Video Transcript
Examining the Reliability and Validity of Teacher Candidate Evaluation Instruments

February 13, 2019

R. MARC BRODERSEN: Good afternoon, everyone. Thank you so much for joining us. We're excited to be hosting this webinar today. I'm Marc Brodersen, a senior researcher from the Central Regional Educational Laboratory or REL Central at Marzano Research. We're very happy to have everyone with us today.

Today, we're going to provide an overview of the need for developing standards for reliable and valid teacher candidate evaluation instruments. We're also going to hear from leadership from several educator preparation programs in North Dakota about their experiences in developing a teacher candidate evaluation instrument and the steps they've taken to examine its reliability and validity. We're very excited to have representatives from the Council for the Accreditation of Educator Preparation and the North Dakota Association of Colleges for Teacher Education (NDACTE).

As I mentioned, this webinar is being hosted by REL Central. REL Central is one of 10 regional educational laboratories that is funded by the Institute of Education Sciences in the U.S. Department of Education. REL Central serves a seven-state region, including Colorado, Wyoming, Kansas, Nebraska, North and South Dakota, and Missouri.

As with all of the RELs, we are charged with providing technical assistance and conducting research on priority topics for stakeholders in our region and nationwide. All of our work is directed by our stakeholders and organized under research partnerships or alliances. This webinar is being conducted under the umbrella of our Educator Pipeline Research Alliance, which currently has projects focusing on educator preparation, evaluation, and mobility.

We very much hope that this presentation is of interest to everyone that is participating today. Now, I'm going to hand this over to Steve Meyer from REL Central, who will introduce our presenters.

STEPHEN MEYER: Hey. Good morning, everyone. This is Steve Meyer. I'm with REL Central and RMC Research. I'd like to welcome you all to today's webinar and am pleased to introduce our producers–our presenters, excuse me.

First, we'll hear from Gary Railsback of the Council for Accreditation of Educator Preparation or CAEP, which as most of you I expect to know is the national organization that accredits

educator preparation programs. Gary is a vice president at CAEP, where he has oversight of the accreditation process and the accreditation team. He has a wide range of experience with accreditation through his work at previous institutions and at CAEP as a trained site–lead site visitor and inquiry brief lead site visitor.

He's performed site visits in 15 states and was a reviewer with the California Commission on Teacher Credentialing and the Oregon Teacher Standards and Practices Commission. Gary is going to tell us about the CAEP guidelines for establishing the validity and reliability of assessments created by educator preparation programs and some of the challenges that programs have faced in meeting CAEP guidelines.

Following Gary, we're going to hear from three presenters who each represent an educator preparation program in North Dakota. So, Sarah Anderson is an associate professor and accreditation coordinator in the Division of Education at Mayville State University. She's a former high school special educator and is now teaching graduate and undergraduate pedagogical courses. Her research interests include teacher appraisal for continual improvement, poor instruction progress monitoring for RTI, and effective instruction and interventions.

Stacy Duffield is a professor in the School of Education at North Dakota State University. She teaches graduate and undergraduate teacher preparation courses. Her main research interests include teacher preparation and assessment, and middle-level education. She's been involved with the work of the Network for Excellence in Teaching since 2010, helping to further the development of valid and reliable instruments for teacher preparation.

And last, Alan Olson is a professor and assessment coordinator in the School of Education and Graduate Studies at Valley City State University in North Dakota. He teaches graduate and undergraduate preparation coursework and he's been involved with common metric assessment efforts through the Network of Excellence in Teaching, as well as the North Dakota Association of Colleges of Teacher Education.

And each of these presenters is very familiar with the CAEP Standards, having served as CAEP reviewers over the years. And, for the past few years, Sarah, Stacy, and Alan have been working as part of a collaboration of the 12 member institutions of the North Dakota Association of Colleges of Teacher Education, or NDACTE, to develop a Student Teacher Observation Tool. And so, they're going to tell us about the process they used to develop the tool and how they examined its reliability and validity.

All right, in the next slide, I'm going to talk about the webinar objectives. These are our objectives for today. The webinar is really designed to familiarize you with the CAEP requirements for demonstrating reliability and validity of teacher candidate evaluation instruments and focusing on those developed by educator preparation programs. And then we'll also be talking about approaches for examining and supporting the reliability and validity of these instruments.

All right, next slide. This is just a set of resources that have informed the presentations today and some things that you may want to refer to later. There's a link to the slides available, I believe, in the chat box. And this is also something that is downloadable via the IES website. So just a quick overview of what we're featuring here. The CAEP Standards themselves are the first document that specify the expectations for CAEP accreditation.

Next is the CAEP evidence guide, which has guidance around the use of data and evidence in educator preparation and also in the accreditation process. And there's also guidance in there related to data quality, data collection, and data analysis–and specific guidance related to the topic we're talking about today, which is the reliability and validity of candidate evaluation instruments.

Next on the list is a report about a teacher evaluation rubric that's used in Texas. And, like the candidate evaluation tools that we're discussing today, this is a rubric-based evaluation tool. And the report that we've linked here illustrates how evidence of reliability and validity may be generated.

And then, the last two bullets on this slide list a couple of resources that were developed by our presenters from North Dakota. These are reports of studies that have been used to examine the reliability and validity of the Student Teacher Observation Tool, which we'll hear about in a little bit later. All right.

So next, we'd just like to take a couple of minutes to open up the Q&A box to hear from you all—this question. So we're interested, just to kind of kick off our thinking about this discussion, in why you all think reliability and validity are important when evaluating teacher candidates?

So, I invite you all to use the Q&A box to weigh in with a response to this question. And we'll take a couple minutes—maybe about a minute or so—to allow you to do so. So again, if you hover at the bottom of your screen, a black bar will appear and you'll have the opportunity to click on the Q&A box and your question.

OK, great. So we've got a few thoughts here. So one response—let me just kind of convey what people have said. So, "Why is reliability and validity important when evaluating teacher candidates?" So one is for quality assurance to know that we're assessing the right things. So that's the idea of validity really is that you're sort of measuring what you intend to measure.

It's important to ensure that we have a clear understanding of how we define "ready to teach" and ensure that we're measuring that and only that. So again, like focusing the content of what you're measuring to be what you intend to measure. Validity is a potentially important predictor of effectiveness after graduation.

Reliability and validity help ensure that everyone is measured the same way. And then knowing it's information that helps know that you're measuring candidates consistently. It helps identify teachers who are most likely to be actually effective in the classroom.

So those are just some of the comments we're getting, I think that are all in the spirit of how we collectively think about reliability and validity. So, equity is another one. Making sure the instrument is measuring the construct accurately. So I think I'm seeing that we're all very much on the same page about where this comes from.

