Measuring Educational Outcomes with Reliability and Validity

Facilitator: Ann Frye

Development Team:
Judy Shea, Karen Richardson-Nassif,
Larry Gruppen, and David Cook
Disclosure(s)
Overview of Today

Reliability
Small group exercise
Validity
Short review/wrap-up
Objectives - Reliability

Reliability is a characteristic of the scores rather than the test

3 types of reliability

Match reliability types with statistical measures

Select the type best for particular study
Reliability

Two inter-related concepts:

• Consistency

• Minimizing error
Reliability as Consistency
Reliability as Consistency
Reliability as Consistency
Reliability as Consistency
Reliability as Reducing Error

Classical test theory
Observed score = true score + error

\[ \text{systematic} \quad \text{random} \]

Reliability = extent to which random error is minimized

• As error is minimized, consistency should increase
3 Types of Reliability

1. Inter-rater (consistency over raters)

2. Test-retest and intra-rater (consistency over time)

3. Internal consistency (consistency over different items/forms)
Case #1: Effect of an Intervention

The issue:

• Students need to acquire good oral presentation skills.
Case #1: Effect of an Intervention

Students in the Medicine clerkship are randomized to 2 groups. One group is given an “oral case presentation” (OCP) booklet. They are to ask attendings/residents to rate/assess them 9 times over course of clerkship. The rating form has 7 items. At the end of the clerkship all students give an oral presentation. The rater, who uses the 7-item rating form, is blinded to Treatment/Control group assignment.
Case #1: Effect of an Intervention

What types of reliability should be assessed?
- Consistency over raters?
- Consistency over time?
- Consistency over items/forms?

Kim S, Kogan JR, Bellini LM, Shea JA. Effectiveness of Encounter Cards to Improve Medical Students’ Oral Case Presentation Skills: A Randomized Controlled Study. *Journal of General Internal Medicine* 2005; 20:743–747. In that study, they refer to a 9-point scale (slide 41) – pdf of article is attached.
Case #2: Survey Research

The issue:

• Identification and treatment of depression during medical school has important learning and behavioral implications.
Case #2: Survey Research

All students in all 4 years at a “private NE medical school” complete an anonymous questionnaire with demographic information, the Beck Depression Inventory, and self-report of treatment for depression and/or other emotional issues.
Case #2: Survey Research

What types of reliability should be assessed?
Consistency over raters?
Consistency over time?
Consistency over items/forms?

Case #3: Education Intervention

The issue:

• Residents need to learn to follow clinical guidelines regarding appropriate primary care prevention, screening, and safety issues.
Case #3: Education Intervention

Interns in medicine residency randomized to 2 groups. All residents had 8-10 charts abstracted for their primary care patients. A mean percentage adherence was calculated for several types of prevention guidelines. Interns in treatment group received feedback in the form of a report card with review from their attending. All interns had 10 charts reviewed at the end of the year (about 7 months later).
Case #3: Education Intervention

What types of reliability should be assessed?

Consistency over raters?
Consistency over time?
Consistency over items/forms?

3 Types of Reliability

1. Inter-rater (consistency over raters)

2. Test-retest and intra-rater (consistency over time)

3. Internal consistency (consistency over different items/forms)
Inter-rater Reliability

Multiple judges code independently using the same criteria

Reliability = raters code same observations into same classification

Examples

• medical record reviews
• clinical skills
• oral examinations
Measures of Inter-rater Reliability

Measures of agreement:
• Total percent agreement
• Cohen’s kappa

Measures of association:
• Pearson correlation coefficient
• Intraclass correlation
• Phi
Percent Agreement

% of agreement in coding between raters

Number of agreements / total number of cases (n)

Starts with a contingency table
## Percent Agreement

<table>
<thead>
<tr>
<th>Rater B</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater A</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES (Occurrence)</td>
<td>5 (A)</td>
<td>2 (B)</td>
<td>7 (G)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO (Nonoccurrence)</td>
<td>1 (C)</td>
<td>2 (D)</td>
<td>3 (H)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>6 (E)</td>
<td>4 (F)</td>
<td>10 (I)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total % Agreement = \( \frac{(A + D)}{I} = \frac{(5 + 2)}{10} = .70 \)
Percent Agreement

