# *2021 NCSBN Midyear Meeting* - Next Generation NCLEX (NGN) Forum Video Transcript



**Event**
2021 NCSBN Midyear Meeting

More info: ncsbn.org/14987.htm

**Presenter**
Phil Dickison, PhD, RN, Chief Operating Officer, NCSBN

Welcome, everyone, to this presentation. I want to give you some updates on the Next Generation NCLEX. You'll all know, many of you will know, it's a journey that started many years ago, around 2012, and we got some exciting updates to give you and a look at what success might look like.

So if you'll come along in the journey with me through this presentation, we'll be giving you some updates on just exactly where we're at right now with our research, what the item development looks like at this point, scoring, how much we're going to score these models, testing, design, what the actual overall exam might look like. And finally, how we're going to do beta testing so that we can assure every one of you, one, that we can launch this safely, and two, that we can provide the best measurement that has become the hallmark of NCSBN and the NCLEX, and we can do that both with psychometric and legal defensibility.

So where are we now? Well, the first place is we've already done this. The green check mark tells you it's completed. We developed a clinical-judgment measurement model, we published a great deal of information on this through 2014, '15, and '16. And during that same time period, we started developing item prototypes.

What could items look like? If you were to remember, the NCLEX currently has multiple response items and multiple choice items. We found early on after developing the model that those particular items, one, could not measure, at its best, the concepts and the various elements of clinical judgment like cue recognition versus hypothesis generation.

And so we have started developing item prototypes. And I'm going to show you where we've ended up on that and what new items might look like as they appear on the NCLEX. We also created a great deal of due diligence around, how usable are the items? And I don't mean just usable from a psychometric perspective or can they actually be scored, we know that too, but could individuals sit in front of a

screen, look at one of these items, and actually know how to answer it because they're a little more complex.

And so we did a significant amount of research, one-on-one with individuals and a talk aloud protocol to get input on what a item should look like, what the instructions should look like, what the screen should look like, and we incorporated all of that in, created a set of items, and started what we're doing right now.

And you can see I've done that check mark right there, this is in progress. And so we're doing item data collection right now. I tell you some very interesting things about item collection that has surprised me and, in some cases, astounded me, but we started item collection 18 months ago, maybe 24 months ago by now, and we've had over 600,000 RNs actually participate in the special research section that we added to the end of the NCLEX.

That really is an amazing number. What it allowed us to do was, then, to get a much larger sample, a lot more data. Now, it also allowed us to develop, what I call, the motivation index in which we knew some of these six hundred-plus thousand nurses, RNs, that took this just wanted to see the items, and they weren't really interacting.

But I can tell you, we know that over 340,000 of those completed every single item, passed what we created as a statistical-motivation test. And so as a minimum, every one of those participated in 10 items at a minimum, at a maximum, 22 items. Some quick math tells you that we have data points to analyze for more than 3.4 million data points.

That is pretty amazing. It's allowed us to do some really great work with the staff, the psychometric team, the content team just have been amazing in the work around this. And so that allowed us to get to measurement research, start thinking about how do you score this, can we score them, what are the impacts around various scoring categories and those sort of things, and we'll talk a little bit about that.

We're also, all of these last three, you can see the check marks are in progress now. We're now building, working with our vendor, Pearson VUE, right now to build the technology that will allow us to launch this exam because we don't want to change anything. The ability to deliver an item in the same amount of time as the old NCLEX items, the ability to deliver these items and simultaneously in the same exam, deliver NCLEX items, the ability to get scoring out of this in real time, so we don't slow any of the processes that have been the hallmark of the NCLEX exam for a number of years.

And finally, we're going to be talking about the alpha/beta testing which hasn't started, but I'll give you a lot more information a little bit later on. The item development, where are we at is that we know a few things about item development. First, the clinical judgment on the new exam on NGN, Next Generation NCLEX, will be measured in two ways.

Two ways are case studies, which are basically real-world scenarios accompanied by different test items. So what I'm getting at, and I'll show you a little bit after this slide, is that we're going to go back and follow a client through a case much like a nurse would do in the real world.

