BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCCCO Assessment Advisory Committee Training
September 17, 2021

Dr. Jessica L. Jonson
Dr. Maria Elena Oliveri

Training Objective

Prepare for the review of ESL assessment applications

- For second-party applications
- Based on CCC Standards (September, 2017)

Three criteria for review:

- Validity
- Fairness (including accommodations)
- Reliability

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Training Structure

Conceptual overview of each criteria

Outlines requirements/expectations for criteria from CCC Standards

Use applied examples to discuss requirements and expectations

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Application Overview

Application Types
Application Statuses
Approval Decisions Types

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Application Types

SECOND-PARTY: TEST/ASSESSMENT DEVELOPED AND MAINTAINED BY EXTERNAL VENDOR (COMMERCIALLY-AVAILABLE)

LOCALLY-MANAGED/DEVELOPED: TEST/ASSESSMENT DEVELOPED LOCALLY OR EXTERNAL TEST MAINTAINED LOCALLY

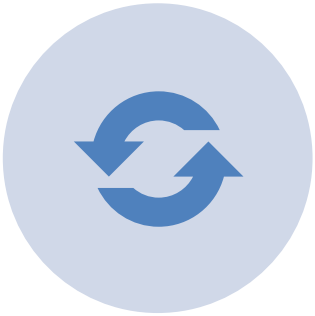
CRITICAL MASS: LOCALLY DEVELOPED/MANAGED INSTRUMENT USED BY MINIMUM OF SIX COLLEGES

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Types of Application Status

NEW



RENEWAL

RESUBMISSION

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Approval Decision Types

Full Approval: All standards met (6 Academic Years (AY) from initial approval)

Provisional Approval: Most standards met but lack some clarifying information (1AY + 2 AY @ probationary)

Probationary Approval: Minimum standards met but missing critical info on other standards (2 AY)

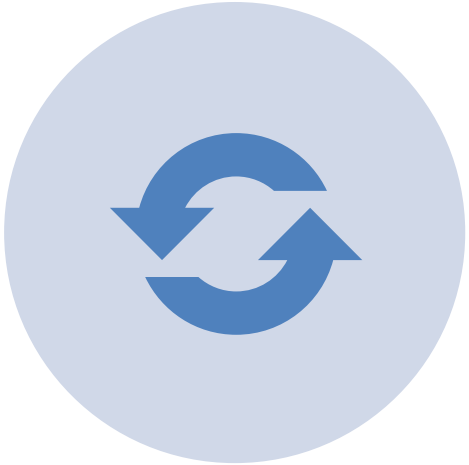
Not Approved: Minimum standards not met (Cannot be used to inform decisions)

Types of Application Status

NEW APPLICATION

- Test not previously reviewed
- Test previously approved but not renewed by 6 years
- Substantive changes to previously reviewed test

Types of Application Status



RENEWAL APPLICATION

- Test previously approved in last 6 years
- No changes in test content, placement courses, or student demographics

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

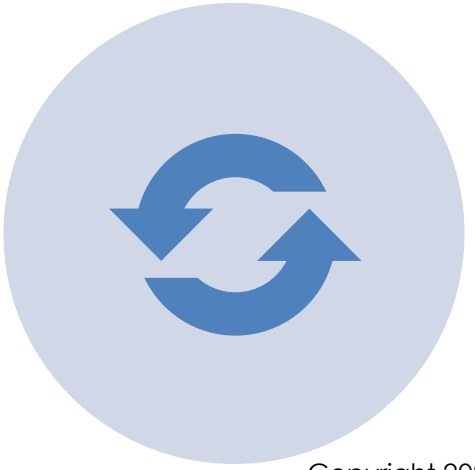
Types of Application Status

RESUBMISSION

- Previously provisional/ probationary approvals seeking full approval
- Previously not approved seeking approval

Minimum Standards by Application Status

NEW APPLICATIONS
Content Validity
Criterion/Consequential Validity
Fairness (including accommodations)



RENEWAL APPLICATIONS
Criterion/Consequential Validity
Fairness (including accommodations)

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Minimum Standards for Probationary Approval

