Past Event: 2026 NCSBN Scientific Symposium - Clinical Judgment Assessment: Unfolding Case Studies and NCLEX Results Video Transcript

Event
2026 NCSBN Scientific Symposium

More info: https://www.ncsbn.org/past-event/2026-ncsbn-scientific-symposium

Presenters
Will Muntean, PhD, Principal Research Scientist, Innovation & Measurement Research, Examinations, NCSBN;
Joe Betts, PhD, EdS, MMIS, Director, Measurement & Testing Examinations, NCSBN

- [Will] Hi, my name is William Muntean, and I'm with the National Council of State Boards of Nursing. This presentation is part of a two-part discussion on measuring clinical judgment on the NCLEX. My portion focuses on unfolding case studies as an assessment format. I'll discuss how these cases are structured and how they support measuring clinical judgment.

Dr. Joe Bett's portion will then broaden the focus to first-year operational results for measuring clinical judgment on the NCLEX. Clinical judgment has long been recognized as essential to safe entry-level nursing practice. Nurses are expected to interpret patient information, identify concerns, make decisions, take appropriate action, and evaluate whether those actions are effective.

However, when reviewing the literature, one of the major findings was that although there was a broad agreement on clinical judgment that it matters, there was less agreement about how to define it, where its conceptual boundaries are, and what observable behaviors represent it, and most importantly, how it should be assessed. Different sources use related terms such as clinical reasoning, critical thinking, decision making, and clinical judgment, sometimes interchangeably and sometimes with different meanings.

This is a critical challenge for high-stakes exams. To measure clinical judgment, we need more than a general recognition that it matters. We need a structured model that defines the construct and supports defensible assessment. That served as motivation to develop a clinical judgment model and ultimately the unfolding case study approach I'll focus on today.

Go ahead and scan the QR code for more information on the literature review. Clinical judgment can be understood as a hierarchy with different levels of abstraction. At the broadest level, the model begins with a client need and clinical decisions. This is the practice context. Nurses encounter client needs and

must make decisions to support safe and effective care. Client needs are addressed through the general construct of clinical judgment, but this is too broad.

It tells us what we care about, but it's not yet specific enough for item writing or exam scoring. When looking deeper at clinical judgment, we discover that there are three main components driving decisions. Forming hypotheses, refining those hypotheses, and then evaluating them. This process is iterative, especially upon receiving new information about the clinical situation.

The first component is forming hypotheses. This is where the nurse identifies relevant clinical information and begins to make sense of it. In the measurement model, this corresponds to recognizing cues and analyzing cues. The nurse determines what information matters and what that information may suggest about the client's condition.

This creates the initial clinical understanding that guides the next steps. The second component is refining hypotheses. After the nurse has interpreted the cues, they must decide what concern is most important and how to respond. This includes prioritizing hypotheses and generating solutions. In other words, the nurse moves from understanding what might happen to determining what matters most and selecting the best clinical response.

The third component is about implementation and evaluation. After generating solutions, the nurse determines how to implement those solutions and how to best evaluate them. Did the intervention produce the expected outcome? For assessment, this is the critical level. These six cognitive functions provide observable targets for item development.

This is the level in which unfolding case study items are written and classified. Go ahead and scan the QR code for more details on the NCSBN Clinical Judgment Measurement Model. So once clinical judgment is operationalized to the measurement model, the next question is how to design an assessment format that could measure these steps in a clinically meaningful way.

That is where unfolding case studies come in. Unfolding case studies are one way to translate the clinical judgment model into an assessment format. Rather than presenting all the patient information at once, unfolding cases reveal information sequentially. This better reflects how nurses encounter clinical situation in practice. Information emerges, patient conditions evolve, and the nurse must continually update their reasoning.

Take this item, for example. Here's an examinee that is presented with an initial presentation of a clinical situation, and they're asked a question about it. On the next question, additional information is added to the clinical situation, and they're asked another question as it relates to the updated info. Likewise, the next question may introduce additional information or update previous information.

The unfolding format only works if the cases are carefully constructed. So not simply six questions attached to the same scenario. They are intentionally designed to balance clinical realism with measurement requirements. The goal is not simply to make items more realistic. The goal is to measure clinical judgment in a structured way while preserving the clinical context in which the judgment actually occurs.

The examples highlight the core features of unfolding case studies. They present a realistic clinical situation that evolves over time with information revealed sequentially rather than all at once. Each case includes six items aligned with the six steps of the Clinical Judgment Measurement Model. The goal is

to assess clinical judgment in context. Candidates must interpret new information and update their reasoning as the patient situation develops rather than responding to isolated decisions.

