Council for the Accreditation of Educator Preparation

CAEP EVALUATION FRAMEWORK FOR EPP-CREATED ASSESSMENTS

For use with: Educator preparation provider (EPP)-created assessments, including subject and pedagogical content tests, observations, projects, assignments, and surveys

For use by: EPPs to evaluate their own assessments and by CAEP site teams to review evidence in self-study submissions

CAEP uses the term "assessments" to cover content tests, observations, projects or assignments, and surveys. All of these assessment forms are used with candidates. Surveys are often used to gather evidence on aspects of candidate preparation and candidate perceptions about their own readiness to teach. Surveys are also useful to measure the satisfaction of graduates or employers with preparation and the perceptions of clinical faculty about the readiness of EPP completers.

Assessments and scoring guides are used by faculty to evaluate candidates and provide them with feedback on their performance. Assessments and scoring guides should address relevant and meaningful attributes of candidate knowledge, performance, and dispositions, aligned with standards. Most assessments that comprise evidence offered in accreditation self-study reports will probably be used by an EPP to examine candidates consistently at various points from admission through exit. These are assessments that all candidates are expected to complete as they pass from one stage of preparation to the next, or that are used to monitor progress of candidates' developing proficiencies during one or more stages of preparation.

CAEP site teams will follow the guidelines in this evaluation tool and it can also be used by EPPs when they design, pilot, and judge the adequacy of the assessments they create.

EXAMPLES OF ATTRIBUTES **BELOW** SUFFICIENT LEVEL	**CAEP SUFFICIENT LEVEL**	EXAMPLES OF ATTRIBUTES **ABOVE** SUFFICIENT LEVEL
– a. Use or purpose are ambiguous or vague. b. There is limited or no basis for reviewers to know what information is given to candidates. c. Instructions given to candidates are incomplete or misleading. d. The criterion for success is not provided or is not clear.	**1. ADMINISTRATION AND PURPOSE** (informs relevancy) a. The point or points when the assessment is administered during the preparation program are explicit. b. The purpose of the assessment and its use in candidate monitoring or decisions on progression are specified and appropriate. c. Instructions provided to candidates (or respondents to surveys) about what they are expected to do are informative and unambiguous. d. The basis for judgment (criterion for success, or what is "good enough") is made explicit for candidates (or respondents to surveys). e. Evaluation categories or assessment tasks are aligned with CAEP, InTASC, national/professional and state standards.	**+** a. The purpose of the assessment and its use in candidate monitoring or decisions are consequential. b. Candidate progression is monitored and information is used for mentoring. c. Candidates are informed how the instrument results are used in reaching conclusions about their status and/or progression.
–	**2. CONTENT OF ASSESSMENT** (informs relevancy) a. Indicators assess explicitly identified aspects of CAEP, InTASC, national/professional and state standards.	**+**

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
a. Indicator alignment with CAEP, InTASC, national/ professional or state standards is incomplete, absent or only vaguely related to the content of standards being evaluated. b. Indicators fail to reflect the degree of difficulty described in the standard. c. Indicators not described, are ambiguous, or include only headings. d. Higher level functioning, as represented in the standards, is not apparent in the indicators. e. Many indicators (more than 20% of the total score) require judgment of candidate proficiencies that are of limited importance in CAEP, InTASC, national/professional, and/or state standards.	b. Indicators reflect the degree of difficulty or level of effort described in the standards. c. Indicators unambiguously describe the proficiencies to be evaluated. d. When the standards being informed address higher level functioning, the indicators require higher levels of intellectual behavior (e.g., create, evaluate, analyze, & apply). For example, when a standard specifies that candidates' students "demonstrate" problem solving, then the indicator is specific to candidates' application of knowledge to solve problems. e. Most indicators (at least those comprising 80% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards. [NOTE: the word "indicators" is used as a generic term for assessment items. For content tests, the term refers to a question. For projects or assignments, it refers to a prompt or task that the candidate is to perform. For an observation, an indicator might be a category of performance to observe or a specific aspect of candidate performance that a reviewer would record. For a survey, an indicator would stand for a question or statement for which a response is to be selected.]	a. Almost all indicators (95% or more of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards.
- a. Rating scales are used instead of rubrics; e.g., "level 1= significantly below expectation" "level 4 = significantly above expectation." b. Proficiency Level Descriptors (PLDs) do not align with indicators. c. PLDs do not represent developmental progressions. d. PLDs provide limited or no feedback to candidates specific to their performance. e. Proficiency level descriptors are vague or not defined, and may just	**3. SCORING** (informs reliability and actionability) a. The basis for judging candidate performance is well defined. b. Each Proficiency Level Descriptor (PLD) is qualitatively defined by specific criteria aligned with indicators. c. PLDs represent a developmental sequence from level to level (to provide raters with explicit guidelines for evaluating candidate performance and for providing candidates with explicit feedback on their performance). d. Feedback provided to candidates is actionable—it is directly related to the preparation program and can be used for program improvement as well as for feedback to the candidate. e. 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.]	**+** a. Higher level actions from Bloom's or other, taxonomies are used in PLDs such as "analyzes" or "evaluates."