So we're calling them case studies but ultimately, real-world scenarios, and we'll be able to measure those as a whole case in the individual item. The other thing that's going to happen is there will be standalone items. Now, standalone items means that we're not able to measure the entire clinical-judgment model, but we're able to measure portions of it throughout the remaining exam. So it just adds more data to what we're looking at in terms of getting a valid and reliable measurement at the end of the exam.

Here's how it kind of looks. When you're measuring clinical judgment, standalone items are going to target one or more of the six layers of the clinical-judgment model. We don't have it here, but over the last several years, I've talked about this clinical-judgment model that had different layers, and we were going to focus on Layer 3. So what I've cut out here, this is Layer 3, and this is what you would call the action model or the measurement model, meaning we need to be able to measure, recognize queues, analyze queues, prioritize hypotheses, generate solutions, take actions, and evaluate outcome.

And what I'm getting at is a standalone item won't measure all of those, but it will measure one or more of those. The case study, it'll target multiple of those, in fact, it will target six of those, maybe all six of those, maybe a repeat of three of those, but the idea is there's a case study, and you imagine in any case, you may have iterations where you go from recognize cues to analyze cues, and then you have to go back to recognizing more cues, but it will be a fidelity situation that can measure any of those six.

Every case study will, in fact, have six items in it. And what would those look like? Well, I've put it in an example in today's presentation, so I can give you an idea. Now, the point of this, when I start showing you this sample, is to show you how the items might look. This is not an item that will ever appear on the exam now because it's compromised by showing it to you, but the point is, I want you to take a close look at how they're designed.

This isn't a time for me to stop and debate the words in the item, but if you want to do that, you can probably pause this video and read through it a little closer, but the point I want to show to you here is what items might look like. And so if you had a case study, you can see this week case study, this would be the first screen of six screens that the candidate is going to see, and we tell them that.

You'll notice that what happens, and like a regular NCLEX item, is we start out and say, the nurse is caring for a client, in this case, a 78-year-old female. And we tell them the context, the environment, they're in the ED. With that, they have the setup. And now, we start giving them a client care report.

So you'll see these tabs. Important thing to point out here, this one has one tab in it because only one thing has happened. This tab could grow. It could include lab results, there'd be a tab for lab results, a tab for vital signs, a tab for… So it's basically a progression of the client's history through the case, and it just continues on through it.

So as you look at this, always on the left side of the screen as you look at it, will be the information about the client or the client care record. On the right side, will be the actions required by the candidate to answer the questions. So they read through the nursing note, they can see it was at 10:00, they know it's in the emergency department, they know it's a 78 female, and they are now asked to select the four findings that require immediate follow up.

So they have to read through this nursing note and define, out of those six, which four required immediate follow up. So basically, you can see this is a close-up view, and I'm going to go into this now. Sorry, I know this is repeat, but I want to go into it because the rest of the screens, I'm not going to show you split screens, I'm going to show it to you like this.

So this is a close-up view of what's on the left-hand side of your screen, we just talked about it. This is what a close-up view of the right hand would see. Now, your candidate sees both of them side by side, but I want you to see them this way. The second one, remember, the tab hasn't changed now. So I'm showing you a close-up view of the second item. And so for the information that was in that nurse's notes, now they're being asked for each client finding, and we give them the finding.

Those were in there. They can read about those. These existed, fever, confusion, body soreness, you see them there. We now ask them that they have to click to specify the findings that are consistent with the disease process of pneumonia, urinary tract infection, or influenza. So now they're evaluating the clues. So they have a fever, they're going to have to suggest to you, is fever indicative of pneumonia, UTI, or influenza?

This is a little different from a regular NCLEX where they could only answer one of those. And here, they can answer any one of those. So each finding may support one or more disease process. So fever, pneumonia, you'd click it, UTI, click it, influenza, click it. So they click all three of those whereas as they go down on some of these other ones, they may only click pneumonia, they may only click UTI.

What they have to do in this one, as you'll note at the bottom, each column must have at least one response. So they got to have one response in each column. They could leave the rows blank. They're not required to answer everything. This item, we would call a matrix item, obviously, because it's a 3 by 5 matrix. They can answer any of the items in that matrix.

This, we call a cloze item. I like to think of this item, really, as not particularly new. It's new in measurement and the ability to measurement but the item type is not. This item, at least for me, came into my world when I started reading highlight magazines and reading…there was always an article on "Goofus and Gallant" and you had to finish the story.

