CAEP Council for the Accreditation of Educator Preparation

Reliability and Validity:

Establishing and Communicating Trustworthy Findings

Spring Conference, 2015

Glenda Breaux, Formative Evaluation Specialist for the Inquiry Brief Pathway

Emerson Elliott, Director, Special Projects

Session Description

- A toolkit of qualitative and quantitative strategies and capacities for faculty to meet CAEP's expectations for reliable and valid evidence.

Welcome!

- What makes you proud of your program completers?
- What convinces you that your pride is justified? What is the evidence?
- Could you convince an outsider that you are right? How would you do it?
- Would the case hold up in a professional journal?

Agenda for the session

- Expectations for evidence:
  CAEP Standard 5 (Component 5.2)
- Reliability
- Validity
- Practical Application

The Context

- Self-studies must include evidence about the reliability and validity of the reported evidence.
- Many strategies are available to establish them. Often, the simpler the strategy, the better.

Common Types of Evidence Used

- Grades (content major, pedagogy, & clinical; course embedded assignments as assessments)
- Scores on Standardized tests (candidates’ entrance, exit, and license scores; completers’ own pupils’ test scores)
- Surveys – pupils, candidates, completers, employers
- Ratings – portfolios, work samples, cases
- Rates – hiring/tenure, certification, graduate study, awards, publications, NBPTS cert, etc.

Common Challenges Faced by EPPs

- Designing observation instruments in ways that maximize construct validity and inter-rater reliability
- Assessing reliability and validity when sample sizes are small.
- Determining which, and how many, of the many possible validity and reliability analyses to run?

Expectations for Evidence: Standard Five

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. AND…

Standard 5: Provider Quality Assurance and Continuous Improvement

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.

Strategies: Quantitative and Qualitative Approaches

- 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, or that quantitative methods must be used to establish validity and reliability.

Issues to Consider: Quantitative Methods

- These methods of establishing validity and reliability are easier to describe briefly
- The standards for judging the results are less subjective, but
- They require statistical literacy, and the results are decontextualized

Issues to Consider: Qualitative Methods

- These methods of establishing validity and reliability depend much more on anticipating and disconfirming a variety of potential doubts that various readers could have
- The process takes more effort, and the reader’s judgment is less predictable
- They require **strong skills in logical argumentation and expository writing**, but
- The results are more contextualized

For Example

- EPPs sometimes use different types of observers or raters to gather information about candidates through different lenses.
- The different lenses through which these experts view the same candidate can provide a broader and well-rounded picture of candidate capacities, but can wreak havoc on inter-rater reliability values.

- Forcing all observers to have exactly the same perspective can undermine the value of having different types of observers participate in candidate evaluation.
- The results of their ratings may be more appropriately treated as multiple sources of data than as multiple ratings. In such cases, triangulation would be a better strategy of establishing reliability than inter-rater correlations, and EPPs have successfully argued this when the argument was **clear, explicit,** and **credible**.

Making the case for dependability and credibility

- The elements of the conceptual frameworks from which each type of observer is operating must be specified.
  - How are they similar and different?
  - In what areas should you expect them to view performances similarly? There should be at least a few that are embodiments of the program's vision?
    - Can they use a rubric similarly to rate those performances? If they display commonality on key universal elements, variation in perspective-based ratings is a less of a threat to reliability. But this requires establishing credibility via calibration on universal items, and parsing the instrument and sufficiently justifying that parsing for other items.

- It also involves documenting why the observers are considered to be experts and competent raters.
  - Do experts from the same category offer the same comments or ratings about instances of candidate performance (whether live or videotaped)?
    - If the results of coding or calibration exercises show this, there is a justifiable basis for accepting future ratings from a single observer or for rescoring only a small sample of measures.
    - If observers compare well within categories, this supports the claim of perspective-based difference and justifies triangulation between categories rather than inter-rater reliability calculations between categories.

- Are their observations stable across candidates and over time?
  - What evidence can you offer that they use reasonably similar criteria to evaluate each candidate, and are not swayed by preferences that are not central to competence and that should be allowed to vary between candidates.
  - What evidence can you offer that they evaluate based on the target behavior and don’t raise or lower the bar for current candidates based on the performance of past candidates?

