most fundamental consideration in assessments. It is essentially investigating and collecting a body of evidence to support the claims that we make about the assessments that we develop.

Often in conferences we're asked, what is the ideal level of validity or the number I should mention that validity is not a number or a level or even a property of a test itself. Validity is the argument that we build to support the inferences that we make based on the assessment outcomes. Always consider validity in the context of the assessment purpose, not by itself.

We can conceptualize validity as a structured argument, and this argument can be thought of as a chain of links. And each link in this chain represents an aspect of the argument. And as we go through these chains, the whole argument is as strong as the weakest link in the chain.

In that sense, validity in traditional assessment design is a post hoc consideration, whereas in evidence centered design, ECD validity is an “A” priority consideration. It is the most important consideration of the whole process. This process begins by thinking about what do I want to claim about the learners using these assessment results?

Once we think about the claims, we need to also think about what evidence do I need to collect to support these claims?

And once we organize ourselves in terms of claims and evidence pairs, we move on to developing tasks. And it's important here to think about what claims am I going to make and what kind of tasks would actually elicit those behaviors I'm looking for? And then we collect data, and then we go back to the beginning. Does the data support our claims?

On-screen: [Diagram with heading Not So Fast... A wheel contains Claim, Evidence, Task, and Data. Text boxes saying 3 Proof-of-concept (POC) studies, and Range of reasoning approaches, are to the side of the wheel.]

Again, this is a very highly iterative process. So when we're developing these pieces, we go back and forth between them and using the data. We might even start from the beginning. Before we start the evidence center design process. However, we need to make three important decisions.

One is, what is it that we want to measure? In our case, it was hypothesis driven information gathering.

The second is, who is the target population? Who is this assessment for? In our case, it was clerkship and post clerkship learners.

Lastly, the purpose of the assessment. How are we going to use these assessment outcomes? For what purpose? This is key for us because we wanted to, as Chris mentioned before, we wanted to assess a process. And for the for the process, we need to think about what aspects of that process do we want to focus on.

On-screen: [

1. Characterization of Chief Concern: Ability to elicit the essential components of each element of the presenting symptom.
2. Curiosity: Ability to identify relevant information prior to the encounter, explore them in depth to change the likelihood or prioritization of possible hypotheses.
3. Agility: Ability to promptly recognize critical information during the encounter and explore them in depth to change the likelihood or prioritization of hypotheses.]

And here are those three aspects. The first one is characterization of chief concern. This is where we look at learner's ability to elicit the most essential information prior and during the encounter.

Curiosity is the ability to identify the critical information prior to the encounter.

And agility is the ability to promptly recognize the information obtained from the patient during the encounter and explore those aspects in depth. To elicit these aspects, we knew we needed a stimuli that had certain properties.

On-screen: [Whole Task Case: The full history-taking visit for a patient presenting with a new medical concern.

Patient Intake Form: Basic patient information (like common ambulatory visit forms) to provide a stimulus for initial differential.

Pivot Points: Key findings elicited during the encounter that would significantly narrow the range of diagnoses.

Case Ambiguity: Following the encounter, at least five diagnoses should be plausible, one or two being clearly most likely.]

For example, we knew that the learner had to go through a full history taking process, and we knew before that process they needed some information. And that information was provided through a patient intake form, which is used commonly in doctor visits.

And then we knew that our cases had to have certain points where the patients share critical piece of information that would significantly narrow the range of diagnoses.

And lastly, we knew that the cases needed to be somewhat ambiguous so that learners could explore more than one hypothesis.

On-screen: [Before Encounter, Identify Key Features, Differential Diagnosis.

During Encounter, Interview the Patient, Characterization of the Chief Concern, Agility to reconsider their differential, Curiosity.

After Encounter, Reconsider Key Features, Differential Diagnosis.]

And here's the task. Before the patient encounter, learners were provided with a patient intake form and asked to identify the critical in pieces of information in the form and come up with a differential.

Next, they interviewed the patient and following the patient encounter, students were asked to reconsider their differential, and the key features.

Of course, this process was not linear, as I mentioned before. As we were developing our claims, we were going to the evidence. And then as we develop our evidence, we were going back to our claims. And as we were developing our tasks, we administered multiple proof of concept studies to ensure that, our assessment is eliciting the behaviors of interest. And this was followed by a large scale pilot, which Thai is going to talk about.

Thai: So just to recap where we are, Chris did a really nice introduction to diagnostic reasoning and why that was the focus of our commute creative community. Sue just discussed the assessment development process, evidence sensor design that we engaged in to develop the assessment. And like Su said, I'm going to talk about the larger pilot that we conducted in fall 2023.