So you'd start reading along and you'd come along some activity that you'd have to pick whether Goofus or Gallant did it, and then you would read along and you'd have to say what happened because Goofus did it, this happened because Gallant did... You got to write the story in some ways, but you had to select one and it impacted the other. So if you've selected Goofus, you needed to actually respond on the outcome based on what would happen with Goofus versus Gallant.

So this is the same sort of thing but we figured out how to measure. So you go, the client is at the highest risk for developing something, and I'm going to show you another slide to get to that, as evidenced by. So you're setting up an analysis of the risk and the evidence that supports that, this, the way clinical thinking works. So in this case, the client has the highest risk for developing, and then you get to select hypoxia, stroke, dysrhythmia, or an embolism, and then they have to select, based on whichever one they picked, what were the client's signs that actually indicated that.

So you can see it's sort of a two part. Now the interesting thing about this, we could have gone on two more or three more, right? So this doesn't have to stop in one risk and evidence. The story could have went on, and they would have to answer more of these, but they're always connected one to two.

All right, so it's a cloze, what we call a cloze, C-L-O-Z-E, item. Now, what happens in the fourth item is this…you can see at the top, the nurse is still caring for the same client in the same place, but it's two hours later and there's more information. The nurse has now reviewed the nurse's note entries for 10 and 12. So it's sort of a cue, you probably ought to have read that, and they're planning the care.

So they need to now talk about for each potential nursing intervention, click whether it's indicated or contraindicated. So here are your potential interventions, and the candidate has to determine is this indicated, would it be appropriate, would it be contraindicated, inappropriate? You go on to the next question and you see that the nurses reviewed the orders for 12:15, so there'll be orders for 12:15, and they're to click on the three orders that they should perform right away.

And so they're told that they have these orders to give, now, which one or what priority are those given in, and they do this in a different way by highlighting. So this is really an interesting item because you're seeing it this way, but you could also just have an open paragraph where they have to highlight sentences actually in the nursing note to answer this.

So there's a variety of way this one can work, but I wanted to show you in a simple way where they simply highlight those answers that, in fact, should be performed right away. And then the sixth and final item looks like this. The nurse has performed the interventions that they just talked about on the previous answer. For each assessment finding, they have to click to say, has it gotten better, there's no change, or it's gotten worse.

This is what we expect of nurses, and we've figured out a way to do this. So the difference between this matrix item, this got a matrix item again, but you'll notice it has circles instead of squares. Because in this item, you look at respiratory rate, 36, it cannot… They can only pick one thing in a row, meaning that it can't improve and have no change. So they can only pick one.

So they're forced along this to pick a single answer. This is what a sample case might look like. So a candidate is definitely going to get six items, it's going to run through either all six of the Layer 3 boxes in the model, or it is going to use an entry approach and perhaps measure a subset of those items and revert back to, but the point is, there will be six in the case study, and it will measure Layer 3 along the action model that I showed earlier.

The items that I highlighted, you saw a highlighted item, a cloze item, a matrix item, extended response, the trend item is one that, that trends along, as I was talking about, it has to do with the case study. It doesn't actually change. And the extended drag and drop includes a bow tie, which I want to show you now. The sample standalone is called a bow tie item.

So take a look at this. This is an item that a candidate would get. It just pops up in the middle of the exam. It doesn't have a case attached to it. The entire case is on the screen right now. So you take a look at this, you should have the nurse, they're in an emergency department, remember we got context again,

and they're caring for, now, a 79-year-old client, they have two tabs they need to read, a nurse's note, a history, and a physical.

And then, the question is on this side. And so basically, if you take a look at this, the nurse is reviewing the client's assessment data, and they're going to be preparing a care plan. So they need to create the diagram. So in the middle, conditions likely…the condition most likely experiencing, they have to pick from Bell's Palsy, hyperglycemia, ischemic stroke, or urinary tract infection and drag that condition up to the top where it says, most likely experiencing.

Then from there, they have to go on the left-hand side, pick two of the actions they would take based on that potential condition, and then take and draw from parameters to monitor what they should monitor based on those actions. So you can see this is a bit more complex, but it measures two or three boxes of the clinical-judgment model in a single item.

I will tell you a little bit more how those are going to get put in the exam later in this presentation. So let's talk about the status of item development. What is going on right now? Where are we at? So clinical-judgment item sets, we began that work in 2017, and I can tell you, we're pretty near launch goal of what we would need to launch, and then we would simply continue to maintain that bank.

So we've been really successful. The PN work began in January, this January of 2020. And even through the pandemic and some of the things, we've been able to continue that work. Our primary focus right now is PN work because, as I said, the RN work is pretty close to launch. Now, we will have to have a maintenance on that but really exciting that we're doing that.

The clinical-judgment, standalone items are in progress. So those have not being completely developed, but we are confident we'll be able to do that and wanted you to see what they might look like. So ultimately, I can tell you, item development is on track. Our project plan is in place.

We haven't missed anything. COVID has not slowed us down, and we are on track to launch in 2023, more importantly, spring or April of 2023. What's the scoring going to look like? So we're going to take a new approach to scoring these items. You can imagine that the way we've scored items before or today, as we're scoring them actually today, we score them as either right or wrong or what we call dichotomous.

You either get everything right or you get everything wrong. And that's worked well for us based on having multiple-choice items with four responses or multiple-response items with up to maybe six. The problem with that is if you start looking at what we were calling matrix items and you take a 3 by 5 matrix, there's a potential of 15 responses in that, and choosing to get all are right or all wrong has some problems with it both in terms of psychometric fairness and legal defensibility.

So what we've looked at is doing something where we call polytomous scoring, which means there are more possible categories. You are not one or zero, either you get it right or wrong, which suggests that on any item, if one of you get it right and I get it right, we have equal ability, and if one of us get it wrong or both of us get it wrong, we have no ability.

And we fundamentally know that's not true, but now with the ability to create larger-scoring categories, we can actually distribute ability across more categories, let's say zero, one, two, three, four, maybe up to as high as eight categories.

So this is actually a great thing for us. In fact, what has it done for us? The ability to partial-credit score has given us higher, what we call inter-item correlations. What does that mean? It means actually, we're getting better correlations with our NGN items to the final score of that candidate than our current NCLEX items are giving us.

So really, probably a very positive thing for us to do, and I'm glad we got there. I want to tell you, the other thing that we're able to do is, even with this scoring, we've been able to maintain the 95% decision accuracy that we've required for the NCLEX, our correlations and reliabilities are staying essentially the same.

So we're gaining a significant amount or, well, we're gaining a certain amount of inter-rater, inter-item correlations to the final score, adds the credibility, defensibility to the exam, and we aren't losing anything related to validity and decision accuracy. So these are really good approaches, and I'll talk just a little bit more how we're going to do partial-credit scoring in this next slide.

So there's different types of partial credit or polytomous scoring, and we're going to use all three of those because our items dictate that. So partial credit can be assigned three ways. You can have a candidate receive a point for a correct response and lose a point for an incorrect response, what we call this is plus-minus scoring.

What this amounts to is allowing a candidate to over or under respond, meaning they can pick as many as they want or as less as they want, and we can still score that item, so we call it plus or minus. So if you look at this, which of these countries is in North America? And let's say the candidate picked France, Mexico, and Canada.

Using plus-minus scoring, here's what happens, they get plus one for Mexico, they get plus one for Canada, and they get a minus one for France. When you add that all up, they get…there's a maximum of three points because it should have been Mexico, Canada, United States, they get a maximum three points, but they would get one point, two minus one, they got two right then minus one, they're going to get a one point.

This has worked really well in our model. It's really stable for those items that allow for over and under responding. Now, you have another model that let's say you can't over-under. We tell you, you have to select a certain amount, and we won't let you select more than the amount. So that's sort of restricting this. We call it zero-plus scoring. So same example, instead of getting you a minus, you get either a zero or a plus one.

So in this case, where we force them or restrict them from over responding or under responding, we use this model, so the candidate earns two points out of this. So there's still a maximum of three points, but there's no minus for France because they weren't free to pick more or less.

They had to pick three. And this has worked out really well for us too. The third type is where we're using tokens and different things of that. The candidate then gets an all or nothing or a cloze. So think about that cloze item that I gave you where you had to pick the disease process and the evidence of that disease process.

