BUROS
CENTER FOR TESTING
TEST REVIEWS • ASSESSMENT LITERACY
PSYCHOMETRIC CONSULTING
WWW.BUROS.ORG

Jessica L. Jonson, PhD
University of Nebraska-Lincoln

IUPUI Assessment Institute - October 28, 2012

CRITICAL TESTING AND MEASUREMENT CONCEPTS: ASSESSMENT PROFESSIONALS

Introduction: Speaker

- PhD in Educational Measurement
- 13 years experience HE Assessment
  - UNL Director of Institutional Assessment
- Commitment: Bridging the gap where it matters

2

Introduction: Buros

About:

- Independent, nonprofit organization (located University of Nebraska – Lincoln)
- Founded 1938 by Oscar Krisen Buros
- Moved to UNL in 1979 after Professor Buros' Death

Mission:

- Improve the science and practice of testing and assessment.
- Special emphases in psychology and education.
- World's premier test review center.
- Outreach through consultation and education.

3

Buros & Education

www.buros.org/assessment

4

Introduction: Participants

- Institution, Position, How Long?
- Why this workshop?

5

Goal

Provide information effective/responsible assessment professional

- Understand why testing/measurement knowledge can be helpful.
- Know what questions to ask when reviewing and evaluating a test or assessment.
- Recognize technical characteristics of quality tests/assessments
- Identify resources that available to identify and select tests/assessments.

6

BUROS

Why?

Testing/Measurement & Assessment Profession

7

BUROS

Diversity of Offices & Titles

Academic Administration
- *Provost/Assoc Provost:* Academic -Accreditation

Institutional Research
- *VP/Director:* Research - Planning -Effectiveness

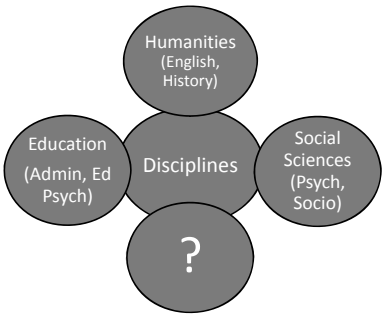
Teaching & Learning Center
* Excellence – Innovation - Effectiveness

Assessment Center/Office
- *Director/Coordinator/Manager:* Assess - Accreditation
- Institutional & College

8

BUROS

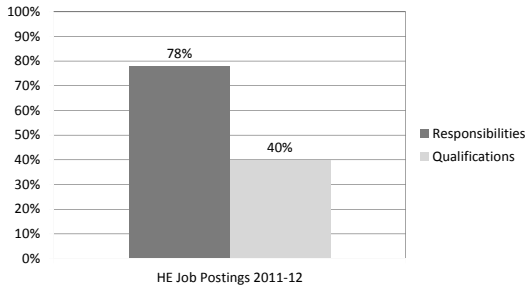
Diversity of Educ Degrees



9

BUROS

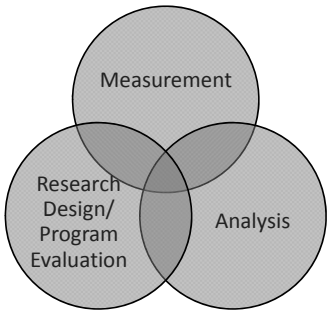
Qualifications & Responsibilities



10

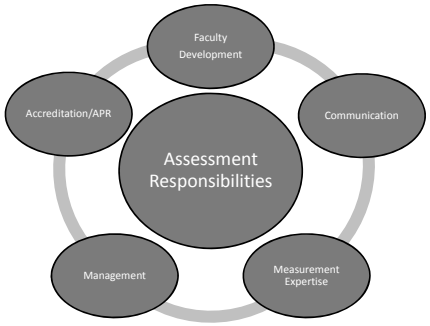
BUROS

Close relationships



BUROS

Responsibilities: Assessment Professionals

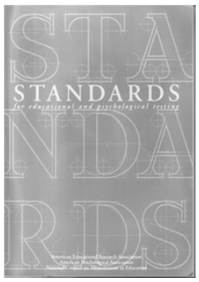


12

Testing Standards

Standards for Educational & Psychological Testing (AERA, APA, NCME 1999)