We conducted a large scale pilot with 76 post clerkship students. These learners took the assessment on a virtual platform called Recollective, and they completed four different cases. And these cases were selected to reflect a wide range of chief complaints that are typically seen by proposed clerkship population, and they varied in patient demographics. So we had 76 students who took four cases. This resulted in 304 pre and post responses and 304 recorded encounters.

We then recruited a total of 24 faculty members, including the 10 faculty members in our creative community to help evaluate the student performance. All the faculty members participated in a two hour standardized training where they learned about the assessment, the constructs, the rubric did several practice ratings and were given personalized feedback, prior to evaluating the student's performance.

We had three specific research questions for our pilot. First, we were interested in evaluating the accuracy of the SPs. Second, we were interested in evaluating learners' performance at the various assessment points that Su talked about. So how do students do at the pre encounter, how do students do during an encounter and the post encounter? And lastly, we were interested in evaluating the psychometric properties of the HDIG scores.

Now for this webinar, and for the sake of time, I'm going to quickly summarize our findings for the first two research questions and then really dig a little deeper into research question number three.

With respect to SP accuracy, we randomly selected 30% of our encounters and had our SP trainers rate the accuracy of those SPs. We found that across these randomly selected encounters, the SPs delivered the patient information correctly 90% of the time.

We wanted to make sure that the learner's performances were not impacted by the SPs that they saw. And were happy to see that this was not the case. Given the high accuracy rate.

With respect to learner performance, we found that learners did vary in the number of key features that they were able to identify at both pre and post encounter. They also varied in their line of questioning with the SPs. And lastly, they varied in the number of times they were able to obtain the correct diagnoses.

So, these findings suggest that at least for the 76 students that we saw on our sample, the case stimuli and the assessment did elicit a variety of learner behaviors leading to variability in both the pre and post responses and the HDIG scores.

Given substantial variability in HDX scores, we wanted to further investigate the psychometric properties of those scores. And specifically, there were two main factors that we looked at.

One, of course, is the reliability of the scores, how trustworthy are the scores? And the second factor was validity. So, can the scores be actually used to predict learner diagnostic success?

I'm going to go through the reliability results and then the validity results. But before we dive into the results, I think it's important to understand how scores in our assessment derived and specifically understand what factors contribute to those scores.

And so I'm going to walk you through an example here. We have learner, a learner A participated in our pilot and completed the four cases. These four encounters were then rated by different rater pairs, one rater pair for each case.

Now the score matrix for literary A looks something like this. You can see there's four rows in the matrix, one for each case. And then you see three columns for the three subdomains, agility, charact, duration of the chief concern and curiosity.

So, this matrix contains the case level HDIG scores for learner A. Now, like I said, we don't just have one learner, we have 76 learner.

So, we can actually look at it in terms of the learner group. Here we have the learner group who completed the same four cases and each learner by case combination was rated by a rater pair. So instead of looking at that smaller matrix, we now have just this larger matrix. And all this larger matrix contained is the case level scores for all of the learners in our pilot.

When you look at this larger matrix, you begin to clearly see the variability in the HDIG scores. For example, if you look at the agility column, you can see there's variability in the agility scores within a learner and between learners.

So, for learner A or for learner Z, you can see the agility score did fluctuate from case to case. Between learners, so from learner A to learner Z, you can see that there's differences in performance as well.

Now these variabilities in the context of our assessment could be due to three related factors.

On-screen: [Variance in HDIG scores due to:

1. Differences in learners' abilities,
2. Differences in case difficulties,
3. Differences in rater strategies.

Can apply Generalization Theory to estimate proportion of variance due to learners' abilities (e.g., reliability) and factors that impact reliability.]

First, the score may vary due to true differences in learners' clinical reasoning abilities. Now, this is what we want. We want the scores to vary because we have differences in student's ability in our pilot.

The scores may vary due to differences in case difficulties. So, we have four different cases in our assessment. Some cases may be more difficult than others, and this could lead to differences in performance.

The scores may also vary due to differences in rate pairs. The rate peers may vary in their strategy or leniency across students, and this could lead to differences in performance as well.

So, you can think of learnability case and rater as potential factors that influence score variability. Now, why is all of this important? Why is it important to think about the factors that impact scores?

We can then apply a statistical model called generalizability theory to evaluate the extent to which these factors contribute to score variability. And more importantly, we can use that information and calculate a measure of reliability known as the phi coefficient. The phi coefficient represents the proportion of total variance that's due to learners or due to true differences in learners' abilities.

