Mississippi Valley State University

Teacher Education Department



Validity and Reliability Plan

2019-2020

Spring 2020 implementation

**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) action ability . 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.

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.

**How to address reliability and validity**

MVSU 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)

**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
* Concurrent: correlates with a known good measure
* 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

Following Dr. Stevie Chepko’s (previous CAEP President) 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 is directly observable

• Can be simple proficiencies

• Can be higher mental process (inductive/deductive reasoning)

• Operational definition – Extent to which overlap exists between

* performance on assessment under investigation, and
  + 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) ne = 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.

**Calculating the content validity ratio**

**(CVR) CVR = (ne – n/2)/ (n/2)**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Panel size** | **N(critical) Minimum number of Experts required to agree on an item essential for inclusion** | **Proportion agreeing**  **Essential** | **CVR critical** |
| **5** | **5** | **1** | **1.00** |
| **6** | **6** | **1** | **1.00** |
| **7** | **7** | **1** | **1.00** |
| **8** | **7** | **.875** | **.750** |
| **9** | **8** | **.889** | **.770** |
| **10** | **9** | **.900** | **.800** |
| **11** | **9** | **.818** | **.636** |
| **12** | **10** | **.833** | **.667** |
| **13** | **11** | **.769** | **.538** |
| **14** | **11** | **.786** | **.571** |
| **15** | **12** | **.800** | **.600** |
| **16** | **12** | **.750** | **.500** |
| **17** | **13** | **.765** | **.529** |
| **18** | **13** | **.722** | **.444** |
| **19** | **14** | **.737** | **.474** |
| **20** | **15** | **.714** | **.429** |
| **21** | **15** | **.714** | **.429** |
| **22** | **16** | **.727** | **.455** |
| **23** | **16** | **.696** | **.391** |
| **24** | **17** | **.708** | **.417** |
| **25** | **18** | **.720** | **.440** |
| **26** | **18** | **.692** | **.385** |
| **27** | **19** | **.704** | **.407** |
| **28** | **19** | **.679** | **.357** |
| **29** | **20** | **.690** | **.379** |
| **30** | **20** | **.667** | **.355** |
| **31** | **21** | **.677** | **.355** |
| **32** | **22** | **.688** | **.375** |
| **33** | **22** | **.667** | **.333** |
| **34** | **23** | **.676** | **.353** |
| **35** | **23** | **.657** | **.314** |
| **36** | **24** | **.667** | **.333** |
| **37** | **24** | **.649** | **.297** |
| **38** | **25** | **.658** | **.316** |
| **39** | **26** | **.667** | **.333** |
| **40** | **26** | **.650** | **.300** |

**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.

**Establishing Validity & Reliability for EPP Key Assessments**

**Directions**

1. Revisit the assessment. Goal #1 is to re-evaluate the assessment.

• Is it appropriately aligned with your purpose?

• Is it current? Was it handed down to you from your predecessor?

• Does it address critical elements required by CAEP? Which ones?

1. Make necessary adjustments. Goal #2 is to make the assessment what it should be.

• Define in very specific terms what should be addressed & assessed

• Align with course objectives, program objectives, CAEP standards

• Add, change, clarify, or omit components to achieve alignment

• Clarify language very carefully to include distinguishable & measurable units

1. Establish content validity. Goal #3 is to agree upon which components are necessary & appropriate for the assessment.

• Submit to colleagues who utilize the assessment & explain any changes you’ve made to it

• Follow Salkind’s (2013) guidelines in chapter 4 for establishing content validity (This could be completed individually or in a group setting with discussion)

• Calculate the responses for each item of the assessment to determine which items will remain • Insert the content validity results on the CAEP sufficiency rubric

1. Establish inter-rater reliability. Goal #4 is to yield similar item scores on each submission regardless of the instructor.

• Practice utilizing the new rubric to score student submissions

• Each instructor should score at least 3 samples independently of one another

• Collect results and calculate the percentage of agreement on each component & submission

• Follow Salkind’s (2013) guidelines in chapter 3 for establishing reliability (This could be completed individually or in a group setting with discussion)

• If scores vary & yield<80% agreement, meet to discuss each item score on each submission

* To reach the 80%, pay close attention to the following:

1. Discrepancies between/among scorers
2. Whether discrepancies are due to language or how items are defined.
3. 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; the group will then follow buts # 3-7 again
* When instructors reach at least 80% agreement, report that data on the CAEP sufficiency rubric.

**References:**   
1. (1) Ewell, P. (2013). Principles for measures used in the CAEP accreditation process.

2. CAEP-Validity-Reliability-Directions.pdf

3. [www.caepnet.org](http://www.caepnet.org)

4. Salkinds, Neil.J (2013); Tests & Measurement for People Who (Think They) Hate Tests & Measurement 2nd Ed. Sage Publication.