

# Examining the Reliability and Validity of Teacher Candidate Evaluation Instruments

February 13, 2019



# Introduction

Examining the Reliability and Validity of Teacher Candidate Evaluation  
Instruments

# Presenters

---

- Council for the Accreditation of Teacher Preparation (CAEP), <http://www.caepnet.org/>
  - Gary Railsback, Vice President, [gary.railsback@caepnet.org](mailto:gary.railsback@caepnet.org)
- North Dakota Association for Colleges of Teacher Education (NDACTE), <http://ndacte.org/>
  - Sarah Anderson, Associate Professor of Education, Mayville State University, [sarah.anderson2@mayvillestate.edu](mailto:sarah.anderson2@mayvillestate.edu)
  - Stacy Duffield, Professor, Teacher Education, North Dakota State University, [stacy.duffield@ndsu.edu](mailto:stacy.duffield@ndsu.edu)
  - Alan Olson, Professor, School of Education and Graduate Studies, Valley City State University, [al.olson@vcsu.edu](mailto:al.olson@vcsu.edu)

# Webinar Objectives

---

- By the end of this webinar, participants will have learned the following:
  - The CAEP requirements for demonstrating the reliability and validity of teacher candidate evaluation instruments.
  - Approaches for examining and supporting the reliability and validity of teacher candidate evaluation instruments.

# Webinar Resources

---

- Links for resources were provided in the registration email.
  - Council for the Accreditation of Educator Preparation. (2013). *2013 CAEP Standards*. Washington, DC: Author. Retrieved from <http://caepnet.org/~media/Files/caep/standards/caep-standards-one-pager-061716.pdf?la=en>
  - Council for the Accreditation of Educator Preparation. (2015). *CAEP evidence guide*. Washington, DC: Author. Retrieved from <http://caepnet.org/~media/Files/caep/knowledge-center/caep-evidence-guide.pdf?la=en>
  - Lazarev, V., Newman, D., Nguyen, T., Lin, L., & Zacamy, J. (2017). *The Texas Teacher Evaluation and Support System rubric: Properties and association with school characteristics* (REL 2018–274). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from <https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=4478>
  - North Dakota Association of Colleges for Teacher Education. (2016). *Validation study for the student teacher observation tool*. Retrieved from <http://ndacte.org/wp-content/uploads/2017/03/STOT-Validation-2016.pdf>
  - North Dakota Association of Colleges for Teacher Education. (2017). *Validation study for the student teacher observation tool*. Retrieved from <http://ndacte.org/wp-content/uploads/2017/07/STOT-Validity-Summary-June-2017-1.pdf>

# Q & A

---

- Tell us: Why is reliability and validity important when evaluating teacher candidates?

# CAEP Guidelines for Establishing Validity and Reliability for Educator Preparation Programs (EPP)- Created Assessments

Gary Railsback, PhD  
Vice President, CAEP  
[Gary.Railsback@caepnet.org](mailto:Gary.Railsback@caepnet.org)



# Reminders About CAEP Guidelines

- CAEP differentiates in the self-study between two major types of assessments:
  - **Proprietary assessments** are developed by another organization and adopted/required by a state (e.g., edTPA), or are purchased by an EPP (e.g., the Danielson framework).
  - **EPP-created assessments** are created by your faculty or staff.

# Reminders About CAEP Guidelines

- CAEP differentiates in the self-study between two major types of assessments:
  - **Proprietary assessments** that a state has adopted/required (e.g., EdTPA) or that an EPP has purchased (e.g., the Danielson framework).
    - The self-study asks you to identify these proprietary assessments but does not require reporting validity and reliability. The instructions say “if available.”

# EPP-Created Assessments

- CAEP does not have a required minimum number of assessments. Most EPPs have 2 to 3, and CAEP discourages EPPs from developing more than 5 to 6 (although no rule or policy prevents it).
- CAEP encourages EPPs to develop these assessments across all **initial-level programs**. The most common are:
  - **Clinical practice observations**: used in early fieldwork, pre-student teaching, and/or the full semester.
  - **Dispositions**: Although many programs are now adopting proprietary assessments for this construct, many EPPs still create their own.
  - **Unit plan**: Developed in a methods course or used in student teaching/clinical practice.

# EPP-Created Assessments, continued

- **For advanced-level programs**, CAEP understands that most EPPs do not have common courses or common assessments across advanced-level programs (although they can do that if they desire).
- EPP-created assessments that will most likely be seen as CAEP fully implements advanced-level site visits in fall 2019:
  - Professional skills addressed in Standard A.1.1.
  - Content specific assessments for Standard A.1.2.
  - Clinical practice for programs with this element for Standard A.2.