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
repeat the language from the standards.		
- a. Description of or plan to establish reliability does not inform reviewers about how it was established or is being investigated. b. Described steps do not meet accepted research standards for reliability. c. No evidence, or limited evidence, is provided that scorers are trained, and their inter-rater agreement is documented. d. Described steps do not meet accepted research standards for reliability.	**4. DATA RELIABILITY** a. 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. b. Training of scorers and checking on inter-rater agreement and reliability are documented. c. The described steps meet accepted research standards for establishing reliability.	+ a. Raters are initially, formally calibrated to master criteria and are periodically formally checked to maintain calibration at levels meeting accepted research standards. b. A reliability coefficient is reported.
- a. Description of or plan to establish validity does not inform reviewers about how it was established or is being investigated. b. The type of validity established or investigated is miss-identified or not described. c. The instrument was not piloted before administration. d. Process or plans for data analysis and interpretation are not presented or are superficial. e. Described steps do not meet accepted research standards for establishing validity. For example, validity is determined through an internal	**5. DATA VALIDITY** a. 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. b. 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. c. If the assessment is new or revised, a pilot was conducted. d. The EPP details its current process or plans for analyzing and interpreting results from the assessment. e. The described steps meet accepted research standards for establishing the validity of data from an assessment.	+ a. Types of validity investigated go beyond content validity and move toward predictive validity. b. A validity coefficient is reported.

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
review by only one or two stakeholders.		
WHEN THE INSTRUMENT IS A SURVEY: **Use Sections 1 and 2, above, as worded and substitute sections 6 and 7, below for sections 3, 4 and 5.**		
- a. Questions or topics are not aligned with EPP mission or standards. b. Individual items are ambiguous or include more than one subject. c. There are numerous leading questions. d. Items are stated as opinions rather than as behaviors or practices. e. Dispositions surveys provide no evidence of a relationship to effective teaching.	**6. SURVEY CONTENT** a. Questions or topics are explicitly aligned with aspects of the EPP's mission and also CAEP, InTASC, national/professional, and state standards. b. Individual items have a single subject; language is unambiguous. c. Leading questions are avoided. d. Items are stated in terms of behaviors or practices instead of opinions, whenever possible. e. Surveys of dispositions make clear to candidates how the survey is related to effective teaching.	+ a. Scoring is anchored in performance or behavior demonstrably related to teaching practice. b. Dispositions surveys make an explicit connection to effective teaching.
- a. Scaled choices are numbers only, without qualitative descriptions linked with the item under investigation b. Limited or no feedback provided to the EPP for improvement purposes c. No evidence that questions/items have been piloted	**7. SURVEY DATA QUALITY** a. Scaled choices are qualitatively defined using specific criteria aligned with key attributes. b. Feedback provided to the EPP is actionable. c. EPP provides evidence that questions are piloted to determine that candidates interpret them as intended and modifications are made if called for.	+ a. EPP provides evidence of survey construct validity derived from its own or accessed research studies.

Criteria listed below are evaluated during the stages of the accreditation review and decisionmaking:

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