"…effective testing and assessment require that all participants in the testing process possess the knowledge, skills and abilities relevant to their role." (p.2)



Standards: User Qualifications

Standard 13.12

"In educational settings, those who supervise others in test selection, administration, and interpretation should have received education and training in testing to ensure familiarity with the evidence for validity and reliability of tests used…"

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

14

Code of Professional Responsibilities in Educational Measurement (1995)

- Goal: Conduct professionally responsible manner
- Public Service – all individuals involved in educational assessment activities
  - K-12 teachers/admin, Policy makers, College faculty/admin
- Applies formal/informal, traditional/alternative, large-scale/classroom scale

15

Code of Professional Responsibilities in Educational Measurement (1995)

- Develop Assessments
- Market and Sell Assessments
- Select Assessments
- Administer Assessments
- Score Assessments
- Interpret, Use, and Communicate Assessment Results
- Educate About Assessment
- Evaluate Programs and Conduct Research on Assessments

16

Code of Fair Testing Practice in Education (2004)

Guidance in four critical areas:

- Developing and Selecting Appropriate Tests
- Administering and Scoring Tests
- Reporting and Interpreting Test Results
- Informing Test Takers

Audience:

- Test developers: *construct tests, set policies for testing programs.*
- Test users: *select tests, administer tests, commission test development services, or make decisions on the basis of test scores.*

Why pay attention to the *Standards* and *Codes*?

- Provide guidance on test selection / development
- Guidance on assessment administration and score reporting
- Help avoid drawing misleading or inappropriate inferences
- Fairness in testing and test use
- Protect students' rights
- Implications for public policy

Are assessment professionals appropriately educated/prepared for their responsibilities?

Workshop Schedule

- Resources for Identifying/Selecting Tests
- Test/Rubric Construction
- Characteristics of Quality Tests
  - Standardization/Norms
  - Reliability
  - Validity

19

Resources for Identifying/Selecting Tests

20

Clarifying Purpose

- What is the specific purpose?
- Who will use the information for what decisions?
- Who will be helped by the assessment?
- What decisions and/or intended uses of the assessment results?
- What are the time and costs considerations?
- How in-depth will the assessment be?

21

Test Information & Reviews

BUROS publications

- Tests in Print (TIP)
- Mental Measurements Yearbooks (MMY)
- Test Reviews Online (TROL)

Other sources

- *PsychTESTS (APA)*
- *ETS Test Collection Database*

22

Test in Print (TIP)

- Comprehensive bibliography commercially available tests
- In 8th Edition,
- test purpose, test publisher, in-print status, price, test acronym, intended test population, administration times, publication date(s), test author(s), and score index
- Reference MMY reviews
- Indexed by title, acronym, subject, publisher
- NEW! *Pruebas Publicadas en Español (Fall 2012)* Listing of commercially available Spanish tests

23

Mental Measurement Yearbook

- Timely, consumer-oriented test reviews
- 18 MMY Editions, 19th Edition in progress
- Minimum tests reviewed in MMYs: 160
- Evaluate 80-100 tests each year
- Over 90% test reviews doctoral-level professionals
  - No conflicts of interest
- Service Agency

Reviews also are available through many libraries via Ovid or EBSCO subscriptions (no cost)

24

Test Reviews Online (TROL)

- Search and access reviews online (under construction)
- Free info on 3,500 commercial tests
- Search alphabetically or by category
- Provides partial TIP info
- Updated every 6 months
- Purchase reviews $15 per test title
- http://buros.unl.edu/buros/jsp/search.jsp

Reviews also are available through many libraries via Ovid or EBSCO subscriptions (no cost)

25

Demonstration

- Test Search
- Accessing Test Reviews

26

Buros Educational Resources

http://buros.org/assessment

1. Guides & Lessons
   - How to use TIP & MMY
   - Use MMY to Evaluate a Test
   - Questions to ask when Evaluating a Test
2. Clearinghouse Online Resources
3. Guidelines, Codes, Standards
4. Assessment Glossaries

