Demystifying Assessment Validity and Reliability

Susan Gracia, PhD
Director of Assessment
Feinstein School of Education and Human Development
Rhode Island College

Session Goals

- As a result of attending this session, attendees will be able to:
  1. Identify critical dimensions of assessment validity and reliability
  2. Identify strategies for collecting key validity and reliability evidence
  3. Consider how to update or improve the ways in which they currently collect validity and reliability evidence.

Validity

- The extent to which an assessment measures what it is supposed to measure, and the extent to which inferences and actions on the basis of assessment scores are appropriate and accurate (CRESST).
  - Validity is the most fundamental consideration in developing and evaluating assessments.
- Reliability=consistency of measurement and is a necessary condition for validity.

Imperatives for Evaluating Validity and Reliability of Assessments

External

- The provider maintains a quality assurance system comprised of valid data from multiple measures… and produces empirical evidence that interpretations of data are valid and consistent (CAEP Standard 5).

Ethical

- If assessments lack sufficient reliability and validity for their intended purposes, there is potential for serious harm. (AERA Position Statement on High-Stakes Testing in Pre-K – 12 Education, 2000)

Building a Case

- “Just as an attorney builds a legal case with different types of evidence, the degree of validity for the use of [an assessment] is established through various types of evidence including logical, empirical, judgmental, and procedural evidence” (CollegeBoard, n.d.).
- Professional judgment guides decisions regarding the specific forms of evidence that can best support the intended interpretation and use. (Standards for Educational and Psychological Testing).
- Assessing validity is not a one-time event.
- A series of validity studies are needed.

Lines of Evidence

- Validity
  - Content-related validity: Do assessment items/components adequately and representatively sample the content area(s) to be measured?
  - Construct validity: Do assessments and the assessment system measure the content they purport to measure?
  - Prediction (Criterion-related validity): How well do assessment instrument predict how well candidates will do in future situations?
  - Fairness: Are all candidates afforded a fair opportunity to demonstrate their skills, knowledge, and dispositions?
  - Utility: How useful are the data generated from assessments?
  - Consequences: Are assessment uses and interpretations contributing to increased candidate achievement and not producing unintended negative consequences? (Linn, 1994)

Lines of Evidence

- Reliability
  - What is the degree of internal consistency in assessments?
    - To what degree do items that propose to measure the same general construct produce similar scores?
    - Are the scores on similar items related (internally consistent)?
  - What is the level of inter-rater consistency among faculty?
    - Do the scorers differ in the levels of severity they exercise, or do they function interchangeably?
    - Are there any inconsistent raters whose patterns of ratings show little systematic relationship to the scores that other raters give?

Strategies for collecting & presenting validity evidence

Content-Related Validity: Alignment

- Provide evidence that assessments are aligned with key standards or learning expectations.
- Create alignment documents linking learning expectations to 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.