At the same time, this format introduces two key measurement considerations. Whether the additional context increases response time and whether the shared scenario creates item dependency. The first issue is efficiency. Unfolding case studies include more shared clinical context. So examinees may need more time for their initial orientation with the patient situation.

In a high-stakes exam, that matters because testing time is a practical constraint. But the shared context could also create efficiencies. Once examinees understand the scenario, they may be able to answer later items more quickly because they are not starting over with a new clinical situation each time. The second issue is item dependency.

Items within an unfolding case share a common patient scenario, and that shared context is intentional. But shared stimulus can lead to residual correlations among items. From a psychometric perspective, this is important because standard measurement models rely on local independence. After accounting for examinee ability, item responses should not be overly dependent on one another.

These considerations led to the following two research questions. Do unfolding cases take more or less time than equivalent standalone items? And do unfolding cases show problematic item dependency? The response time design started with selecting six well-performing case studies, each containing six clinical judgment items.

Then each item was cloned into a standalone version. So the clinical content was held constant while the presentation format differed. Across 25 experimental forms, examinees received either the unfolding version or a standalone version of a corresponding item. Because the standalone items were matched to unfolding items, we were not comparing unrelated questions.

We were comparing equivalent clinical judgment content presented either within an involving case or as a standalone independent item. The final response time sample included nearly 55,000 examinees. Go ahead and scan the QR code for more information on the study design and results. The results I present now are the median response times as a function of the clinical judgment step for both items presented within an unfolding case and when presented as individual standalone items.

For the standalone items, the response times are all pretty stable, hovering around 60 seconds. These serve as the baseline response times for items when administered independently. For the unfolding case study items, we see an initial increase in response time on the first item relative to the standalone counterpart.

Examinees spend more time on that first item. This likely reflects a deeper encoding of the initial information as they recognize they're entering an unfolding case study. After that initial investment, the pattern reverses. For the remaining items in the case, response times are substantially lower than the equivalent standalone items, and that reduction is sustained across the rest of the case sequence.

Overall, the unfolding format produced a net reduction in total response time. So while there is an upfront time cost, that cost is more than offset by the efficiency gained once examinees are working within an established clinical context. For item dependency, we used unfolding cases administered through the NCLEX experimental pool.

These were unscored pretest items embedded within the operational exam. The analysis included 111 unfolding case studies and 90,000 examinees. We used Yen's Q3,* statistic to evaluate whether items within the same case showed residual dependence after accounting for examinee ability. If two items are independent after accounting for ability, those residuals should not be strongly correlated.

The goal was to determine whether the shared clinical scenario created problematic item dependency. The data you will see is the cumulative percentage of the item sets as a function of Yen's Q3. Eighty percent of the case studies reveal residual correlations less than 0.16, with 90% of the case studies less than 0.24, and all cases with correlations less than 0.3. The median correlation across all cases is 0.11.

So returning to the first research question, do unfolding cases take more or less time than their equivalent standalone items? Well, the answer depends on where candidates are in that case sequence. On the first item, unfolding cases took longer than the standalone version. Candidates spent about 23.5% more time on that initial item.

That likely reflects the time needed to orient to the patient scenario, process the clinical information, and recognize that they are entering a multi-item case. But after that investment, the pattern reversed. For the remaining items, candidates completed the unfolding case item substantially faster than the equivalent standalone items, reducing response times as much as 44%.

So across the full set of six items, unfolding cases were not more time-intensive. In fact, they produced a net reduction in total response time of about 27%. In other words, the shared clinical context required more upfront processing, but then supported more efficient responses throughout the rest of the case.

The second research question was whether unfolding cases show problematic item dependency. This is an important issue because items within an unfolding case share a common clinical scenario. The shared stimulus material can sometimes create residual statistical dependence among items. To evaluate this, we examined item dependency using Yen's Q3, and the results showed minimal dependency across all the unfolding cases.

These low values suggest that items within unfolding cases are not overly dependent. That is important because it supports the local independence assumption used in standard psychometric models. In practical terms, the cases were clinically connected, as intended, but they did not show problematic statistical dependence. Combined with the response time findings, this suggests that carefully constructed unfolding case studies can both be efficient and psychometrically sound.

They preserve the authenticity of an evolving clinical scenario while still meeting the expectations of a high-stakes exam. This addresses the functioning of the unfolding case format itself. Next, Dr. Betts will broaden the focus on how clinical judgment items performed operationally on the NCLEX.

