Meeting the CAEP Requirements for Assessments

Margie Crutchfield, Ph.D.
CAEP Consultant
mdc1501@yahoo.com
October 27, 2016

What we are going to do today:

- Discuss the CAEP expectations for assessments
  - Proprietary vs EPP created
  - Documentation Expectations
  - Early Instrument Review
- Examine the CAEP Assessment Rubric
  - Use it to evaluate several different assessments
- Discuss CAEP's expectations for validity and reliability
- Discuss how to adapt program-level (or SPA) assessments to meet CAEP expectations for EPP-wide assessments

Introduction

- CAEP's perspective on the need for quality assessment tools (Evidence Guide, foreword and pp. 6 – 7).
  - The responsibility lies with the EPP to provide valid (and reliable) evidence,
  - CAEP's commitment to stronger preparation and accreditation data, and
  - CAEP's belief that current evidence is less than ideal.
  - The profession needs evidence that is
    - intentional, purposeful, and addresses deliberately posed questions of importance.
    - entails interpretation and reflection,
    - integrated and holistic,
    - qualitative and quantitative, and
    - direct and indirect

- At the Initial Level
  - Provide data from EPP-wide assessments
  - Typically seeing less than 10
  - Disaggregate data for each licensure/program area
- At the Advanced Level
  - Not required to have common assessments across the EPP
  - 3 – 5 assessments per program
- ALL EPP created assessments used in the CAEP review must meet the Sufficient level on the CAEP Instrument rubric

Proprietary vs. EPP-Created

- Proprietary
  - Developed by some other entity (usually state or national)
  - Examples: Praxis Series, EdTPA, State-Required Assessments
  - Not required to meet CAEP Instrument Rubric
  - You will provide links to validity and reliability information if available

- EPP-developed
  - Any assessment you developed and are using in the CAEP review
  - Any proprietary assessment that you modified
  - If you are using a proprietary assessment with local evaluators, then you will need to address inter-rater reliability

What will you submit?

- You submit all evidence in the on-line Evidence Room
- Allowed up to 90 uploads (will be increased for EPPs submitting Advanced Programs)
  - Currently each upload limited to 5 mgs, but can be increased to 20mg upon request
- You can combine all the documents for one assessment into a single file

For one assessment…

- You will submit
  - A copy of the assessment
  - Data chart(s)
  - Instructions to candidates
  - Can include a page on how you addressed validity and reliability or you can simply respond to the 5 questions
  - (optional) Can include an analysis of data for this instrument—or this may be in your Self Study Report Narrative
  - Response to five questions in AIMS

Questions to be answered for each submitted EPP created assessment

1. During which part of the candidate's experience is the assessment used? Is the assessment used just once or multiple times during the candidate's preparation?

2. Who uses the assessment and how are the individuals trained on the use of the assessment?

- What is the intended use of the assessment and what is the assessment purported to measure?

- Please describe how validity/trustworthiness was established for the assessment.

- Please describe how reliability/consistency was established for the assessment.

Want to see what this looks like in AIMS?

- URL: aims.caepnet.org
- User login: 24319
- Password: caep
- Go to “Visit Reports” then “Self Study Evidence”
- Go to bottom of page and click on “Add” and follow the prompts, selecting “EPP-Created Assessment” to upload an assessment (use a dummy)
- Click on the paper icon to find the five questions

Early Instrument Review

- You can submit up to 10 assessments for review
  - Should all be submitted at the same time
- Can be submitted up to 3 years prior to your visit
  - The sooner the better!
- A team of two trained CAEP Assessment Reviewers (now over 100) will review each of your assessments
- Use the CAEP Assessment Rubric
- Provide you with substantive feedback on each assessment (not just a rating)
- Submitted in AIMS—must request shell from CAEP

CAEP Instrument Rubric

- Current version is June 2016
- Used by Assessment Reviewers and Site Visitors to evaluate any EPP-created assessments
- Six sections

1. Administration and Purpose Sufficient Level

- The point or points when the assessment is administered during the preparation program are explicit
- The purpose of the assessment and its use in candidate monitoring or decisions on progression are specified and appropriate
- Evaluation categories or assessment tasks are tagged to CAEP, InTASC or state standards

Example- Administration and Purpose:

- **CAEP Sufficient Level**
  - Point or points when the assessment is administrated during preparation are explicit
  - Purpose of the assessment and its use in candidate monitoring or decisions on progression are specified and appropriate
  - Evaluation categories or assessment tasks are tagged to CAEP, InTASC or state standards (*This criterion will not be used to score this example – use only the first two criteria*.)

- Completed by Cooperating Teacher at the end of Student Teaching; assessment completed once for each student teacher; Cooperating Teacher attends online workshop and University Supervisor "trains" Cooperating Teacher in use of rubric; purpose of assessment is to evaluate EPP program in relation to candidate's preparation and performance as well as to evaluate program in relation to standards. This is a summative assessment that determines if the candidate satisfactorily completes the final student teaching experience.

Rating?

- Point or points when the assessment is administrated during preparation are explicit
- Purpose of the assessment and its use in candidate monitoring or decisions on progression are specified and appropriate

Score: Sufficient (except for tagging to standards)

- Let's look at the Sample handout (no number)
  - When is it administered?
  - What is its purpose?
  - How is it used in candidate monitoring?
  - How is a decision on progress made?
  - Are indicators tagged to CAEP, InTASC or state standards?
    - A note about this…..

2. Informing Candidates

- The candidates who are being assessed are given a description of the assessment's purpose
- Instructions provided to candidates about what they are expected to do are informative and unambiguous
- The basis for judgment (criterion for success, or what is "good enough") is made explicit for candidates
  - Let's look at another example

Example- Informing Candidates

- **CAEP Sufficient Level**
  - Candidates who are being assessed are given a description of the assessment's purpose
  - Instructions provided to candidates about what they are expected to do are informative and unambiguous
  - The basis for judgment (criterion for success, or what is "good enough") is made explicit for candidates.

Candidates are assessed four times during their program, corresponding roughly to their freshman, sophomore, junior and senior years. Teacher candidates complete a self-assessment using the Dispositions Survey each year, and are assessed by their cooperating teacher and university supervisor in their methods block (junior year) and during student teaching (senior year). Faculty review the survey results to affirm that responses are consistent with desired dispositions. TCs expressing/demonstrating inappropriate dispositions are counseled out or referred for remediation.

Based on the rubric – How would you score this submission on Informing Candidates?

No description of purpose is provided.

No specific expectations are set for candidates.

No benchmark is set on the level candidates need to achieve.

Score: Below minimal level of sufficiency

3. Content (Indicators) Sufficient Level

- Indicators assess explicitly identified aspects of CAEP, InTASC or state standards
- Evaluation indicators reflect the degree of difficulty or level of effort described in the standards
- Indicators unambiguously describe the proficiencies to be evaluated
- When the standards being informed address higher level functioning, the indicators require higher levels of intellectual behavior (e.g., create, evaluate, analyze, & apply).
- Most indicators (at least those comprising 80% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards

- Read through the indicators for Sample #1. Do they meet the Sufficient Level for Content on the CAEP Rubric?

InTASC Standards

- Standard #1: Learner Development. The teacher understands how learners grow 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.

- Standard #7: Planning for Instruction. The teacher plans instruction that supports every student in meeting rigorous learning goals by drawing upon knowledge of content areas, curriculum, cross-disciplinary skills, and pedagogy, as well as knowledge of learners and the community context.

- Standard #8: Instructional Strategies. The teacher understands and uses a variety of instructional strategies to encourage learners to develop deep understanding of content areas and their connections, and to build skills to apply knowledge in meaningful ways.

For Sample #1: Content

- Indicators assess explicitly identified aspects of CAEP, InTASC or state standards
  –
  - Yes, the indicators align with the identified InTASC Standards
- Evaluation indicators reflect the degree of difficulty or level of effort described in the standards
  - Yes, indicators do require demonstration and application of knowledge, skills, and dispositions.
- Indicators unambiguously describe the proficiencies to be evaluated
  - Yes, indicators use action verbs specific to content
- When the standards being informed address higher level functioning, the indicators require higher levels of intellectual behavior
  - Yes, higher levels are measured by way of "integrate, reflect, use, plans and delivers, etc.
- Most indicators (at least those comprising 80% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards
  - Yes, most indicators judge consequential attributes

Now let's look at Sample #2 for Content

Sufficient Level

- Indicators assess explicitly identified aspects of CAEP, InTASC or state standards
- Evaluation indicators reflect the degree of difficulty or level of effort described in the standards
- Indicators unambiguously describe the proficiencies to be evaluated
- When the standards being informed address higher level functioning, the indicators require higher levels of intellectual behavior (e.g., create, evaluate, analyze, & apply).
- Most indicators (at least those comprising 80% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards

Sample #2– Answers for 3. Content

- Categories were used in lieu of indicators
  - Learning standards, assessment strategies
- Indicators were tagged to multiple InTASC subcomponents
- CAEP Standards were not specifically identified for each indicator
- Since categories were used it was difficult to determine the degree of difficulty or alignment of the indicator to criteria
- No indication of higher level function since general categories are used