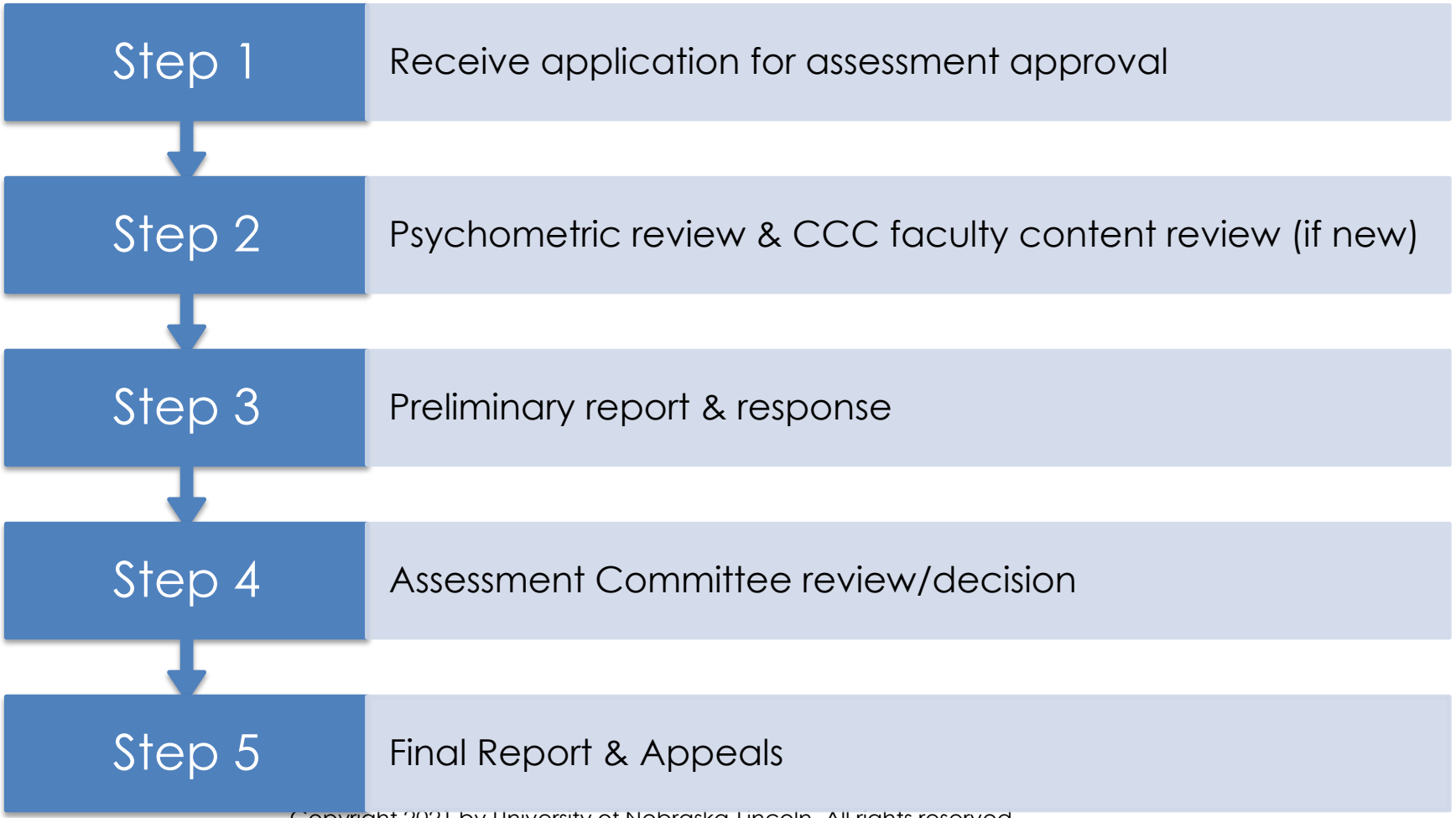
Test Content: Objectives, Specifications, and Scores

Farness: representative panel review & empirical study

Criterion/Consequential validity: study with CCC or similar students for placement

ADA Accommodations: availability

Application Review Cycle (Section 3, Step 4 Diagram)



BUROS CENTER FOR TESTING

University of Nebraska Lincoln | College of Education & Human Sciences

Why preliminary and final review?

Preliminary review: psychometric review

Final review: contextual review

Both are needed for a comprehensive review of quality, relevance, and appropriateness.

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Training Topics & Structure

Conceptual overview for validity (content, criterion/consequential), fairness, and reliability

Outlines requirements and expectations for each criterion from CCC Standards

Use applied examples to discuss requirements and expectations

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

QUESTIONS?

BUROS
CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

BUROS
CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Validity Overview

Content Validity

Criterion/Consequential Validity

What Is Validity?

Validity is…

extent to which evidence shows that an instrument or procedure appropriately measures the construct of interest (e.g. construct = English language proficiency) for a particular interpretation and use (e.g. course placement)

Construct = related knowledge, skills, or other attributes (KSA) a test is intended to measure

Validity: An Analogy

Provide a "snapshot not a film" of an individual's functioning that "describes a moment frozen in time, described from the viewpoint of the psychologist" (p. 637).