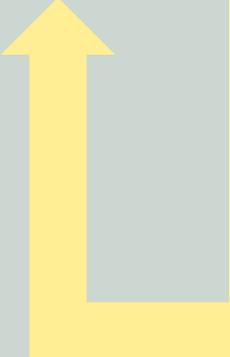
# CAEP Definition of Validity

- **The extent to which a set of operations, test, or other assessment measures what it is supposed to measure.** Validity is not a property of a data set but refers to the appropriateness of inferences from test scores or other forms of assessment and the credibility of the interpretations that are made concerning the findings of a measurement effort.
  - Source: *CAEP Handbook: Initial-Level Programs 2018*, p. 126.

# Using the CAEP Evaluation Framework

Attributes <b>BELOW</b> Sufficiency	Attributes <b>AT</b> Sufficiency	Attributes <b>ABOVE</b> Sufficiency
<p>The CAEP Evaluation Framework has three levels:</p> <ul style="list-style-type: none"><li>• Attributes <b>BELOW</b> Sufficiency (left column)</li><li>• Attributes <b>AT</b> Sufficiency (center)</li><li>• Attributes <b>ABOVE</b> Sufficiency (right column)</li></ul>		

# Using the CAEP Evaluation Framework: Validity Criteria – Accreditation Process

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
<div style="display: flex; align-items: center;">  <div style="background-color: #ffffcc; padding: 10px; border: 1px solid black;"> <p>Attributes from the left or <b>below sufficiency</b> column:</p> <ul style="list-style-type: none"> <li>Plan to establish validity does not inform reviewers whether validity is being investigated or how.</li> <li>The instrument was not piloted prior to administration.</li> <li>Validity is determined through an internal review by only one or two stakeholders.</li> <li>Described steps do not meet accepted research standards for establishing validity.</li> </ul> </div> </div>		

# Using the CAEP Evaluation Framework: Validity Criteria – Accreditation Process, continued

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
<p data-bbox="359 654 1378 696">Attributes from the center or <b>at Sufficiency</b> column:</p> <ul data-bbox="359 711 2209 1210" style="list-style-type: none"><li data-bbox="359 711 2209 811">• A description or plan is provided that details steps the EPP has taken or is taking to ensure the validity of the assessment and its use.</li><li data-bbox="359 825 2209 925">• The plan details what types of validity are under investigation or have been established (construct, content, concurrent, predictive, etc.) and how they were established.</li><li data-bbox="359 939 2209 982">• If the assessment is new or revised, a pilot was conducted.</li><li data-bbox="359 996 2209 1096">• The EPP details its current process or plans for analyzing and interpreting results from the assessment.</li><li data-bbox="359 1110 2209 1210">• The described steps generally meet accepted research standards for establishing the validity of data from an assessment.</li></ul>		



# Using the CAEP Evaluation Framework: Validity Criteria – Accreditation Process, continued

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
<p data-bbox="333 743 2163 1083">Attributes from the right or <b>above Sufficiency</b> column (not required):</p> <ul data-bbox="333 853 2163 1083" style="list-style-type: none"><li data-bbox="333 853 2163 915">• A validity coefficient is reported.</li><li data-bbox="333 939 2163 1083">• Types of validity investigated go beyond content validity and move toward predictive validity.</li></ul> 		

# Q & A

---

- Tell us: What are some challenges you think EPPs face when developing teacher candidate evaluation instruments and examining their reliability and validity?

# Examples of EPP-Created Assessments Used in the Accreditation Process (Initial & Advanced)



# Using the CAEP Evaluation Framework: Validity Criteria – Accreditation Process

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
 <p>An example from the left or <b>below sufficiency</b> column is:</p> <ul style="list-style-type: none"><li>• <b>Curricular validity</b>, which refers to the extent to which the content of the assessment matches the objectives of a specific curriculum as it is formally described.<ul style="list-style-type: none"><li>• <i>Course grades</i></li><li>• <i>GPA, courses specific P–12 learner</i></li><li>• <i>End-of-course/program assessments (without validity &amp; reliability)</i></li></ul></li></ul>		

# Using the CAEP Evaluation Framework: Validity Criteria – Accreditation Process, continued

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
<div data-bbox="445 585 637 949"></div> <div data-bbox="637 682 2303 1178"><p>Another example from the left or <b>below sufficiency</b> column is:</p><ul style="list-style-type: none"><li>• <b>Face validity</b>, which refers to the extent to which items in an assessment appear to measure particular constructs, in view of examinees.<ul style="list-style-type: none"><li>• <i>Dispositional data (qualitative with no analysis)</i></li><li>• <i>Candidate interviews (without instrument, no analysis)</i></li><li>• <i>Portfolios (without instrument, qualitative with no analysis)</i></li></ul></li></ul></div>		