When we hear from Gary Railsback from CAEP, we're going to hear a lot of these priorities echoed in his presentation. And I think that's where, in any measurement, you know, we're trying to make sure the measures are doing what we set out for them to do.

So with that, I think I will turn this over to Gary from CAEP. And as I said, he's going to talk about the guidelines for establishing the reliability and validity of assessments created by educator prep programs and some of the challenges they faced. And we'll be hitting more on the importance of these topics in his talk as well. So let me turn that over to Gary. And Gary Railsback from CAEP.

GARY RAILSBACK: So thank you all for participating, and especially to Marc and Steve for setting this up. Validity and reliability are really important as EPPs (educator preparation programs) make recommendations for states for who's going to get a credential and who's going to be in a classroom.

So I just wanted to start off with some general reminders. When EPPs start working on their self-study, we ask them to think about these kind of assessments–performance assessment–in really two broad categories. One is a proprietary assessment. So that would be one that's been developed by somebody else. Maybe it was purchased by an individual or the state. And the validity and reliability should have already been done on that, but we don't ask people to provide that. The self-study just says "if available."

So what we're going to focus on today are really the larger category, where you feel like those proprietary assessments don't cover all of the standards and you want to do some things that are customized for your campus. They're usually developed by faculty. So we've got some examples here of proprietary edTPA, Danielson Framework. And we're going to spend most of our time on the latter category—EPP-created assessments.

So I get lots of calls about this, and people want to know what is the magic number. There is no minimum. There's not really a maximum, but having been a site visitor, I would keep that number single-digit. A lot of EPPs only have two to three. So especially on the initial programs, we're not looking for you to develop, like the old NDACTE policies, which said six to eight.

The ones that are most common are a clinical practice observation, sometimes dispositions, possibly a unit plan. It could be a portfolio. Those are the ones that are really most common. Starting fall of '19, we're also going to be moving into doing site visits for advanced-level programs. And if you're not aware of that yet, I'd encourage you to look at our website to see if your programs actually fall into the scope.

But essentially what people are doing at the advanced level is they're developing some instruments that might be related to Standard A.1.1, which are the six broad professional skills, maybe content-specific. And then if it's in a full advanced program and it has clinical practice, maybe something for there. They could also be doing something in dispositions.

So, as we think about validity, we want to make sure that whatever assessment we're using is actually measuring what it's supposed to measure—that it's not measuring something else and then we're using it for some other purpose. So I did cite here out of our handbook, our definition. So "it's an operation, a test, or some other kind of measure that we're really sure that's what it's measuring."

And so, to help people think about that, CAEP developed an instrument that's called the CAEP Evaluation Framework. Sometimes, it's called the CAEP Evaluation Tool. Sometimes, it has a whole bunch of words in the title beyond that, but essentially we're going to go through that and kind of look at what CAEP is looking for.

The way the rubric is put, the left column is below sufficiency, the middle is sufficient, and the right is for the eager beavers that want to get the highest grade, but it's not required. So that's above sufficiency. So, when we look at the CAEP Evaluation Framework, there's seven different sections of it. And section four and five are the two we're going to look at for validity and reliability.

So, for validity below the sufficient, we're finding places that there was no plan at all to establish any kind of validity. And we accept multiple kinds of validity, and this will be talked about later in the presentation as North Dakota gives their example. But we need to have some validity. Is it content validity? Whatever. We want to make sure that the instrument before it was actually used was piloted.

And we want to make sure that it was done by more than one or two people. So the common way this used to be done is two or three faculty teaching in an area might develop an instrument, and then they just run with it. And so we're trying to move people to the middle of the sufficiency and involving some external stakeholders. And then the last one is to develop some steps to do this research. And if it's on the below sufficiency, there wasn't evidence of that in the site—the self-study, or any other materials given.

So now, as we move toward the middle of the column here in the sufficient, we find that there was a plan. We know how the EPP developed validity. We know what kind of validity it's looking at. Most people do content validity, but that's not a requirement. If it was new or revised, there was some kind of a pilot, and that's clearly delineated in the description. The EPP details its current process or plans for analyzing and interpreting the results. And then whatever they did follows generally accepted research methods.

Now in the right column, some people look at that and think, well, that's what I have to get. And that's just saying this is like way beyond or above, so you don't have to do these things. One of them is reporting the validity coefficient using certain types of validity, which are a little

more complex, like predictive validity. But again, those are above sufficiency and they're not required. OK, so we'd like to take a question here.

R. MARC BRODERSEN: All right, yeah. So we're going to open up the Q&A box again and so we'd like to hear from you. What do you think are some challenges you think EPPs face when developing teacher candidate evaluation instruments and examining the reliability and validity? OK, so we're starting to get a few responses here.

So some of the challenges that are being raised here—one common—is that time constraints can be a constraint.

GARY RAILSBACK: Absolutely.

R. MARC BRODERSEN: We have another as having sufficient number of faculty to help with the process. Another challenge might be to define the constructs that are to be measured. And I think related to that, we have another comment here about measurement without the content and maybe how those two can be aligned. We have another one that's funding and time for field testing.

And I think we'll be addressing a lot of these issues as we move forward in the presentation. How about we'll do this last one here. "Getting stakeholders from the public schools involved might be a challenge." So thank you everyone for your comments. And we're going to hear some more from Gary.

GARY RAILSBACK: All righty. So what we're going to go through now are some examples of EPP-created assessments that have been used in the accreditation process on both initial and advanced. So again, we're looking at that evaluation framework. And on the left side on below sufficiency, one of the things that we look at is curricular validity. It "refers to the extent to which the content of the assessment matches the objective of a specific program."

So you know, as a site reviewer, I saw a lot of places trying to give course grades. And there's just not a lot of strength in a course grade. So I would encourage people not to use course grades. Even if it's a specific course—that's somewhat problematic. End-of-course assessments might be a portfolio, it might be a term paper, or whatever, but there's no discussion about validity and reliability. So those are some examples that come out on the below sufficiency.

Some other examples—face validity. It might be a dispositional data. A lot of EPPs develop a dispositions instrument where they ask candidates, or faculty, or the clinical faculty out in the schools to rate a candidate. And it might be very qualitative or it might have no analysis. Another one is candidate interviews, which is a great thing to do. But they're often done without an instrument and there's no analysis. Same thing for portfolios.

In the middle column, the ones that usually reach the sufficiency. Most people are doing content validities to make sure that whatever is in your rubric that you're assigning and you're using for assessment. And the common ones—lesson plans. A lot of teacher ed programs have

a unit plan, teacher work samples, portfolio assessments, observation, capstone thesis. Or the better one would be a problem-based project that would be in partnership with a school district.