Pros

• Frequently used
• Easy to calculate
• Interpretation is intuitive

Cons

• Does not account for chance agreements
  ▪ HUGE point
Kappa

Controls for the problem of inflated percent agreement due to chance

Ranges from +1.00 to -1.00

• +1.00 = 100% of the agreement above chance possible

• 0 = no agreement above that expected by chance

• -1.00 = 100% of the disagreement below chance possible
# Kappa

<table>
<thead>
<tr>
<th>Rater A</th>
<th>YES (Occurrence)</th>
<th>NO (Nonoccurrence)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YES (Occurrence)</td>
<td>5 (A)</td>
<td>2 (B)</td>
<td>7 (G)</td>
</tr>
<tr>
<td>NO (Nonoccurrence)</td>
<td>1 (C)</td>
<td>2 (D)</td>
<td>3 (H)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>6 (E)</td>
<td>4 (F)</td>
<td>10 (I)</td>
</tr>
</tbody>
</table>

Observed agreement = .70

Chance agreement = \((E/I \times G/I) + (F/I \times H/I)\) = .54

Kappa = \((\text{Obs.} - \text{Chance}) / (1 - \text{Chance})\)

Kappa = \(.70 - .54) / (1 - .54) = .35\)

= 35% of the improvement possible above chance
## Kappa – example #2

<table>
<thead>
<tr>
<th>Rater B (Occurrence)</th>
<th>YES (A)</th>
<th>NO (B)</th>
<th>TOTAL (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES (Occurrence)</td>
<td>50</td>
<td>35</td>
<td>85</td>
</tr>
<tr>
<td>NO (Nonoccurrence)</td>
<td>15</td>
<td>900</td>
<td>915</td>
</tr>
<tr>
<td>TOTAL</td>
<td>65</td>
<td>935</td>
<td>1000</td>
</tr>
</tbody>
</table>

Observed agreement = \( \frac{50 + 900}{1000} = .95 \)

Chance agreement = \( \frac{E \times G}{I} + \frac{F \times H}{I} = .86 \)

Kappa = \( \frac{\text{Observed} - \text{Chance}}{1 - \text{Chance}} \)

= \( \frac{.95 - .86}{1 - .86} = .64 \)

= 64% of the improvement possible above chance
Kappa

General interpretation guidelines:

• 0 - 0.2 slight
• 0.2 - 0.4 fair
• 0.4 - 0.6 moderate
• 0.6 - 0.8 substantial
• 0.8 - 1.0 almost perfect
Limitations of Kappa

Sensitive to prevalence rates

• Higher kappas more likely when prevalence is near 50%, lower kappas more likely when prevalence is either high or low

Difficult to compare kappa across studies
Correlation Coefficients

Are not influenced by the number of coding categories

Indicate the direction/sign of the association
  • - sign...as one goes up, the other goes down
  • + sign...as one goes up, the other also goes up

Indicate the size of the association
  • $-1 = \text{perfect negative relationship}$
  • $+1 = \text{perfect positive relationship}$
Correlations

\[ r = .80 \]
positive
strong correlation

\[ r = - .60 \]
negative
moderate correlation

\[ r = 0 \]
neither positive nor negative
no correlation
Note that a perfect 1.0 correlation coefficient does not indicate perfect rater agreement.
Intraclass Correlation

Is a measure of changes in both magnitude and order:

• Magnitude: a change in mean value
• Order: a change in the order of data

Pearson’s correlation measures ordinal changes only
3 Types of Reliability

1. Inter-rater (consistency over raters)

2. Test-retest and intra-rater (consistency over time)

3. Internal consistency (consistency over different items/forms)
Test-Retest (& Intra-rater) Reliability

Give a test (make a rating - the rater as the instrument)

Allow time to pass

Give another test (make another rating)

Correlate the two test scores (ratings)
Test-Retest

Change in scores across test administrations is treated as error.

If trait being measured is stable, a change in score must be due to either:

• Measurement error
• Trait instability
Test-Retest Time interval

If too short, people may remember their responses (ratings)

If too long, the trait being measured may in fact have changed

A time interval of 2-4 weeks is generally recommended
3 Types of Reliability

1. Inter-rater (consistency over raters)

2. Test-retest and intra-rater (consistency over time)

3. Internal consistency (consistency over different items/forms)
Internal Consistency Estimates

Measures of internal consistency

Only requires one testing session

2 kinds:

1. Cronbach’s alpha (α)
   assesses homogeneity of *continuous* items

   assesses homogeneity of *dichotomous* items
Cronbach’s Alpha (\(\alpha\))