# Using the CAEP Evaluation Framework: Validity Criteria – Accreditation Process, continued

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
<p data-bbox="293 635 1439 692">Attributes from the center or <b>at Sufficiency</b> column:</p> <ul data-bbox="293 706 2254 1206" style="list-style-type: none"><li>• <b>Content validity</b> refers to the appropriateness of the content of an assessment: Does it measure/assess what examinees must demonstrate?<ul style="list-style-type: none"><li>• <i>Lesson/unit plans (rubrics)</i></li><li>• <i>Teacher work samples (rubrics)</i></li><li>• <i>Portfolio assessments (rubric)</i></li><li>• <i>Observation Instruments</i></li><li>• <i>Capstone/thesis/action research/summative project</i></li><li>• <i>Problem-based project in conjunction with a school or district partner(s)</i></li></ul></li></ul>		

# Using the CAEP Evaluation Framework: Validity Criteria – Accreditation Process, continued

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
<p>Attributes from the right or <b>above Sufficiency</b> column (not required):</p> <ul style="list-style-type: none"> <li>• <b>Predictive validity</b>, which refers to the extent to which performance on an assessment is related to later performance, that assessment was designed to predict.               <ul style="list-style-type: none"> <li>• <i>Pre-service measures of candidate impact</i></li> <li>• <i>Comparisons of candidates in education program and other IHEs</i></li> </ul> </li> </ul>		

# Validity

- Can be supported through evidence of the following:
  - Agreement among reviewers of narrative evidence.
  - Expert validation of performance or artifacts.
  - Expert validation of the items in an assessment or rating form.
  - A measure's ability to predict performance in a future setting (predictive validity).

# Approaches to Developing Content Validity

- CAEP does not require or disallow any research method used by EPPs to determine content validity.
  - The most common method used by EPPs is the Lawshe method.
  - CAEP does not require the Lawshe method of developing content validity.
  - CAEP does not disallow the use of the Lawshe method.

$$\text{CVR} = (n_e - N/2) / (N/2)$$

# Developing Content Validity Requires a Content Evaluation Panel

- **Selection** of content experts to serve on a panel representing different perspectives:
  - P–12-based clinical educators
  - Faculty members (content)
  - P–12 administrators/leaders/partners
  - Candidates/completers
  - Parent advisory boards
- Ask to do the following:
  - Rate the statements as “essential,” “useful but not essential,” or “not necessary.”
  - Statements must be aligned with the construct being measured.

# Another Reminder...Quantifying Consensus

- Any statement (indicator) which is perceived as “essential” by more than half of the content experts has some degree of content validity
- The more panelists (beyond 50 percent) who perceive the statement (indicator) as “essential,” the greater the extent or degree of its content validity.
- Utilize the Lawshe article to determine the content validity ratio (CVR).

$$\text{CVR} = (n_e - N/2) / (N/2)$$

# CAEP Definition of Reliability

- **The degree to which test scores for a group of test takers are consistent over repeated applications** of a measurement procedure and hence **are inferred to be dependable and repeatable** for an individual test taker.
  - Source: *CAEP Handbook: Initial-Level Programs 2018*, p. 119.

# Reliability Criteria

Attributes BELOW Sufficiency	Attributes AT Sufficiency	Attributes ABOVE Sufficiency
<div data-bbox="410 572 621 933"> </div> <div data-bbox="621 682 2252 1196"> <p>Another example from the left or <b>below sufficiency</b> column is:</p> <ul style="list-style-type: none"> <li>• Plan to establish reliability does not inform reviewers whether reliability is being investigated or how</li> <li>• Described steps do not meet accepted research standards for reliability</li> <li>• No evidence, or limited evidence, is provided that scorers are trained and their inter-rater agreement is documented</li> </ul> </div>		

# Most Common Type of Reliability Claimed for EPP-Created Assessments

## Inter-Rater/Inter-Observer Reliability

- Used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon.
- Agreement measures how frequently two or more evaluators (e.g., faculty) assign the same rating.
  - *Candidate interviews*
  - *Lesson/unit plans (rubrics)*
  - *Observation instruments*

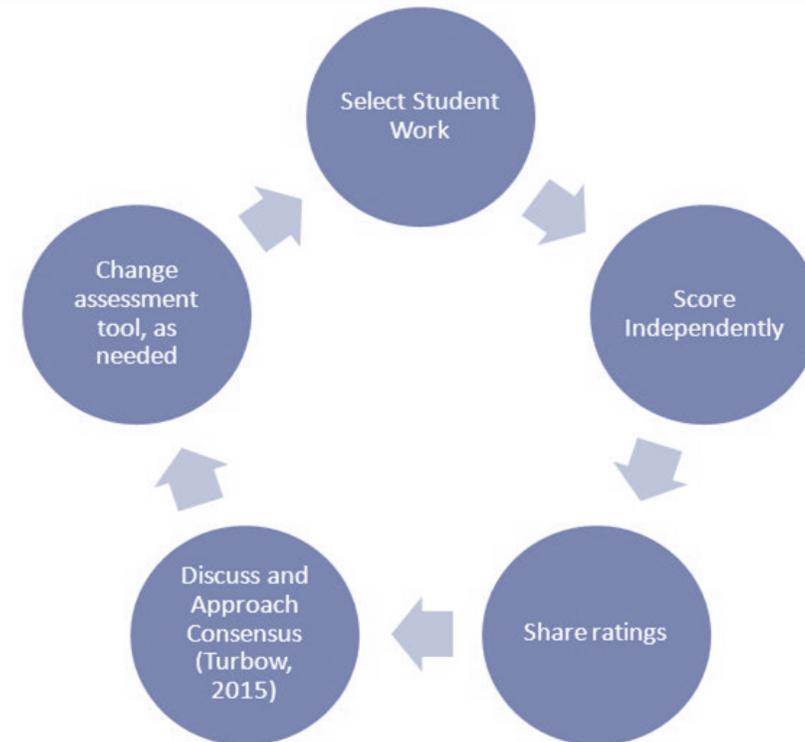
# Reliability, continued

- Can be supported through evidence of the following:
  - Agreement among multiple raters of the same event or artifact (or the same candidate at different points in time).
  - Stability or consistency of ratings over time.
  - Evidence of internal consistency of measures.

# Instrument Clarification...Supporting Reliability/Validity

- Content expert process/instrument feedback
- Feedback/clarification – evaluate quality of measure
- Content experts/developers of measure – review and seek feedback on what was learned
- Pilot/implement – examining the data

## Inter-rater Reliability Process



(CEHD/STEL)

RADFORD UNIVERSITY

# Common Challenges EPP Staff Face in Meeting CAEP Standards



# Questions?

# North Dakota Association for Colleges of Teacher Education (NDACTE)

Dr. Sarah Anderson, Mayville State University

Dr. Stacy Duffield, North Dakota State University

Dr. Alan Olson, Valley City State University

# Development of the Student Teacher Observation Tool

(STOT)

- Tell us: What do you think are the pros and cons of EPP staff developing their own instruments versus purchasing or adapting an existing instrument?

# Primary Reason for Developing an EPP-Created Assessment

---

- North Dakota had a common Exit Survey, Completer Survey, and Employer Survey.
  - NDACTE Common Metrics assessments grew from Network for Excellence in Teaching (NExT) efforts funded by the Bush Foundation in Minnesota, North Dakota, and South Dakota.
- NDACTE representatives believed they could use their collaborative expertise and experiences to design a student teacher assessment instrument that would work well for teacher candidates and cooperating teachers in our state as well as meet our EPP expectations for accreditation.
- The EPP-created student teacher observation tool fulfilled a **need** for a valid and reliable assessment instrument that met CAEP sufficiency levels.

# The Benefits and Challenges of EPP Collaboration

Benefits	Challenges
A variety perspectives contribute the vision and outcome.	Collaborative work can take more time and effort.
Increased potential to add skill and experience in research, statistics, assessment, practicality and/or field expertise.	Loss of some autonomy with a probable need for making some concessions.
The establishment of common language for compiling aggregate data and identifying educator preparation areas of strength and improvement that can lead to meaningful conversations and actions.	
Improved communication and networking for future collaboration.	
Mutually beneficial outcomes.	
Collaboration with P–12 partners for shared resources or expertise.	
Common instrument for EPPs and cooperating teachers.	

# EPP-Created Assessments: Pros and Cons

Pros	Cons
Work can begin with standards, but proceed with more freedom and independence.	Increased time and efforts to create, pilot, and assess the assessment.
Greater opportunity to develop assessments that are practical and meaningful to an EPP.	Increased responsibility for meeting reliability, validity, and addressing accreditation expectations compared to proprietary assessments.
Items can be aligned with or complement other EPP assessment instruments.	The creation process can cost money and take time.
Assessment can be validated to the population using the instrument.	