Below the minimal level of sufficiency

Now let's look at Sample #3 for Content

Sample #3– Answers for 3. Content

- Multiple subcomponents are identified for each indicator
- Mix of categories (e.g., Self-reflection) and specific indicators (e.g., Engages in random acts of collaboration (trustworthy, just, respectful, fair, equitable, honest)
- Some criteria for indicators do not align with or assess the same level of complexity
- While some indicators may identify consequential attributes, criteria are often related to the frequency of the behavior
- Vague terms that are open to multiple interpretations are often used as criteria

Below the minimal level of sufficiency

4. Scoring: The Rubric Levels Sufficient level

- The basis for judging candidate work is well defined
- Each proficiency level is qualitatively defined by specific criteria aligned with indicators
- Proficiency level descriptions represent a developmental sequence from level to level (to provide raters with explicit guidelines for evaluating candidate performance and candidates with explicit feedback on their performance)
- Feedback provided to candidates is actionable
- Proficiency level attributes are defined in actionable, performance-based, or observable behavior terms. NOTE: If a less actionable term is used such as "engaged", criteria are provided to define the use of the term in the context of the indicator

Let's look at Sample #1

- Does this assessment reach the Sufficient level for Scoring?

For Sample #1 – Scoring

- The basis for judging candidate work is well defined
  - No, generic definitions are provided for each level that lacks specificity aligned with the indicator.
- Each proficiency level is qualitatively defined by specific criteria aligned with indicators
  - No, there are no qualitative definitions specific to the indicator; in some cases the criteria have little relationship to the indicator
- Proficiency level descriptions represent a developmental sequence from level to level
  - No, there is no sequence identified specific to the indicator
- Feedback provided to candidates is actionable
  - No, since a generic definition is used for each level, no specific actionable feedback is provided
- Proficiency level attributes are defined in actionable, performance-based, or observable behavior terms. NOTE: If a less actionable term is used such as "engaged", criteria are provided to define the use of the term in the context of the indicator
  - No, since no specific criteria are identified for each indicator

Now let's look at Sample #2

- Does this assessment reach the Sufficient level for Scoring?

Sample #2– Answers for 4. Scoring

- Basis for judging candidate work is defined
- Proficiency levels are qualitatively defined with specific criteria
- Developmental sequence is evident
- Feedback provided to candidate is generally actionable
- Proficiency level criteria are performance-based and observable

- Met CAEP Sufficient Level

Sample #3– Answers for 4. Scoring

- Basis for judging candidate work is not well defined or specific
- Often the proficiency levels are differentiated by the frequency of the behavior without assessing quality
- While four levels are identified, often the frequency of the behavior is the only difference from level to level
- Often feedback provided to candidate is vague or relies on a count. The feedback provided by the assessment varies widely from indicator to indicator.
- Overall, more specific criteria for each indicator is needed
- Below minimal level of sufficiency

5.a. Data Validity Sufficient Level

- 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
- The plan details the types of validity that are under investigation or have been established (e.g., construct, content, concurrent, predictive, etc.) and how they were established
- If the assessment is new or revised, a pilot was conducted.
- The EPP details its current process or plans for analyzing and interpreting results from the assessment
- The described steps generally meet accepted research standards for establishing the validity of data from an assessment

CAEP's Definition of Validity

- 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.

Instrument Validity

- CAEP's Working Definition: Degree to which an assessment measures what it purports to measure and how the results will be used –
  - A description or plan is provided
    - Describes the steps to be used for determining content validity
    - Research was used in the development of the plan
    - Pilot was completed prior to administration
    - Steps meet accepted research standards/protocols
  - Submitted assessments should be at Sufficient level or above based on the assessment rubric

Establishing content validity: there are various ways to do this

- Lawshe's method (CVR) in general…..
  - Uses a panel of experts (CAEP's case this would include P-12 teachers)
  - Ask the following question for each item:
    - "Is the skill or knowledge measured by this item 'essential', 'useful, but not essential,' or 'not necessary' to the performance of the construct?"
    - If half of the panelists rate the item as essential, that item has at least some content validity.

Content Validity

- Another Method of Establishing Content Validity
  - Conduct a job-task analysis to identify essential job tasks, knowledge areas, skills and abilities
  - Link job tasks, knowledge areas or skills to the associated test construct or component that it is intended to assess
  - Use subject-matter experts
  - Document that the most essential knowledge areas and skills were assessed and explain why less essential knowledge and skills were excluded
- EPPs need to describe some systematic review of the items on EPP created assessments

- Requires a degree of agreement among “experts”
  - Requires the use of recognized subject matter experts
  - Based on the judgment of subject matter experts
  - Relies on individuals who are familiar with the construct such as –
    - Faculty members
    - EPP based clinical educators
    - P-12 based clinical educators
  - Ask the fundamental question – “Do the indicators really assess the construct to be measured?”

