Validity and Reliability in Medical Education Assessment: Current Concepts

Congreso Nacional De Educacion Medica
Puebla, Mexico
12 January, 2007

Steven M. Downing, PhD
Department of Medical Education
University of Illinois at Chicago
sdowning@uic.edu

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

- How are these numbers properly interpreted?
- Many questions to answer in order to understand what these numbers mean
- We need much more information

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

- What number scale?
- Are these test scores?
  - Counts?
  - Percent-correct?
  - Ranks?
  - Standard scores?
  - Percentiles?
- Scores on what exam?
  - Exact content tested?
- What type of test?
  - Cognitive achievement
  - Standardized performance
  - Observation of clinical performance?

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

How can these numbers be properly interpreted?

<u>What must be known</u>

- **Test scores**
- **Percent-correct scores**
- **Final MCQ exam in pathophysiology**
- **250 total MCQs**
- **Cumulative course content**
- **Items/test developed by instructors**
  - **Used systematic sampling plan for content**
  - **Sampled all instructional objectives**
  - **Emphasized higher cognitive levels**

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

But… MORE INFORMATION NEEDED

- **How trustworthy are these test scores?**
- **How reproducible are these scores?**
- **What is average difficulty of MCQs on this test?**
- **What is average discrimination of MCQs on this test?**
- **Quality of MCQs?**
  - **Well written, edited?**
  - **Evidence-based principles?**
  - **Content review, revision?**

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

STILL MORE INFORMATION MAY BE NEEDED

- **How do scores on this test relate to scores on similar/different tests?**
- **Sensible, expected relationships?**
- **Fit to theory?**
- **Evidence of a single achievement or ability construct?**
- **Any unexpected relationships?**

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

YET MORE INFORMATION

- **What is the passing score? Grade levels?**
  - **How was cut score established?**
  - **How defensible is cut score?**
  - **Is pass score acceptable?**
- **Consequences of failing this test?**
  - **To students?**
  - **Faculty?**
  - **Schools?**

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

Answers to these types of validity questions provide some scientific evidence concerning the meaning or the proper interpretation of assessment data

What do these numbers mean?

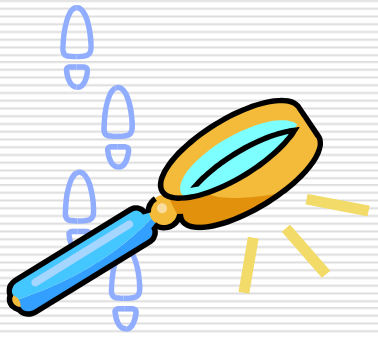
80
55
47
99
94
39
68
71
79
56
88
93
86
88

Validity research searches for evidence, like a detective

What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

Many different sources and types of scientific evidence to support or refute specific interpretations of assessment data



What do these numbers mean?

80
55
47
99
94
39
68
71
79
56
88
93
86
88

Validity concerns inferences, interpretations, and meaning associated with assessment scores

Validity

"Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores and other modes of assessment."

Messick, 1989

Validity

"To validate a proposed interpretation or use of test scores is to evaluate the claims being based on the test scores. The specific mix of evidence needed for validation depends on the inferences being drawn and the assumptions being made."

Kane, 2006

Overview

- Modern views of test validity
  - Scientific evidence needed to support test score interpretation
    - Cronbach, Messick, Kane
    - *Standards of Educational & Psychological Testing (1999)*
  - Some theory, key concepts, examples
- Reliability as part of validity

Validity

- Validity = Scientific evidence, using theory and research, to help explain interpretation of scores
- Essence of all assessment in education
  - Assessments derive meaning only through validity evidence
  - Measurement in social sciences: Little or no intrinsic meaning
- Nearly all topics in measurement fall under the broad rubric of validity

Contemporary View of Validity

- **All validity is construct validity**
- **Validity as hypothesis**
  - Scientific method applied to assessments
  - Theory, hypothesis, observation, analysis, results, conclusions: Repeat

Validity Principles

- Validity research: more or less evidence for or against specific uses of assessment scores
  - Purpose, intended interpretation, meaning
  - Multiple sources of scientific evidence
  - Higher the stakes, the more evidence required

Validity and Science

"A proposition deserves some degree of trust only when it has survived serious attempts to falsify it."