Over on the right side again, this is above sufficiency—so not required, but it's predicting. And that might be doing something in preservice, trying to predict whether that person is going to be a qualified teacher a year or two down the road. And then comparisons of candidates in education programs with either other higher ed institutions, or it could be even across campus if it's a content course.

So for us, validity is that it can be supported through evidence of the following. There's some agreement among reviewers that the description or the narrative about it is measuring the same thing. There is expert validation of performance, or looking at artifacts, expert validation of the items in an assessment or rating form, and then the measure's ability to predict performance in a future, and that's predictive, which is on the right side.

So there's a number of different ways that EPPs are going about developing content validity for some of these common instruments that we've talked about. My first couple of weeks at CAEP, I got some interesting emails from people. And one email said, does CAEP require the Lawshe method? And I said, "no." And then a week or so later, I got another one. And somebody said "CAEP does not allow you to use the Lawshe method."

And I don't know where these rumors started. We have used the Lawshe method at our CAEP conferences and a lot of other places, but we don't require you to use a certain kind of methodology. The Lawshe method has been used a lot because it asks you to go out and have external stakeholders evaluating whether this is beneficial, but it's really not. In the Lawshe article, the little equation there—if algebra scares you, that's just the equation that's used to conduct the content validity ratio.

So, and I actually did this at Azusa Pacific before I left last year. You take some P-12-based educators. If it's content, maybe faculty members, P-12 administrators, leaders, candidates, partners, parent advisory boards. In other words, you're getting people that are familiar with what you're doing and you're asking them.

And then you can do this in a survey. You can send it out electronically. And you're just asking people, is this task that we're asking somebody to do, is it essential for whatever the task is of the job—classroom teaching, school counselor, school administrator, whatever. Is it essential, is it useful, but not essential, or not necessary?

And then you just determine how many people that you asked. And if you have a lot of panelists, anything beyond 50 percent is most likely going to get you a content validity ratio. So you don't have to go out and hire a psychometrician or have somebody from your psych department do this. It's just going out and working with your advisors, your advisory council, or whatever that might be.

So our definition of reliability is "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement and are inferred to be dependable and repeatable for an individual test taker." Now, if EPPs are developing two to three assessments, one person may not take it more than once. So we're not saying that you have to do it, but as you look at it over the three cycles, you might be able to tell a little bit about that.

The next slide talks about reliability. And so again, we're going back to the CAEP Evaluation Framework. On the left side, what makes it below sufficiency is the "plan to establish reliability does not inform reviewers whether reliability is being investigated or how." So the reviewers don't really know what was going on. The "described steps do not meet accepted research standards for reliability or that there's actually just no evidence."

What most people are doing is, when we go to the middle column, is they're doing inter-rater reliability. So let's say it's an observation instrument. They're having a candidate observed by a classroom teacher and a faculty member. And they're looking at those two scores to determine what the inter-rater reliability is. It can be supported by multiple raters of the same event. And it doesn't have to be live. One of them could have been on video. We're looking for stability or consistency of ratings over time and evidence of internal consistency of those measures.

So here's a little example that we put together to kind of think about this supporting the reliability. First of all, you start with content expert—some kind of feedback on that instrument. Then you get some feedback. You clarify it. You may change it a little bit. You get content experts, developers of the measure to review and seek feedback. Then you pilot and you examine that data.

So common challenges EPPs' staff face in meeting the CAEP Standards. So if you look at the CAEP Standards, the component we're looking at is 5.2, where it asks that your instruments are valid and reliable. So most of the time, as I've said, there's going to be two or three. And they might be for Standard 1 about content or pedagogy. They might be in Standard 2 about something about clinical practice. They might be in Standard 3 about dispositions or something that's non-academic.

And so what our teams do is they look at all those instruments—two or three of them. And then they say, do the majority of them—the new handbook starting fall of '18 moved to 75 percent—do 75 percent of them meet the sufficiency criteria? So if you go back to my slides in that middle section, that's really what we're looking for.

And what we find in most cases is they haven't done content validity, or it was done by two or three people, or we don't really know how they did it. And as far as the inter-rater reliability, which is the most common method, we don't find much evidence of that. And that's a change from both NCATE and TEAC to actually say, you're making a consequential decision. How can you demonstrate that? The other thing I'd say about Standard 5 of the five standards, it's the one that most institutions had trouble meeting.

OK, I'm going to turn it over to the North Dakota group.

STEPHEN MEYER: Great. Thanks so much, Gary. So I think we'd like to take a few minutes here to see if folks have questions related to Gary's presentation before we move on to North Dakota. So again, I'll invite you all to use the Q&A box to submit some questions.

So one question, Gary, was around the last thing that you said about Standard 5. Apparently, there was an audio problem for at least one person.

GARY RAILSBACK: OK. So Standard 5 is the quality assurance standard that ask essentially for an EPP to have partners—stakeholders. You define what that is. And having them involved with creating your instruments, evaluating, looking at data, and then using that for program improvement in the future. So that kind of cycle that the regional accreditors are also pushing EPPs to work toward—that's the one that more institutions have gotten either stipulations with or on probation because they don't fully meet that standard. And the big part of that is component 5.2, which looks at validity and reliability of the instruments that you're using in your quality assurance system.

STEPHEN MEYER: OK, great. Thanks. Another question is around whether there is an expectation that assessments used for operational assessment will need to have reliability or validity?

GARY RAILSBACK: So operations like on the unit?

STEPHEN MEYER: That's a good question. Maybe I'll ask the questioner to clarify that question.

GARY RAILSBACK: Yeah, I need some more clarification on that.

STEPHEN MEYER: OK, so another question in the meantime. How is CAEP building consistency among chairs and teams to look for and rate validity and reliability?

GARY RAILSBACK: Well, that's a very good question. So I've been working with training volunteers the last two summers. And we're now doing it regionally. So I did one in West Virginia, and one in Tennessee this fall, and in a couple of weeks, I'm going to Mississippi. And so that's what we're working with new trainees.

The CAEP Evaluation Framework has been out there, I think, since about 2015. So it's not really new. But getting those 500 volunteers that were trained several years ago back in, we haven't had a chance to do that yet. And that's why we do a lot of webinars. We try to get people to go to CAEPCon.

But even if it doesn't happen at the team level, it goes to the accreditation council. And I've also been doing the training for them to make sure that there is inter-rater reliability between the individuals and the different panels. And that's really where the major decision is made.