Cates, J. A. (1999). The art of assessment in psychology: Ethics, expertise, and validity. Journal of Clinical Psychology , 55, 631-641.

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Validity: An Analogy

- "Good photos" represent subjects as they are.
- Even good photos more appropriate for some purposes than others.
- Good tools (tests) and good techniques (test givers, test administration) essential for a "good picture" of test takers.

Validity: What Is Being Validated?

Don't validate: instrument or its scores

- ESL test or Overall language score

Do validate: inferences for particular test uses or applications

- ESL placement decisions for students at CCC

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Validity: Types of evidence

Test content: Analysis test content and measured construct

Response processes: Analysis of individual responses

Internal structure: Relationship test items and test components

Relations with other variables: Test score inference and important external variables & generalize to new situations

Consequences of testing: Soundness of proposed interpretations and intended uses & unintended consequences

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Validity: Types of Evidence

Test content: Analysis test content and measured construct

Response processes: Analysis of individual responses

Internal structure: Relationship test items and test components

Relations with other variables: Test score inference and important external variables & generalize to new situations

Consequences of testing: Soundness of proposed interpretations and intended uses & unintended consequences

BUROS CENTER FOR TESTING

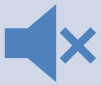
University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Validity Evidence Expectations and Review

Primarily responsibility test developer

Evidence that supports the particular interpretation and use for which test will be used

Evidence described in detail and rationale for how supports interpretation/use



Is evidence sound and sufficient for the purposes in which test will be used?

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: Content-related Validity Evidence

Clear definition of the content domain

- Test objectives and rationale
- Test specification/blueprint
- Scores reported
- Operational test form

Committee review

- Are test objectives/rationale appropriate for CCC
- Is content relevant and representative of ESL courses and curriculum
- Enough details for a local to evaluate the alignment of the content
- Are scores appropriate for the decisions to be made

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: Content-related Evidence

- Performance assessment (e.g. writing)
  - Prompts
  - Scoring rubrics/algorithms
- Computer-based/adapted
  - Describe item bank
  - CAT algorithms (rules/restrictions)

CCC Standards: Criterion-related Validity Evidence

Extent to which inference from a test score relates to alternate measure of construct or an outcomes measure.

Correlation between test score and criterion must be .35 or higher.

Typical criterion measures:

- Student rating of ability to meet course requirements
- Instructor rating of students' abilities to meet course requirements
- Midterm grades or test scores
- Final course grades or test scores

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: Consequential-related Validity

Examination of the intended and unintended consequences of test use

EXAMPLE: After the first few weeks of a course (4th-6th week), how do students/instructors evaluate the class placement?

- Students: Have you been placed in the correct level or should you have been placed at a higher or lower level?
  - At least 75% students affirm

- Instructor: How ready is each student placed in the class based on the test score to undertake the course material?
  - At least 75% students are considered properly placed by instructors

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: Criterion-/Consequential-related Validity Studies

- Samples demographically representative of students served by CCC
  - At least 30 students from each course
- Data from at least 3 different districts (Probationary approval)
  - 4 districts (Provisional) & 6 districts (Full)
  - Majority in CCC (temporary exemption)
- Can aggregate results across colleges but must report findings for each college separately
- Validity evidence for each reported scores (e.g. total and subscores)

EXAMPLES
Criterion- and Consequential-related Validity

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

BUROS
CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Fairness Overview

Logical and Empirical

What is Fairness?

Construct = related knowledge, skills, or other attributes (KSA) a test is intended to measure

A test that is fair

- reflects the same construct for all test takers
- scores have the same meaning when interpreted and used for all individuals in the intended population

A test that is unfair

- occurs when aspects of the test or testing process produces construct-irrelevant variance in scores
- systematically raises or lowers the scores for an identifiable group of test takers

Fairness as Absence of Bias: Potential Sources of Bias

Cognitive sources:

- Knowledge/skill needed to respond to an item but is not measured (e.g. unnecessarily difficult language)
- Possession of unrelated knowledge/skill groups differ on (e.g. sports)

Affective sources:

- Offensive/sensitive topics (e.g. suicide)
- Inappropriate labeling of groups (e.g. retarded)

Physical sources:

- Formatting (e.g. small or decorative fonts)

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Attributes of a Fair Test: More than Bias

Accessibility – all test takers have an unobstructed opportunity to demonstrate their standing on the construct of interest

- Universal Design approach seeks to maximize accessibility for all intended examinees