27

American Psychological Association (APA)

- FAQ/Finding Information on Tests

http://www.apa.org/science/programs/testing/find-tests.aspx

- "Finding the Right Tools" for one's research (2007)

http://www.apa.org/gradpsych/2007/01/tools.aspx

28

PsycTESTS (APA, 2012)

http://www.apa.org/pubs/databases/psyctests/index.aspx

- Focuses primarily on unpublished tests
- Provides descriptive information, not reviews
- Information on 2,200 tests, ready access to 1,500 tests
- Information from peer-reviewed journals and books

29

Educational Testing Service (ETS) Website

ETS Test Collection database

http://www.ets.org/test_link/about

- Database of > 25,000 tests/measures
- Provides information on standardized tests and research instruments
  - Information is similar to that provided in *Tests in Print* (which parallels descriptive information in MMY, when available)
  - Does not offer expert reviews of tests, as does the MMY series
- From 1900s to present

30

BUROS

Tests/Rubric Construction

31

BUROS

Rubric Construction

Kansas State Dept of Education:
http://www.k-state.edu/ksde/alp/module7/

What: Set of rules for consistent rating/scoring performance assessment using learning criteria and performance descriptors.

Types: Generic & Task-specific

Rubric Scoring: Holistic & Analytic

32

BUROS

Test Construction

Steps:
1. Content analysis/blueprint
2. Item writing
3. Item review
4. Scoring plan
5. Piloting
6. Item Analysis (Review & Modify)
7. Final Test

UNESCO
- *www.unesco.org/iiep/PDF/TR_Mods/Qu_Mod6.pdf*

Indiana-Bloomington – Center for Teaching
- http://www.teaching.iub.edu/wrapper_big.php?section_id=assess#s2_7_assess_03_tests.shtml

33

BUROS

Characteristics of Test Quality

34

BUROS

Desirable Properties of Measurement Instruments

What are characteristics of a good test?

1. Standardization/Norms
2. Reliability
3. Validity

Why important?

- Contribute to the interpretability of scores

BUROS

Standardization

What does it mean for a test to be standardized?
- Uniform procedure: Administration and Scoring

How do standardization procedures across different measures vary?

- Examples: Multiple-choice & IQ

Questions: Test Administration

- What specific qualifications are needed to administer the test? What qualifications are needed to interpret results?
- Will the test administrator understand precisely what is expected of them?
- Do the test administration procedures parallel the conditions under which the test was validated and normed?

37

Existence of Norms

Why are existence of norms important for measures especially measure of latent traits?

- Latent Constructs: no inherent metric
- Norm group: Gives meaning

(NRTs) vs. (CRTs)

- Norm Referenced Tests (NRTs)
  - Goal: Discriminate among individuals
  - Scores: compared relative to others
- Criterion Referenced Tests (CRTs)
  - Goal: Assess mastery on set of tasks
  - Scores: compared to standard of performance within a domain

CRTs are used to determine "… *what test takers can do and what they know, not how they compare to others*." (Anastasi, 1988, p. 102)

Test Content - Items

- NRTs
  - Domain: Broadly-defined
  - Item selection goal: maximal discrimination
  - # of items: fewer (touch upon)
- CRTs
  - Domain: Narrowly-defined
  - Item selection goal: match to outcomes
  - # of items: more (thoroughly address)

Test Interpretation (NRT)

- NRTs
  - Interpretation: level of performance
  - Relative to: other individuals/groups (norm group)
    Example: 34th percentile rank
- CRTs
  - Interpretation: mastery decisions
  - Relative to: achievement of educational objectives
  - Example: mastered 70% material

Norm Reference Group: Factors & Types

- Factors
  - Recency "Time Bound"
  - Representativeness "Sampling Error"
  - Relevance "Comparable"
- Types
  - National Norms
  - Special Group Norms
  - Local/State Norms
  - School Norms

Norm Referenced Test Scores

Scale Type	Interpretation
Percentiles	Percentage of norm group surpassed by examinee
Standard scores	Performance expressed relative to the mean
Grade norms	Performance matched to that of students in a given grade
Age norms	Performance matched to that of students at a given age

Questions: Appropriate Sample

- What methods were used to select samples for test development, validation, and norming? Are these samples representative of the intended population?
- Was the sample large enough to develop stable estimates of test statistics? If subgroups are present, are there enough examinees in each (generally > 100)?
- Is the range of scores sufficient to provide an adequate basis for making validity claims and for norming?
- Do scores vary enough to make distinctions among examinees?
- How are test results reported? Are the scales used in reporting results conducive to proper test use?