Again, we're looking for meeting all—if you look at that evaluation framework, there are five sections on a performance assessment, and then Section 6 and 7 are about surveys. So if you

just look at the first five, we're looking for a summary statement that the EPP-created assessments are at that essential category. It's not, well, you got one wrong or you got two wrong. It's a summary statement.

STEPHEN MEYER: Great. OK, thanks, Gary. Another question is around what is the expected inter-rate reliability coefficient? Can you say a little more about that?

GARY RAILSBACK: Yeah. So that actually hasn't been in the evaluation framework, but it is in the handbook. And the suggestion is 0.8, but it's one of five factors. So it's not like if you don't have 0.8 then it's going to sink the whole ship. That's one of five factors.

STEPHEN MEYER: Great. OK. And I think just one last question. So this is an individual who says their program has three distinct pathways that use three different clinical field evaluations. Would they be expected to come together and create one evaluation tool if they plan to use the clinical experience for a tool for validity and reliability?

GARY RAILSBACK: That's a good question. So if you're talking about an initial program that has three pathways—and I'm certainly familiar with that—that's a choice you need to make if you want to have three different instruments. What CAEP says is any performance assessment you're using to make the case your program meets the CAEP Standards needs to be valid and reliable. So you can either do each one of those three and make sure they're all valid and reliable or you could take the best of them, put them together in one, and make sure that it was. But that would be a campus choice. You decide if those programs are really different, then that's a decision that you get to make.

STEPHEN MEYER: Great. All right, thanks. Thanks, Gary. And thanks, everyone, for submitting questions. We are going to shift over to our presenters from North Dakota—Dr. Stacy Anderson—or sorry, Sarah Anderson, Stacy Duffield, and Alan Olson. They will be talking about a candidate evaluation tool that they collaboratively developed called the Student Teacher Observation Tool.

And before we get into their conversation, we're going to take one more moment to have a quick Q&A to just kind of kick off this conversation with a discussion of "What do you think are the pros and cons of educator preparation programs staff developing their own instruments versus purchasing or adapting an existing instrument?"

So as [AUDIO OUT]—that are [AUDIO OUT] administered, some states have tools that are administered statewide. But I just wanted to—kind of before we get into this conversation of North Dakota's work—to think collectively about what some of the pros and cons are of doing local development of an instrument versus adapting something that's already out there. So please consult your Q&A box once more and share your thoughts.

OK, so one comment is that a locally-developed educator prep program developed instrument can be tailored to address a local mission. So there may be local priorities or emphases that you can better reflect on a locally-developed tool. It's more customized. One advantage is that

there are more opportunities to test the reliability and validity in a local context. So that's certainly an important aspect is that you can kind of demonstrate reliability and validity for the candidates who you're serving. And if you have a locally-developed tool, the opportunity to do that may be better. There is also the potential of less costs with a locally-developed tool.

A downside—here, we have a comment—is that you have to actually go through the process yourself of evaluating the tool's reliability and validity. You know, you can't draw upon some existing study that may be out there for a widely-available tool. Of course, you can tailor the tool to exactly what you want to measure.

OK, a couple other comments. You know, one advantage, which we're going to hear about from our North Dakota colleagues, is that a locally-developed instrument may give you the opportunity to partner with other institutions in your locale to put together something that kind of makes sense for the candidates who you serve in a particular area. So if you're looking to meet standards for a given locale, like a state or a district, and you're trying to meet certain priorities, you have the opportunity to partner with other folks. And there are other advantages of partnership that we'll hear about.

So maybe I'll just mention one more. One of the cons of adopting a widely-administered instrument like a commercial instrument is it may not align particularly well with your state standards or your program values. So thanks a lot for sharing all those things. There are lots of factors to weigh in making this decision about developing something locally versus, you know, adopting something that's already out there.

So I'm excited to turn over control to our North Dakota colleagues. Alan Olsen is going to start out this conversation. And they're going to talk about their work in developing the Student Teacher Observation Tool, and also the work they've done to demonstrate its reliability and validity. So over to you, Alan.

ALAN OLSON: As representatives from the North Dakota Association of Colleges for Teacher Education, we'd like to share our experiences about creating a valid and reliable assessment instrument. My name is Dr. Alan Olson from Valley City State University. I'll tell you about some of the benefits and challenges, pros and cons that we experienced while we developed an EPP-created assessment, and a little bit about our collaborative process.

Dr. Sarah Anderson, from Mayville State University, will discuss the process that we used for developing the instrument. And then Dr. Stacy Duffield, from North Dakota State University, will explain the NDACTE efforts to establish reliability and validity of the assessment instrument.

Our experience began when North Dakota EPPs were transitioning from NDACTE to CAEP. And we had a common need. We had assessments, but they didn't all have actionable descriptors and they weren't quite at the depth we needed to have. And so the need was to develop valid, reliable, practical assessments in order to gather data that we could analyze for program

improvement and meet our accreditation standards. We wanted quality assessments, so we could get quality data. And we all needed to make those types of improvements.

So the idea for collaboration came about because it was mutually beneficial for all of us. We all had a similar need and we could share expertise among our colleagues, across the campuses to fulfill that need. We're going to talk about sharing among the institutions, but even within our own institution then we have to share on our own campuses and find colleagues with different talents and different expertise to help us if it's only our EPP. But in this case, we'll be talking a little bit more about collaborating across institutions.

The concept of having common assessments is something that we came about because we wanted to meet the CAEP sufficient levels. And we were used to the work that was being done by the Network for Excellence in Teaching Efforts. It was funded by the Bush Foundation in Minnesota. And our State of North Dakota received permission from the NExT—this Network for Excellence in Teaching—for their exit survey, completer survey, and employer surveys.

And we liked this idea of some common assessments that institutions didn't have to use in our state, but could use in our state in order to have valid and reliable instruments. And we believed that we could develop a student teaching instrument that would meet the CAEP sufficient levels and become a quality assessment that each of us could use. We pursued a grant from the AACTE, and that was the beginning of our collaborative process.

Some of the benefits and challenges that we found in collaborating with each other. Some of the benefits—getting a variety of perspectives really was helpful to us. And often, we had ideas from multiple stakeholders that ended up making a difference for us. And the idea that one of us can have an idea but with the help of others, sometimes that idea can be much greater than anything we'd have thought of individually on our own. We involved our stakeholders to increase the potential of skill and expertise for research, or assessment, or field experience that could help us.

And also it gave us a common language. When we were talking to each other across universities. We had the same assessment, and so we had the same types of data, so we could talk apples and apples with each other—compared to the days when we all had our own separate student teaching instruments. And even the same topic of assessment, the questions might be different, or planning lessons, the questions might be different.

