Meeting CAEP's Requirements for EPP-Created Assessments: Putting It All Together

CAEP Conference
Friday, March 25, 2016
Hilton Bayfront
San Diego, California

Presented by Margaret Crutchfield and Deborah Eldridge, CAEP advisors

Agenda for the Session

- Introduction to the session
- Answering the evidence questions for EPP-created assessments: Do's and don'ts
- Critiquing an assessment using the CAEP assessment rubric
- Using the reviewers' assessment checklist to critique your own assessment
- Debrief with Q&A
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.

Finding the 5 evidence questions for EPP-created assessment tools

• Aims.caepnet.org  Log-in = 24319  Password = caep
• Go to visit reports, then self-study evidence
• Scroll down to click on “add” and follow the prompts, selecting EPP-created assessment to upload your assessment tool.
• Click on the paper icon to find the five questions to answer.
Answering the 5 evidence questions:

**Question 1** (2000 characters)

- *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?*

  - *Sample response:* The observation rubric is introduced to candidates during Summer 1. It is used here for the instructors and the candidates to assess candidates' teaching in SSI. Not all of the standards on this rubric are appropriate in assessing the SSI teaching, so only certain standards are used.

- *Guidelines:*

  - Answer the question in full!
  - When referring to specific occurrences of any assessment, put in terms of semesters of the program. Fall, first semester or summer of first year. To cite occurrences as during “Summer I” does not provide enough context for reviewers who are not familiar with the program in its entirety.
  - Introduce acronyms when first used in each document. There is no guarantee the same person will be reading every submission or read each one in order.
  - Don't include statements that will raise unwanted questions!

---

**The rewrite**

- The rubric is used multiple times throughout the program. The rubric is introduced to candidates during their first summer in the program by faculty in the summer science course, which includes a lesson planning component. Candidates, in teams of three, are required to plan and teach one three-hour lesson to high school students who are participating in a three-week Museum summer program. The rubric is first used (by faculty and peers) to rate the teaching experiences, thus gaining familiarity with the tool.

- The rubric is primarily used by the program faculty who teach academic courses and supervise candidates in their school placements twice each month during the ten month residency (September through June) during visits to observe candidates in their school residency placements. Thus each candidate is assessed with the rubric approximately 10 times in the fall residency and 10 times in the spring residency.
Answering the 5 evidence questions:

Question 2 (2000 characters)

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

  *Sample response:* The assessment is used by supervising faculty and classroom mentors in their observations of candidates in their residency teaching. It is also used by course instructors. In a sense, then, the faculty trained themselves, as they were, fundamentally, the main developers of the rubric.

- Guidelines
  - Use the same name for an EPP-created assessment tool throughout the responses.
  - Describe roles (not specific people) when first mentioned (just like acronyms) in each document. Classroom mentors?
  - Be consistent question to question.
  - Don’t include statements that will raise unwanted questions!

---

The rewrite

- The assessment is used by the two supervising faculty in their observations of candidates in their school placements. Initial training on the use of the assessment occurred through a series of sessions that involved watching three videos of classroom teaching, determining together scores they would give to the teachers in the videos on particular standards, and sharing what evidence they saw in the videos to justify each score. This activity served to norm the understanding among the supervising faculty of the rubric standards. They came to consensus through this analysis of video as to what each rubric item on the Observation Rubric would look like in the classroom, including what basic, proficient, and accomplished performance levels would entail. When new supervising faculty come into the program, the norming process is repeated including the video analysis and joint observations of candidates’ teaching.
Answering the 5 evidence questions:

Question 3 (2000 characters)

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

  Sample response: The main intention of the observation rubric is to provide structure for assessing candidates' progression in their learning to teach over time. For this purpose, the assessment is used to follow candidates' development in planning, instruction, assessment and professionalism (becoming part of the school culture) across the academic year of their school-based residency.

- Guidelines:
  - Avoid vague terms and provide examples or specificity as often as possible.
  - Make sure that what you claim is what the tool is designed to do.

The rewrite

- The main intention of the Observation Rubric is to assess the development of candidates' abilities over time to plan, instruct, assess, and behave professionally according to each of the program standards. The rubric measures candidates' abilities to implement instruction, recognize and attend to students' needs, collaborate with colleagues to support students' learning, and reflect upon their own learning and teaching.

