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 educator preparation providers (EPPs) 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.

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
– • Use or purpose are ambiguous or vague	**1. ADMINISTRATION AND PURPOSE** (informs relevancy) • The point or points when the assessment is administered during the preparation program are explicit • The purpose of the assessment and its use in candidate monitoring or decisions on progression are specified and appropriate • Evaluation categories or assessment tasks are tagged to CAEP, InTASC or state standards	**+** • Purpose of assessment and use in candidate monitoring or decisions are consequential
• Limited or no basis for reviewers to know what information is given to respondents • Instructions given to respondents are incomplete or misleading • The criterion for success is not provided or is not clear	**2. INFORMING CANDIDATES** (informs fairness and reliability) • The candidates who are being assessed are given a description of the assessment's purpose • Instructions provided to candidates about what they are expected to do are informative and unambiguous • The basis for judgment (criterion for success, or what is "good enough") is made explicit for candidates	• Candidate progression is monitored and information used for mentoring • Candidates are informed how the instrument results are used in reaching conclusions about their status and/or progression
• Category or task link with CAEP, InTASC or state standards is not explicit • Category or task has only vague relationship with content of the standards being informed • Category or task fails to reflect the degree of difficulty described in the standards	**3. CONTENT OF ASSESSMENT** (informs relevancy) • Indicators assess explicitly identified aspects of CAEP, InTASC or state standards • Evaluation indicators reflect the degree of difficulty or level of effort described in the standards • Indicators unambiguously describe the proficiencies to be evaluated • 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 category or task is specific to students' application of knowledge to solve problems	• 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

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
• Evaluation categories or 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	• Most indicators (at least those comprising 80% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards	
• Rating scales are used in lieu of rubrics; e.g., "level 1= significantly below expectation" . . "level 4 = significantly above expectation". • Levels do not represent qualitative differences and provide limited or no feedback to candidates specific to their performance. • Proficiency level attributes are vague or not defined, and may just repeat from the standard or component	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 indicators • 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 indicator	• Higher level actions from Bloom's taxonomy are used such as "analysis" or "evaluation"
• Plan to establish validity does not inform reviewers whether validity is being investigated or how • The instrument was not piloted prior to administration • Validity is determined through an internal review by only one or two stakeholders. • Described steps do not meet accepted research standards for establishing validity. • Plan to establish reliability does not inform reviewers	**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 • If the assessment is new or revised, a pilot was conducted. • 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 RELIBILITY** • 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	• A validity coefficient is reported • types of validity investigated go beyond content validity and move toward predictive validity • A reliability coefficient is reported • Raters are initially, formally calibrated to

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
whether reliability is being investigated or how. • Described steps to not meet accepted research standards for reliability. • No evidence, or limited evidence, is provided that scorers are trained and their inter-rater agreement is documented.	consistency, etc.) and the steps the EPP took to ensure the reliability of the data from the assessment • Training of scorers and checking on inter-rater agreement and reliability are documented • The described steps meet accepted research standards for establishing reliability	master criteria and are periodically formally checked to maintain calibration at levels meeting accepted research standards
WHEN THE INSTRUMENT IS A SURVEY: **Use Sections 1 and 2, above, as worded and substitute 6.a and 6.b, below for sections 3, 4 and 5.**		
• Individual item are ambiguous or include more than one subject • Items are stated as opinions rather than as behaviors or practices • Dispositions surveys provide no explanations of their purpose • Scaled choices are numbers only, without qualitative description linked with the item under investigation • Limited or no feedback provided to candidates • No evidence that questions are piloted	**6.a. SURVEY CONTENT** • Questions or topics are explicitly aligned with aspects of the EPP's mission and also CAEP, InTASC or state standards • Questions have a single subject; language is unambiguous • Leading questions are avoided • Items are stated in terms of behaviors or practices instead of opinions, whenever possible • Surveys of dispositions make clear to candidates how the survey is related to effective teaching **6.b DATA QUALITY** • An even number of scaled choices helps prevent neutral (center) responses • Scaled choices are qualitatively defined using specific criteria aligned with key attributes identified in the item • Feedback provided to the EPP is actionable • EPP provides evidence that questions are piloted to determine that candidates interpret them as intended and modifications are made, if called for **Criteria listed below are evaluated on site:** • *EPP provides evidence that candidate responses are compiled and tabulated accurately* • *Interpretations of survey results are appropriate for the items and resulting data* • *Results from successive administrations are compared (for evidence of reliability)*	• Scoring is anchored in performance or behavior demonstrably related to teaching practice • Dispositions surveys make an explicit connection to effective teaching • EPP provides evidence of survey construct validity derived from its own or accessed research studies

CHECKLIST

Item Category	Below Adequate	CAEP Adequate Level	Above Adequate	N/A
1. ADMINISTRATION AND PURPOSE: 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.				
2. INFORMING RESPONDENTS: Information given to respondent before and at the administration of the instrument (informs fairness and reliability); basis for judging candidate performance is explicit.				
3. CONTENT OF ASSESSMENT: 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).				
4. SCORING: 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).				
5. a. DATA VALIDITY: 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. DATA RELIABILITY:** 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.				
WHEN THE INSTRUMENT IS A SURVEY: **Use sections 1 and 2, above, and substitute 6.a and 6.b for sections 3, 4 and 5.** **6. 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 the EPP 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:				