# Interest in Student Teacher Observation Tool from Institutions in 17 Other States (*so far*)

---

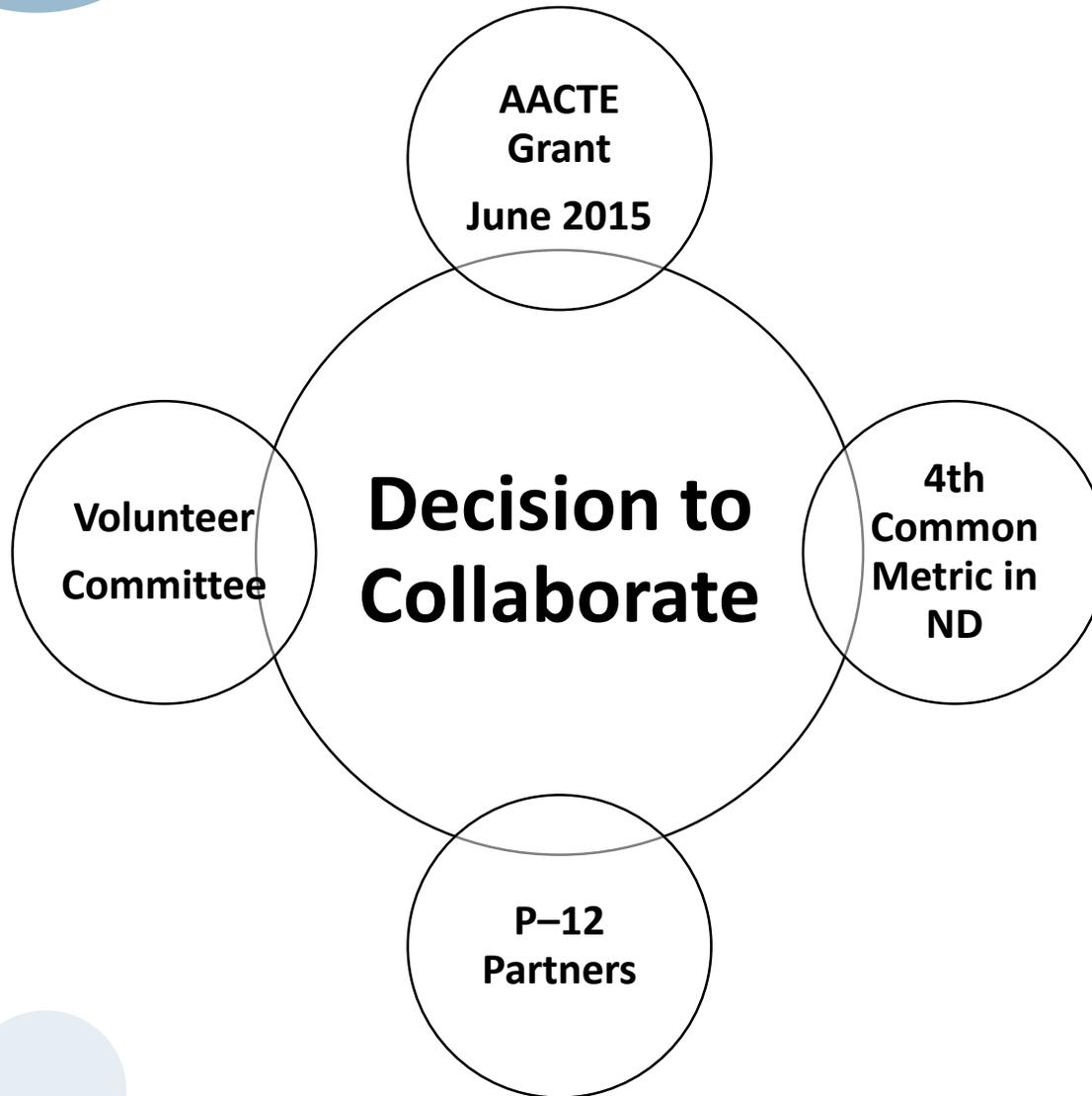
- Alaska
- Arizona
- Connecticut
- Florida
- Illinois
- Indiana
- Kentucky
- Maryland
- Minnesota
- Montana
- New York
- Pennsylvania
- Rhode Island
- South Dakota
- Tennessee
- West Virginia
- Wisconsin





## Committee Members

- Dr. Meghan Salyers, University of Mary
- Dr. Teresa Delorme, Turtle Mountain Community College
- Kim Marman, MEd, University of Mary
- Dr. Lisa Borden, King-Minot State University
- Dr. Stacy Duffield, North Dakota State University
- Dr. Alan Olson, Valley City State University
- Dr. Sarah Anderson, Mayville State University