Cronbach, 1980

Classic View of Test Validity

- Traditional trinitarian view of validity
  - Content
  - Criterion-Related
    - Concurrent
    - Predictive
  - Construct
- Tests were “valid” or “invalid”
- Reliability was a separate test trait

Five Sources of Evidence

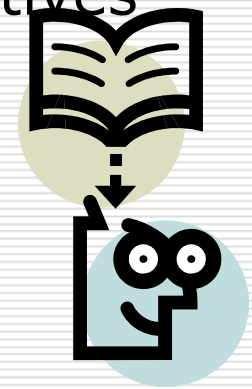
1. **Test Content – Task Representation → Construct Domain**
2. **Response Process – Item Psychometrics**
3. **Internal Structure – Test Psychometrics**
4. **Relationships with Other Variables – Correlations**
   - **Test-Criterion Relationships**
   - **Convergent and Divergent Data**
5. **Consequences of Testing – Social context**

Standards for Educational and Psychological Testing, 1999

Sources of Validity Evidence: Test Content

D*etailed understanding of the content sampled by assessment and relationship to content domain*

- Content-related validity studies
  - Exact sampling plan, specifications, blueprint
  - Representative sample of items/prompts → Domain
  - Appropriate content for instructional objectives
    - Cognitive level of items
    - Match to instructional objectives
  - Content expertise of item/prompt writers
  - Expertise of content reviewers
  - Quality of items/prompts



Sources of Validity Evidence: Response Process

Fit of student responses to hypothesized construct?

- Basic quality control information – accuracy of item responses, recording, data handling, scoring
- Statistical evidence that item measures intended construct
  - Achievement items measure intended content and not other content
  - Ability items predict targeted achievement outcome
  - Ability items fail to predict a non-related ability or achievement outcome

Sources of Validity Evidence: Internal Structure

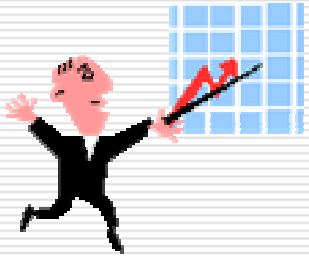
Statistical evidence of the hypothesized relationship between test item scores and the construct

- **Reliability**
  - Test scale reliability
  - Rater reliability
  - Generalizability
- **Item analysis data**
  - Item difficulty and discrimination
  - MCQ option function analysis
  - Inter-item correlations
- **Scale factor structure**
- **Dimensionality studies**
- **Differential item functioning (DIF) studies**

Sources of Validity Evidence: Relationship to Other Variables

Statistical evidence of the hypothesized relationship between test scores and the construct

- Criterion-related validity studies
  - Correlations between test scores/subscores and other measures
  - Convergent-Divergent studies



Sources of Validity Evidence: Consequences of Testing

Evidence of the effects of tests on students, instruction, schools, society

- The Big Picture
  - Consequential validity
  - Social consequences of assessment
- Effects of passing-failing tests
  - Economic costs of failure
  - Costs to society of false positive/false negative decisions
- Effects of tests on instruction/learning

Reliability

Reliability – One aspect of validity

- Reliability is one important type of validity evidence
  - Assessment data can be properly interpreted only if data are "reliable," scientifically reproducible
  - Without reliability, there can be no validity
- "Reliability is a necessary but not sufficient condition for validity."

Sources of Validity Evidence: Internal Structure

Statistical evidence of the hypothesized relationship between test item scores and the construct

- **Reliability**
  - Test scale reliability
  - Rater reliability
  - Generalizability

Reliability

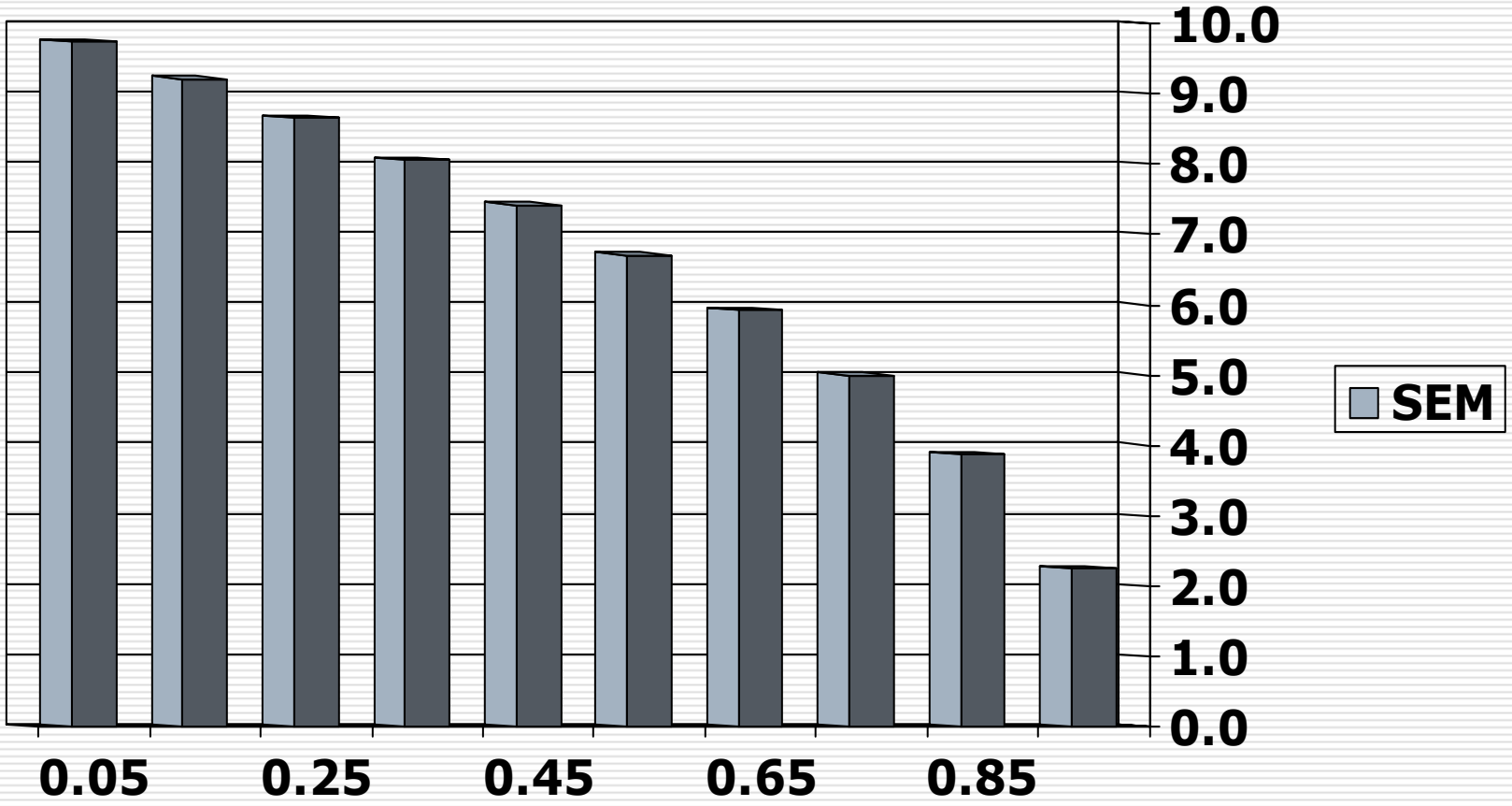
- Reproducibility of assessment data
  - Science requires reproducible experimental data
  - Assessments are mini-experiments
- Evidence from reproducible data
  - Trustworthy
  - Consistent
  - Interpretable
  - Few random errors of measurement

Reliability – Precision

Index of Measurement Precision

- Low random errors of measurement = high reliability
- Statistical estimates of random error
  - Index: 0.0 to +1.0
  - High value better than low value
- Standard error of measurement (SEM)

Standard Error of Measurement as function of reliability



Reliability – Various Types

- Different types of assessments require different kinds of reliability
- Written MCQs
  - Scale reliability
  - Internal consistency
- Written CR—Essay
  - Inter-rater agreement
  - Generalizability Theory

Reliability – Various Types

- Oral Exams
  - Rater reliability
  - Generalizability Theory
- Observational Assessments
  - Rater reliability
  - Inter-rater agreement
  - Generalizability Theory
- Performance Exams (OSCEs)
  - Rater reliability
  - Generalizability Theory

Reliability – How high?

- How high must reliability be?
  - Higher the better! Always.
  - Depends on purpose of test
    - Very high-stakes: > 0.90 +
      (Licensure tests)
    - Moderate stakes: at least ~0.75
      (Classroom test, med school OSCE)
    - Low stakes: >0.60
      (Quiz, test for feedback only)

How to increase reliability?

- For Written tests
  - Use objectively scored formats
  - At least 35-40 MCQs
  - MCQs that differentiate high-low students
- For performance exams
  - At least 7-12 cases
  - Well trained SPs
  - Monitoring, QC

How to increase reliability?

- Observational Exams
  - Lots of independent raters (7-11)
  - Standard checklists/rating scales
  - Timely ratings

Summary

- Validity = Meaning
  - Evidence to aid interpretation of assessment data
    - Higher the test stakes, more evidence needed
  - Multiple sources or methods
  - Ongoing research studies
- Reliability
  - Precision of the measurement
  - One aspect of validity evidence
  - Higher reliability always better than lower

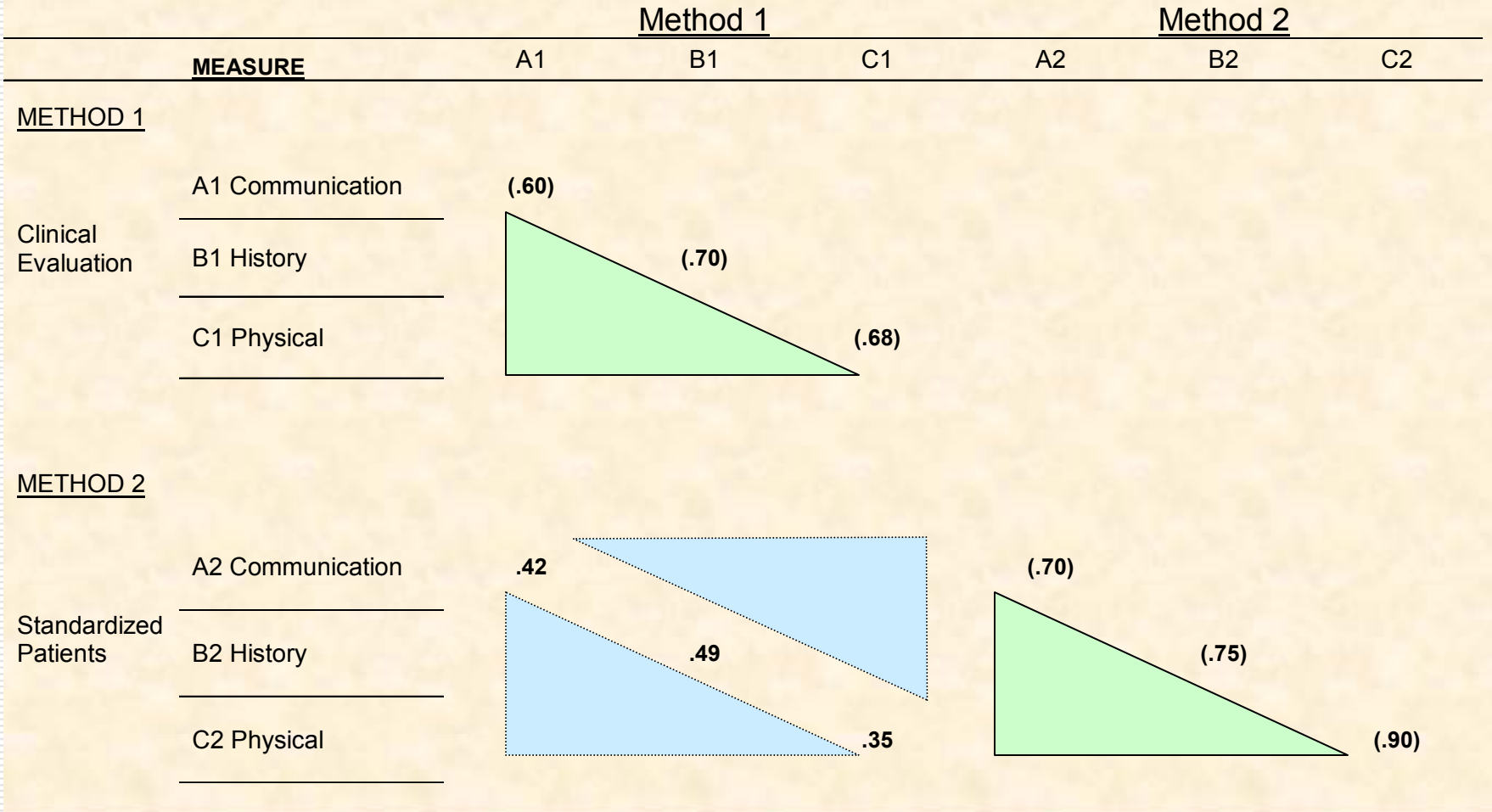
Classic Construct Validity Design: Convergent and Discriminant Studies

- Campbell and Fiske, 1959
- Empirical methods and procedures to collect, analyze data
  - Multiple methods, multiple measures
- Triangulation of Meaning and Interpretation
  - Rule in
  - Rule out

Classic Construct Validity Design for Two Clinical Performance Methods and Three Measures

EXAMPLE:

MULTITRAIT-MULTIMETHOD MATRIX



Based on design by Campbell and Fiske, 1959

Threats to Validity

- Two Major Sources of Validity Threats (Messick, 1989)
  - Content Underrepresentation (CU)
  - Construct-irrelevant variance (CIV)

CU: Content Underrepresentation

- Non-representative sample
  - Test fails to adequately sample population
  - Incorrect inferences to domain possible
- Examples
  - Too few essays (SR), oral prompts, MCQs, or OSCE cases to reliably sample domain

CIV: Construct-Irrelevant Variance

- Reliable measure of unintended construct
  - Good measure of an irrelevant construct
    - Anatomy essay test which measures writing skill more than anatomy
    - Written Psychiatry test which measures reading proficiency better than Psychiatric content
    - Internal Medicine performance test more associated with personality than patient communication competence
    - Oral exam in Pathology which is a better measure of student "stage presence" than understanding of path
  - Variable that interferes with intended interpretation or test score use