And now, we had a common language, common assessments that we could share with each other. It improved our communication and networking for future collaboration. And it also—we talked about having mutually beneficial needs and how this was going to help us. And it did help all of us feel the need for what we needed to do for assessment.

It also helped us have some common assessments, so that if somebody was going to have—a cooperating teacher was going to have a student teacher from Valley City one year, and Mayville the next year, and NDSU the next year, and the University of North Dakota the next year, they'd still have the same instrument. And principals could be used to seeing our similar

completer surveys, employer surveys. And our cooperating teachers could be used to similar student teaching instruments. And that was helpful for our cooperating teachers and our state as well. And then the collaboration of using—working with other partners helped share resources and expertise as well.

Some of the challenges would be that it does take more time and effort to get together, figure out schedules, and get to meet is a challenge. And there has to be some concessions. So some loss of autonomy was a challenge for collaboration.

Some of our pros and cons of having an EPP...created assessment, and the ones that were discussed that Steve was able to gather in the chat—they were excellent—about tailoring the instrument to be more exact, or having a local instrument to meet a local mission, or being able to be more customized—those were all great points that people mentioned.

For us, we talked about how we worked with standards, to begin with, but then we could proceed with a little bit more freedom, like the InTASC (Interstate Teacher Assessment and Support Consortium) wording. That is pretty lengthy. And all the InTASC items gets to be pretty lengthy. And we had a little more freedom and autonomy to make our own decisions about how we would use those. And then a greater opportunity to develop assessments that were practical and meaningful to our EPPs and our state.

And items could be aligned or complement other assessment instruments. We had an exit survey, a completer survey, an employer survey that we liked. We wanted to have some alignment with not only InTASC, but with our other surveys so we could look across instruments and across perspectives of cooperating teachers, student teachers, first-year teachers, and employers of our first-year teachers.

And then also an assessment that can be validated to the population using the instrument. And I think the other comments that came in from people on the chat were excellent.

Some of the cons—increased time and effort, more responsibility for the process, and some of us would lack confidence in our validity or reliability statistical analysis type things. And that's where we need to seek some expertise, or as Gary mentioned, something like the Lawshe method isn't quite as complicated to find ways that we can find validity and reliability that maybe aren't quite as complicated or to seek some expertise from some partners to help us. And then the creation process can cost money and take time.

We've been able to share about our student teaching assessment instrument at some different NDACTE and CAEP conferences before and at the state level. And we've had some people from other states have interest in our assessment tool. And this is an idea of some of the other states that people from different institutions have shared some interest in the observation tool that we created to be valid and reliable for use in our state.

SARAH K. ANDERSON: All right, thank you, Al (Alan). So this is Sarah Anderson from Mayville State University chiming in. And I'm going to share a little bit more about our process and outline what that looks like.

So our decision to collaborate on the common performance evaluation was really prompted, as Al mentioned, by that chapter support grant from the American College Association of Colleges for Teacher Education. We were awarded that chapter grant way back in June of 2015. And subsequently at our state chapter monthly meeting that September, representatives from our NDACTE organization then made the collaborative decision to select what would become our fourth statewide common metric.

So our discussion then led to the selection of our preservice teacher performance measure that would again accompany our exit survey transition to teaching, which is administered one year postgraduation, as well as our employer surveys. So this decision then really became a part of the discussion that was occurring across a number of our educator preparation programs with their school partners. As about 5 percent of our local EPPs were in conversation with three of our largest school districts in our state about how could we simplify processes, how could we make field experience and student teaching placements more of a common process. How could we identify what some of those best practices were and really make it easier for our cooperating teachers that are working with teacher candidates coming in from a variety of institutions? That really made this work very timely when we were thinking about our partner schools and what their needs were as well.

So I volunteered to lead that work along with six of my committee colleagues from NDACTE. We ensured that we had representation on that subcommittee across the different types of institutions in our state that included research institutions, regional, private, as well as our tribal teacher preparation programs.

So with that, we were able to take a look at that purpose statement of why were we doing this in the first place. And Al explained that very well, that we did have this need. It was evident. We had a very clear statement from that CAEP sufficiency rubric that told us that we need to look at our validity and reliability in a very different way. And so that purpose was pretty clear at the beginning of what our end goal needed to look like.

So the steps then to start gathering the information from the tools really first began with taking a look at what was already in place. We, at the 12 universities, did have all of our own performance assessments, which did serve the purpose of evaluating preservice teaching skills. Each of our tools did have alignment, to begin with, within task and our CAEP standards, but as Al mentioned, none of really which completely met those sufficiency requirements for validity and reliability.

So we really did start with some good basis for developing a common tool. Those tools, as we collected them from all of our partner institutions, had a variety of rating levels, whether that was from a one to four or one to three. We had both quantitative and qualitative formats that

were in place, again varying levels of actionability of the descriptors in those statements. And a number of elements for rating, ranging from very small numbers to very intensive numbers.

So as our institutions then were working through how are we going to move forward, not only did we take our own instruments and take a look at those, but we also took a look at our P12 partners to see what was happening in our schools. We really wanted to make sure that a preservice evaluation instrument was also aligned with what was happening with teacher supervision and evaluation in our state.

And so, across our institutions, we found that the most common ways teacher evaluation was being conducted was through the Marzano Teacher Evaluation Model, the Danielson Framework for Teaching, and the Marshall Teacher Evaluation Rubrics. So these were brought to our subcommittee also as part of our starting point for instrument development.

So the committee also then took a look at other literature, doing essentially a literature review, to see if there were other developed models beyond these ones that we could use also to inform our process. So, as we gathered each of these tools, we pulled them together, and then came back to our basis of our InTASC Standards, which when we took a look at the standard's progressions, actually had 166 individual constructs to measure the skills of teaching, which if we're going to put that into the hands of a cooperating teacher, is not exactly well-received in terms of their time, energy, and effort. And so we knew that we needed to make this feasible, also looking at who would be filling this out.

We also then specifically looked at the CAEP Standards and took a look at what the STOT would be used for as evidence within our programs. So, through that process, we acknowledged that we really would be addressing our candidate knowledge skills and dispositions of Standard 1.1. We would be collecting evidence for state and SPA reports in Standard 1.3. We would also be addressing our partnerships with our P12 schools as explained in CAEP Standard 2. We also really needed to investigate how the tool would provide evidence for standard three and selectivity during preparation and at completion. And so, we knew that we would be developing this tool with those end goals in mind.

We also wanted to ensure that was in the final product that we were able to address the crosscutting themes of diversity and technology. And so, the instrument was developed to make sure that we could address each of those items. So coming back to some of the steps, our subcommittee, then, as we pull together all of these resources, which was really in that development phase, we had to, of course, collaborate in-person.

That was one of our greatest assets was the ability to come together because of funds by that chapter grant from AACTE to do the work face to face. And that did have a huge taxation on resources in time for us, as we full know what a faculty's workload can be in a typical week or month. So we had about four work sessions where we were face to face three to four hours and one full day.

We had a number of web-based meetings and smaller work in pairs on each of those 10 impacts, standard, and the associated items on the thought. We had a number of times where drafts were sent out to committee members for review at each institution and, again, to touch base on expertise in certain areas. That brought us to a period of refinement.

So, as we move through with the refinement, the seventh draft was actually the first draft that we brought to the whole NDACTE statewide group for review and institutional feedback. We also, at that point, gathered some information from CAEP site visit reviewers. We also brought our seventh draft to the CAEP staff for some additional guidance and ensuring that we were on the right track.

And we ended up with 12 drafts before we were ready for an initial pilot, which I think speaks to the complexity of the task that we were given and also to the quality of the end result. So we really did consider at this point the use of the half-point scale, whether or not those would be included. Those were based on Marzano's Proficiency Scales. And we did take a look during some pilot exploratory factors if our cooperating teachers were choosing to use the half points. And they really did.

A large number of them chose to use those half points based on that Marzano Proficiency Scale. So we ended up with a seven-point scale. We had multiple discussions about the performance levels and the labels on what those scores would be—again, referring back to our 12 institutions. And, as Al mentioned, we had to make some concessions.

At our institutions, we might be using levels and labeled differently than the statewide tool to look at, and it was one of the pieces that we did have to compromise sometimes that our institutions on that we were changing those indicators. We also, at this point, consider the use of "does not language" within the actionable descriptors. And we ensure that the order was from the distinguished category or a Level four that was listed first next to the actual item.

You can also see in this example from InTASC Standard, one with our two elements that are being evaluated that we also made sure that we had very clear connection back to which InTASC Standard was being measured. That then led us to the next portion of our work, which was in our first pilot. And so, that first pilot was a voluntary participation of an exploratory factor analysis in May 2016.

We, again, leveraged our funding from the AACTE grant and North Dakota State University. And DSU conducted that analysis. And my colleague, Stacy, will describe little bit later on here the pilot and validation process, and what those results look like. From that first exploratory factor analysis, we look towards some instrument refinement and in what would end up being our final version.

And we took those results, the recommendations—we took those, especially, the double-barreled items—we had a lot of items that had multiple indicators being measured. We worked through those together. We looked at the language, and how they were loading on alternate factors to direct them more towards what they were intended to measure. That refinement

process really worked specifically with the verbiage of those actionable descriptors. And, again, a collective and collaborative process to reach that end goal.

We also at this point, with the first pilot had some stakeholder feedback through our Institutional Advisory Board through graduate students in the educational leadership program. And we also held a presentation at our department of Public Instruction State Conference where we could solicit feedback from across the state from administrators and teachers in the field. We brought that back to our NDACTE committee. And that resulted in six more drafts. Those recommendations continued to come in.

And pilot two, then, the second exploratory factor analysis occurred in December of 2016. This was completed with our 18th version of the stock. We administered that to a cooperating teachers working with student teachers from 11 of our 12 institutions. And results showed some pretty impressive commonalities, which we'll share with you in a few moments here.

So, from that second pilot, we ended up coming to a final version, which we know is draft 20. And this is what you will find available on the NDACTE website. And that is available for download at this time. And simultaneously, to implementing the 20th draft, which was the first to go statewide, we were also the recipient of a second AACTE chapter support grant, which would help fund the development of inter-rater reliability training modules, that again could be used statewide.

So the first time that the STOT went statewide in North Dakota was the 2017/18 academic year. We really wanted to ensure that there was institutional flexibility in the administration. What systems it would be put into to collect that data breach institution could be used with whatever system was in place. And again, how that institution chose to use data for decisions within their own programs was maintained.

So, for example, at Mayville State University, we operate with the Taskstream Operational System. And all of these items can be put into that system for us to use. We use it formidably during the clinical experience with acceptable targets for performance that are set by our own faculty. And we use that for making continuance decisions within our program for our candidates.

We also use it summatively during the student teaching experience. And we have that completed by the students for self-evaluation reflection, as well as by the cooperating teacher and the university supervisors so that we do have those multiple perspectives. So each institution is left with that flexibility of how they will use the STOT within their frames of reference.

So the Inter-rater Reliability Training Module is now available. So again, at the ndacte.org website, that was released in October of 2018. You also have that thought and validity information that was mentioned at the beginning. And you have links to at the start of this presentation. That inter-rater reliability training involves panels of experts from across three of our North Dakota institutions.

Again, some teaching videos that were used to practice evaluating and using the STOT and some training, as well as on bias, and what that looks like when ratings are being given. We also utilize some examples from student capstone portfolios of some of the items that were used for training. So, at this stage, we are actually working on the confirmatory factor analysis. We had enough changes between the first and the second pilot that the confirmation occurred during our first full use of statewide data.

So the process for doing that is that here at North Dakota State University the analysis is run. We take that information and results from our own institutional system and we place that into a common data template. And that data is then all processed together. So we have both an aggregate score from state of North Dakota and we also then have our institutional numbers, which for smaller regional institutions, this is one of the greatest advantages is that we can have a bigger N to work with in terms of doing some of the statistical work, as well.

So NDSU has really served as the lead research university conducting that work. And we are at the point now, as we will be seeing those results very shortly to do a second round of revision and refinement of that tool, as we ourselves engage in the continuous improvement process.

And the exciting part of all this work and the collaboration that's occurred amongst our institution is that this confirmatory analysis work of the STOT is actually occurring alongside our fifth statewide assessment, which is the disposition evaluation. And an exploratory factor analysis has been completed with that, and we are piloting that across the state this semester.

R. MARC BRODERSEN: All right, thank you so much for all that great information. Before we move forward and talk about the approaches that have been taken to look at the reliability and validity of the STOT, we want to open up a poll and get your thoughts about, can an instrument be unreliable but valid? So, if you have a chance, you should see the poll results up in front of you. And so, just take a moment, and then we will move forward.

OK, so they're starting to slow down. What we have so far is 51 percent say—of respondents say, "Yes, the instrument can be unreliable but valid;" 38 percent say, "No, it cannot be unreliable and valid;" and 11 percent say that "It depends." Can we go to the next slide, please? So the answer is no. For a test to be valid, it must be reliable.