Example

- Let's look at the Sample Assessment handout (no number)
  - What information do they provide for validity?
  - Does this meet the CAEP sufficient level for data validity?

Sufficient Level:

- Describes the steps to be used for determining content validity
- Research was used in the development of the plan
- Pilot was completed prior to administration
- Steps meet accepted research standards/protocols

Based on the Assessment Rubric

- While a summary was provided on a process used for development of the assessment, it was unclear what specific research-based process was completed.
- There are other references made to a wide review of the assessment by various stakeholders, but it is unclear if a panel of experts were used to establish content validity.
- A pilot was conducted
- Sufficient level not met

5.b.: Data Reliability Sufficient Level

- A description or plan is provided that details the type of reliability that is being investigated or has been established (e.g., test-retest, parallel forms, inter-rater, internal consistency, etc.) and the steps the EPP took to ensure the reliability of the data from the assessment
- Training of scorers and checking on inter-rater agreement and reliability are documented
- The described steps meet accepted research standards for establishing reliability

Instrument Reliability

- Degree in which an assessment produces stable and consistent results –
  - Ask the question - Can the evidence be corroborated?
  - Criteria
    - A detailed description or plan is provided
    - Training of scorers and checking on inter-rater reliability are documented
    - Steps are described that meet accepted research standards for establishing inter-rater reliability

Inter-Rater or Inter-Observer Reliability –

- Used to assesses the degree to which different raters/observers give consistent estimates of the same phenomenon
  - Calculate the correlation between the ratings of the two or more observers viewing the same clinical experience at the same time
  - Could be consistent with master raters determined by calibration
  - Hold "calibration" meetings
    - Watch a clinical experience with a group
    - Talk about how ratings were determined and what each reviewer noted
    - Come up with rules for deciding what represents a "3" or "4" on the instrument
  - Can use percentage of agreement

Describing the Process

- EPPs must provide information on how they will attempt or are currently checking for inter-rater reliability
  - Could provide a description of training protocols used for each assessment
  - Could provide description of calibration meeting and protocols for those meetings
  - Could use video analysis to determine inter-rater reliability
- Process must be described and be consistent with research standards for the establishment of inter-rater reliability.

Example of Reliability:

The University Supervisor visits, personally, with the Cooperating Teacher about content of the Disposition criteria prior to the Cooperating Teacher's use of the instrument. There is, then, opportunity for stakeholders (i.e., Advisory Board, EPP faculty) to compare responses of the Cooperating Teacher, University Supervisor and Student Teacher (since the Student Teacher uses the same rubric two times during student teaching to self-evaluate) in order to assess inter-rater reliability. (sample)

Sufficient Criteria:

- A detailed description or plan is provided
- Training of scorers and checking on inter-rater reliability are documented
- Steps are described that meet accepted research standards for establishing inter-rater reliability

Response

- Training of scorers and checking on inter-rater reliability are not fully documented
- Described steps do not meet accepted research standards
- Type of reliability is identified
- While the type of reliability is identified and some evidence of training of scorers is provided, the steps are informal and fall short of research standards.

Below Sufficient Level

Surveys

- Not required, but likely to be used for Components 4.3 and 4.4
- Expectations are very different from other EPP created assessments.

6.a.: Survey Content Sufficient Level

- Questions or topics are explicitly aligned with aspects of the EPP's mission and also CAEP, InTASC or state standards
- Questions have a single subject; language is unambiguous
- Leading questions are avoided
- Items are stated in terms of behaviors or practices instead of opinions, whenever possible
- Surveys of dispositions make clear to candidates how the survey is related to effective teaching

6.b. Survey Data Quality Sufficient Level

- An even number of scaled choices helps prevent neutral (center) responses
- Scaled choices are qualitatively defined using specific criteria aligned with key attributes identified in the item
- Feedback provided to the EPP is actionable
- EPP provides evidence that questions are piloted to determine that candidates interpret them as intended and modifications are made, if called for

Criteria listed below are evaluated on site:

- *EPP provides evidence that candidate responses are compiled and tabulated accurately*
- *Interpretations of survey results are appropriate for the items and resulting data*
- *Results from successive administrations are compared (for evidence of reliability)*

Program Level vs EPP-wide Assessments

- Many of you have multiple program-level (or SPA) assessments
- Can these be used as EPP-wide assessments?
  - Yes, but. . . . . .
- How would you modify, adapt, add to program-level assessments so that they can be used for EPP-wide data?

An example of feedback from the Early Assessment Review Process

Questions?