# NDACTE Partners



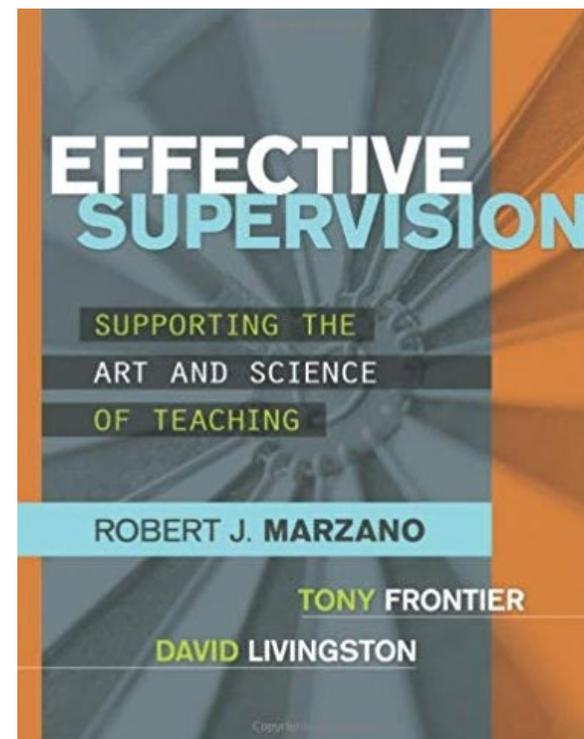
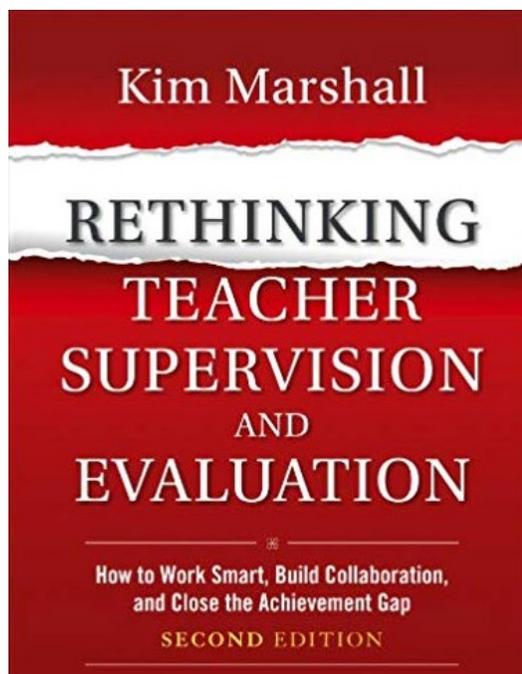
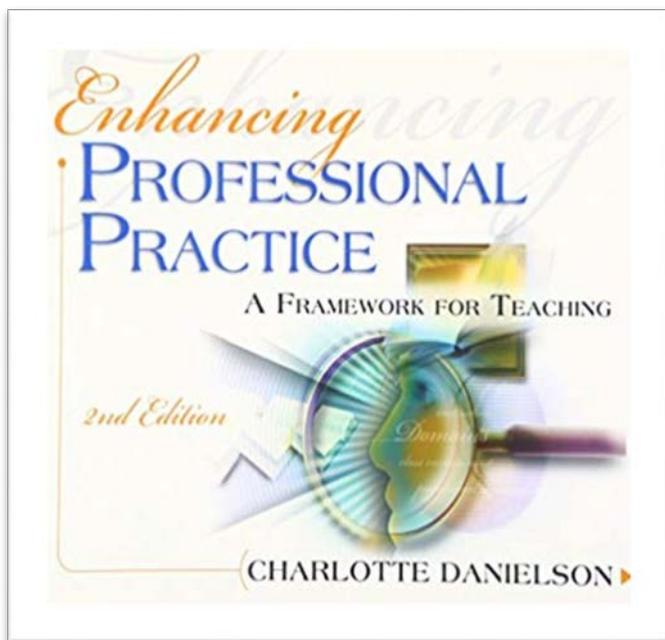
# STOT Instrument Development Process

---

1. Decision to collaborate
2. Establish purpose
3. Gather instruments from institutions
4. Instrument development phase
5. Feedback and refinement
6. Pilot #1 and exploratory factor analysis – May 2016
7. Instrument refinement
8. Pilot #2 exploratory factor analysis – December 2016
9. Review of results and instrument refinement – May 2017
10. Statewide use in North Dakota – 2017/18 academic year
11. Inter-rater reliability training module development
12. Confirmatory analysis – spring 2019