For continuous items

Preferred method of calculating internal consistency

Easy to interpret

The proportion of a scale’s total variance that is due to the true score on the measure -- as opposed to variance which is due to error

Ranges from 0 - 1
Interpreting α

General guidelines:

• .70 is adequate (although lower alphas are sometimes reported)

• .80 - .85 is good

• .90 or higher indicate significant overlap in item content -- scale can probably be shortened
# Summary of Reliability

<table>
<thead>
<tr>
<th>This reliability…</th>
<th>assesses this error…</th>
<th>and estimates…</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inter-rater</td>
<td>• rater/scorer</td>
<td>• rater reliability</td>
</tr>
<tr>
<td>2. Test-retest &amp; intra-rater</td>
<td>• individual changes over time or administration</td>
<td>• stability</td>
</tr>
<tr>
<td>3. Cronbach’s alpha</td>
<td>• sampling</td>
<td>• internal consistency</td>
</tr>
</tbody>
</table>
Factors Influencing Reliability

Test length
• Longer tests give more reliable scores

Group homogeneity
• The more heterogeneous the group, the higher the reliability

Objectivity of scoring
• The more objective the scoring, the higher the reliability
Concluding Remarks

Assess your reliability:

• Compute appropriate measure(s) of reliability

• Decide if reliability is adequate for your research goals

• Always report reliability coefficient for your particular sample—even with established measures
Remember

You never really know the ‘true score’

Reliability is an estimate

Speak of reliability of the scores of an instrument when applied to certain population
Objectives - Reliability

Explain that reliability is a characteristic of the scores rather than the test.

Identify 3 types of reliability.

Match types of reliability with appropriate statistical measures.

Select the type that is best for your particular study.
Exercise
Case #1: Effect of an Intervention

The issue:

- Students need to acquire good oral presentation skills.
Case #1: Effect of an Intervention

Students in the Medicine clerkship are randomized to 2 groups. One group is given an “oral case presentation” (OCP) booklet. They are to ask attendings/residents to rate/assess them 9 times over course of clerkship. The rating form has 7 items. At the end of the clerkship all students give an oral presentation. The rater, who uses the 7-item rating form, is blinded to Treatment/Control group assignment.
Case #1: Effect of an Intervention

What types of reliability should be assessed?

How?

Case #2: Survey Research

The issue:

• Identification and treatment of depression during medical school has important learning and behavioral implications.
Case #2: Survey Research

All students in all 4 years at a “private NE medical school” complete an anonymous questionnaire with demographic information, the Beck Depression Inventory, and self-report of treatment for depression and/or other emotional issues.
Case #2: Survey Research

What types of reliability should be assessed?
How?

Case #3: Education Intervention

The issue:

• Residents need to learn to follow clinical guidelines regarding appropriate primary care prevention, screening, and safety issues.
Case #3: Education Intervention

Interns in medicine residency randomized to 2 groups. All residents had 8-10 charts abstracted for their primary care patients. A mean percentage adherence was calculated for several types of prevention guidelines. Interns in treatment group received feedback in the form of a report card with review from their attending. All interns had 10 charts reviewed at the end of the year (about 7 months later).
Case #3: Education Intervention

What types of reliability should be assessed?
How?

Validity
Objectives - Validity

Explain that validity is a characteristic of the interpretation of the scores rather than of the test.

Describe the relationship between reliability and validity.

Describe five sources of validity evidence.

Select the best type of validity evidence for your particular study.
Reliability and Validity
...In Reality
Validity

Degree to which a test or instrument (e.g., scale, rating) measures what it was intended to measure (a construct) or operates as expected

A property of the interpretation given to the results, NOT a property of an instrument or even the scores, per se

Most scores on most measures are never perfectly valid or invalid
What is a construct (and why should I care)?

"An intangible collection of abstract concepts and principles"
What's the construct?

USMLE Step I
USMLE Step II
Beck Depression Inventory
CAGE questionnaire
Lung Cancer Symptom Scale
APACHE II
Kolb Learning Style Inventory
Why does this matter?

1. All instruments and assessment procedures are intended to measure a construct (inference)

2. All validity is construct validity
   • How well do instrument scores measure the intended construct
   • As applied to specific purpose (use)
Validity and Error

Classical test theory
observed score = true score + error

Systematic error threatens validity
(Recall that reliability was concerned with random error)

Systematic error comes from many sources
Threats to Validity

Construct under-representation

Construct-irrelevant variance

Assessment, measure or score

Both

Construct
Validity: Old Framework

Different types of validity

• Face
• Criterion
• Predictive
• Construct
Validity: Unified Framework

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests”.