- [Dr. Betts] Thank you, Will. My name is Joe Betts, and I'm going to be presenting on some longitudinal trends for the NCLEX with respect to clinical judgment. So we all know that clinical judgment is a necessary skill within the first year of practice. Not only is it important and necessary throughout the lifespan of nursing, it's important the minute the nurses step onto the floors for work.

Now, the other things that we know from research is that clinical judgment must be assessed and taught. So it's something that can be taught throughout schooling, and this is one of the reasons why NCLEX measures clinical judgment directly, because, one, it's needed for entry into the practice, and also number two, is because we want to be able to evaluate those skills for a minimal level of competence.

So some of the implications for evaluating entry-level. Practice analysis that you've heard about before is something that we do consistently here, and that helps us identify what types of skills and tasks are needed for the entry-level nurse. What we find consistently over time, for at least the past 5, almost 10 years, is that whenever we do this type of research, clinical judgment has always been identified as really roughly being necessary for underlying roughly 46% to 49% of the entry-level tasks.

So about 50% of the tasks that nurses need to be capable and competent when they're coming onto the floors are basically underlined by clinical judgment. Now, what we do know is that clinical judgment is not evenly distributed across a number of the duty areas. For instance, there are some areas such as managing care and providing care to patients, where it's vitally necessary.

It's a vital component for clinical judgment, whereas there's other ones, such as working with clients regarding their medical care, for educating them, and also operating medical tools and instruments. It's not as necessary for those. One of the other things we know from research is that case-based reasoning and evolving scenarios have been used for quite a long time when it comes to evaluating and assessing clinical judgment and clinical reasoning. So the implication for the NCLEX was to include clinical judgment in a way that stays true to that type of teaching, the educational component, the teaching, and the assessment component.

And as you heard from Will earlier, we use what's called the Clinical Judgment Measurement Model at NCSBN. And basically, this is a model, as Will has discussed earlier, that moves basically from recognizing cues all the way through evaluating outcomes. We've got a number of psychometric studies that confirm that clinical judgment integrates really well with the historical NCLEX content.

And throughout this presentation, we'll refer to the historical NCLEX content as legacy content, knowledge content, or something of that liking, just to separate it from clinical judgment for this discussion. One of the other things we know about clinical judgment and implications for NCLEX was we had to expand the scoring model. The scoring model needed to be expanded from a dichotomous model, where everything is either right or wrong, to a partial credit model where you can get partial credit for being able to answer challenges and questions.

So, for instance, I'll give you an example. Under the old regime, the recognizing cues item, one would have had to recognize all of the cues to be able to get just one point for that question. So what would happen is if somebody knew, say, two out of three or one out of three, they would be lumped together with all of the individuals that scored zero. And so what we find is that with the partial credit scoring model, we're able to differentiate those individuals who know one, two, or three of the different important recognizing cues identities.

One of the other things that we look at, now that we've been measuring clinical judgment in the NCLEX, is every year we go out and we ask those newly licensed nurses, those nurses who've just made it out into the field and have been practicing for less than a year. What we do is we ask them questions to validate the use of the clinical judgment evolving scenarios that we place in the exam.

And every year, we consistently find from the research that nurses continually say that the clinical judgment content that they get on the NCLEX is very similar to the type of work that they're doing every day on the floor and in the milieus that they work in. So it really gives us a lot of good information about the validity of these. So we feel that the clinical judgment items and the evolving cases themselves are quite fidelis. So what we'll talk about now is some of the longitudinal NCLEX results.

What we'll do is we'll provide some results for clinical judgment and for the knowledge or the legacy types of items to compare the results over time. We'll also give you some idea about the score precision for the exam, and then we'll compare some of the clinical judgment measurement elements longitudinally. Now this slide right here is a box and whiskers slide, where you can see each one of the major programs: the RN program, the PN program, and RN-1, of course, which is the Canadian program.

We've got them grouped together by quarters, and you'll see in the left-hand side there's a red dotted line. That's the dotted line that demarcates the time shift between the old NCLEX and the newer NCLEX, where we've specifically added clinical judgment. Now on the y-axis is the standard error of measure. And for the standard error of measure, lower is better.

When we have lower error of measurement, we get higher reliability for our scores. And as you can see with the new model, across the board and over time, there's been about a 20% drop in the standard error of measure, which means there's much more reliability in the exam than there was previously. Secondly, you'll notice that the new exam has a very tight band. Those boxes that represent the interquartile range, you can see they're much more compact than previously.