# Instrument Development



# STOT Example: Interstate Teacher Assessment and Support Consortium (InTASC) Standard 1 Performance Skills

InTASC Standard (1)	Distinguished (4)	(3.5)	Proficient (3)	(2.5)	Emerging (2)	(1.5)	Underdeveloped (1)	Rating
<i>The teacher candidate...</i>								
<b>Supports student learning through developmentally appropriate instruction</b>	implements challenging learning experiences that recognize patterns of learning and development across cognitive, linguistic, social, emotional and physical areas.	In addition to rating "3" performance, partial success at rating of "4"	implements developmentally appropriate instruction that accounts for learners' strengths, interests and needs	In addition to rating "2" performance, partial success at rating of "3"	implements grade-level appropriate instruction, but does not account for individual learners' differences.	With assistance, partial success at rating of "2"	implements instruction that exceeds or does not match a developmentally appropriate level for the students.	
<b>Accounts for differences in students' prior knowledge</b>	accesses student readiness for learning and expands on individual students' prior knowledge.		accounts for individual differences in students' prior knowledge and readiness for learning.		addresses students' prior knowledge as a class, but individual differences are not considered.		does not account for differences in students' prior knowledge.	
<i>*The overall rating will be calculated as an average of the ratings for this standard.</i>								<b>*Rating</b>
<b>Standard #1: Learner Development.</b> The teacher understands how children learn and develop, recognizing that patterns of learning and development vary individually within and across the cognitive, linguistic, social, emotional, and physical areas, and designs and implements developmentally appropriate and challenging learning experiences.								

# North Dakota STOT Development Timeline

<b>August 2016</b>	Validation study report for Pilot #1
<b>September 22, 2016</b>	Report reviewed and Draft #16
<b>October 13, 2016</b>	Draft #17 distributed for review
<b>October 20, 2016</b>	Draft #18 used for Pilot #2
<p>December 2016</p> <p>Pilot #2 Exploratory Factor Analysis</p>	

# NDACTE Website



North Dakota Association  
of Colleges for Teacher Education

- Home
- About Us
- Membership
- Events
- News Room
- Resources
- AACTE Resources
- search...

Home | **Student Teacher Observation Tool (STOT)**

## Student Teacher Observation Tool (STOT)

### Update October 2018:

Student Teacher Observation Tool (STOT) Training Module

### Update June 2017:

20 InTASC\_Rubrics for Student Teaching Final 062917

STOT Validity Summary June 2017 (1)\_

### CAEP Presentation 2017:

CAEP\_NDCATE\_Presentation 2017

STOT Validation (2016)

CAEP\_NDCATE\_Presentation 2017

18 InTASC\_Rubrics for Student Teaching\_Oct 18 - Copy

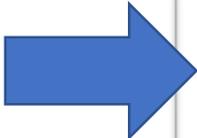
## News on Teacher Preparation Policy

The farmer's lens: ND farmers lose big ally - Grand Forks Herald

Stories of 2018 include triumph, tragedy - Grand Forks Herald

Tim Walz sworn in as Minnesota's new governor. Here's what he said - St. Paul Pioneer Press

This information is updated continuously from Google News.



# STOT Instrument Development Process

---

1. Decision to collaborate
2. Establish purpose
3. Gather instruments from institutions
4. Instrument development phase
5. Feedback and refinement
6. Pilot #1 and exploratory factor analysis – May 2016
7. Instrument refinement
8. Pilot #2 exploratory factor analysis – December 2016
9. Review of results and instrument refinement – May 2017
10. Statewide use in North Dakota – 2017/18 academic year
11. Inter-rater reliability training module development
12. Confirmatory analysis – spring 2019



# Poll: Reliability and Validity Question

---

- Tell us: Can an instrument be unreliable but valid?
  - Yes
  - No
  - It depends

# Poll: Reliability and Validity Answer

---

- Tell us: Can an instrument be unreliable but valid?
  - No, for a test to be **valid** it must be **reliable**.

# Reliability and Validity of the STOT

# Development Began with Validity & Reliability in Mind

---

- Face validity
- Content validity
- Construct validity
- Internal reliability
- Inter-rater reliability

# Face Validity

---

- Engages users.
- Asks if the instrument meets its intended purpose.
- Is established through a pilot and feedback from cooperating teachers and university supervisors.
- Is useful and important, but is not enough.

# Content Validity

---

- Asks if items measure what they are intended to measure.
- Began with the InTASC standards.
- Alignment with previously validated instruments (NExT surveys).
- Relied on the knowledge of experts familiar with the content being measured:
  - Representatives from NDACTE institutions with strong knowledge of student teacher evaluation as well as the InTASC standards participated in STOT development.
  - Supervisors and cooperating teachers who have deep knowledge of the knowledge and skills of teaching reviewed the instrument.
  - Feedback from these experts was collected and analyzed during development and revision.

# Construct Validity

---

- Measured through a factor analysis.
- Initial exploratory factor analysis (EFA) used for revision.
- Because revisions were substantial, a second EFA was conducted.

