From CAEP

The principles for measures used in the CAEP accreditation process include: (a) validity and reliability, (b) relevance, (c) verifiability, (d) representativeness, (e) cumulativeness, (f) fairness, (g) stakeholder interest, (h) benchmarks, (i) vulnerability to manipulation, and (j) actionability. CAEP requires valid and reliable assessments to demonstrate candidate quality and that various stakeholders must contribute to the validity of the assessments. Validity and reliability are two of the most important criteria when assessing instruments. Reliability means consistency and a test is valid if it measures what it is supposed to measure.

Before we start, here is a reminder of CAEP Standard 5 and component 5.2.

<table>
<thead>
<tr>
<th>Standard 5: Provider Quality, Continuous Improvement, and Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>The provider maintains a quality assurance system comprised of valid data from multiple measures, including evidence of candidates’ and completers’ positive impact on P-12 student learning and development. The provider supports continuous improvement that is sustained and evidence-based, and that evaluates the effectiveness of its completers. The provider uses the results of inquiry and data collection to establish priorities, enhance program elements and capacity, and test innovations to improve completers’ impact on P-12 student learning and development.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Components: Quality and Strategic Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2 The provider’s quality assurance system relies on relevant, verifiable, representative, cumulative and actionable measures, and produces empirical evidence that interpretations of data are valid and consistent.</td>
</tr>
</tbody>
</table>

According to CAEP’s Evidence Guide, the responsibility lies with the EPP to provide valid (and reliable) evidence. CAEP is committed to stronger preparation and accreditation data. The profession needs evidence that assessment is intentional, purposeful, and addresses deliberately posed questions of importance. Such reporting entails interpretation and reflection; measures need to be integrated and holistic; and approaches to its assessment can be qualitative and quantitative, and direct and indirect.

ALL EPP created assessments used in the CAEP review must meet the Sufficient level on the CAEP Instrument rubric.

---

1 Ewell, P. (2013). *Principles for measures used in the CAEP accreditation process.*
Submission for EPP-Created instruments

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

How to address reliability and validity

MSU EPP self-studies needs to include evidence related to the reliability and validity of the reported data. Reliability and validity are frequently measured quantitatively.

EPP’s quantitative approach to assess the reliability of instruments can involve four facets:

1. Supervisor (e.g., inter-rater reliability, internal consistency, bias)
2. Candidate (e.g., distribution of ratings)
3. Item (e.g., variability of items)
4. Time (e.g., variability of candidate performance across time)

Before we review quantitative ways of assessing validity, we should consider that describing the reliability and validity of EPP created instruments has not been an easy task for institutions completing their CAEP self-studies. CAEP has recommended the inclusion of both quantitative and qualitative approaches when describing reliability and validity. The information provided next comes from the Breaux and Elliot’s presentation: Reliability and Validity: Establishing and Communicating Trustworthy Findings (CAEP Spring Conference 2015).

“Although the terms validity and reliability are traditionally associated with quantitative research, CAEP does not mean to imply that only quantitative data are expected or valued.”
Differences between quantitative and qualitative methods when establishing validity and reliability:

<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>These methods of establishing validity and reliability are easier to describe briefly.</td>
<td>These methods of establishing validity and reliability depend much more on anticipating and disconfirming a variety of potential doubts that various readers could have.</td>
</tr>
<tr>
<td>The standards for judging the results are less subjective.</td>
<td>The process takes more effort, and the reader’s judgment is less predictable.</td>
</tr>
<tr>
<td>They require statistical literacy, and the results are decontextualized.</td>
<td>They require strong skills in logical argumentation and expository writing, but the results are more contextualized.</td>
</tr>
</tbody>
</table>

**Strategies Used in CAEP Self-Studies: Reliability**

1. Quantitative studies explain how they manage subjectivity using common terminology, standard procedures, and uniform formats for reporting results. We need to make sure the correct procedures are selected, conducted properly, interpreted validly, and communicated clearly.

Focus on key reliabilities. Use inter-rater correlations for large. Use rater agreement for small samples.

<table>
<thead>
<tr>
<th>The instrument’s reliability section on CAEP’s Instrument Rubric addresses the following:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Degree in which an assessment produces stable and consistent results</td>
</tr>
<tr>
<td>• Ask the question - Can the evidence be corroborated? o Criteria</td>
</tr>
<tr>
<td>o A detailed description or plan is provided</td>
</tr>
<tr>
<td>o Training of scorers and checking on inter-rater reliability are documented</td>
</tr>
<tr>
<td>o Steps are described that meet accepted research standards for establishing interrater reliability</td>
</tr>
</tbody>
</table>

**Strategies Used in CAEP Self-Studies: Validity**

It is not necessary to establish every form. Some of the processes are qualitative and involve demonstrating alignment. Others are quantitative and involve calculating values.

Quantitative methods to assess validity;

Focused on key validities:

- Content: all relevant elements of the construct are measured
- Construct: measures intended attribute
• Criterion: measures of attributes predict target behaviors
   o Concurrent: correlates with a known good measure
   o Predictive: predicts score on a significant future measure

• Convergent: measure correlates with related measures

The types of validity that are needed are judgment calls that have to be justified in the self-study’s rationale. However, content validity and construct validity should be included.

Example

When cooperating teachers and university supervisors rate a candidate, we need to show that the assessment is a valid measure of the construct or constructs; and that both raters understand the items and overall intent in the same way.