- The information gathered using the observation rubric over several months is used in the two Academic Progress meetings (January and June) at which candidates' progress through the program as a whole is evaluated. Using data from the most recent 10 observations, the Academic Advisory team (made up of three faculty) inform the candidate of her or his strengths in reference to the Rubric standards, Rubric standards that s/he has shown growth on, and those standards that s/he should focus on improving. Candidates are expected to average Basic across the rubric components by January and to have improved to Proficient by June.
Answering the 5 evidence questions:
Question 4 (2000 characters)

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

  Sample response: This observation rubric has “face validity” because all of the stakeholders (supervising faculty, other faculty, mentors and principals) agreed, after modifications, that it was measuring good teaching.

- Guidelines:
  - Describe in detail the process of alignment, training, development, validity and reliability. Whenever the word “how” appears, describe the process, not the outcome, in as great and specific detail as possible.

The rewrite

- Supervising faculty, Program Co-Directors, the program’s external evaluators, and other faculty were involved in development of the Observation Rubric. Principals in the partner residency schools also provided feedback that was incorporated into changes made between the first and second iterations.

- A preliminary version of the Observation Rubric was used in the fall of 2012. It was then revised substantially and the new version was piloted in the second semester (spring of 2013). Slight revisions were made after this pilot with additional input from supervising faculty and partner school administrators, and the resulting rubric has been used in its current form with subsequent cohorts.

- Because the rubric itself is grounded in two well-researched rubrics (the Danielson Framework for Teaching and the National Science Teachers Association preservice standards), it is well positioned for construct validity; the items in the rubric have been demonstrated to relate to teacher effectiveness.
Answering the 5 evidence questions:

Question 5 (2000 characters)

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

  *Sample response:* Have only two people using – most ever had was 4 people; we talk a lot. We are collecting data from rubric – 20 observations a year – into a data base – a semester, a year – we have examined these data for one year – judging the reliability – only way to really test is if supervising faculty watched same person at the same time. This shows development over time, though – but still! Doing pretty well.

- **Guidelines:**
  - Whenever possible, use professional language to describe processes: “piloted”, “established benchmarks”, “reviewed professional rubrics”, “came to consensus”, etc.

---

The rewrite

- **The assessment has clearly delineated performance levels intended to support uniform application over time. In addition, supervising faculty were trained to use the assessment in a consistent way and regularly communicate to check that their individual implementations remain consistent. Subsequent conversations with external evaluators, and selected observations in the field, have confirmed that their processes continue to be comparable.**

- **We have not yet determined inter-rater reliability for this assessment. But plan to conduct a series of faculty meetings in which faculty observe videos of candidates in their teaching residency, complete the observation rubric individually and then discuss differences in ratings. Inter-rater reliability will be determined and differences discussed in order to ensure consistency across raters.**
Feedback and Question Pause

CAEP Assessment Rubric

- Recently revised—now posted on the CAEP website/Accreditation and Program Review/CAEP Accreditation Resources
- Aligned with Section 6 of the Evidence Guide
Let's look at an example

- Read over the example

Administration and Purpose

- When is it administered? Is it administered more than once?
- What is the purpose of the assessment?
- How is it used in candidate monitoring?
- Is it tagged to standards? Which standards?
Informing Candidates

- When are candidates given a description of the assessment?
- Are instructions informative and unambiguous?
- Do candidates know the basis for how the assessment will be rated?

Content of Assessment

- Do the tasks assess explicitly identified aspects of standards?
- Does the task reflect an appropriate level of difficulty?
- Do most of the tasks required judges to consequential attributes of candidate proficiencies in the standards?
Scoring

- Do the proficiency level descriptions represent a developmental sequence from level to level?
- Are proficiency level attributes defined in actional, performance-based or observable behavior terms?
- Does the rubric provide feedback to candidates that is actionable?

Data Validity

- Is there a description or plan that details the steps the EPP is taking to ensure validity?
- Does the detail the types of validity that are or will be under investigation or established?
- Was the assessment piloted prior to administration?
- Does the EPP detail its current process or plans for analyzing and interpreting results?
- Do these generally meet accepted research standards?
Data Reliability

- Is there a description or plan that details the steps the EPP is taking to ensure reliability?
- How is the EPP training scorers and checking on inter-rater reliability
- Do these generally meet accepted research standards?

A word about surveys

- See the Assessment Rubric and pp. 25-27 of the Evidence Guide
- Expectations for reliability are substantially different
  - "Results from successive administrations are compared"
Evaluating your own assessment

• Using the CAEP Instrument Rubric, rate your assessment on at least two of the categories.
• Be ready to discuss the process and raise any issues that need further discussion or clarification.