Attributes of a Fair Test

Considered in all steps in the testing process

- Not just design and development
- Administration, scoring, test use and interpretation

Potential fairness threats

- test content
- test context
- test response
- opportunity to learn

CCC Standards: Fairness

Logical review: Involves qualitative judgment of a panel of whether items/prompts/tasks are unfair (biased, offensive, or sensitive)

AND

Empirical review: involves (statistical) analysis comparing different groups' responses on each item/prompt

Ideally, other considerations such as the use of universal design in development and the fairness review of administration, scoring, recommended interpretation and use

CCC Standards: Fairness Logical Review

Qualitative review by representative panel of whether items/prompts/tasks are unfair (cognitive, affective, and physical)

Must describe:

- Panel qualifications/background described (represent CCC protected classes 2% or more)
- 2 panelists from each protected group
- Independent of item writers and test developer
- Review training, guidelines, and procedures
- Results of panel decision & response (removal, revision, or retention of items)

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: Fairness Logical Review (CBT/CAT)

Fairness review for all items in pool

Ensure lack of familiarity with technology unfairly impact students

- Training/assessment for examinees

Fairness review of test instructions for examiners and examinees

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Example
Fairness Logical Review

BUROS
CENTER FOR TESTING

University of Nebraska Lincoln
COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: Empirical Fairness Study

Empirical (statistical) study comparing the performance of different groups of students at item or test level

Must include:

- Sample a sufficient number of students from CCC protected classes
- Data collected in last 3 years
- Methodology and results described
- Detail follow-up of flagged items (removal, revision, or retain)

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Empirical Fairness review: Differential Item Functioning (DIF)

Reasons why students from different groups systematically respond differently to a test item, reasons could be:

- Item may be biased
- groups are different on the measured knowledge or skill

To resolve - match groups based on expected performance (level of construct)

DIF Studies

If students of equal standing on construct perform differently on item, item flagged by DIF procedure.

- Usually test score is used for matching
  - Score should be found to be reliable & valid
  - DIF result only helps spot potential statistical differences in standing

DIF Studies: Mantel-Haenszel Statistic

Frequently reported (but not solely) – Needs large samples sizes

Focal group: protected group (Hispanic)
Reference group: non-protected (white)

Results in terms of direction:

- Negative values: item more difficult for focal
- Positive values: item more difficult for reference

Results magnitude:

A – little to no difference (typically retained)

B – Small to moderate difference (removal/retention depends)

C – greatest difference (item is dropped or revised unless content is critical)

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: CAT and Performance

- Computer-Adapted Tests: Use different methods for analyzing group differences at item and test level
- Performance assessments: comparing scores of two groups across multiple points.

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Example Empirical Fairness Study

BUROS CENTER FOR TESTING

UNIVERSITY OF Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

BUROS
CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

RELIABILITY
Overview

Overview

What is Reliability?

Reliability is…how consistently "test" measures a given construct (accuracy or precision).

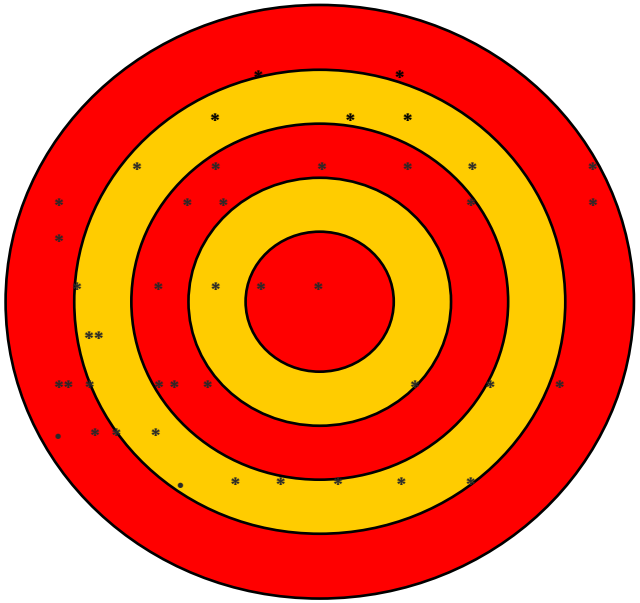
Validity is … degree scores/interpretations from a "test" are relevant.