44

Application

Examples of:
- Criterion-referenced test
- Norm-referenced test

Select sample for norm-referenced test:
- Size
- Demographics

Determine score for reporting:
- Criterion-referenced test
- Norm-referenced test

45

Reliability

46

Reliability

Why is reliability important characteristic of measure?
- Documents consistency – accuracy & precision

What are different definitions of reliability for measures?
- Consistency: within test, across time, across alternate forms, across raters

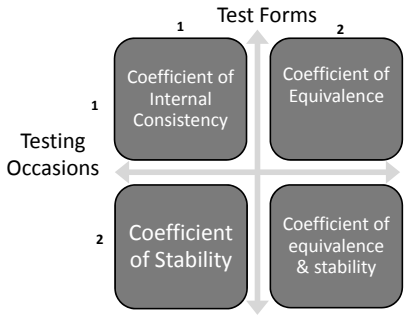
Reliability & Validity

- Reliability: accuracy or precision
- Validity: relevant to the inferences
  - Scores must be reliable before they can be valid

Hypothetical Example:
Child's Height
Measure 10 days
Measure vary 57.80 & 58.20

- True Score: constant (actual) height
- Observed Score: 10 height measurements
  - Observed score = True score + error

Forms for estimating reliability



Tests Are Not Reliable

- A test is not reliable it is the scores from the test.
  - Depends on circumstances test is given
  - Always report the reliability estimate from your sample

Reliability Summary Matrix

	Reliability Estimate		
Type of Error	Test-Retest	Alternate Forms	Coefficient Alpha
Content differences across test forms		√	
Temporal stability	√	?	
Within-test inconsistency			√
Differences among judges ratings			

Questions: Reliability

- How have estimates been computed?
- If test-retest reliability is reported, are time intervals reported? If so, are they appropriate given the stability of the trait being assessed?
- Are reliability estimates sufficiently high (typically > .90) to support using test scores as the basis for decisions about individual examinees?
- Are estimates provided for (a) different groups of examinees and (b) all scores for which interpretations are indicated?

52

Application: Reliability

A test is to be used to assess current levels of academic achievement and to predict academic achievement after one year. What evidence of reliability is appropriate to support each test use?

53

Interrater Agreement and Reliability

- Used scores from subjective judgments
- Consistency in scores from two or more raters are consistent
  - **Percentage of Agreement**
  - **Cohen's Kappa**
  - Intraclass Correlation (ICCs)
- Sources of error
  - Unclear criteria
  - Inadequate rater training

Percentage of Agreement

- Quantifies the number agreements (raters or occasions)
- Scores: All Types of data
- Drawback: Does not consider random chance

Percentage Agreement Example

- The percentage of agreement is 77%

		Rater 2: Pass	Rater 2: Fail	
Rater 1	Pass	57	10	67
	Fail	13	20	33
		70	30	100

Cohen's Kappa (1960)

- Level of agreement relative to chance agreement
  - Values > 0 agreement exceeds chance (-1 to +1)
  - Values = 0 agreement due to chance
- Values greater than **zero** indicate that agreement exceeds chance level
- Apply categorical data only

Cohen's Kappa Formula

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

- P_o = proportion of observed agreement
- P_c = proportion of chance agreement

$$P_c = \sum (P_{i.})(P_{.i})$$
$$= (P_{row1})(P_{col1}) + (P_{row2})(P_{col2}) + \ldots + (P_{rowk})(P_{colk})$$

- $P_{i.}$ And $P_{.i}$ = proportion in each row and column

Cohen's Kappa Example

- The proportion of observed agreement is .77

		Rater 2: Pass	Rater 2: Fail	
Rater 1	Pass	57	10	.67
	Fail	13	20	.33
		.70	.30	100