Debrief and Q & A
CAEP INSTRUMENT RUBRIC

CAEP EVALUATION TOOL FOR EPP-CREATED ASSESSMENTS
USED IN ACCREDITATION

For use with: assessments created by EPPs including observations, projects/assignments and surveys
For use by: EPPs, CAEP assessment reviewers and Site Visitors

EXCERPT from the CAEP ACCREDITATION HANDBOOK on “Optional Early Instruments Evaluation”
Early in the accreditation process, providers can elect to submit to CAEP the generic assessments, surveys, and scoring guides that they expect to use to demonstrate that they meet CAEP standards... The purpose of this review is to provide EPP’s with formative feedback on how to strengthen assessments, with the ultimate goal of generating better information on its candidates and continuously improving its programs... This feature is a part of CAEP’s specialty/license area review under Standard 1.

<table>
<thead>
<tr>
<th>EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL</th>
<th>CAEP SUFFICIENT LEVEL</th>
<th>EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use or purpose are ambiguous or vague</td>
<td>1. ADMINISTRATION AND PURPOSE (informs relevancy)</td>
<td></td>
</tr>
<tr>
<td>• The point or points when the assessment is administered during the preparation program are explicit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• The purpose of the assessment and its use in candidate monitoring or decisions on progression are specified and appropriate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Evaluation categories or assessment tasks are tagged to CAEP, InTASC or state standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Purpose of assessment and use in candidate monitoring or decisions are consequential</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited or no basis for reviewers to know what information is given to respondents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructions given to respondents are incomplete or misleading</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The criterion for success is not provided or is not clear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited or no basis for reviewers to know what information is given to respondents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructions given to respondents are incomplete or misleading</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The criterion for success is not provided or is not clear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited or no basis for reviewers to know what information is given to respondents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructions given to respondents are incomplete or misleading</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The criterion for success is not provided or is not clear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category or task link with CAEP, InTASC or state standards is not explicit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category or task has only vague relationship with content of the standards being informed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category or task fails to reflect the degree of difficulty described in the standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation categories or tasks assess explicitly identified aspects of CAEP, InTASC or state standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation categories or tasks assess explicitly identified aspects of CAEP, InTASC or state standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation categories or tasks reflect the degree of difficulty or level of effort described in the standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation categories or tasks unambiguously describe the proficiencies to be evaluated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>When the standards being informed address higher level functioning, the evaluation categories or tasks require higher levels of intellectual behavior (e.g., create, evaluate, analyze, &amp; apply). For example, when a standard specifies that candidates’ students “demonstrate” problem solving, then the category or task is specific to students’ application of knowledge to solve problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Almost all evaluation categories or tasks (at least those comprising 95% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation categories or tasks assess explicitly identified aspects of CAEP, InTASC or state standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation categories or tasks reflect the degree of difficulty or level of effort described in the standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation categories or tasks unambiguously describe the proficiencies to be evaluated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>When the standards being informed address higher level functioning, the evaluation categories or tasks require higher levels of intellectual behavior (e.g., create, evaluate, analyze, &amp; apply). For example, when a standard specifies that candidates’ students “demonstrate” problem solving, then the category or task is specific to students’ application of knowledge to solve problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Almost all evaluation categories or tasks (at least those comprising 95% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

14
### CAEP INSTRUMENT RUBRIC

#### EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL
- Tasks not described or ambiguous
- Many evaluation categories or tasks (more than 20% of the total score) require judgment of candidate proficiencies that are of limited importance in CAEP, InTASC or state standards

#### CAEP SUFFICIENT LEVEL
- 80% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards

#### EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL

4. **SCORING** (informs reliability and actionability)
- The basis for judging candidate work is well defined
- Each proficiency level is qualitatively defined by specific criteria aligned with the category (or indicator) or with the assigned task
- 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 category or indicator

#### 5.a DATA VALIDITY
- 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
- The assessment was piloted prior to administration
- 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

#### 5.b DATA RELIABILITY
- 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

#### EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
- Higher level actions from Bloom’s taxonomy are used such as “analysis” or “evaluation”
<table>
<thead>
<tr>
<th>EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL</th>
<th>CAEP SUFFICIENT LEVEL</th>
<th>EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>whether reliability is being investigated or how.</td>
<td>reliability of the data from the assessment</td>
<td>master criteria and are periodically formally checked to maintain calibration at levels meeting accepted research standards</td>
</tr>
<tr>
<td>Described steps to not meet accepted research standards for reliability.</td>
<td>• Training of scorers and checking on inter-rater agreement and reliability are documented</td>
<td></td>
</tr>
<tr>
<td>No evidence, or limited evidence, is provided that scorers are trained and their inter-rater agreement is documented.</td>
<td>• The described steps meet accepted research standards for establishing reliability</td>
<td></td>
</tr>
</tbody>
</table>

**WHEN THE INSTRUMENT IS A SURVEY:**

Sections 1 and 2, above apply as worded. Substitute the following for sections 3, 4 and 5.