# Construct Validation: Pilot I EFA

---

- Exploratory factor analysis.
- 80 respondents completed all 35 assessment items in Spring 2016.
- Computed the general measure of factorability (KMO: result of .940).
- 2 factors (coefficients greater than .35 in absolute value):
  - Factor 1: Combination of learner and learning, content knowledge, and instructional practice.
  - Factor 2: Professional responsibility.
- Some errant and cross-loading items.
- Revision to remove double-barreled wording and replace ambiguous wording.

# Construct Validation: Pilot II EFA

- Exploratory factor analysis.
- 139 respondents completed all 34 assessment items in fall 2016.
- Computed the general measure of factorability (KMO: result of .960).
- Four common factors (coefficients greater than .35 in absolute value):

Construct	Number of Items	Mean	Min	Max
Learner, learning, and diversity	8	.665	.541	.777
Content knowledge	7	.670	.607	.730
Instructional practices	12	.653	.504	.731
Professionalism	6	.651	.548	.785

# Example Results from Pilot II

Item	4	3.5	3	2.5	2	1.5	1	Mean	SD	n
1. Supports student learning through developmentally appropriate instruction	31	57	77	9	1	0	0	3.31	.43	175
2. Accounts for students' prior knowledge	33	44	84	9	4	0	0	3.27	.47	174
3. Uses knowledge of students' socioeconomic, cultural, and ethnic differences to meet learning needs	28	44	89	13	3	0	0	3.23	.45	177
4. Exhibits fairness and belief that all students can learn	62	45	62	5	2	0	0	3.45	.48	176
5. Fosters a safe and respectful environment that promotes learning	58	50	62	5	3	0	0	3.44	.48	178
6. Structures a classroom environment that promotes student engagement	41	48	64	19	3	0	0	3.30	.51	175
7. Clearly communicates expectations for appropriate student behavior	47	41	67	18	4	0	0	3.29	.56	177
8. Responds appropriately to student behavior	42	50	59	17	7	0	0	3.29	.54	175

# Reliabilities of Subscales: Pilot II

Subscale/Construct	Number of Items	Cronbach's Alpha
Learner, learning, and diversity	8	.930
Content knowledge	7	.929
Instructional practices	12	.952
Professionalism	6	.902

# Inter-Rater Reliability

---

- Found videos of teachers aligned with STOT indicators.
  - Used short excerpts to focus viewing, making accuracy of rating more likely.
- Expert panels of university faculty.
  - Needed to be expert at the level of classroom in featured video.
  - At least five members in each panel.
- Rating process:
  - Raters independently rated teacher performance, providing rationale.
  - Synchronous meeting during which raters presented ratings and rationale to panel members.
  - Raters re-rated video and arrived at consensus rating with evidence.

# Sample Independent Rating

Rating	Rationale
2.5	The teacher uses various forms of communication—shows students as well as telling them. Uses mnemonic devices to help students remember. Media of clay was used, but no technology. Listens to student answers for correctness, but it is mostly one-way communication.
3	The teacher uses written, nonverbal, and oral communication to connect with students—I saw gestures from both her and students, use of manipulatives, use of an overhead projector, listens and respectively articulates thoughts and ideas.
3	There was no technology present or written feedback that I saw. She does verbally articulate her thoughts and ideas, though.
3.5	Communication is respectful. The concepts are communicated in multiple modes. The pace is not too fast or too slow. Directions are given in steps. Nonverbally, there is smiling and laughter. Verbally, there is praise. The teacher respects and enjoys teaching these students. There are multiple checks for understanding.
3	Technology, using the overhead to show model of worm layers. Great evidence of verbal with the singing and repetition. Uses nonverbal with actions that are used as mnemonic devices. Listens to students; has good eye contact; good expression in her voice; gestures to support verbal communication.
3	When I look at the standard itself, the teacher appears proficient at helping students “to develop deep understanding of content areas and their connections, and to build skills to apply knowledge in meaningful ways.” The selected clips do not show written communication or use of digital technology, but that does not provide disconfirming evidence. The content selected is problematic in that it lends itself to memorizing and lower-order thinking, but this standards is not asking us to evaluate the appropriateness of the content selected, only the effectiveness of communication about it.

# Questions?

# Thank You

**Please visit our website and follow us on Twitter**  
for information about our events, priorities, and research alliances,  
and for access to our many free resources.

**[ies.ed.gov/ncee/edlabs/regions/central/index.asp](https://ies.ed.gov/ncee/edlabs/regions/central/index.asp)**

**@RELCentral**

or contact us at

**[RELCentral@marzanoresearch.com](mailto:RELCentral@marzanoresearch.com)**

This presentation 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. of Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.