So we're getting a lot of people with the same level of reliability, even though we're providing a computerized adaptive test where people will get a different number of items. This next slide will show the difference between how students are doing on clinical judgment and how they're doing on the regular knowledge or legacy type items from the content areas. And what you'll see is there's a red line that demarcates, again, where we changed from the old milieu of the dichotomous exam to directly measuring clinical judgment in this exam.

What you'll see is that the top part of this, with the circles, are the indicators for the trends for clinical judgment, and the ones with the boxes will show the trends for those knowledge-type items. One of the things that you can notice is because there are these types of items, which the exam was completely made up of before we changed over, you can see that over the period of time from a year before all the way up until currently, you can see students are continuing to score similarly as they have previously on this area.

So the knowledge items have pretty much stayed similar to where they were before. You also notice that there's a cyclical variation in there. That's very normal with our data, with higher scores in Q2 and Q3 and lower scores in Q1 and Q4. Now one of the big things, the really interesting things, is this part of the graph here. This shows the clinical judgment, and this is when we started directly measuring the clinical judgment with the evolving scenarios.

And what you'll see here is when we started off, the first time out the door, students were doing really well at this, but they were also very highly variable. Look at how spread out those circles are up there. But what you can see is, over time, they've slowly regressed back towards where they are with the knowledge items today. They're still, though, doing very well on the clinical judgment items. As you can see in the far right-hand corner, they're, on all the exams, still outperforming on the clinical judgment.

What these graphs will show is the breakdown of the clinical judgment model by the clinical judgment components or elements. And what you'll see here is up in the upper left-hand corner, we've got the trend circled. That's for the PN. And what you'll see is there's a slight decrease on the PN, as we noticed

before, just in the overall clinical judgment, but there's the same type of downward trend for all of the clinical judgment component elements.

One of the other things you'll notice is that as early on the first time points over on the left-hand side, there's quite a bit of variability in how students were scoring across the different elements. And we've seen slight shrinkage in that over time, to where we are now, where there's a little less difference between students scoring on each of the elements. However, one of the interesting things is now looking at the RN.

So we have the RN for the U.S. in the upper right-hand corner and the RN for the Canadians down in the left-hand corner. And you can see both of these show very similar trajectories. There's a much more marked decrease over time in the scores between them. And one of the other things you'll notice that's really interesting is on both of the exams, students are scoring less well on the prioritized hypothesis.

Now if you look at the PN up in the left-hand corner, you'll notice that there's somewhat of a similar trend. There are a few quarters where prioritizing hypothesis was not the most difficult, but you see in general it was also one of the more difficult for the PN. So the interesting thing here is that both RN for the U.S.

and the RN for Canada show that prioritizing hypothesis is a very difficult element for them. Another thing to notice with the two nursing exams is you'll see there's almost a funneling effect here, is there's great variability across those elements when we first started providing the assessment. And now, over time, you can see, with the exception of prioritizing hypotheses, you can see whereas there was a lot of variability in the beginning, it's now basically funneled down into a very tight window.

So what this means is that students today are scoring very similarly across all of the clinical judgment elements, excluding, of course, prioritizing hypotheses. So in conclusion, we know that clinical judgment is vital for practice. And so to measure it in the NCLEX, we use evolving case studies, which have been shown to be very effective both in education and in assessment. We follow up with the students as they go out into the world of work after they pass the exam and they begin working that first year.

And we get very good feedback that the types of things and the type of challenges that they see on the exam are what they're seeing in their everyday work. We show that candidates still show a strong ability on the clinical judgment question and evolving case studies. And what we've also seen is that over time, there's been a decrease in the variability between those elements. However, the one caveat for that is that students continue to score or have a more difficult time with the prioritizing hypothesis element.

And so I'll provide you here with some references so you'll be able to take a look at those references. All right. And I want to thank you all for coming to the Scientific Symposium and sitting in for this presentation. And I guess since I'm the last person in line here, I will wrap it up and move over to the question and answer for the Q&A for everybody. Thank you very much.

- [Jason] Okay. Welcome, Joe.

Welcome, Will. We only have about six minutes in front of us, so I'll encourage you to apply some parsimony to your responses. Joe, I'm going to start with you. What have the first few years of operational results shown in terms of measuring clinical judgment at scale?

- Oh, all right. A lot. First of all, that we could do it. As you could see from the numbers, we've done over a million assessments here. The nice thing about it is I think it shows that these evolving case studies and the process by which we build those items, which is outlined in one of the references that we provided, has really been able and really been scalable in the sense of being able to build out a bank of items.