So, as we discussed, there's different definitions or approaches for validity. But, basically, validity is, are you measuring what you think your measuring. Or are the conclusions based upon this data, the assumptions you're making about this, can those be supported? If a test is unreliable and that it generates different results when you use it, the items inside of it are not consistent with each other.

You do not know exactly what it is that you're measuring. It might be a measure of a teacher content knowledge in an area. But, if the test is unreliable, it's also picking up something else. So, generally speaking, a test must be reliable as a requirement for validity. So thank you everyone for your responses. And I think we're going to learn a little bit more about reliability and validity as it was addressed with the STOT.

STACY DUFFIELD: So I'm Stacy Duffield. I'm from North Dakota State University. And I'm going to talk just a little bit about the process that we underwent to ensure validity and reliability of the Student Teacher Observation Tool. And I want to repeat what Sarah said. Just kind of getting us started that we did have some grant support that enabled us to do a lot of this work that may not have been possible had we not had some of that funding support from AACTE for our chapter grant.

And we also had, through our experiences with the network for excellence in teaching, had a protocol and a great deal of experience in doing this work through the development and validation of the networks common metrics. And so we did have a bit of an advantage in network.

So I'm going to walk through the five types, the five things that we did for our validity and reliability process. And so I think when we look at this process, it's really important to start at the beginning of the STOT development process. And Sarah just went through a great deal of that. But I'm going to parallel that with the work we did with the validity and reliability work. We really follow it, I guess, the Wiggins and McTighe, *Understanding by Design Process*, where we began with the end in mind. And we knew that we wanted to end with a valid and reliable instrument.

And so with that, we recognized that we needed to be very mindful of what needed to happen along the way to ensure that we had a highquality instrument through that development process. It wasn't something we think about at the end. So we work toward that quality with three types ability that included content, face, and construct validity.

So, as Gary pointed out, we did exceed what CAEP would require for sufficiency. We also wanted to ensure full of internal and inter-rater reliability. And I'm just going to walk through kind of an overview, I guess, of how we did that work to ensure validity and reliability for the STOT.

So first, face validity—face validity doesn't mean that an instrument really measures what it's supposed to measure. But rather that in the judgment of the raters, it appears to do so. So it's a fairly low bar. But, at the same time, that's the place you need to start. So, while it's the least sophisticated, it's so important because it engages users and asks them if the instrument seems to meet its intended purpose.

We established face validity through that pilot and feedback process that Sarah described with the cooperating teachers and university supervisors. We also, as part of that, had added an open-ended feedback item to the pilot at the end. And we invited the university supervisors and cooperating teachers to give us qualitative feedback as they were filling it out.

So, just kind of in summary, while face validity is useful and it's important as a beginning part of the process when you're developing an instrument, it's just not enough. And so we went forward from there. So with content validity with our next type of validity, we looked at—and

while it's similar to face validity, it really does take a more formal approach and often uses some lower level statistics, like Gary talked about with the Lawshe method.

It does involve experts in the field. And those experts judge the questions on how well they cover the material or the content that you want to assess. We wanted to make sure that the STOT items actually measured what it was we wanted them to measure. And so, when I'm composing the items, as Sarah described, we began with the InTASC Standards, and those other measures that were being used in the field and we aligned to STOT, with previously validated instruments, the next surveys that were mentioned.

Content validity is often measured by relying on the knowledge of people who are familiar with the construct. And these subject matter experts are provided with access to the tool. And they're asked to provide feedback and we did those processes.

But then the next step is really important. And that was a formal systematic analysis of that feedback that we received. And it informed our decisions, as Sarah talked about, with the different versions that as the team went through, as we developed, what we would share with our stakeholders.

So with construct validity, again this does exceed what CAEP would require necessarily, but we knew we had the funding. We knew we had invested people in the process. And we knew that we did have the resources and expertise. And so we did want to make sure that we did get that internal validity.

So construct validity is "the extent to which the instrument captures a specific theoretical construct or trait." And we did this through a factor analysis. The initial exploratory factor analysis was then used for revision. But because the revisions that we did were fairly substantial, we weren't able to move forward with just a CFA or a confirmatory factor analysis. We really did need to run a completely separated exploratory factor analysis (EFA).

So in the end we've had, the instrument has undergone two exploratory factor analyses. And we're trying to get our hypothetical construct to wash out in the factors.

And I'm just going to tell you a little bit about the work that we did. We began with four hypothesized factors, as Sarah said, we use the InTASC Standards. Our hypothesized factors were the categories that InTASC uses, so the learner and learning, content knowledge, instructional practice, and professional responsibility. And we wrote our items with those constructs in mind, wanting in the end for them to bear out in the factor analysis. So the first internal factor analysis, the first EFA that we did, we only ended up with two factors.

So the first three constructs—learning and learning, content knowledge, and instructional practice—all putting together in one large unwieldy construct. And we had a second factor that did represent the professional responsibility. And so, as Sarah said, there was a pretty lengthy revision process. And we want back to work on the items and found that we had some double barreling. We had some unclear wording. We found that we were using words that were probably making people think about instructional strategies instead of, perhaps, content

knowledge. And so we just really went back to work and thought about how the wording was impacting the way people were interpreting the items.

And we do have those validity and reliability analysis for view out on the NDACTE website. So if you are one of those statistical people that would really like to read about the oblique rotation–oblique solution, and the KMO (Kaiser-Meyer-Olkin) and all of those things, you are very welcome to go out and take a look at that. But I won't go that deep for this. But you can take a look at the chart that we have on the screen and you can see that we actually ended up with some really nice numbers, with 0.35 kind of being the threshold. You can see that we're well above that.

All right, so the next slide. The second pilot exploratory factor analysis we had taken the assessment instrument down to 34 items, which we thought was pretty manageable for a cooperating teacher. We also did some work with timing, like how long would it take a cooperating teacher to complete this, and really thought carefully about wanting to make sure that we got—everybody was completing the instrument, and that we were getting good data. And they weren't just rushing through.

And so we were pretty mindful about all those sorts of things as we were creating the instrument. What you're seeing on the screen right now are example results from the second pilot. And in addition to the more sophisticated statistical analysis doing the exploratory factor analysis, we also just did some basics with frequencies with the responses. I'm looking at means and standard deviation because it helped us to see things—like, was our scale working?

Were people able to use those half-point items? What did they think of those? And in some of our feedback, and some of them are qualitative data that we gathered along the way, we heard from those cooperating teachers and university supervisors that those half points were critically important. And what we found is that we were getting more accurate ratings, because when someone was between scores—say, it was a little bit better than a proficient level, but they weren't quite distinguished—people were choosing the top rating because they wanted to acknowledge that.