Quantitative Strategies Used in CAEP Self-Studies: Reliability

- Quantitative studies explain how they manage subjectivity using common terminology, standard procedures, and uniform formats for reporting results.
  - This reduces the narrative burden, but still requires that the correct procedures are selected, conducted properly, interpreted validly, and communicated clearly.

For example

- There are multiple methods for calculating correlations or associations.
- Pearson's *r* may or may not be the appropriate calculation for the situation. It is important to attend to context and the assumptions of the test. It may be that Spearman's *rho*, Kendall's *tau*, one of the *kappa* calculations, or an intraclass correlation is more sensible.
- All of these processes will produce a value between -1 and 1, but that doesn't necessarily mean the same thing or have the same implications.

Quantitative Strategies Employed in CAEP Self-Studies: Reliability

- Focus on key reliabilities
  - Inter-rater correlations (large samples)
    - e.g., reported a Pearson's *r* or Spearman's *rho* for multiply scored measures at or above .80
  - Rater agreement (small samples)
    - e.g., reported percentage of exact agreement and percentage of agreement within 1 point on an ordinal scale ≥ 80%, AND percentage (≥80%)of score pairs leading to the same practical decision (pass/fail)

Audience Participation: Challenges and Questions

- Before we wrap up with the discussion of reliability, are there any questions or comments about:
  - Reliability, or
  - CAEP's perspective on it?

Validity

- As with reliability, validity has several features and there are several ways to establish it
  - It isn't necessary to establish every form
    - Some of the processes are qualitative and involve demonstrating alignment
    - Others are quantitative and involve calculating values

Strategies Employed in CAEP Self-Studies

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

Please Note:

- Many other types of validity exist.
  - See, for example, http://www.southalabama.edu/coe/bset/johnson/lectures/lec8.htm
- It is not necessary to assess every type.
  - Which types to assess, and how many, are judgment calls that need to be justified in your rationale, but content validity and construct validity are good places to start.

The question about evidence:

How can the program faculty be certain that their measures are reliable and that their interpretation of results is valid?

How do you get there?

Be clear on your ***rationale***:

For each measure you use,

- what is it meant to show, and
- why does the faculty believes this particular assessment is an appropriate and meaningful measure of student performance (or program performance?)

How do you get there?

Know how good is good enough:

For each measure, establish <u>your</u> success criterion:

- Establish a reasoned and empirical basis for the standard or criterion of success (the cut score)
- Test that criterion empirically—how do you know it's right?

Example: Performance Rating

Imagine that interns are evaluated by their mentor teacher and a faculty supervisor. How might we show that:

- The rating form gets at the right stuff? (Is it a valid measure of the construct or constructs?)
- Both raters understand the items and overall intent in the same way? (Do independent raters use the instrument consistently?)

Example: Performance Rating

Does our rating form get at the right stuff?

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

Is the instrument used consistently and does it produce consistent results reliably?

- Inter-rater agreement (**perhaps** correlation)
- Calibration exercises (let’s watch a video…)

Handle data purposefully!

- All measures for individual candidates / completers should be linked.
- Arguments for the validity of interpretations are enhanced by convergent or predictive validity, as the case requires
- Linking measures enables identification of “pressure points” and learning from actions
- Data from any single measure has limited “reach”

Who does what?

- The good news: there's nothing here we don't already know how to do
- The challenge: distributing responsibility without creating "silos" of expertise
- A first step: setting an 'assessment agenda' with clear goals and priorities
- A final step: describing the process with respect to the perspective of someone who wasn't involved.

- Glenda Breaux, Formative Evaluation Specialist for the Inquiry Brief Pathway

  glenda.breaux@caepnet.org

- Emerson Elliott, Director, Special Projects

  emerson.elliott@caepnet.org

Feedback Opportunity

- Engaged feedback is vital to CAEP. You will have an opportunity to complete a survey at the end of the conference. Surveys will be sent via email on Friday, April 10. We encourage your participation.

Thank You!