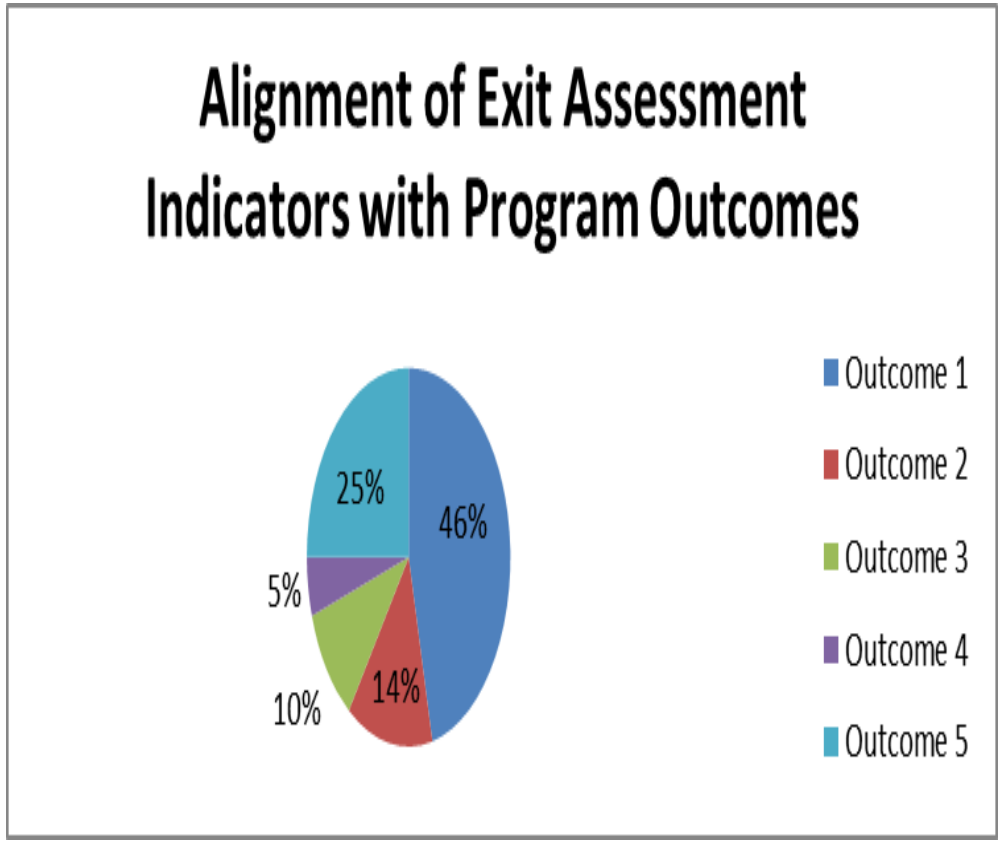
Table 1: Sample Alignment Matrix

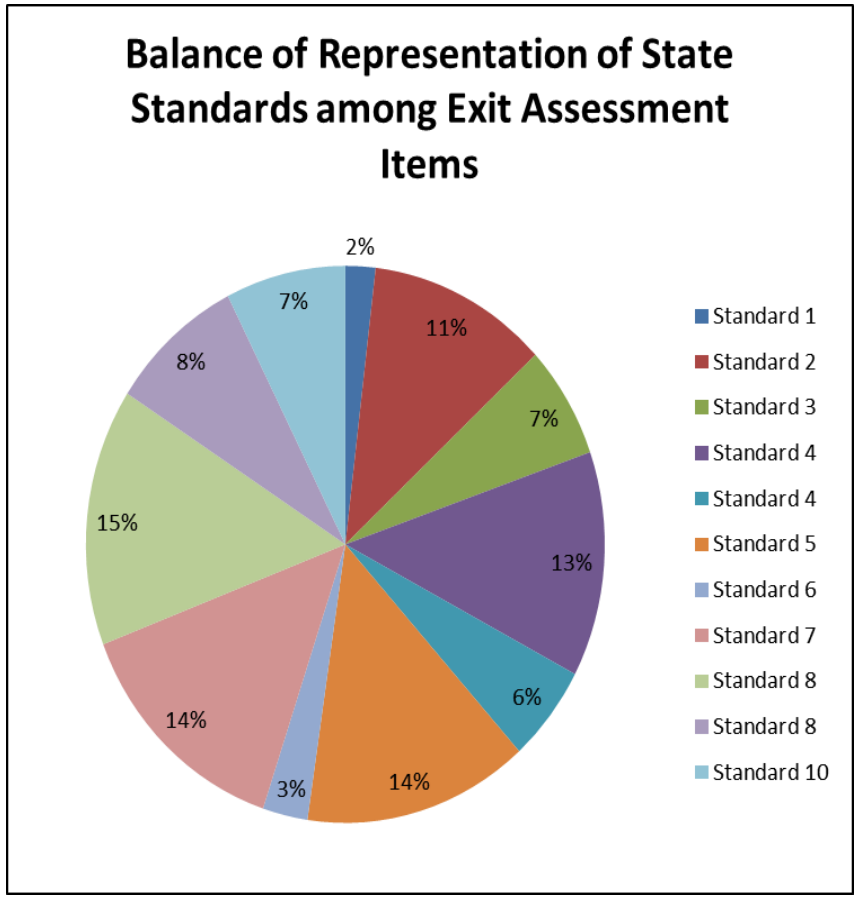
ALIGNMENT OF WORK SAMPLE RUBRIC WITH STATE STANDARDS, INSTITUTIONAL GOALS, AND PROGRAM OUTCOMES				
Rubric Dimension	**Criterion**	**State Standard**	**Institutional Goal**	**Program Outcome**
Rubric Dimension 1	Knowledge of District, Community, School and Classroom Factors	1	KNOWLEDGE	OUTCOME 1
Rubric 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

Content-Related Validity: Balance of Representation

- Is the content of the assessments in the system balanced, or on the other hand, weighted to represent the relative importance of the learning targets?
- Analyze alignment documents.
- Tally # of times a learning target is referenced in an alignment document and divide by total number of items in the assessment to calculate the proportion of assessment items that are aligned with a particular standard.

Examples: Balance of Representation





Content Validity: External Evidence

- Provide documentation that assessments are designed based on best practice in relevant literature and on the professional knowledge, experience, and consensus of faculty, many of whom are developers and definers of best practice in their professional areas.
- If an assessment system includes or is heavily based on externally developed, previously validated assessments, findings from content validity studies conducted by others can be used to lend support to the content validity associated with particular assessments.
- Input from employers of graduates and graduates themselves also helps establish the content validity of assessment measures.

Construct Validity: Factor Analysis

- An assessment has construct validity if it accurately measures a theoretical, non-observable construct or trait.
- Conduct factor analysis of assessment data to examine whether the theoretical framework of an assessment matches the factor representation yielded by factor analysis.
  - Factor analysis is a statistical technique that identifies the smaller number of factors/constructs/dimensions that underlie a larger set of variables (most of which are correlated to each other).

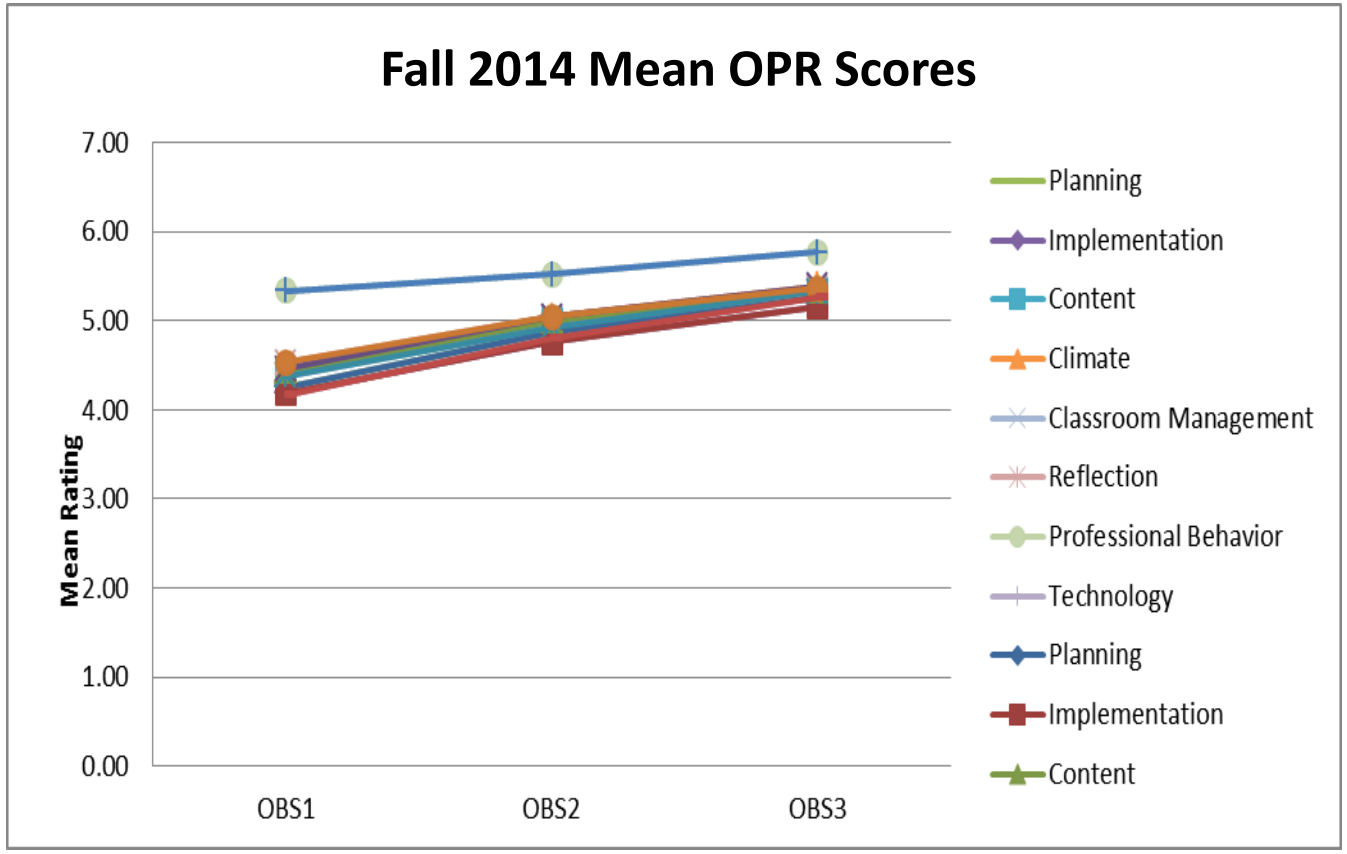
Factor Analysis Example

- The Teacher Candidate Work Sample assessment designed to measure seven skills. Seven separate scores calculated through use of seven rubrics, for a total of approx. 49 criteria.
- Factor analysis of candidate scores (n=253) revealed a seven factor solution accounting for 75% of the variance in the data.
- All rubric criteria except for one loaded on the appropriate, hypothesized factor. (The fit of this single criterion needs to be reviewed.)
- The results of this study provide evidence to support seven teaching process construct on which the TCWS is constructed.

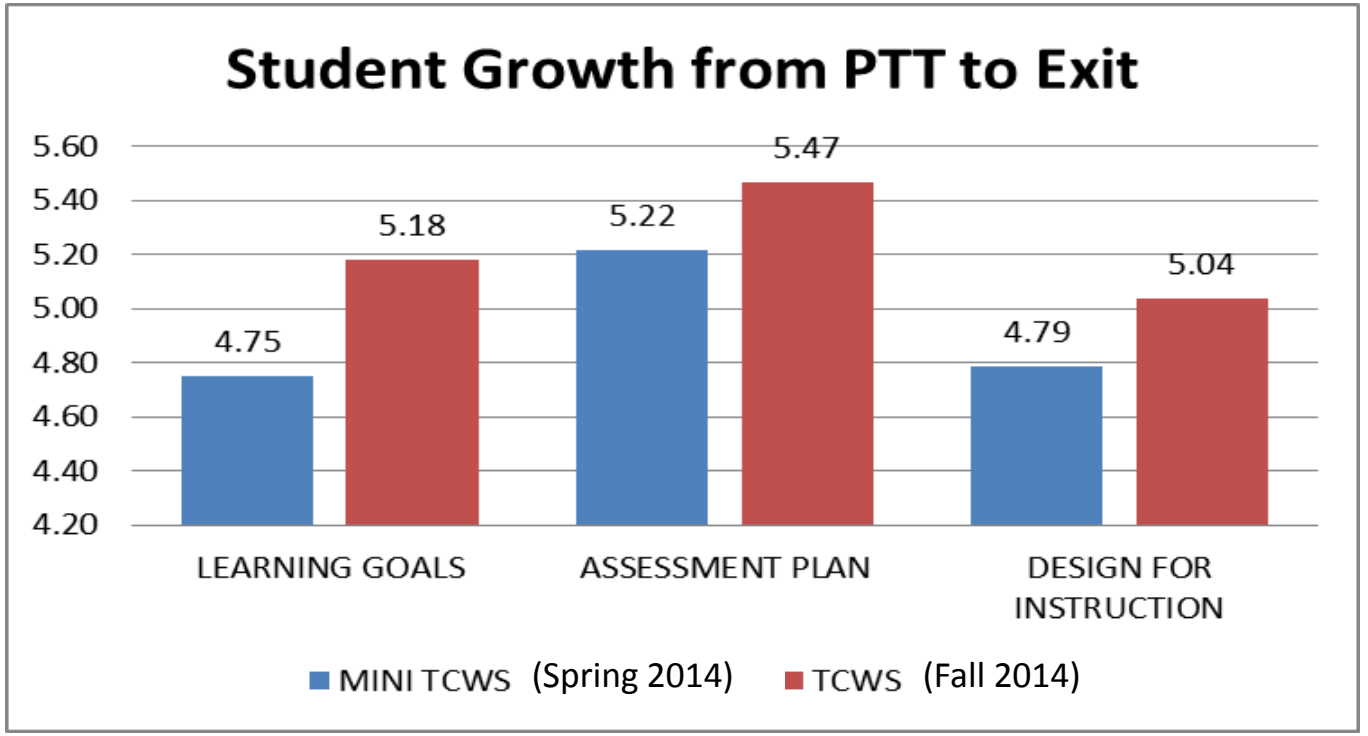
Construct Validity: Developmental Changes

- Assessments measuring certain constructs can produce evidence of construct validity if the scores on the assessments show predictable developmental changes over time.

Developmental Change Example 1



Developmental Change Example 2



Prediction (Criterion-related validity)

- Evaluate the degree to which scores on one assessment are able to predict something they should theoretically be able to predict.
- Relationship is reported as correlation—i.e., “validity coefficient”
- For instance:
  - Scores on the SAT predict freshman GPA.
  - Scores on Teacher Candidate Work Sample might predict candidate success in the field.
  - Can you think of others?

Prediction Examples

- We found Millers Analogy Test scores were highly, significantly correlated with graduate student GPA
  - Validity coefficient (r) = .89
- We found that scores on state mandated teacher preparation program admissions tests had low, statistically insignificant correlations with subsequent GPA.

N=687	MATH (PPST)	READING (PPST)	WRITING (PPST)	SUM(PPST)	GPA
MATH (PPST)	1.00				
READING (PPST)	0.48	1.00			
WRITING (PPST)	0.44	0.42	1.00		
SUM(PPST)	0.85	0.80	0.73	1.00	
GPA	0.24	0.31	0.31	0.35	1.00

Fairness

- Freedom from bias: The language and form of assessments must be free of cultural and gender bias.
  - Document process by which design, format, wording, and presentation of assessments are reviewed by internal and external constituents, as well as changes that have been made to minimize unintentional bias.
- Transparency of expectations: Assessment instructions and rubrics must clearly state what is expected for successful performance.
  - Document key efforts to keep faculty and candidates up to date and informed on all aspects of the assessment system.
  - Ensure that rubrics and prompts are highly descriptive and provide detailed guidance to the candidate and evaluators about assessment expectations and process.
  - Implement and document faculty training on assessment unit assessments.
  - Ensure that program handbooks and web pages clearly delineate program expectations, as well as program and unit assessments.
  - Regularly hold orientation meetings and information sessions for candidates at each transition point.

Fairness

- Opportunity to learn: All candidates must have had learning experiences that prepare them to succeed on an assessment.
  - Engage faculty in curriculum mapping and other processes to examine what is taught at different time points to ensure that candidates have opportunities to learn and succeed at the content and skills inherent in unit assessments. Document this process.
  - Ask candidates in exit survey if they felt that the program prepared them to succeed in the assessments
- Accommodations: Candidates with documented learning differences must be afforded accommodations in instruction and assessment.
- Multiple opportunities: Candidates must have the opportunity to demonstrate their learning in multiple ways and at different times. (Smith & Miller, 2003)
  - Utilize a variety of assessment formats.
  - Offer candidates opportunities to retake or redo all or parts of assessments.

Multiple Opportunities Example:

Table 1: Initial Programs Assessment System Blueprint

KEY **Methods** SR=selected response/short answer; CR=constructed response; PA=performance assessment; OC=observation/ personal communication

Level I=individual course; P=program; U=unit; SN=state or national

INITIAL TEACHER PREPARATION PROGRAMS

Transition Point	Assessment		Method	Level(s)
Admission	B- or better in FNED 346		SR, CR, PA, OC	I
	2.5 GPA		SR, CR, PA, OC	I
	Completion of RIC Writing and Math requirements		SR, CR, PA, OC	I
	Successful completion of the Reading, Mathematics, and Writing sections of the Pre-Professional Skills Test of the PRAXIS I or SAT or ACT		SR, CR	SN
	Supervisor Reference Form: Assessment of Candidate Dispositions in Field Settings		PA, OC	I, U
	Faculty Reference Form: Assessment of Professional Dispositions in College Classroom		OC	I,U
	Technology competency		SR	U
	B or better in Writing 100		SR, CR, PA, OC	I
	Other, program-specific requirements		SR, CR, PA, OC	P
Preparing to Teach (Formative)	2.5 GPA		SR, CR, PA, OC	I
	Passing scores on PLT, Praxis II, and/or Content tests		SR , CR	SN
	Implemented Lesson Plan		PA, OC	P,U
	Mini Teacher Candidate Work Sample		PA	P,U
	Assessment of Candidate Dispositions in the College Classroom		PA, OC	I, U
	Assessment of Candidate Dispositions in Field Settings (derived from ILP and MTCWS scores)		PA, OC	U
	Assessment of Candidate Cultural Competence (derived from ILP and MTCWS scores)		PA, OC	U
	Community service		OC	U
Exit (Summative)	Teacher Candidate Work Sample		PA	P,U
	Observation and Progress Report		PA, OC	P,U
	Assessment of Candidate Dispositions in Field Settings (derived from OPR and TCWS scores)		PA, OC	U
	Assessment of Candidate Cultural Competence (derived from OPR and TCWS scores)		PA, OC	U
	Other, program-specific requirements		SR, CR, PA, OC	P
	(Used for unit & program assessment, not to evaluate candidate)	Supervisor Evaluation of Cooperating Teacher	PA, OC	U, P
		Cooperating Teacher Survey	PA, OC	U, P
		Teacher Candidate Exit Survey	PA, OC	U, P
Post Graduation	Graduate follow up survey		OC	P,U
	Employer survey		OC	P,U

Utility

- Seek feedback from key stakeholders on the utility, user friendliness, and quality of assessments, particularly new/revised ones.
- Conduct surveys, interviews, focus groups.

Utility Example

- Feedback from assessment users
  - *Candidates really learned from doing the TCWS. It helped them look at their practice in new ways (especially the assessment piece).*
  - *The TCWS has a legitimacy in that candidates teach a unit, talk about and describe it, and take it apart*
  - *Candidates liked the cohesiveness of the TCWS; it seemed interconnected, linked across pieces*
  - *Clear expectations, relevant. Allowed faculty anchor discussion on candidates' specific work.*
  - *Having to design pre and post assessments, and write up the assessment and instructional decision-making pieces were meaningful for students, unlike the Exit Portfolio, which is more like busy work*
  - *TCWS engendered rich conversation with candidates about how they knew their students had learned; prompted a lot of reflection*
  - *TCWS was easier for candidates to complete than the Exit Portfolio. It's more cohesive and things are spelled out more clearly. The Exit Portfolio is not as clear to students and evaluators*
  - *The TCWS has cohesiveness. It gives candidates the opportunity to connect the dots. The Exit Portfolio has rich artifacts, but candidates don't typically see a connection among them*
  - *Components of the Exit Portfolio are in the TCWS but the TCWS asks candidates to step back. It requires a different level of reflection.*
  - *TCWS pointed out places in the program to do things better, emphasize more*
  - *It is vastly superior to the Exit Portfolio. The tasks are much clearer to candidates, although they find the project quite onerous*

Consequences

- “It is not enough to provide evidence that the assessments are measuring intended constructs. Evidence is also needed that the uses and interpretations are contributing to enhanced student achievement and, at the same time, not producing unintended negative consequences.” (Linn, 1994, p. 8)

Consequences

- Negative, unintended consequences could include narrowing of curriculum, increased candidate drop out, etc.
  - Track and document this through Institutional Research data, alumni surveys, student surveys.
- Be open to positive, unintended consequences, too.