So you think about just the millions of people and the million people that have taken the exam. We've also been able to build 600, 800, 1,000 of these items using that type of a pipeline. So it is possible to build these items reliably and at scale. So it's really been fun. So just the idea of well-designing unfolding case studies and being able to scale them up in just the production of them and then the distribution.

And I'd say finally, the biggest thing is every year, like we've said, we've had the responses from entry-level nurses when we go out with practice analyses. And they've responded that these really do mimic what they see in the field. And so I think that's a big shout-out to all of the nurses that have come in and all of the educators that have come in to help us write these items and make them as having a great level of fidelity.

So if all of you who are listening helped us out with that, thank you very much. We really appreciate it because there's a lot of good feedback that you're doing a really great job.

- Joe, I feel like that's a perfect note for me to pivot to this question for Will. You talked about the clinical realism. Will, how is it that these case studies achieve clinical realism while still meeting psychometric requirements? Are there trade-offs there, or does it just sort of naturally happen?

- Yeah, I think the balance comes from treating realism as part of the measurement design, not just as decoration. So as the case unfolds in a clinically realistic way, each item is deliberately tied to our Clinical Judgment Measurement Model step.

And that gives us a structure with observable targets, both for item writing and for scoring. So empirically, we find that that format does not tend to increase total response time but does produce or does not produce problematic item dependency as well. So I think that's where the balance comes from, is really treating it as part of the measurement design.

- All right. Thank you. And Joe, I'm going to start this one with you. But Will, you can certainly chime in if you like. Joe, I think you had mentioned that graduates were performing maybe the worst in prioritized hypotheses. Do you have any hypotheses for why that might be the case?

- I wish I did. Not being a nurse myself, it's very difficult to make that leap about why they might be performing worse on that. So I wish I had an answer for you. But I think it's more along the nursing knowledge line that I'm not really competent to answer. Sorry about that.

- That's okay. That's all right. I have another question for either of you. Has anyone explored data on the new NCLEX test plan and patient outcomes?

- Not that I know of. Will, do you know of anybody who's done that directly? I'm not aware of anybody.

- Not directly. I don't think so. For us, our main priority is that nurses who pass the NCLEX practice safely. So that's where our focus is.

- Will, I'm going to keep you on the mic for this one. How do you score clinical judgment scenarios?

- Yeah, so we use a partial credit model. So we award partial credit for showing some level of knowledge on each item. So this really supports some of the nuanced measurements of decision making, especially across complex item types and evolving scenarios.

Specifically, we have three rules. We have the zero-one scoring rule, where points are earned for each correct response. We also have the plus-minus scoring rule. This is applied when an examinee can select as many options as needed to accomplish some objective. It's often in select-all-that-apply type questions.

So points are earned for correct responses, but they are deducted a point for incorrect responses as well. So there's a balance there. There's no negative scores allowed. So they're truncated at zero. And then we also have the rationale scoring rule. And points are earned here when an examinee can show competency when connecting two tightly coupled units of information.

So typically, cause-and-effect type questions. You might know the effect, but you need to know why it occurs. So when you know both units of information, you earn a point.

- Thank you, Will. And Joe, I'm going to put you on the hot seat with a speed round here as I'm looking at the clock. What do you think are contributing factors to clinical judgment item responses decreasing over time?

- Can you repeat that item? Responses decreasing...

- Yes, so the question that came in: what do you think are contributing factors to clinical judgment item responses decreasing over time?

- I'm not sure. I'm not sure I understand what decreasing over time means. I mean, the regression back towards the mean for the scores and the clinical judgment pulling back towards the content areas? Is that the gist of the...

- I think that may be it. That may be it if you've got something for that.

- Yeah. My first thought as a statistician and psychometrician is that that's just a normal sort of return back to the mean. There's a lot of impetus and a lot of really great work that went into the buildup to getting individuals ready to take the new clinical judgment in 2023. There was also a lot of people who took the older test before they decided to come in to take the new test. So there was a big rush of people using the old test design.

So I think, just statistically, a lot of it is just kind of regression back to the mean. There was a lot of preparation that went into it. Things shot up, and now they're just sliding back down slowly over time. I don't know if Will has another opinion on that or not.

- Yeah, I think it is. Yeah, statistical property, I think. Yeah.

- All right. Well, gentlemen, we also just heard from the audience member that indeed that was the question about regression. So well done.

- Did we answer it?

- Well done. So with that, this concludes the session. So thank you very much, guys. And I do want to announce that we're now starting on a break. At 2:30 Central, please come back to the lobby, choose one of two concurrent sessions, and we will see you again soon. Thank you very much.