And so we were ending up with inflated ratings. And when we were able to give that half point that—above a proficient could be recognized without going to the next category and creating those inflated ratings. And so you can kind of see here how they broke down. You can see, especially with a 3.5 and a 2.5, those half ratings were heavily used.

And so then looking at reliability, what was just pointed out in the question before this section, you have to have reliability if you're going to have validity. And so we did run an internal reliability using a Cronbach's Alpha, which is probably something that is pretty familiar to most of us. And we did get very good numbers. You can see that they're all 0.9 and above.

And in some cases, what you'll hear is that above 0.9 is too high. But that's relative. And it's in context. And so we went back to look. And it told us that we may have some repetitiveness in the instrument. But going back and looking, we really found that we had trimmed it down to 34

items. And that every item we had in there really was necessary. And so we didn't feel that looking at the items, themselves, that we were worried about being above 0.9. We were happy with the results and felt we needed to keep the items that were there.

And then finally, Sarah talked a little bit about inter-rater reliability. And we know that you can have a great instrument. It could have incredible internal validity and reliability. But if it isn't being used properly, and you don't have inter-rater reliability, your data isn't worth much. So your interpretations aren't going to be high quality. And so it was very important to us to take that next step and work on inter-rater reliability. And we have the second chapter grant and were able to work with the expert panels. So we found videos that were teachers teaching—kind of everyday teachers from classrooms.

And we would just use short excerpt to focus the viewing on particular indicators, making accuracy of the reading more likely. We used expert panels. And we set a minimum of at least five members in each panel. And the person that was on the expert panel needed to be an expert at the level of the classroom featured in the video, and for secondary, needed to have some expertise in that content area, as well.

We used a rating process. And in fact, this whole process, we really need to give a lot of credit to Erica Brownstein from Ohio State and the work that she did. We really borrowed heavily from the process she used. We spoke to her. We went to her sessions at different conferences and really found that useful. And the rating process that we used, first, we had the expert panel rate independently.

So they watched the video. And they rated without any influence from anyone else, and then determined what evidence they were using for that rating. And so we had them jot down notes on a recorded document. They're thinking around why they gave that rating. Then the group spoke together about the ratings and presented their rationale to each other. And after that conversation, then the raters re-rated the video and arrived at a consensus rating based on the evidence.

And so that became the foundation for the training modules that you'll see out on our NDACTE website. And you'll see that you have the expert rating panel. And we have the videos in there. And then along with what Sarah described at the beginning, just some basic training on removing bias from their rating.

And so finally, I just wanted to share this slide. This is just a sample of the independent rating and how it was collected. In this particular case with this expert panel, we used *Qualtrics*, a survey platform and loaded the video up on a link. And the independent raters sat by themselves before coming to the meeting. And they did their independent rating and entered their feedback. And so when they gave different levels of feedback, they entered their evidence so that we could see that beforehand.

And it was all recorded and documented so that we have that kind of systematic and formal approach to doing this inter-rater reliability training. And then we met together to do the

consensus work, and we had a facilitator for each of those groups that did the documentation and collected that evidence. So we have all of that very much like a research project, to be honest.

And that ends the end of what I want to talk about. But if you have any more questions because it was just a survey, you can absolutely reach out to us or ask questions right now and we'll try to do more clarification or add more detail. And we just thank you for this opportunity to share our experience and, perhaps, give you some ideas for some of the work you might want to do on your own campuses and with collaborators across your state.

R. MARC BRODERSEN: All right, well, thank you so much. It's all very interesting. So we are near the end of our webinar. However, we do want to open this up if there is any remaining questions. We're trying to address those as we could, but we received a lot. So it's hard. We can't address every single one.

But if there's anything that has not been addressed or you have questions for any of our presenters, please use the Q&A box. And we'll try to address a few before we have to close out. So I have one question here is, "Have you considered doing a consequential validity study?"

STACY DUFFIELD: We have not yet, actually. Next on the docket is we want to do with some predictive validity, and probably with the next survey instruments. But we have not put that on our list of things to do yet. But if you use the STOT, and you want to partner with us, shoot us an email and we would definitely be interested.

R. MARC BRODERSEN: And I think that addresses another question. We have one person that was asking if this instrument is available for other institutions to use?

SARAH K. ANDERSON: Yes, it is available on the ndacte.org website. And we just ask that you give copyright back to NDACTE and that is indicated on the instrument itself. So it is available and ready for use. Again, we just appreciate hearing if you are using it so that we can begin to kind of keep track of where it is being implemented.

R. MARC BRODERSEN: And I think you might have already addressed this, but another question we have is, "You mentioned student teachers use the STOT to self-evaluate as a learning tool. Have you, or you considering looking at validity studies for those self-evaluations?"

STACY DUFFIELD: We have not yet. We started out with just the cooperating teachers and university supervisors. But again, that's certainly something that we can put on our to-do list. And if Joshua wants to partner, we are very interested in doing this kind of work.

R. MARC BRODERSEN: Great.

SARAH K. ANDERSON: And at the institutional level, a lot of what we're working with is the STOT is the formative process. And so the self-evaluation component for a student–we do use

that data to take a look at how it relates to the university supervisor as well as the cooperating teacher's rating.

But the final ratings in terms of continuance in the program and what we're really looking for in terms of providing data on our completion for our candidates is really on the supervisor, the cooperating teacher, not the student. Which is why we brought that up as a reflective opportunity for them and to get to them to make comparison and then set some growth goals, as well.

R. MARC BRODERSEN: OK, maybe one more here. "Can the STOT be used for the first and third years of full-time teaching at a self-evaluation and supervisor evaluation by the building principal for measuring student impact?"

STACY DUFFIELD: There are a couple questions in there. So it's not on the InTASC Standards, which are professional teaching standards not just for pre-service but for in-service, as well. So the intention is there for the instrument to be used for teachers. We have not validated it on in-service teachers. But I know that there are institutions in North Dakota that are using it for their CAEP Standard 4 with their in-service teachers. And we need to get enough numbers. You really want 100 or more if you want to do a good internal factor analysis. And so we're just waiting for some of those numbers.

And then the second question is about measuring student impact. It really isn't that sort of an instrument. It's an observation tool. And it doesn't measure student learning...directly.

R. MARC BRODERSEN: All right, well thank you very much for that. I think with that, we'll come to an end. So first, I want to give a big thank you to Gary, to Stacy, Sarah, and Alan. Thank you so much. It's been a pleasure working with you all to prepare for this presentation. So with that, thank you all so much for joining us. And we really appreciate your time.

This agenda was prepared under Contract ED-IES-17-C-0005 by Regional Educational Laboratory Central, administered by Marzano Research. The content does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.