Strategies for collecting & presenting reliability evidence

Internal Consistency

- Cronbach's alpha is the most common measure of internal consistency.
- The internal consistency reliability coefficients for teacher-made assessments generally range from .60 to .85 (Linn & Gronlund, 2000).
- Reliability on standardized tests of achievement and aptitude tend to fall between the .80s and low .90s (Salvia & Ysseldyke, 1998).
- The required level of internal consistency reliability for assessment increases as the stakes attached to the assessments increase (i.e., when assessment-based decisions are important, permanent, or have lasting consequences) (Linn & Gronlund, 2000).
- Salvia and Ysseldyke (1998) specify a minimum reliability of .90 for assessments that are used for tracking and placement.
- It is important to routinely study the internal consistency of assessments.

Internal Consistency…and Construct Validity

- High internal consistency can help establish construct-related validity evidence.
- If an assessment or scale has construct validity, scores on the individual items/indicators should correlate highly with the total assessment score.
- This is evidence that the assessment is measuring a single construct.

Internal Consistency Example

- Internal consistency Teacher Candidate Work Sample components was examined over 4 semesters: Fall 2013/Spring 2014, n=48; Fall 2014, n=120; and Spring 2015, n=253.
- Estimates of internal reliability (coefficient alpha) during these time periods for the seven TCWS constructs was:
  - Contextual Factors, α=.89, .93, .94;
  - Learning Goals & Objectives, α=.83, .96, .94;
  - Assessment Plan, α=.75, .96, .94;
  - Design for Instruction, α=.91, .94, .91;
  - Instructional Decision Making, α=.87, .94, .95;
  - Analysis of Student Learning, α=.87, .96, .94;
  - Candidate Reflection on Student Teaching Experience, α=.61, .94, .87.

Inter-Rater Reliability

- The extent to which two or more raters obtain the same result when using the same instrument /criteria to assess a student.
- Addresses the consistency of the implementation of a rating system.
- (At least) 3 types of reliability :
  - Correlation
  - Percent exact agreement
  - Cohen's Kappa statistic-takes into account the amount of agreement that could be expected to occur through chance

Inter-Rater Reliability Example

OPR Sections	Inter-Rater Reliability of CT and CS Ratings Spring 2014 (n=233 to 239)
Planning	.73
Implementation	.73
Content	.69
Climate	.74
Classroom Management	.75
Reflection	.59
OVERALL	.71

Reflection

- Which of the strategies presented today are you currently using to evaluate assessment validity and reliability?
- What are some additional or new strategies that you heard about today that you may be able to utilize?

CAEP Assessment Rubric-Level 4 (Demonstrates Target Criteria)

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 of an instrument
- ✓Validity coefficient is reported

Instrument Reliability

- ✓EPP describes type of reliability investigated/established and steps taken to ensure/evaluate reliability
- ✓Described steps meet accepted research standards for establishing reliability
- ✓Training of scorers and checking on inter-rater reliability are documented
- ✓Reliability coefficient is reported

Next steps: Planning

- See worksheet: VALIDITY and/or RELIABILITY STUDY PLAN

Wrap Up

- Review of session:
  - Critical dimensions of assessment validity and reliability
  - Strategies for collecting key validity and reliability evidence
  - Consideration of how to update or improve the ways in which they currently collect validity and reliability evidence
- Questions?
  - Feel free to contact me at sgracia@ric.edu

References

- American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Authors.
- Baker, E., L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability (Policy Brief 5).* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Center for the Study of Evaluation & National Center for Research on Evaluation, Standards, and Student Testing. (1999). *CRESST assessment glossar*y. Los Angeles, CA: CRESST/UCLA. Available: http://cresst96.cse.ucla.edu/CRESST/pages/glossary.htm
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and evaluation in teaching* (8th ed.). New York: Macmillan.
- Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23* (9), 4-14.
- *Measured measures: Technical considerations for developing a local assessment system*. (2005). Augusta, ME: Maine Department of Education.
- Smith, D. & Miller, L. (2003). *Comprehensive local assessment systems (CLASs) primer: A guide to assessment system design and use*. Gorham, ME: Southern Maine Partnership, University of Southern Maine.
- Stiggins, R.J. (2001*). Leadership for Excellence in Assessment: A Powerful New School District Planning Guide*. Portland, OR: Assessment Training Institute.
- Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin.
- Webb, N. L. (2005). *Issues related to judging the alignment of curriculum standards and assessments*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.
- Wiggins, G. (1998). *Educative assessment*. San Francisco, CA: Jossey-Bass.