Psychometrically, we think of scores as being more reliable or more trustworthy when the phi coefficient is closer to one. A phi coefficient of one indicates that all the score variance is due to true differences in student's ability. Anything less than one means that the scores were influenced by other factors like cases or rate of peers, in our assessment context, which we've conceptualized as error.

And for simplicity, I'm going to refer to the phi coefficient as the reliability estimate for the rest of the webinar. And like I said, a phi coefficient one indicates that all of the

score variance is due to learners' abilities, but oftentimes obtaining a phi coefficient one or reliable estimate of one, is unrealistic.

With working with OSCEs, we know there are a number of factors that contribute to the scores. And so most OSCEs produce scores with reliability estimate of around 0.5 to 0.8. So that's the target goal of what we want to see in our assessment.

On-screen: [Generalizability Theory.

Individual HDIG scores were, on average, moderately reliable (phi coefficient = ~.50).

When combined, composite scores were more reliable than individual HDIG scores (phi coefficient = ~.70).

Reliability estimates are in line with other OSCE assessments.]

So, we calculate the phi coefficient for the three sub domains, agility, curiosity, and characterization of chief concern. And we found that the scores to be moderately reliable with an average reliability estimate of around 0.5. When the HDIG scores were combined into a single score for each learner those composite scores tend to be more reliable than the individual HDIG scores with a reliable estimate of around 0.7.

Again, like I said, the range that we were looking for was about 0.5 to 0.8. So we were really happy that we saw our reliable estimates within that range. The reliable estimates again represented the total variance in the HDIG scores was due to the learners, due to true differences in learners' abilities. And that was about 50% in our sample.

So, the question is, what about the remaining variants? Well, we found that only a small portion of the remaining variants was due to our raters. This means that HDIG scores were not heavily influenced by our raters. In other words, rater peers were consistent in their scoring across learners.

In contrast, we found that a large proportion of the main variance was actually due to cases. This means that the HDIG scores were heavily influenced by the case content. In other words, learners varied substantially in the performance from case to case.

Given these findings, we wanted to know, well, what's the impact of increasing the number of raters or impact of increasing the number of cases on the reliability estimates? And what we found was an increase in the number of raters from perhaps two to four to six per student didn't actually improve the reliability estimate that much, but increasing the number of cases did.

On-screen: [Graph of Factors that Influence HDIG Reliability. Increasing the number of cases substantially improves reliability.

The y-axis is the phi coefficient or reliability estimate from 0.00 to 1.00, and the x-axis is the number of cases from 0 to 15. Three lines are shown: Agility, CCC, and Curiosity.

4 Cases marks the pilot.

The agility line is at .30 at 1 case, .63 at 4 cases, and .80 at 14 cases.

The CCC line is at .18 at 1 case, .38 at 4, and .75 at 14.

The curiosity line is at .15 at 1 case, .30 at 4, and .68 at 14.

Data is approximated.]

So this plot here shows the relationship between the phi coefficient or the reliability estimate and the number of cases. On the vertical axis is the reliability estimate values. And then the horizontal axis is number of cases. The three lines represent the predicted reliability estimate for the three HDIG sub domains.

You can see that as we increase the number of cases from four, which is a current OSCE design to eight or even 12, the reliability estimate increases substantially. So this means that perhaps in future iterations of the assessment increasing the number of cases may be a viable option for maximizing the HDIG reliability estimates.

In addition to exploring reliability and the factors that influence reliability of the HDIG scores, we also examined from a validity perspective whether the HDIG scores when combined together into a composite score, are those useful, particularly in predicting learners diagnostic success?

Since the HDIG scores represent early diagnostic reading reasoning processes, we would expect learners to demonstrate good HDIG behaviors to have greater diagnostic accuracy.

On-screen: [Graph of Predictive utility of HDIG ratings.

Composite scores statistically significantly predicted learner's diagnostic accuracies ($p < .05$).

Line on graph going from .20 to .50 on the y-axis of Probability of Correct Diagnosis, and from 1 to 4 on the x-axis of Composite Score (Theoretical Weighting). The line has a slight upward curve.]

And that's exactly what we found as theorized. The composite scores were statistically significantly related to diagnostic accuracy. So this plot here shows the predictive probability of getting the correct diagnosis by various levels of composite score, and it illustrates both that positive relationship and the practical significance of this effect.

So, in summary, we talked about the reliability estimate. We found that the HDIG scores were moderately reliable and the composite scores were even more reliable than the individual scores. We talked about from a validity perspective, the predictability of those scores. Again, those are just several pieces of validity that Su mentioned as potentially useful to creating that validity argument that Su discussed earlier.