Cohen's Kappa Example

- The proportion of chance agreement is .568

$$P_c = \sum (P_{i.})(P_{.i})$$
$$= (.67)(.70) + (.33)(.30) = .568$$

		Rater 2: Pass	Rater 2: Fail	
Rater 1	Pass	57	10	.67
	Fail	13	20	.33
		.70	.30	100

Jessica L. Jonson, PhD
Buros Center for Testing
University of Nebraska – Lincoln

Kappa Computations

- Kappa is subsequently computed as .4676
- K > .70 (Interrater reliability is satisfactory)

$$\kappa = \frac{P_o - P_c}{1 - P_c} = \frac{.77 - .568}{1 - .568} = .4676$$

Weighted Kappa

- Drawback Kappa: too strict w/ ordinal judgments

Example: Mild, Moderate, Severe

Mild & Severe greater difference than moderate and severe

- Weighted Kappa: credit for partial agreements

Validity

63

Validity

What information does validity provide about the characteristics of a measure?

- Clarity of score interpretations relevant to inference

How is validity established?

- Pattern of relationships consistent with theory

Definition of Validity

- Standards for Educational & Psychological Testing – AERA, APA & NCME(1990)

Validity refers to the degree to which evidence and theory supports the ***interpretations*** of test ***scores*** entailed by the proposed ***uses*** of the ***test***

Nature of Validity

- Distance between measuring and interpretation
- Scores (numbers) produced are useful for intended purpose
- Examples:
  - Tape Measure (measuring length & college admission)
  - Reading Comprehension (score inference)
  - Emotional Adjustment (quality of behavior)

Traditional View of Validity

- Three "types" of validity
  - Content
    - How well does the test represent the trait?
  - Criterion-related
    - How well does the test predict performance?
  - Construct
    - What do scores on the test mean?

Important Aspects of Validity

Validity is …

- matter of degree
- Inferences/actions not test or score
- ongoing, accumulates
- examination of consequences (actual and potential)

Content Validity

- Haynes et al. (1995) defined content validity as follows:
- "Content validity is the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose." (p. 238)

Defining Test Content

- Two purposes:

1) Write Items
2) Determine if test matches domain of interest

Content Validity: Application

- Table of specifications
- Considerations

Establishing Content Validity

- Multiple judges/content experts
- Rate relevance, specificity, representativeness, clarity, etc.
- Quantitative rating scales (e.g., 5-point scales)

Criterion-Related Validity

- Correlation between test score and chosen criterion
- Two types:
  - Predictive validity (future performance)
  - Concurrent validity (current performance)

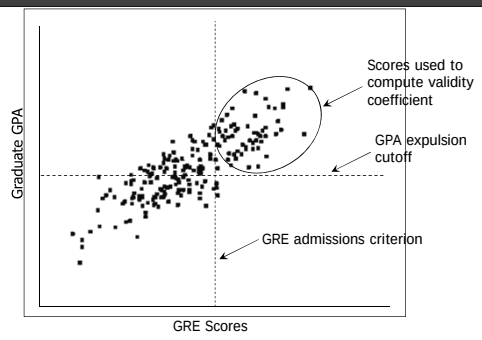
The Criterion Problem

- Major obstacle: suitable criterion
  - relevance, specificity, freedom from bias, reliability, and validity
- Criterion scores subject rigorous validation

Range Restriction

- Variability of test & criterion scores restricted
- Attenuates correlation coefficient

Graphic Representation



Application: Criterion Validity

What are the advantages and disadvantages of using freshman grade-point average as a criterion measure for validating a college admission test?