<table>
<thead>
<tr>
<th>SURVEY CONTENT</th>
<th>DATA QUALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Individual item are ambiguous or include more than one subject</td>
<td>• Scoring is anchored in performance or behavior demonstrably related to teaching practice</td>
</tr>
<tr>
<td>• Items are stated as opinions rather than as behaviors or practices</td>
<td>• Dispositions surveys make an explicit connection to effective teaching</td>
</tr>
<tr>
<td>• Dispositions surveys provide no explanations of their purpose</td>
<td>• EPP provides evidence of survey construct validity derived from its own or accessed research studies</td>
</tr>
<tr>
<td>• Scaled choices are numbers only, without qualitative description linked with the item under investigation</td>
<td>• An even number of scaled choices helps prevent neutral (center) responses</td>
</tr>
<tr>
<td>• Limited or no feedback provided to candidates</td>
<td>• Scaled choices are qualitatively defined using specific criteria aligned with key attributes identified in the item</td>
</tr>
<tr>
<td>• No evidence that questions are piloted</td>
<td>• Feedback provided to candidates is actionable</td>
</tr>
<tr>
<td></td>
<td>• EPP provides evidence that questions are piloted to determine that candidates interpret them as intended and modifications are made, if called for</td>
</tr>
<tr>
<td></td>
<td>• EPP provides evidence that candidate responses are compiled and tabulated accurately</td>
</tr>
<tr>
<td></td>
<td>• Interpretations of survey results are appropriate for the items and resulting data</td>
</tr>
<tr>
<td></td>
<td>• Results from successive administrations are compared (for evidence of reliability)</td>
</tr>
</tbody>
</table>
### CAEP INSTRUMENT RUBRIC

**CHECKLIST**

<table>
<thead>
<tr>
<th>Item Category</th>
<th>Below Adequate</th>
<th>CAEP Adequate Level</th>
<th>Above Adequate</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>ADMINISTRATION AND PURPOSE:</strong> Point when instrument is administered in the program; its purpose, and standards addressed (informs relevance). Evaluation categories or assessment tasks are tagged to CAEP, InTASC or state standards</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. <strong>INFORMING RESPONDENTS:</strong> Information given to respondent before and at the administration of the instrument (informs fairness and reliability); basis for judging candidate performance is explicit.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. <strong>CONTENT OF ASSESSMENT:</strong> Evaluation categories explicitly linked with standards, reflect degree of difficulty in standards, and unambiguously describe proficiencies to be evaluated; when standards include higher level functioning, the evaluation categories explicitly require higher levels of intellectual behavior; most evaluation categories require judgment of consequential candidate proficiencies (informs relevancy).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. <strong>SCORING:</strong> Basis for judging candidate work is well defined; each proficiency level is qualitatively defined by criteria aligned with the category; proficiency descriptions represent a developmental sequence from level to level and are defined in actionable, performance-based or observable behavior terms; feedback for candidates is actionable (informs reliability and actionability).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. a. <strong>DATA VALIDITY:</strong> EPP provides a description or plan that details steps to ensure validity of the assessment and its use; assessment was piloted prior to administration; EPP details process or plans for analyzing and interpreting results. b. <strong>DATA RELIABILITY:</strong> EPP provides a description or plan that details steps to ensure reliability of the assessment; training of scorers and checking inter-rater agreement and reliability are documented.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**WHEN THE INSTRUMENT IS A SURVEY:**
Sections 1 and 2, above apply as worded. Substitute the following for sections 3, 4 and 5.

a. **SURVEY CONTENT:** survey items explicitly aligned with EPP mission and CAEP, InTASC or state standards; questions have a single subject, use unambiguous language; leading questions are avoided; items stated as behaviors or practices rather than opinions.

b. **DATA QUALITY:** Even number of scale choices prevents neutral responses; scaled choices are qualitatively defined using criteria aligned with key attributes identified in the item; feedback provided to candidates is actionable; questions are piloted to ensure intended interpretation; interpretations of results appropriate for items and data.

**OVERALL – How would you rate this assessment?**

Provide a rationale for your overall rating: