Georgia Professional Standards Commission

Protecting Georgia's Higher Standard of Learning

2016 Technical Assistance Conference

Validity and Reliability of EPP Assessments

April 19, 2016

Nate Thomas and Beverly Mitchell

Topics

1. Approval Standard expectations
2. Expectations of Instrument Validity
3. Expectations of Instrument Reliability
4. Approaches to Establishing Validity & Reliability

Georgia Professional Standards Commission

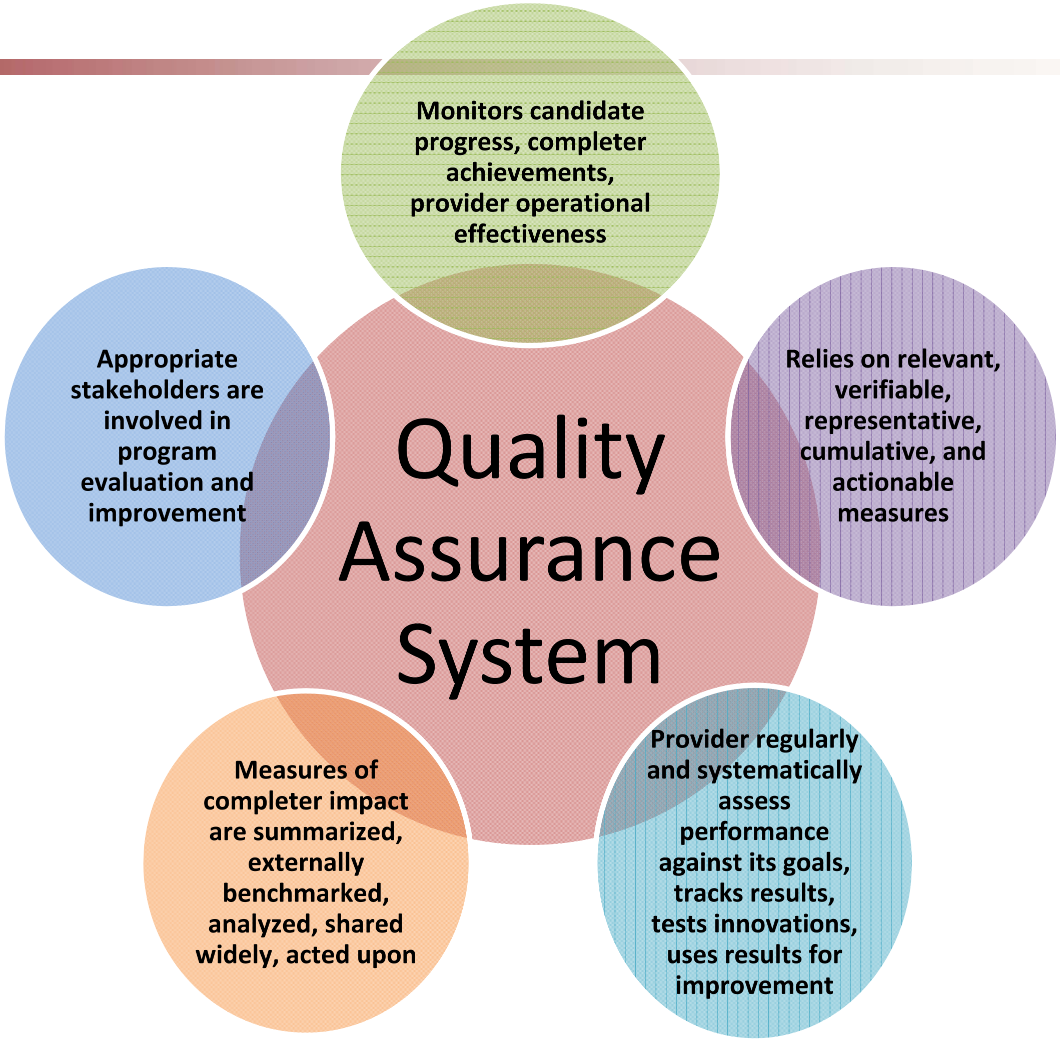
Protecting Georgia's Higher Standard of Learning

Standard 5

Provider Quality Assurance & Continuous Improvement

Standard 5:

Provider Quality Assurance & Continuous Improvement



Quality Assurance System

Monitors candidate progress, completer achievements, provider operational effectiveness

Relies on relevant, verifiable, representative, cumulative, and actionable measures

Provider regularly and systematically assess performance against its goals, tracks results, tests innovations, uses results for improvement

Measures of completer impact are summarized, externally benchmarked, analyzed, shared widely, acted upon

Appropriate stakeholders are involved in program evaluation and improvement

Quality Assurance System: Characteristics of Measures

- **Relevant**: Evidence is related to standard and assesses what it is claimed to assess
- **Verifiable**: Accuracy of sample of evidence in data files
- **Representative**: Samples are free of bias and typical of completed assessments, or limits to generalizability are clearly delineated
- **Cumulative**: Most assessment results are based on at least 3 administrations
- **Actionable**: Analyzed evidence is in a form that can guide the EPP decision-making
- **Valid and Consistent:** Produces empirical evidence that interpretations of data are valid and consistent

Georgia Professional Standards Commission

Protecting Georgia's Higher Standard of Learning

Standard 5

Instrument Validity

Quality Assurance System: Validity

- The extent to which an assessment measures what it is supposed to measure
- The extent to which inferences and actions on the basis of assessment scores are appropriate and accurate

Reference: CRESST – National Center for Research on Evaluation, Standards, and Student Teaching

Instrument Validity

- Instrument content and format are research-based
- Instrument was piloted before use
- EPP describes steps it has taken or will take to ensure validity of assessment
- Plan details types of validity investigated/established and results
- Investigations/plans meet accepted research standards for establishing validity
- inter-rater reliability or agreement is at .80 or 80% or above (except for surveys)
- Surveys align to standards

Reference: CAEP Accreditation Handbook, 2016

Instrument Validity: Standard Alignment

- Provide evidence that assessments are aligned with national, state, and institutional standards

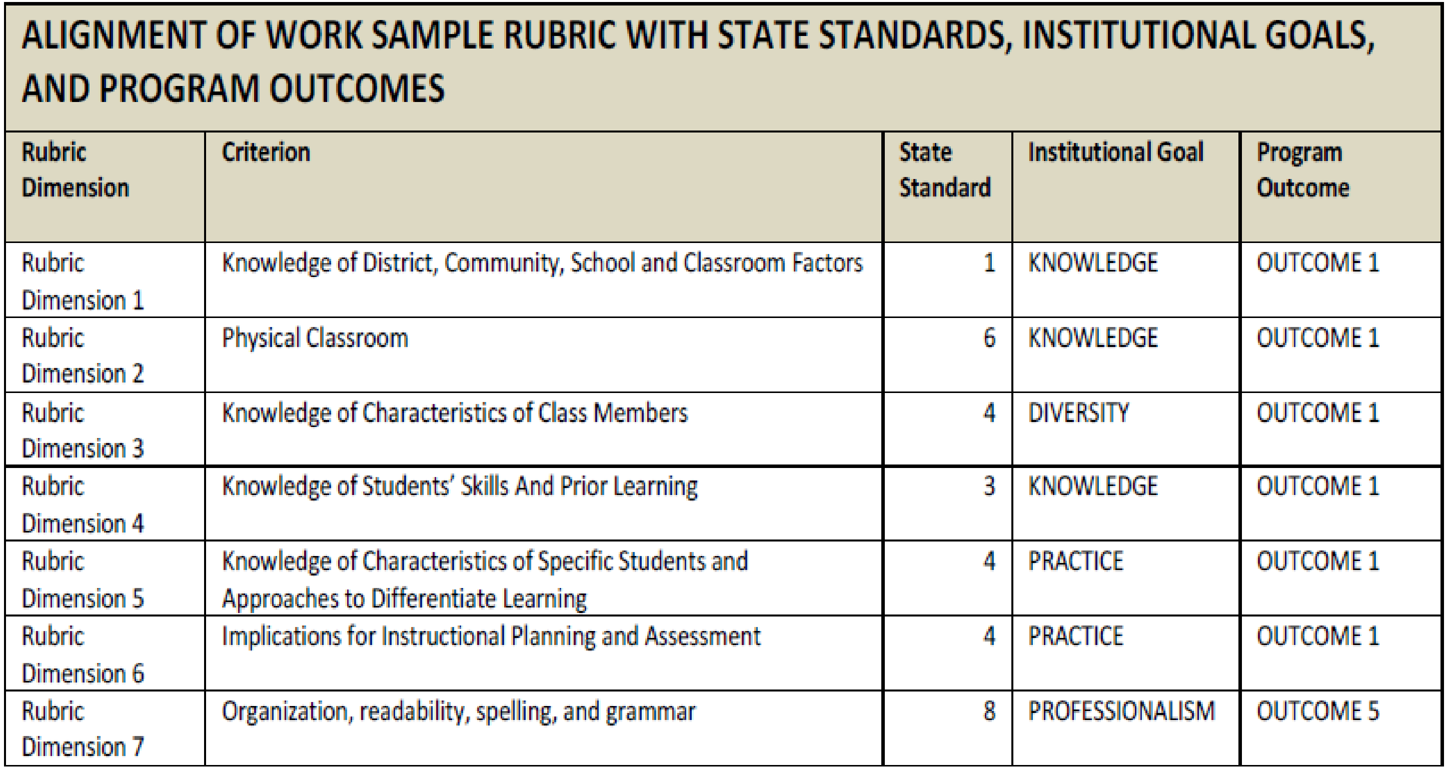
- Create alignment documents linking the standard to assessment items (i.e., test questions, rubric dimensions, indicators)

- Determine whether learning expectations are adequately and representatively sampled within and/or among assessments in the system

Reference: Garcia, S. (2016). Demystifying Assessment Validity and Reliability. CAEP Conference Presentation. San Diego, CA.

Instrument Validity:
Example of Standard Alignment

Table 1: Sample Alignment Matrix



ALIGNMENT OF WORK SAMPLE RUBRIC WITH STATE STANDARDS, INSTITUTIONAL GOALS, AND PROGRAM OUTCOMES				
Rubric Dimension	**Criterion**	**State Standard**	**Institutional Goal**	**Program Outcome**
Rubric Dimension 1	Knowledge of District, Community, School and Classroom Factors	1	KNOWLEDGE	OUTCOME 1
Rubric Dimension 2	Physical Classroom	6	KNOWLEDGE	OUTCOME 1
Rubric Dimension 3	Knowledge of Characteristics of Class Members	4	DIVERSITY	OUTCOME 1
Rubric Dimension 4	Knowledge of Students' Skills And Prior Learning	3	KNOWLEDGE	OUTCOME 1
Rubric Dimension 5	Knowledge of Characteristics of Specific Students and Approaches to Differentiate Learning	4	PRACTICE	OUTCOME 1
Rubric Dimension 6	Implications for Instructional Planning and Assessment	4	PRACTICE	OUTCOME 1
Rubric Dimension 7	Organization, readability, spelling, and grammar	8	PROFESSIONALISM	OUTCOME 5

Reference: Garcia, S. (2016). Demystifying Assessment Validity and Reliability. CAEP Conference Presentation. San Diego, CA.

Instrument Validity:
Balance of Representation

- Analyze alignment documents to determine the frequency and proportion each standard is addressed
- Ensure a balance based on the relative importance of each content standard item

Reference: Garcia, S. (2016). Demystifying Assessment Validity and Reliability. CAEP Conference Presentation. San Diego, CA.

Instrument Validity: Fairness

- Assessments are reviewed by internal and external stakeholders to ensure language and form of assessments are free of cultural and gender bias
- Assessment instruments and rubrics clearly state what is expected for successful performance
- All candidates have had learning experiences that prepare them to succeed on an assessment

Reference: Garcia, S. (2016). Demystifying Assessment Validity and Reliability. CAEP Conference Presentation. San Diego, CA.

Georgia Professional Standards Commission

Protecting Georgia's Higher Standard of Learning

Standard 5

Reliability of Results

Reliability

Degree in which an assessment produces stable and consistent results

- EPP describes type of reliability investigated/ established and steps to ensure/evaluate reliability
- Described steps meet accepted research standards for establishing reliability

Reference: CAEP Assessment Rubric, 2015

Instrument Consistency: Inter-rater reliability

Used to assess the degree different raters/observers give consistent estimates of the same phenomenon

- Calculate the correlation between the ratings of the two or more observers viewing the same clinical experience at the same time
- Hold "calibration" meetings
  - Watch a clinical experience with a group
  - Talk about how ratings were determined and what each reviewer noted
  - Come up with rules for deciding what represents a specific rating on the instrument

Georgia Professional Standards Commission

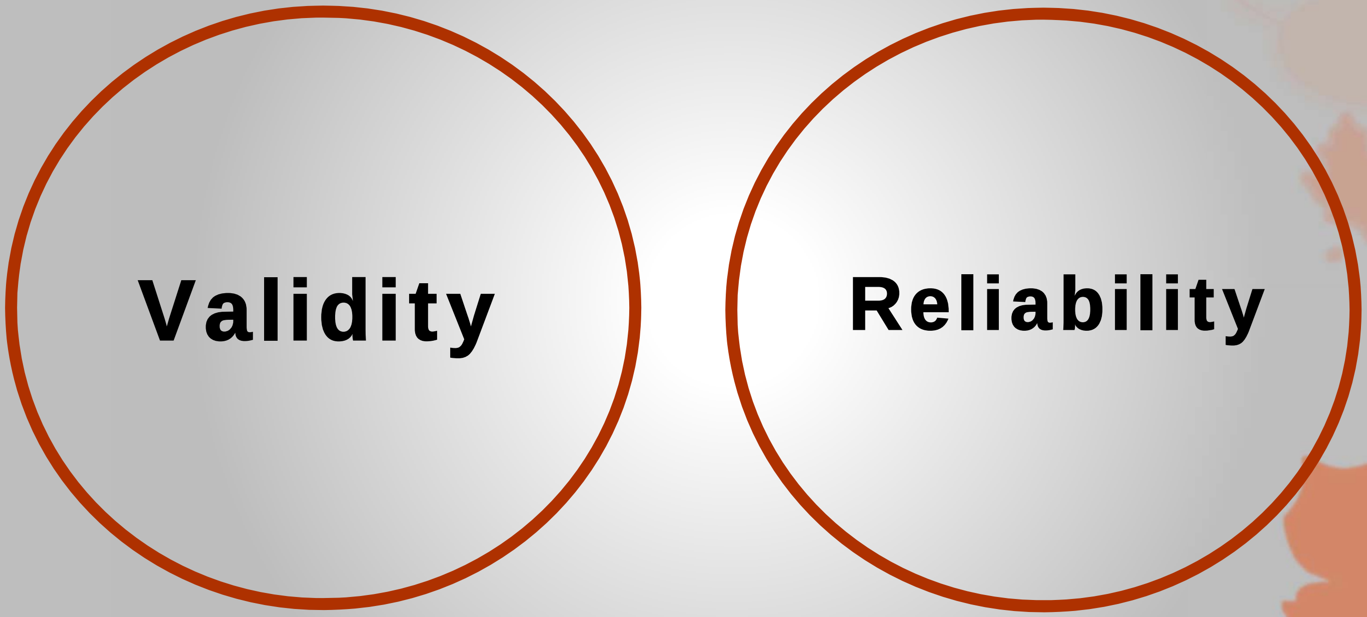
Protecting Georgia's Higher Standard of Learning

How to Improve Assessments

Validity / Reliability Matters

Really?

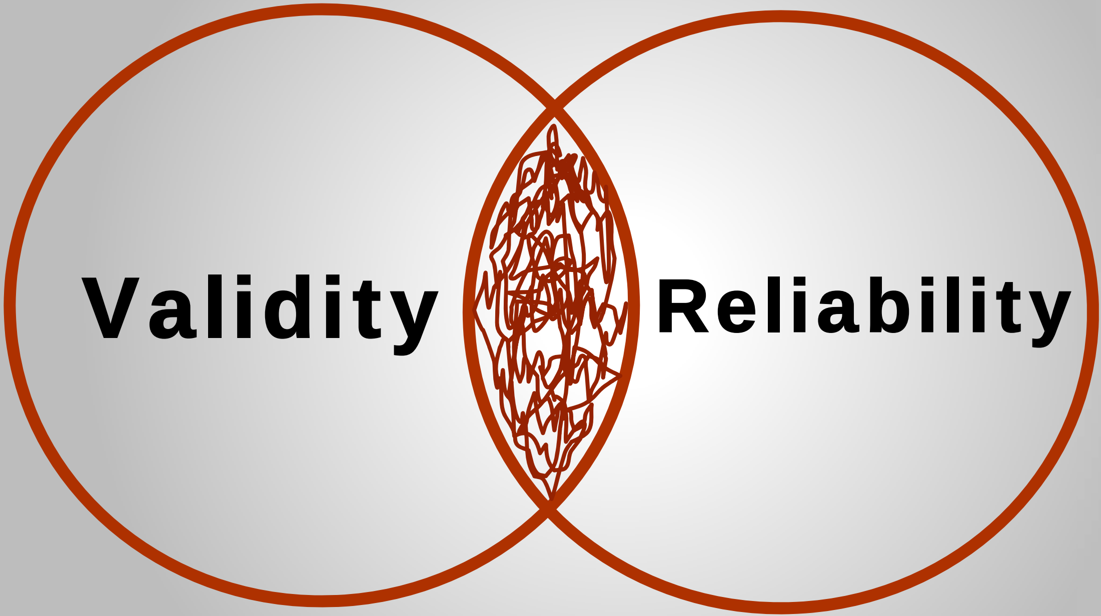
Beverly Mitchell, Kennesaw State University



Validity

Reliability

Beverly Mitchell, Kennesaw State University



Validity

Reliability

Beverly Mitchell, Kennesaw State University

Two Types of Assessments

1 - Proprietary

2 – EPP/Program

Beverly Mitchell, Kennesaw State University

Components of an Assessment

Conditions for Use

Purpose

Content

Rubric or Scoring

Instructions or Procedures

Parts/Sections

Measures

Data Generated

Interpretation of Results

Analysis

Experience & Training of Assessors/Users

Beverly Mitchell, Kennesaw State University

	Instrument	Data Interpretation	Users
Validity	√	√	(Trained)
Reliability	√	√	-Trained-

Establishing Content Validity

- Know standards or know content
- Determine purpose & conditions
  - Invite key players
  - Review language, items, scoring
  - Obtain collective, collaborative feedback
  - Compare with other similar measures
- Generate drafts – note “drafts”
  - Send out for review – stakeholder feedback
  - Conduct pilot
- OR, Measure against another known/validated test

Establishing Content Validity

- Warnings:
  - Adapting from…
  - Aligning with…

Reliability

(Estimating) Agreement

Among Assessors

Survey

n=10 adults
Scale: Agreement dimension
Purpose: Solicit opinion about eating habits

Item	Strong Agree	Agree	Disagree	Strong Disagree
1	30%	70%	0%	0%
2	100%	0%	0%	0%
3	20%	30%	30%	20%

Item 1 = 70% with “Agree”
Item 2 = 100% with “Strong Agree”
Item 3 = Mixed responses

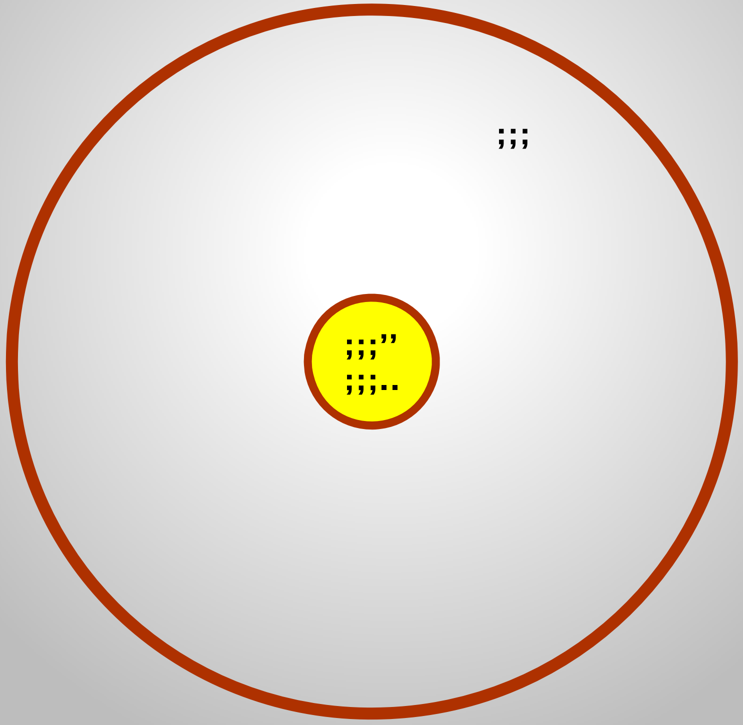
Observation

N=10 university supervisors
Scale: 4-level rubric
Purpose: Evaluate quality of candidates

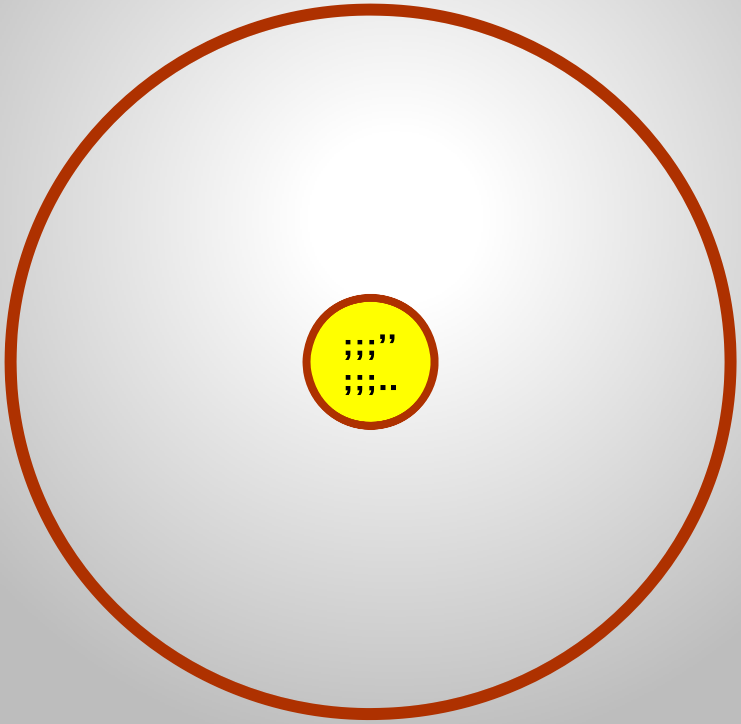
Item	Exceptional 3	Good 2	Approaching 1	Poor or Not Evident 0
1	30%	70%	0%	0%
2	100%	0%	0%	0%
3	20%	30%	30%	20%

Item 1 = 70% with “Good”
Item 2 = 100% with “Exceptional”
Item 3 = Mixed responses

Item 1 = 70% agreement with rating of 2 (Good)

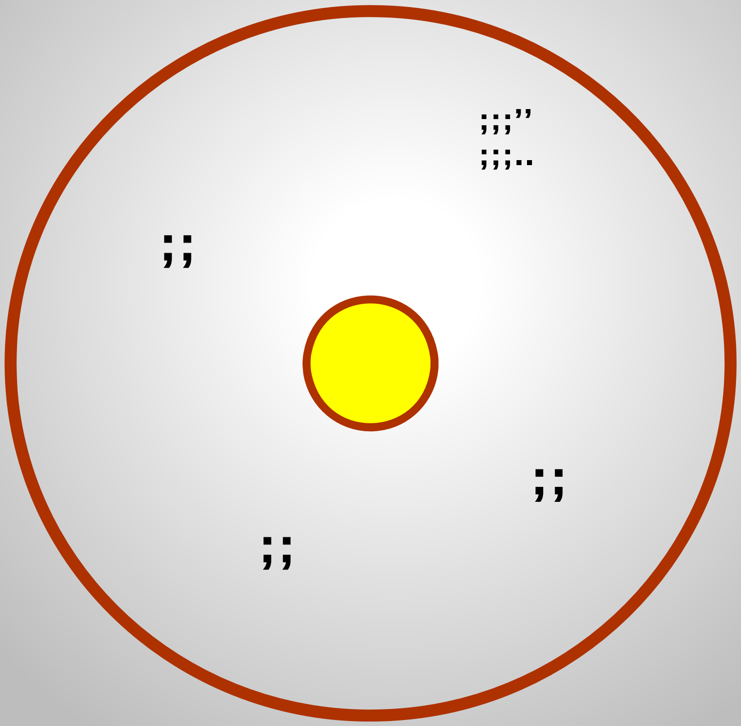


Item 2 = 100% agreement with rating 3 (Exceptional)



Beverly Mitchell, Kennesaw State University

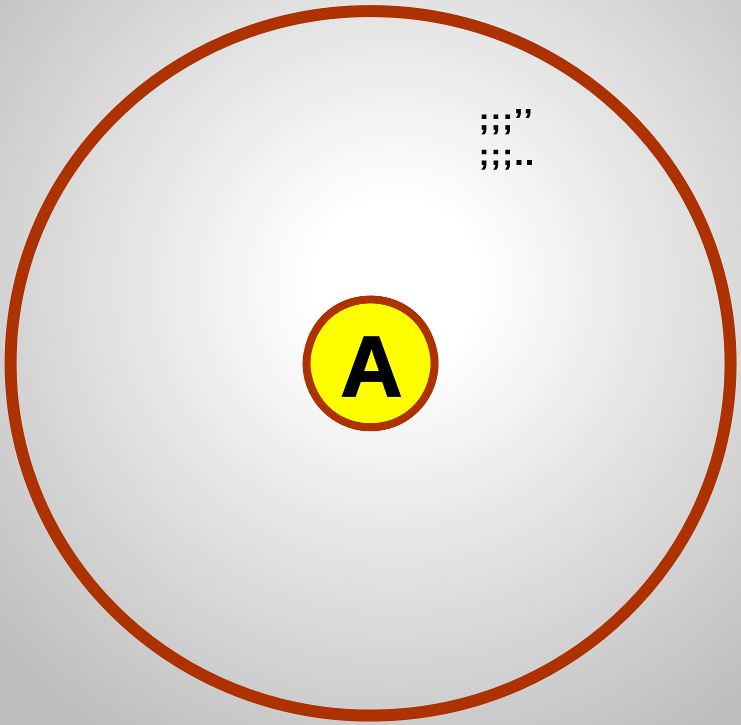
Item 3 = Mixed and highly variable



Beverly Mitchell, Kennesaw State University

Item 2 = 100% agreement with rating 3 (Exceptional)

But…what if the “true/ expert validated ” rating were 2…”approaching”???



Validity

Design Issues

Beverly Mitchell, Kennesaw State University

Design:
Creating the Instrument

1-Inference

2-Complexity

Inference

- Judgment
- Conclusions
- Surmise

Inference

Low

High

Beverly Mitchell, Kennesaw State University

High Inference

- In absence of guidance and/or fact-based language, it requires:
  - Thinking, reflecting, comparing, contrasting, depth of analysis, "Wondering"
  - Training to use properly, reliably, & to agree with others
  - Conceptual

Beverly Mitchell, Kennesaw State University

Low Inference

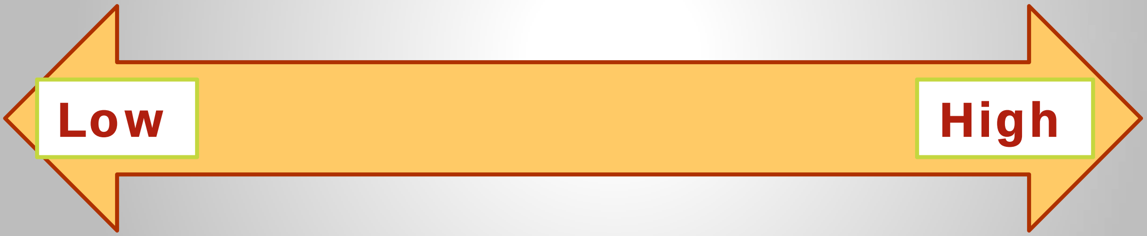
- Straightforward
  - Language = precise & targeted, fact-based
- Clear – no competing interpretations of words
- No doubt as to what point is being made

Complexity

- Complicated
- Intricate
- Comprised of interrelated parts or sections
- Developed with great care or with much detail

Beverly Mitchell, Kennesaw State University

Complexity



High Complexity

- Developed with much detail
  - Attend to sequence, pre-requisite parts
  - Could require much coordination for implementation
  - Complicated scoring

Low Complexity

- Simple
- Straightforward
- Unsophisticated
- Few Parts
- Little integration or connections among parts

Beverly Mitchell, Kennesaw State University

Design: Keeping Inference & Complexity Under Control

Inference

- Rubric:
  - High: e.g., general rubric (much guessing about performance)
  - Low: e.g., analytic rubric (describes performance in detail in relation to criterion)
- Language:
  - Precise
  - Avoid “wondering” words, e.g., some, often,
- Scoring levels:
  - More versus fewer
  - I.E., 5 versus 3

Complexity

- Instrument:
  - Focused versus kitchen sink
  - Shorter versus longer
- Parts:
  - Fewer versus many
- Instructions:
  - Detailed but concise
- edTPA good example of C
  - Lengthy, detailed, many parts,
- Implementation
  - Little versus much coordination

APPENDIX G – Assessment Rubric

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 EPP’s 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

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
• 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 • 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	3. **CONTENT OF ASSESSMENT** (informs relevancy) • Evaluation categories or tasks assess explicitly identified aspects of CAEP, InTASC or state standards • Evaluation categories or tasks reflect the degree of difficulty or level of effort described in the standards • Evaluation categories or tasks unambiguously describe the proficiencies to be evaluated • When the standards being informed address higher level functioning, the evaluation categories or tasks 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 • Most evaluation categories or tasks (at least those comprising 80% of the total score) require observers to judge consequential attributes of candidate proficiencies in the standards	• 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
• 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 the category (or indicator) or with the assigned task • 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 category or indicator	• Higher level actions from Bloom's taxonomy are used such as "analysis" or "evaluation"

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
• 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.	**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 • The assessment was piloted prior to administration • 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	• A validity coefficient is reported • types of validity investigated go beyond content validity and move toward predictive validity
• Plan to establish reliability does not inform reviewers 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.	**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 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	• A reliability coefficient is reported • Raters are initially, formally calibrated to 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	**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	• Scoring is anchored in performance or behavior demonstrably related to teaching practice • Dispositions surveys make an explicit

EXAMPLES OF ATTRIBUTES BELOW SUFFICIENT LEVEL	CAEP SUFFICIENT LEVEL	EXAMPLES OF ATTRIBUTES ABOVE SUFFICIENT LEVEL
• Dispositions surveys provide no explanations of their purpose		connection to effective teaching
• 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.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 • 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)	• EPP provides evidence of survey construct validity derived from its own or accessed research studies