AERA, APA, NCME, 1999
Validity: Unified Framework
The Validity Hypothesis

Validity is a hypothesis

Sources of validity evidence contribute to accepting (or rejecting) the hypothesis

The required amount of evidence varies with the type of assessment

Downing SM. Med Educ. 2003; 37: 830
Validity: Unified Framework

Not a dichotomous “valid” or “invalid” decision

NOT different types of validity for the measure

Different types of evidence for validity of judgments made on the basis of the scores
Sources of Validity Evidence

Content
Internal Structure
Relations to Other Variables
Response Process
Consequences


APA and AERA. Standards for Psychological and Educational Testing. 1999

Downing SM. Med Educ. 2003; 37: 830
Validity Evidence: Content

Themes, wording, and expert review

A description of steps taken to ensure items represent the target construct
Example of Content Evidence

Assessment criteria (constructs) were developed by faculty input and literature review.

Items were created to represent the defined constructs.

Experts revised the final item set.

Validity Evidence: Internal Structure

Degree to which items fit the underlying construct. Often measured using:

• Factor analysis, which identifies item clustering within constructs

• Internal consistency reliability, which demonstrates inter-item correlations
Example of Internal Structure Evidence

Factor analysis of clinical teaching assessments revealed 3 predicted domains of clinical teaching.

Internal consistency and interrater reliability were high.

Validity Evidence: Relations to Other Variables

The relationships between scores and other variables (criteria) relevant to the construct being measured

Can be determined using correlation coefficients, regression analysis, etc.
Example of Relations to Other Variables Evidence

Student-on-teacher assessment scores (a proxy for instructional quality) were significantly related to clerkship grades and choice of medical specialty

James PA, Osborne JW. Fam Med. 1999; 31: 263
Validity Evidence: Response Process

Examining the reasoning and thought processes of learners

Systems that reduce the likelihood of response error
Example of Response Process Evidence

Residents and Students, using the same rating forms and assessing the same teachers, yield significantly different scores

Implication: resident and student perceptions of teaching substantially differ

Validity Evidence: Consequences

Assessments have intended (often implied) consequences:

- Desired effect
- Intended purpose

Analyzing consequences of assessments support validity or reveal unrecognized threats to validity.
Example of Consequences Evidence

Clinical educators who performed poorly showed improvements in teaching after reviewing their assessment scores.

Implication: the assessment had consequences that impacted on validity

Remember

Speak of validity of the judgments made from the scores of an instrument when applied to certain population

NOT the validity of the instrument
Objectives - Validity

Explain that validity is a characteristic of the interpretation of the scores rather than of the test.

Describe the relationship between reliability and validity.

Describe five sources of validity evidence.

Select the best type of validity evidence for your particular study.
Exercise
Case #1: Survey Research

The issue:

• Identification and treatment of depression during medical school has important learning and behavioral implications.
Case #1: Survey Research

All students in all 4 years at a “private NE medical school” complete an anonymous questionnaire with demographic information, the Beck Depression Inventory, and self-report of treatment for depression and/or other emotional issues.
Case #1: Survey Research

What types of validity evidence should be assessed? How?

Case #2: Instrument Development

The issue:

• Students need to acquire strong clinical evaluation skills, including history-taking and physical examination.
Case #2: Instrument Development

Students in the Medicine clerkship were given a “mini-clinical evaluation” (mini-CEX) booklet. They were instructed to ask attendings/residents to rate/assess them 9 times over course of clerkship. The rating form has 9 items. At the end of the clerkship all students turned in the booklets.
Case #2: Instrument Development

What types of validity evidence should be assessed? How?

Case #3: Education Intervention

The issue:

• Residents need to learn to follow clinical guidelines regarding appropriate primary care prevention, screening, and safety issues.
Case #3: Education Intervention

Interns in medicine residency randomized to 2 groups. All residents had 8-10 charts abstracted for their primary care patients. A mean percentage adherence was calculated for several types of prevention guidelines. Interns in treatment group received feedback in the form of a report card with review from their attending. All interns had 10 charts reviewed at the end of the year (about 7 months later).
Case #3: Education Intervention

What types of validity evidence should be assessed? How?

References


References


References