Reliability is a necessary precondition for validity

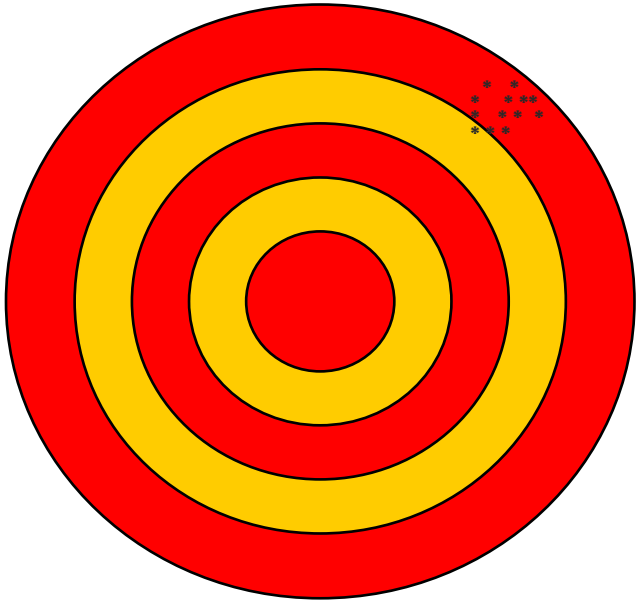
- Scores must be reliable before they can be valid
- Consistency allows for confident inferences/interpretation of scores

BUROS CENTER FOR TESTING

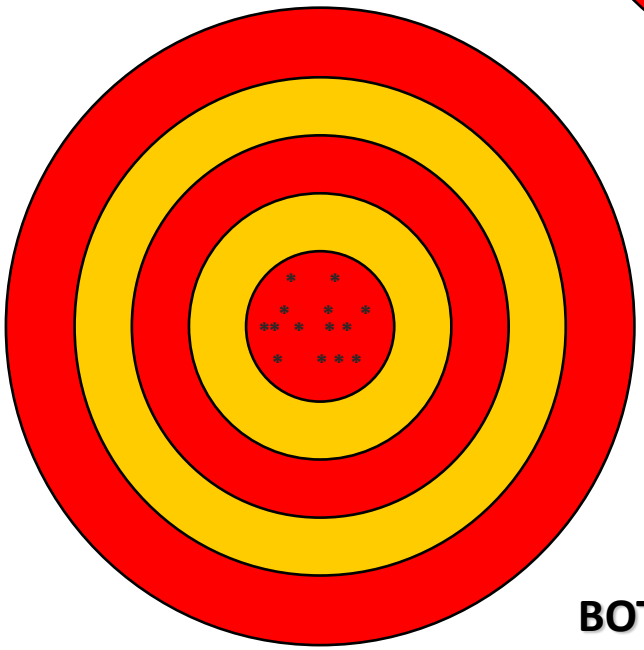
University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES



NEITHER
Reliable nor Valid



Reliable
but NOT Valid



BOTH Reliable and Valid

BUROS
CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Tests Are Not Reliable

A test is not reliable – reliability evidence is attributable to the **scores** from the test

- Depends on circumstances test is given
- Also sample dependent

Ideally…

Looking for reliability evidence from a sample of students similar to EL population at CCC in a context in which placement decisions will be made.

Types of Reliability Estimates

Test-Retest/Stability(1 form, 2 occasions)
Compare scores across time from 2 occasions for same group

Internal consistency (1 form, 1 occasion)
Compare responses across items within test from 1 occasion

Equivalent forms (2 forms, 1 occasion)
Compare scores across two measures same construct with different items/prompts/tasks but similar difficulty

Human scoring
Compare scores across multiple raters

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

CCC Standards: Reliability

Responsible all relevant error sources

- Test-retest (stability) particularly important for placement tests

Sample representative of CCC students (Minimum 50)

Report reliability estimate and standard error measurement(SEM) for all reported scores

If corrected estimates are reported, uncorrected should be as well.

CCC Standards: Reliability Estimate Minimums

Test-retest = .75 or higher

Internal consistency = .80 or higher

Equivalent form/inter-prompt = .75 or higher

Human Scorers (depends on estimate)

- Interscorer correlation = .70 or higher
- Percent agreement = 90% or higher (1 point difference)
- Cohens Kappa = .40 or higher
- Report how resolve inconsistencies between scorers

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

What is Standard Error of Measurement (SEM)?

Index related to reliability estimate

Degree of precision of test score

- Lower values preferred (less error)
- Based on standard deviation and reliability

EXAMPLE: Student score = 30

- SEM 3.5 30 +/- 3.5 = 26.5 to 33.5
- SEM 10 30 +/- 10 = 20 to 30

Important to know SEM across score distribution especially at consequential cut points

BUROS CENTER FOR TESTING

University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

Examples
Reliability

BUROS
CENTER FOR TESTING

University of Nebraska Lincoln
COLLEGE OF EDUCATION & HUMAN SCIENCES