So I'm going to put this in not those terms but let's assume that you have to pick…you get these tokens, it's a cloze item, so the capital of, and you have to pick from that list, France is Paris, the capital of Egypt is Japan. We know that, that bottom one is incorrect, but the way this one gets scores is the, can only earn a single point for the pair.

So if they get the capital right but the nationality wrong, they don't get any points. So it has to be connected. So you can't get one part of it right and one part of it wrong. Has to be connected. So per line, there is one point on this. So we're going to use that type of scoring model. Again, this has been really stable, as added credibility, and allowed us to have the appropriate reliability across scoring categories, very stable.

The benefits of partial-credit scoring are numerous but importantly is partial-credit scoring has allowed for more precision of measurement. Remember that inter-rated reliability, inter-item correlation I'm talking about, this has allowed that because it's really tied to the complexity of items but having multiple ways of earning partial credit also reduces the impact of random guessing or gaming the items.

So both of those have come about because partial-credit scoring. The other thing in partial-credit scoring obviously it has done is given us a great deal more item information. So for instance, partial-credit scoring in case studies have allowed us, per item, to gain probably three to four times more information per item than we have on the regular NCLEX, so another large benefit for us.

We can make our decisions in a quicker way, get more information on the candidate in a much quicker way using partial-credit scoring. The other thing, though, to keep in mind, partial-credit scoring does not change our ability to use the CAT exam. We will continue to use that.

It doesn't change our scaling. So the scores going forward with NGN will be comparable to the scores 2 years ago, 10 years ago. So we haven't broken that connection between following scores over time on the NCLEX. We've been able to do that for 20 years, and given what we've been able to do with partial credit, we will be able to do that long into the future, very positive thing for defensibility and credibility of the NCLEX exam.

What is the test going to look like? Exactly what does it mean when I say test design? There's a lot of things that go into that, but I want to be clear what I'm talking about. How long is the exam? How long is the, you know, how long are the exam items, meaning not just in hours and minutes, but how many items are going to be on it? What's the mix of current knowledge items versus clinical-judgment items, and how will the items and cases be selected for delivery?

How are we going to distribute them across the exam? That's an important piece as well. So I wanted to give you this information, where we're at. The length of the exam will vary by candidate. Remember, I told you that we were able to stay with the CAT, which means that we'll still have a variable-length

CAT, nothing changes for the candidate in that regard. So the actual examination experience that has been the hallmark of NCLEX since it went to CAT will continue to be the same thing.

What will we see that's a little bit different? So the minimum-length exam, the candidate with a very low or very high ability, remember, that's where the exam shuts off at the least amount of items. What will that entail? Well, that will entail in a minimum-length exam, the candidate will get three scored case studies. Remember, I already described the case studies.

Those are going to be 6 items, thus, a total of 18 items that are tied together following a client care plan. The other one is that they will get 52 scored-knowledge items. So you can do some quick math here. That means the total minimum-length exam on the NGN will be 70 score items and 15 unscored item. Now, that's what we do today in the unscored items so that's no change.

You'll see a little bit of a change in the scored items. We use 60 today, so there'll be 10 additional items when we add the NGN items. So the minimum number of items will go from 60 to 70. What is the maximum length? That means the total number in time that an individual can sit for the exam. And this always happens for individuals where their ability estimate, that we're measuring at this point in time, as they go through the exam is too close to their CAT score for us to have reliability meaning there's a little more error around it, and we need to give them more questions.

So what are they going to get? Just like minimum length, they will get 3 scored case studies, another 18 items. So they won't get additional 18 item, they'll get the same 3 items in terms of distribution that the minimum links will get. They will get 117 other scored items.

Here's where things change. So they'll get more other scored items. Most of those will be knowledge items. So remember, up top there was 52, now there's going to be 117. Most of that 117 will be knowledge items, but approximately 10% of those items will be the clinical-judgment, standalone items, that bow tie item that I was showing you earlier, and there'll be about 10% of those.

So what does that mean in total? That the maximum-length exam will be 135 scored items and 15 unscored items for a total of 150 items. And that candidate will have five hours on a standard exam to complete the exam, more if that accommodations…if they need accommodations but the general is that.