77

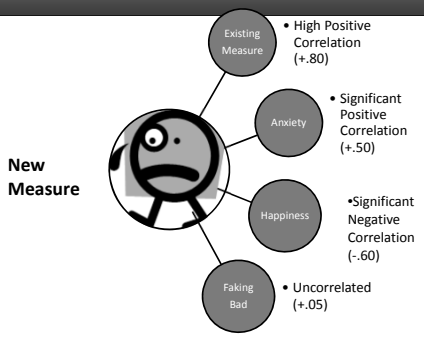
Construct Validity

- Construct: attribute seek to measure
  - social sciences, constructs latent traits (e.g. IQ)
- "What do scores on this test mean?"
- Theory based (e.g. academic achievement)

Nomological Network

- Construct validity involves "nomological network"
  - Relationship among constructs
  - What construct is, but also what it <u>is not</u>

Depression Example



Establishing Construct Validity

- Different tasks
  - "Gold standard" instrument
  - Factor analysis
  - Convergent and discriminant validity
  - Experimental studies
  - Structural equation modeling

Factor Analysis Example (1)

- 2 clusters, or dimensions
  - F1 = B1-B3
  - F2 = C1-C3

	B1	B2	B3	C1	C2	C3
B1	1.0					
B2	.55	1.0				
B3	.50	.60	1.0			
C1	.20	.15	.10	1.0		
C2	.20	.25	.15	.75	1.0	
C3	.25	.20	.25	.80	.70	1.0

Factor Analysis Example (2)

- Reveal only a single cluster
- Validity evidence would not be established

	B1	B2	B3	C1	C2	C3
B1	1.0					
B2	.40	1.0				
B3	.50	.45	1.0			
C1	.43	.49	.51	1.0		
C2	.47	.43	.46	.42	1.0	
C3	.42	.46	.52	.44	.45	1.0

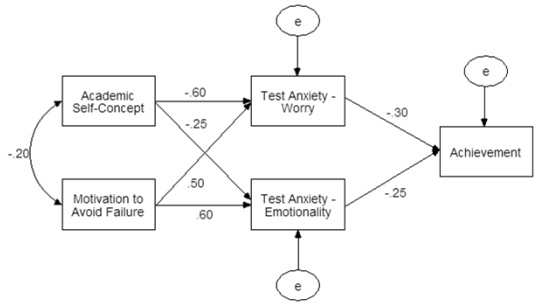
Depression Example

- Hypothesize correlations based on theory
  - Convergent
    - A strong positive with anxiety measure
    - A strong <u>negative</u> with happiness measure
  - Discriminant
    - Minimal (or zero) correlations with physical health and faking bad

Depression Example Correlation Results

	New Depression	Beck Depression	Anxiety	Happy	Health	Fake Bad
New	1.0					
Beck	.80	1.0				
Anxiety	.65	.50	1.0			
Happy	-.59	-.61	-.40	1.0		
Health	.35	.10	-.35	.32	1.0	
Fake Bad	.10	.14	.07	-.05	.07	1.0

SEM:Test Anxiety Example



Threats to Construct Validity

- Scores imperfect indicators
- 2 failures
  - Construct underrepresentation (excludes)
  - Construct-irrelevant variance (includes)

Questions: Validity

- What evidence exists to support claims that test content matches test specifications? Were expert panels used? If so, what was the composition of these panels?
- Is there a clear statement of the universe of knowledge, behaviors, and skills represented by the test?
- What criterion measure was used to provide validity evidence and what rationale supports its selection?

Continued…

88

Questions: Validity (con't)

- Is the psychometric quality of the criterion measure adequate?
- Are scores on the criterion and target test sufficiently distributed to avoid problems re: range restrictions?
- What is the overall predictive accuracy of the test? How accurate are predictions for examinees who score near the cut point(s)?
- Are experimental studies well designed and are the conclusions consistent with the findings?

89

Application: Validity

If an organization was developing a math test for the following general education learning outcome, what steps might the developer take to validate that test?

Use mathematical, computational, statistical, or formal reasoning to solve problems, draw inferences, and determine reasonableness.

90

BUROS

Other Topics

- Test construction (M-choice & Essay)
- Item development & analysis
- Test/Item Bias
- Rubric Design & Admin
- Other topics of interest?

91

BUROS
CENTER FOR TESTING
TEST REVIEWS • ASSESSMENT LITERACY
PSYCHOMETRIC CONSULTING
WWW.BUROS.ORG

Jessica L. Jonson, PhD
jjonson@buros.org