To show the assessment is a valid measure:

• Expert judgment: what do university supervisors and cooperating teachers say?
• Alignment with relevant standards
• Agreement with logically-related measures
• Is there sufficient variance in the evidence?

To show the assessment is a reliable measure:

• Inter-rater agreement

More on Content Validity

Following Dr. Stevie Chepko’s view, there are three important components to establish content validity:

1. Determining the body of knowledge for the construct to be measured. The agreement among “experts” requires the use of recognized subject matter experts and it is based on their judgment. It also relies on individuals who are familiar with the construct such as faculty members, EPP based clinical educators, and/or P-12 based clinical educators. The key is having them answer the fundamental question: “Do the indicators assess the construct to be measured?”

2. Aligning indicators to construct. Indicators must assess some aspects or segment of the construct and indicators must align with the construct.

3. Using Lawshe’s Content Validity Ratio
4. Lawshe’s Content Validity Ratio (CVR)

Performance domains:

- Behaviors that are directly observable
- Can be a simple proficiencies
- Can be higher mental process (inductive/deductive reasoning)
- Operational definition – Extent to which overlap exists between (1) performance on assessment under investigation, and (2) ability to function in the defined job
- Attempts to identify the extent of the overlap

The Content Evaluation Panel is composed of persons knowledgeable about the job, and it is most successful when it is a combination of P-12 based clinical educators, EPP based clinical educators, and faculty. Each panel member is given the list of indicators or items independently and are asked to do the following:

- Rate the item as “essential”, “useful but not essential”, or “not necessary.”
- Items/indicators must be aligned with the construct being measured

To quantifying consensus, any item/indicator which is perceived as “essential” by more than half of the panelists, has some degree of content validity. The more panelist (beyond 50%) who perceive the indicator as “essential,” the greater the extent or degree of its content validity. Calculating the content validity ratio (CVR)

\[
CVR = \frac{n_e - \left( \frac{N}{2} \right)}{\frac{N}{2}},
\]

where:
- \(n_e\) = number of panelists indicating “essential”
- \(N\) = total number of panelists

Compare answer with CVR chart to determine CVR value based on the number of panelists. CVR is calculated for each indicator, and minimum value of the CVR is based on the number of panelists and is on a CVR Table. Keep or reject individual items based on the table results.

CVR values range from -1.0 to +1.0. The more panelists, the lower the CVR value. For example,

- 5 panelists requires minimum CVR value of .99
- 15 panelists requires minimum CVR value of .60
- 40 panelists requires minimum CVR value of .30
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

The instrument’s validity section on CAEP’s Instrument Rubric addresses the following:

- 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
- Lawshe’s method (CVR)

Questions to Be Answered for each Submitted EPP-Created Instrument

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?
3. What is the intended use of the assessment and what is the assessment purported to measure?
4. Please describe how validity/trustworthiness was established for the assessment.
5. Please describe how reliability/consistency was established for the assessment.

MSU EPP Homegrown Instruments

We currently use seven EPP-created instruments to assess education programs at MSU. We need to determine the adequacy of these measures for the accreditation process. The measures are:

1. Experiential Log
2. Candidate Professional Disposition Traits
3. Missouri Educator Evaluation System (MEES) Rubric
4. Diversity Proficiencies
5. Comprehensive Exam Assessment Rubric for Advanced Programs
6. Research Rubric for Advanced Programs
7. Student teaching exit survey (based on a proprietary survey)
8. EDC345 Multiculturalism Lesson Plan

Next Steps

1. Revisit the instruments.
   - Does it address critical elements required by CAEP? Which ones?

2. Make necessary adjustments.
   - Define in specific terms what should be addressed and assessed
   - Align with program objectives, CAEP, InTASC and/or state standards
   - Clarify language make sure we include distinguishable and measurable statements in rubrics

3. Use CAEP’s Assessment Rubric as a guide
4. Establish content validity.
   - Create Content Evaluation Panels for each instrument composed of P-12 based clinical educators, EPP based clinical educators, and faculty.
   - Through a survey, ask them to rate the items as “essential”, “useful but not essential”, or “not necessary.”
   - Calculate the Lawshe’s Content Validity Ratio (CVR) for each item of the assessment to determine which items will remain.
   - Focus groups to discuss the content of each instrument

5. Establish inter-rater reliability.
   - Use spring 2017 data as a “pilot” of the instruments
   - When possible, instructors within a program should score at least 3 samples independently of one another
   - Collect results and calculate the percentage of agreement on each component & submission
- If scores vary and yield <80% agreement, meet to discuss each item score on each submission
- Pay attention to:
  - Discrepancies between/among scorers
  - Whether discrepancies are due to language or how items are defined
- Resolving discrepancies with clarified language, rearranging items, or other changes
- Make note of these changes and revise the assessment as necessary
- If substantial changes are necessary, each instructor should score at least 2 work samples independently of one another; until instructors reach at least 80% agreement
- When possible triangulate (compare cooperating teacher and supervisor results)
- When possible, compare results over time (cohort and panel-wise)
- When summarizing reliability, try to include data on the following:
  - Supervisor (e.g., inter-rater reliability, internal consistency, bias)
  - Candidate (e.g., distribution of ratings)
  - Item (e.g., variability of items)
  - Time (e.g., variability of apprentice performance across time)