So I wanted to give this to you first, but I thought it'd be nice if I provided a comparative chart. So let's walk through this. The time allowed today for the NCLEX is five hours. The time allowed, minimum time and maximum time for the exam is five hours. So it's five hours, no matter what.

Case studies, current NCLEX doesn't have any. The case studies in the minimum-length exam will be three, and then the maximum-length exam will be three. The clinical-judgment standalones, there are none today in the NCLEX, in the minimum exam, there will be none and in a maximum exam, there's going to be approximately seven. I said 10%, that's about where you would end up on that.

So knowledge items, on the NCLEX today, there are 60 to 130, because that's all there are today. On the new exam, for a minimum-length exam, there would be 52 knowledge items and for a maximum-length exam, some are around 110.

Because remember, about seven of those would be bow tie items. So total scored items today on the NCLEX is 60 to 130 on what it will be in the future is 70. You can see that on a minimum length. That's where the 6…so you compare 60 to 70, it will increase by 10. And the maximum length will increase from 130 to 135.

Unscored items don't change and the CAT doesn't change either. So I did put some asterisks to say something about this and that is that items within case studies are static, they're not adaptive. So remember, there's always six items in a case study. And in that case study, those are static, meaning that you pick the case study and it doesn't change.

The items within that case study don't adapt, they stay exactly the same. So I wanted to make sure I was clear on that, those are not adaptive, the case study is static once it is picked. So then what is beta testing look like and what does it mean? So I want to tell you, beta testing, I want to define what we define as beta testing here. And beta testing is what we call end-to-end testing of all elements of the NGN prior to launch.

So lot of times when you do beta testing in psychometrics, you're just testing, do the items work. For us, beta testing is much bigger than that. So it's the registration and scheduling, it's the test publication, it's launching the exam, administering the exam, then it is actually looking at the function of the items and the cases.

It's the algorithm, does it work, does the stopping rules work, the pass-fail decisions, and the data, and reports both internally for the council and externally for the regulators. That follows basically the alpha test. That means that we've built all of…most of this stuff that I just said, we've already built, but we're now alpha testing each of those as they are built and are completed to make sure that they individually work, and then we have to put them all together.

So I want to talk about the beta testing, how we think that's going to work. We think that it should be two phases, and we're calling the first one "Friends and Family." We think that that one will start around April of 2022, and it will work in a way that is sort of high level in this regard.

We, the NCSBN, are going to select the participants in this, and they will include regulatory board staff and some other stakeholders but no students are candidates. And we're going to administer this actually at the PPC similar to the way member board reviews occur. The point of this is that you, if you participate in this, would actually review a maximum-length exam.

You'll get to see how those exams questions flow, how case studies flow, how the screens look, the test centers, how that operates. So it's basically friends and family, you're going to have an opportunity and probably a requirement to provide us responses to some set questions. We want to know about your experiences, you went through there to provide any concerns or enhancements that we might be able to make.

That's what we're going to call "Friends and Family." Now, the other one I'm going to call "Live" beta test. Live beta test means exactly what it means and that is that we're going to actually test real nursing students who are expecting to graduate after April of 2023. We're going to start this in December of '22, and we're going to get these individuals to actually take a fully functioning exam that we will run all of

the psychometrics on, all of the scoring algorithms, but we will not provide the scoring or the results, it will not count for this candidate, but they will be interacting with it in a real time real way.

Reports will be generated for the NCSBN QC, but they will not be reported for licensure or other results outside of NCSBN. We think this is the best way to do this because you need real, live candidates to take it in such a way that you can determine that all things in a real situation are working well, but we also need those users of the results.

So that's the friends and family. The regulatory bodies are the users of the results and we need them to take it and provide us with as much information as possible so that when we get results, they can not only trust those results, we can defend those results, but those results and how we report them are usable to those. So this is our approach to the…the two-phase approach to beta testing.

And with that, I have actually come to the end of this. I will be hosting a question-answer session, if you will, at the midyear, at the upcoming midyear, and I wanted to actually give you an opportunity to see more in-depth information on this video presentation so that we can spend our time at midyear in a valuable way, you asking specific questions and me helping to provide specific answers.

So I hope this has been helpful, and I look forward to seeing you at midyear. Thank you.