And so we found ultimately that the HDIG scores did show promising psychometric properties. I personally do think there's value in pursuing and exploring more of HDIG criteria specifically for the purpose of writing feedback to learners on diagnostic reasoning processes.

So with that, I'm going to turn it back to Chris.

Chris: Thank you, Thai. That was a great summary of the results. So I think our approach really represents a significant shift, to a more nuanced and effective assessment of diagnostic reasoning skills.

On-screen: [Summary:

Creation of a process-oriented assessment, insight into cognitive processes.

Initial validity evidence for assessment, G-theory provides direction for future development.

Successful co-creation between NBME and UME institutions.

Effective use of ECD framework for a complex construct.]

And this process oriented assessment can really advance the way we think about as assessments for learning. We've been able to develop some initial validity evidence, and as Thai mentioned, we have a direction for future enhancements of this type of assessment.

I think the success of this project was in large part due to one, the co-creation process and really the combined expertise that we are able to bring together. And second, the process we followed that Su discussed. That evidence center design process really

enabled us to focus in on the critical aspects and ensure that what we were measuring was meaningful.

All of this work addresses a critical need in competency-based medical education, really the development of more specific skill related assessments and assessments for learning.

Now, there are certain challenges that I think we experienced. you know, so if we look at the number of resources that were necessary to create this, this assessment, they were significant. Both in case development as well as in the implementation using the recollective platform. And not to mention scoring. So having two faculty members observe, 304 encounters is, is very resource intensive. And we have to admit that even though we had a large sample with 76 students, a larger sample may find some different results.

So, as we think to next steps, there are definitely some things that we believe should be taken. Number one, we likely need to make some continued revisions to the material. As Su mentioned, this is a very iterative process. And what we learned from our proof of concept studies, we learned even more from our pilot testing. So once again, thinking and considering those cases and the difficulty of those cases and how they impact student performance and continuing to refine our, our HDIG rubrics to achieve higher levels of reliability.

Second, we still have to create that feedback model and that feedback model is critical because that's really the ultimate goal of our work, to provide specific targeted feedback to learners on how they can improve these types of skills.

And then last, there's definitely additional piloting we need to do. There's platform issues, as we mentioned before, to effectively deliver this. And we want to explore automated scoring models because ultimately we believe that a formative diagnostic reasoning tool that generates conversational AI avatars and is able to provide immediate feedback to learners as our ultimate goal.

So that concludes the prerecorded portion of our webinar. Thank you for your attention.

On-screen: [Live Q&A.]

Chris: And we're back. So good to see everyone again. Just a reminder that you can continue to use the q and a tool in Zoom to submit your questions. We'll do our best to

try to answer all the questions that have been submitted, but we may not have time to get to all of them, but we'll definitely try.

So let me now introduce our panel. Hopefully you recognize, Su, Thai and myself from the video, in our starring roles, but let me focus on our other three members of our creative community and our panelists today. I'm just going to have them each introduce themselves and then we'll get to your questions on Analia. On Analia, you're on mute.

Analia: Hi, I am on mute. I apologize. Hello everybody. Good afternoon, and thank you for joining us. My name is Analia Castiglioni and I'm the Assistant Dean for Clinical Skills and Simulation at UCF College of Medicine in Orlando, and the Medical Director for Simulation Center.

Thank you, David.

David: Hi, everyone. David Gordon. I am a emergency physician and associate dean for student affairs at Duke University.

And Matt.

Matt: Hello everybody. Matthew Kelleher. I'm a clinician educator, med-peds hospitalist in Cincinnati, Ohio at the University of Cincinnati College of Medicine.

Chris: Alright, well, thank you all, let's get to the question. So I want to start with one question, just so that we're all on the same page.

And that first question is about USMLE's Step 2 CS and how this work is related to that, and quite simply, it's unrelated, right? So, outside of the importance that we feel towards clinical reasoning, right? So this is not, related to, a licensure exam. This work, as we talked about in the video, was really centered upon a focus on clinical reasoning and more specifically diagnostic reasoning skills.

And then how can we use that to provide better feedback to our learners? And just really understanding that in that CBME model, our learners really need much more feedback, right? More formative, more assessments for learning than of learning and really trying to advance that, and I think as a hopeful outcome of that improving their diagnostic skills and eventually patient outcomes.

So, let me move on to the next question. Su, perhaps you can talk about how do we meaning meaningfully assess clinical reasoning or other skills, particularly using OSCEs? You know, they're used at most schools, but I think some general guidance on how do you make sure that they're really meaningfully assessing?

Su: Thank you, Chris. So, I just want to take a step back and acknowledge how difficult it is to measure skills. And by that I mean soft skills, which is not a term I like, but there is not, I haven't heard any good alternatives, let's say soft skills or 21st century skills.

These are constructs that usually don't have universally agreed upon definitions and also very context dependent constructs. And in medicine that becomes particularly challenging because medicine itself is a complex system. And then knowledge in interacts with the skills and it's hard to disentangle. So psychometrically, it's a hard problem.

That's why we need a principle approach such as ECD to tackle this. ECD provides the framework and the guidance, and makes sure that every component and every question or every task serves a purpose in creating defensible and valid assessments.

Thai: You're on mute, Chris.

Chris: Matt, maybe you can talk a little bit about how a formative OSCE can provide feedback to students and how you see that tool being used.

Matt: Sure. Thanks, Chris. I think a lot of us are familiar with, you know, using simulation and using OSCE encounters to directly observe our learners. And, it's a powerful experience, excellent for assessment.

I think a lot of times, unfortunately, those are focused on outcomes and often have more of a summative lens. And so, I think what was really unique about this project and the opportunity was when we used the ECD framework, our focus was specifically on feedback. It was specifically on what is the process that learners go through to reach an outcome.

And so ideally, by focusing on that process, then you can imagine you can define or try to identify what's the difference between, you know, the expected place that a learner should be and where they are currently, which sets up nicely for actionable feedback and things that can help someone grow.

Chris: Thank you for that, Matt. Analia, perhaps you can talk a little bit about the current state of OSCE nationally and how the assessment of clinical skills and perhaps clinical reasoning… is there a standard, how, how is that going nationally?

Analia: Well, nationally it's very variable. It's highly variable across schools. Schools are shifting or trying to shift towards more competency based, medical education and OSCE are now being used more as part of a progress of assessment and using more

assessments for learning, like you were alluding earlier and for the purpose of feedback and developmentally. There's also schools coming together and forming consortia and most of you probably are familiar with the Mid-Atlantic, the California now our new, Florida clinical skills are collaborative and we're trying to identify standards, best standards, for this performance based assesses assessments such as OSCEs.

Chris: Great. Great. So, David, I know you were very involved in some of those hypothesis driven information gathering skills. There's a question about how they relate back to communication skills, and perhaps is there a distinction between communication skills and clinical reasoning skills when you think about those HDIG areas?

David: Yeah, great question. There's overlap, but also distinction. We really focused with our primary goal being to assess clinical reasoning. We really focused on the questions that were asked and the information that was obtained, but we didn't try to measure in this project how well they were asked or how humanistic the questionings were. Not to say that's not important, that's tremendously important. It's just measuring in our minds a different construct.

So our focus was on whether or not they obtained the information, the questions they asked, and the sequencing of those questions, but not necessarily how humanistic those were. Those would have to be measured in, in a different complimentary model.

Chris: Now I think this gets back to kind of our evidence center design framework, right? That we had to be very, very targeted in what we were trying to get out of the assessment, but thank you for that, David.

When we think about the assessment itself, I think there's some questions about where do we see this going, when is it going to be available, what are the costs? How is it going to be administered? All sorts of programmatic kind of systemic type issues. Thai, do you want to provide some context for that?

Thai: Sure. Thanks Chris. Those are all very good questions. I think what we view this work to be is foundation to upcoming work or revisions.

And so as Chris alluded to in the video, we're continuing to build off this work and thinking of innovative ways to reduce that cognitive load on the readers potentially and including some sort of integration with AI to help kind of reduce the additional sources that we experienced during the rating, particularly with faculty time.

So I don't have a definite answer yet for those questions, but, ultimately our goal is to be able to, offer an assessment for learning to schools that can be administered without substantial additional resources like faculty time.

Chris: Very good, Thai. Thai you mentioned generative AI, and as I think everyone can imagine, there are a few questions in the chat, with regard to generative AI. Su, perhaps you provide some context on how you see generative AI supporting this work, and what are some of the benefits and challenges to using generative for this type of work?

Su: Thank you. This is a very hard problem again, that we're working on currently. One way that is usually very obvious to people for content generation, right? We can use AI for after we create a good conceptual framework, we can use AI to create content, or at least get the help of AI and the subject matter experts to create content.

The other one is more complicated. The faculty it goes back to our beginnings, right? The faculty simply do not have time, enough time to give the feedback or observe the students to interact with patients and so forth.

So, if we can actually harness AI to portray a patient and then also use it to provide feedback and hopefully unload some of that responsibility from the faculty's shoulders, that could be a really, really good solution for medical education.

Chris: Matt, you've had some experience with different AI models and seeing how they could be used in the assessment. How do you see AI fitting into the assessment of clinical reasoning?

Matt: Yeah, great question, and I think I'd echo some of what Su said. I guess just from my own experience, I think we've all seen AI and are excited to see where it's going and what it will do.

And so, I think in my own experience, I've seen AI probably mostly in clinical reasoning, just provide a stimulus for deliberate practice for students to practice cases over and over. That obviously begs a lot of questions around validity though, right? Like, how well did the case actually create a stimulus and represent a patient and introduce potentially new complications or errors, or even worse, potentially give feedback to a learner that could take them, in a wrong direction.

And so we've also seen AI, I think a lot of people have been using AI to try to analyze, you know, written work and things like that to try to give feedback. And I think it's an exciting world, but it, it is definitely, in my experience at least lacking, validity right now

at this point. And I think that's where a lot of us are hoping to kind of marry our current work with AI.

Chris: Yeah, no, very important, right? Thai?

Thai: I do, Yeah, thanks Chris. Matt, to go along with that validity point that you made, I do think that we can use AI, but we got to tell AI what to do. And so what we developed was this HDIG rubric that tells essentially AI exactly what to look for, what to look for within a patient encounter, then provide the necessary feedback. And so I think what we've done is really the foundation to any AI work in terms of validity.

Chris: Yeah, no, I think that's a great point. We have to be able to train AI appropriately, right? So, I think that corpus of knowledge and experience that we're giving it is so important. And so, thinking about the model that we're developing to have AI do the automated scoring.

So a lot of questions in the chat about longitudinal assessment. And I think as we consider assessments for learning, there's definitely that longitudinal aspect to it. So Thai, maybe you could tackle this. How do you think clinical reasoning can be adequately assessed longitudinally, and tracked throughout someone's medical education?

Thai: Yeah, thanks Chris. I think, like Su mentioned, clinical reasoning is complex, it's multidimensional. And so what I really want to encourage people to do is considered what elements within clinical reasoning you want to assess, and keep those elements consistent across time, and make sure those elements mean the same way for a different learner population. And so that those, those comparison across time is meaningful rather than comparing apples to oranges.

Chris: And David, I know with Duke, I believe you all have been challenged with this kind of longitudinal assessment approach. And, and any thoughts from, from your experiences?

David: I'll just add to what Thai said, that, you know, assessing clinical reasoning is really about having a program of assessment that I think uses multiple instruments and multiple venues.

And so I think some of the concepts we've taken in this standardized structure format can also be applied to workplace-based assessments, where we've built uncertainty into

our stories, but nothing matches the uncertainty and messiness of an actual clinical scenario.

So, I think over the years it's, it's you picking your elements, being familiar with what type of assessments are particularly good at assessing those elements and having a mixture of both standardized structured assessments, but also workplace-based assessments, and getting multiple points of data.

Chris: Oh, very helpful. So also, several questions, reflecting the context specificity, and I know that's something that this group has discussed, frequently, right? About how do we get that. Analia, what are your thoughts? We know that performance differs across cases, so how do we address that issue?

Analia: Well, that's challenging and it's always been challenging for our performance based assessments in general. And I think you, you try to tackle that without increasing the number of cases, sampling from different organ systems, different scenarios, different complexity, and what David and Thai were talking about doing it, longitudinally and developmentally appropriately for the learner, for the level of learner, and mixing types of assessments that can assess clinical reasoning in different ways, and in different environments and from different patient populations and complexity.

And I think when you triangulate that or you aggregate or look at the data altogether in that program of assessment, it will kind of give you that picture of that learner's clinical reasoning over time.

Chris: Yeah. So I think there's also been some comments in the chat about what's the ideal number of cases, and Su, Thai, I'll refer that to you. Thai, you talked a little bit about the number of cases in the video and any comments that you'd like to make about context specificity, the number of cases?

Thai: Sure. I think in the video I showed that increase in the number of cases to perhaps eight to 12 did increase reliability, roughly about 0.8, which is where we want to be. I think it's important to consider kind of what the ideal reliability estimate should be versus the practicality of adding cases within an OSCE for this assessment.

And so as we think about how many cases should be in our assessment, I think there should be enough to provide a breadth of content for that specific construct or clinical reasoning that we want to assess, but also consider the practical implications of it.

And perhaps, depending on the psychometric, the measure of scientists, more or less, some of us are comfortable with less number of cases that may get a reliability of 0.7, 0.8 or 0.6 even, given the use of the assessment scores.

Su: I also want to add, this is one and another really hard problem in measurement the content or context specificity. And it's not something that we understand fully. There's been many studies, but we don't really know what goes into it. And maybe more theoretical approach to study what's going on underneath could help everyone in medical education as well.

Chris: No, that's helpful. Particularly when we consider we're focused on an assessment for learning, I think we've got some questions about assessment of learning. So, as we think about those, is there a difference in how we should approach those? Or how do you take an assessment for learning and develop an assessment of learning? So, taking something from a formative space into a summative space, Thai, Su.

Su: So those two things have completely different psychometric let's say problems. One, assessment of learning is you build the assessment to be psychometrically as precise as possible. So precision is what we're optimizing on, whereas assessment of learning completely ignores that. The focus is teaching or learning, whether you learn or what is the best way and how many times do you need to repeat to get competency.

So those are very different goals and therefore they're built very differently. But the concepts of learning can help inform development of assessment of learning, in my opinion.

Chris: So, I think it's important. I think you brought up a really important aspect though, that they really are different types of assessment and perhaps we're making different claims about a student because of the different types of assessment that they are. Thai, other thoughts?

Thai: I just want to echo what Su said. I think I totally agree that their needs of different purposes, but I do think that the way that we think about consistency, reliability, the scores or the feedback or validity of the scores, the feedback is similar that of course, we want our feedback to be consistent and to be valid or accurate.

And so we still consider validity and reliability in the same way. We're just applying it to whether it's a feedback to a student versus a set of scores. And in both instances, we want those to be consistent and accurate.

Chris: Very, very helpful. Very helpful. When we think about some of the skills that we're assessing, how do you see, and maybe this is more for Analia, David, Matt, how do you see this type of assessment impacting, how we teach and assess on clerkships? How would you see this impacting that space?

David: What I found very helpful in this exercise is it gave me a conceptualization and a vocabulary to use for things that I'm seeing in clinical space. So I'll hone in on curiosity and agility, and just the emphasis it places on how students respond to information. Not just recognize information, but what is the questioning and are they asking questions in a meaningful way?

But now if I have, sometimes you, the first thing you recognize is that's not going well, when you are observing a learner. But now I feel like these terms allow us to diagnose the learner, so to speak, and to say, I think that's a curiosity problem, or, I think that's an agility problem. And so I really appreciate the language that that came out of this project.

Analia: I also would add that it's helped me and I think all of us, really think about teaching clinical reasoning and teaching this constructs and this concepts to the students. So not only assess them and give them feedback, but also creating a longitudinal clinical reasoning curricula across our, our programs. So we can assess them on elements that we taught them before and not assume that they're learning them just naturally. And also give us, like David said, a new set of vocabulary that, talking about hypothesis that we're talking about in hypothesis generation and this process and problem representation. I think all of us now have a shared mental model of those concepts and can both teach and assess along.

Matt: I would echo the same thing as David and Analia. I just want to emphasize or underline, I think like most of us, this project really taught me to emphasize the process in clinical reasoning rather than the outcome. And I think that permeates literally everything I do and I teach students and learners now is, it's not what's your differential, it's how did you get to the differential, right? Or those kinds of questions where you're trying to drive at the process.

Chris: Yeah, I think that's so important in feedback, right? And so, we've got some questions in the chat regarding feedback. And reflecting on my own journey in medical education, I think we oftentimes give task level feedback, right? In this scenario, you should ask about this condition, or that you should ask about this type of problem, or

you should have delved more deeply into this area. How do you all see that difference between task and then process, in terms of giving feedback and as a corollary to that, who is in the best position to provide that feedback? Because we have some questions in the chat that frequently in OSCE standardized patients are giving that feedback or faculty members are giving that feedback. So how do you see that balance?

See, this is a tough question. Our panelists are pondering.

Analia: Let me clarify. Are you talking about the balance between how much feedback should come from the patient and how much feedback should, at any single assessment, how much feedback should be on clinical reasoning? Should we be doing both together?

Chris: Well, I think part of it, I think those are all elements. What I'm seeing from some of the questions is I think we have focused on process feedback rather than necessarily just task feedback. And so if that is the goal, who can provide that type of feedback? Just recognizing that oftentimes at our universities, our feedback is oftentimes given by standardized patients, perhaps with some faculty involvement.

Analia: Yeah, I personally think that this type of feedback on clinical reasons should not be coming from the patient. The patient can provide unique feedback on interprofessional, interpersonal skills, communication skills, how they felt as a patient, as a recipient of the care.

But their reasoning, should come from faculty or faculty in collaboration with AI. But I think it needs to come from another clinician or somebody who can understand that process and can read into that process.

Matt: Yeah, I would add, I think I agree with Analia. I think ideally the feedback is coming from a clinician and ideally it's a clinician who has observed an encounter. I don't know that that's always the case. So, I do think the question begs other questions around, like many of us have set with learners where we're trying to share data, right? And there's a lot of complexity around how do you share that data, how do you explain what you're seeing in the multiple encounters or things like that.

And so, I'm not sure that that's an easy answer, but I think there's a lot of thought that needs to go into the way we explain the direct observations and the data that we're obtaining about their clinical reasoning skills.

David: Yeah, I agree with Matt and Analia as far as what you'd expect to be the primary source of feedback. I will say though, one thing that I do think is different with concepts like curiosity and agility is they also reflect a mindset. So they can be independent of content. Someone just generally could say, even a standardized patient, there was an opportunity here to explore more and be more curious about this.

And it may not be as detailed or as clinically nuanced as from feedback that can come from an experienced clinician. But I do see those things as generalizable in a way as a way to approach pre-encounter information, as a way to approach a general mindset and openness during an encounter.

So there is some room there, but I think specifically, an experienced clinician would be the one to provide most of the feedback.

Chris: And I think that's a challenge we know that faculty time is at is precious. We don't have a lot of it. And I think as Su mentioned before, thinking about how we can automate this and can generative AI once again with a human in the loop, right? We're not talking about eliminating faculty, but how can they facilitate this process?

We have a question here about, well, how do we ensure that students develop solid hypothesis driven behaviors and differential diagnosis consideration? And I'll provide a little bit of context, which is with the 76 students in our pilot, we actually received some very interesting feedback.

And part of that feedback was that they thought they were learning just by taking the assessment, which we weren't expecting because we didn't have feedback for them. But many of them were marked on just the way the cases were designed and how they rolled out how educational it was. And I think these are things that we need to challenge ourselves with.

Because one of the comments that was made several times by students is, I had to change my differential diagnosis. And I think that's partly a reflection on their experiences with OSCEs at their own institutions. In fact, some directly brought that up, that oftentimes when we're thinking about these types of behaviors, if you go in the room and you already know the diagnosis, how much hypothesis driven in differential diagnosis do you need to develop?

So, I think case design is really essential and it's going to be something that we really need to think more deeply about. I think Su mentioned before, it's also an area where

we don't have a lot of knowledge, right? Like, what makes a difficult case, how do we challenge our students appropriately?

Well, I just want to wrap up. I think we have one final question, that I see, which is really about how do we teach clinical reasoning, right? So we've talked a lot about this, but as you think about the frameworks that we've developed, how does this impact instruction, particularly earlier in the curriculum? We're focusing kind of at the end of clerkship. So Analia, David, Matt, do you all have thoughts about perhaps just participating in the creative community, how this has changed your approach?

David: Something I wish I had more in medical school was more understanding the metacognitive aspect of clinical reasoning. Understanding the process itself. What HDIG problem representation?

Because I know as I went through medical school and residency when things did not go right, that language would've been very helpful for me to understand. It wasn't a knowledge problem, it was a reasoning problem.

So, in addition to cases, everything like that, I think just talking about the concepts and constructs, to learners would be helpful.

Matt: Yeah, I would agree. And I think giving that language and those constructs to students actually allows them to ask experts. I think we've all been around expert clinical reasoners, and sometimes it's hard to know how they got to that diagnosis or how they got to the answer.

And I think giving students the ability to ask the right questions, to think about the process, allows them to even ask for feedback or to ask how someone arrives at a diagnosis in a different way.

Chris: Right. And Analia any parting thoughts?

Analia: Emphasizing, like you said, the teaching, doing it small bits at a time or in a developmental level, building blocks focusing on those different components of the process of clinical reasoning and giving them plenty of feedback and eventually role modeling like I think Matt was saying, also role modeling as they go into the clinical, into the clinical setting.

Chris: Yeah, so I think again, we would probably emphasize that process feedback, right? And I think one of the things that we saw was how important it is. Not just what

information you get, but how you ask those questions and the order you ask those questions and what you're doing with that information to feed your differential diagnosis.

Alright, well we are past time, so I'd like to thank our panel for bringing their insights, and answering all the questions in the webinar. In follow up to this presentation, everyone who registered can expect to receive a video link as well as a written summary of the work that we discussed today. As we close out this webinar, there will be a very short survey, two questions that pops up at the very end. We'd really appreciate you giving us that feedback. Thank you for your time and have a wonderful day.

On-screen: [NBME, Copyright 2024 NBME. All rights reserved.]