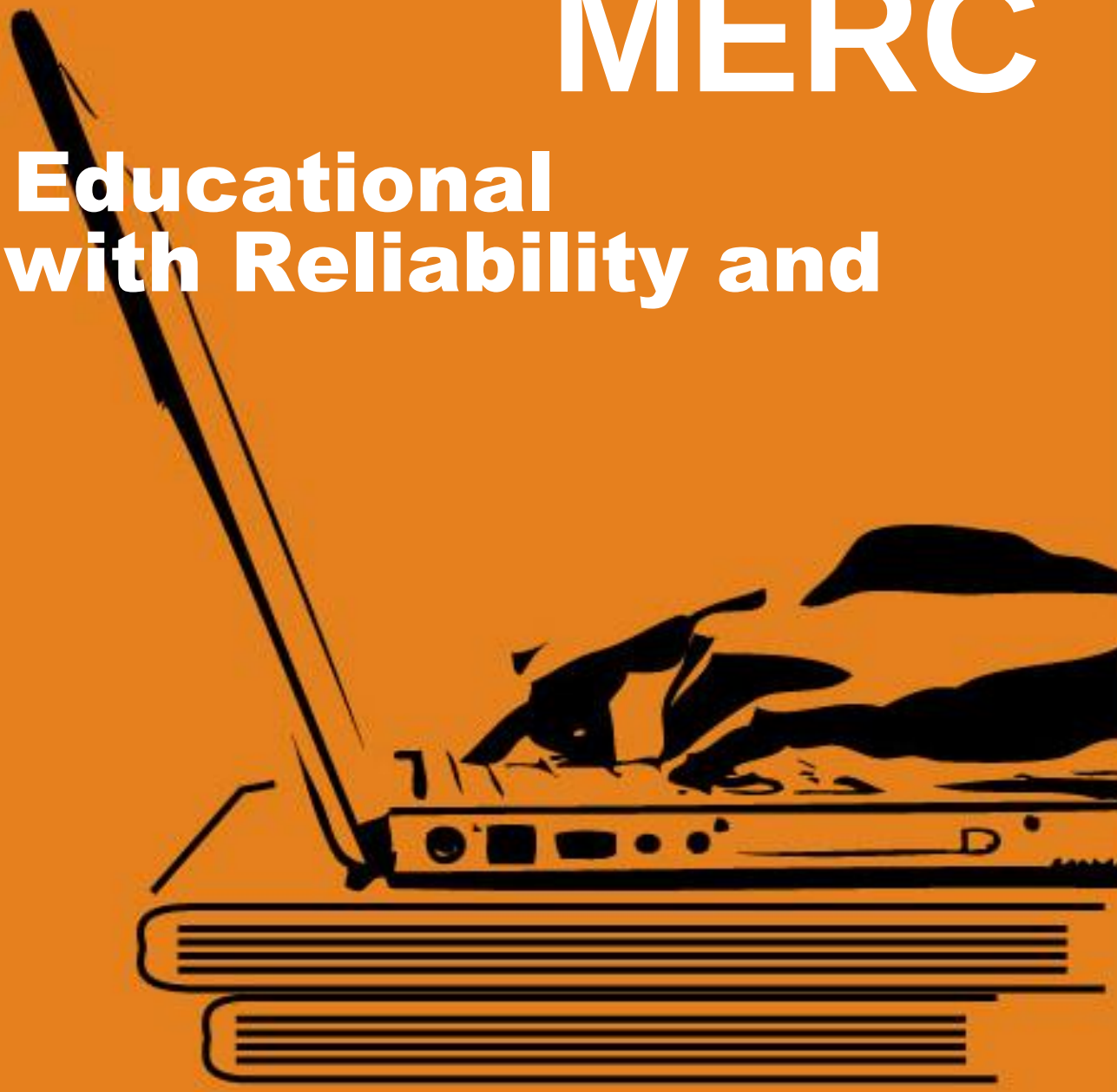
MERC

Measuring Educational Outcomes with Reliability and Validity



AAMC

Tomorrow's Doctors, Tomorrow's Cures®

Development Team: Judy Shea, Karen Richardson-Nassif, Larry Gruppen, and David Cook

Last reviewed date: May 2015

Reviewed by: Judy Shea, Anne Frye, Larry Gruppen

Objectives

Identify three types of reliability (inter-rater, test-retest, and internal consistency)

Match types of reliability with appropriate statistical measures

Describe the relationship between reliability and validity

Describe multiple forms of evidence for validity

Select an approach to reliability and validity assessment for a particular study

Overview of Today

Reliability

Small group exercise

Validity

Short review/wrap-up

Objectives - Reliability

Know that reliability is a characteristic of the scores rather than the test

Be familiar with 3 types of reliability

Match reliability types with statistical measures

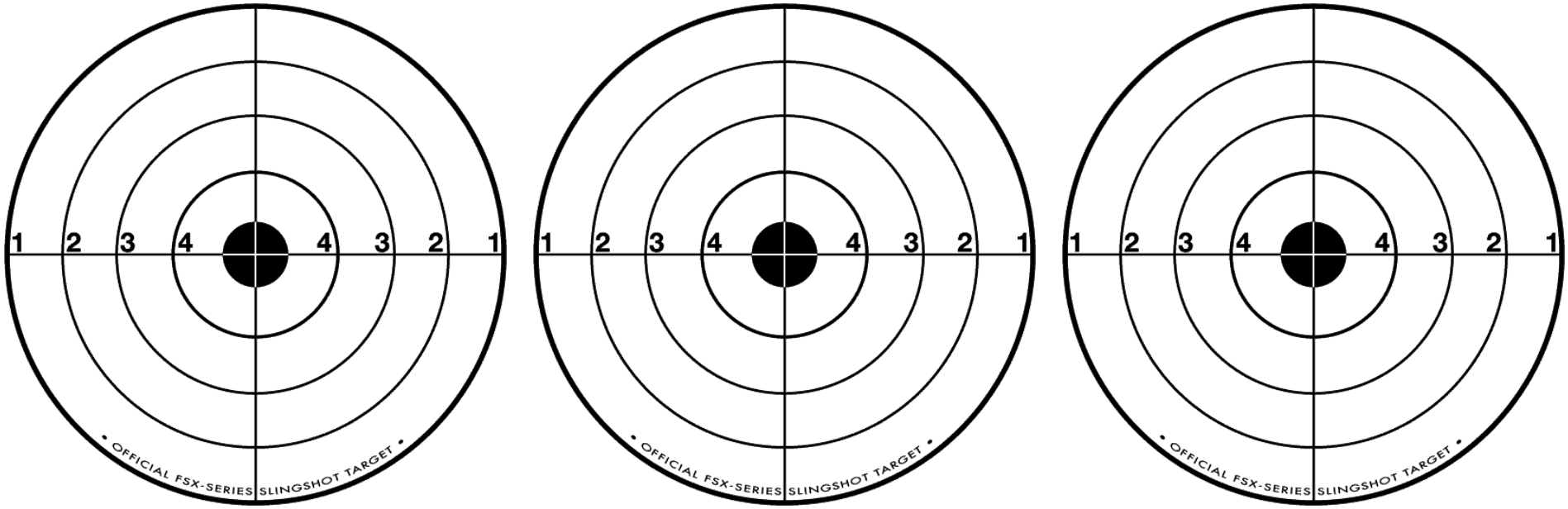
Select the best type for particular study

Reliability

Two inter-related concepts:

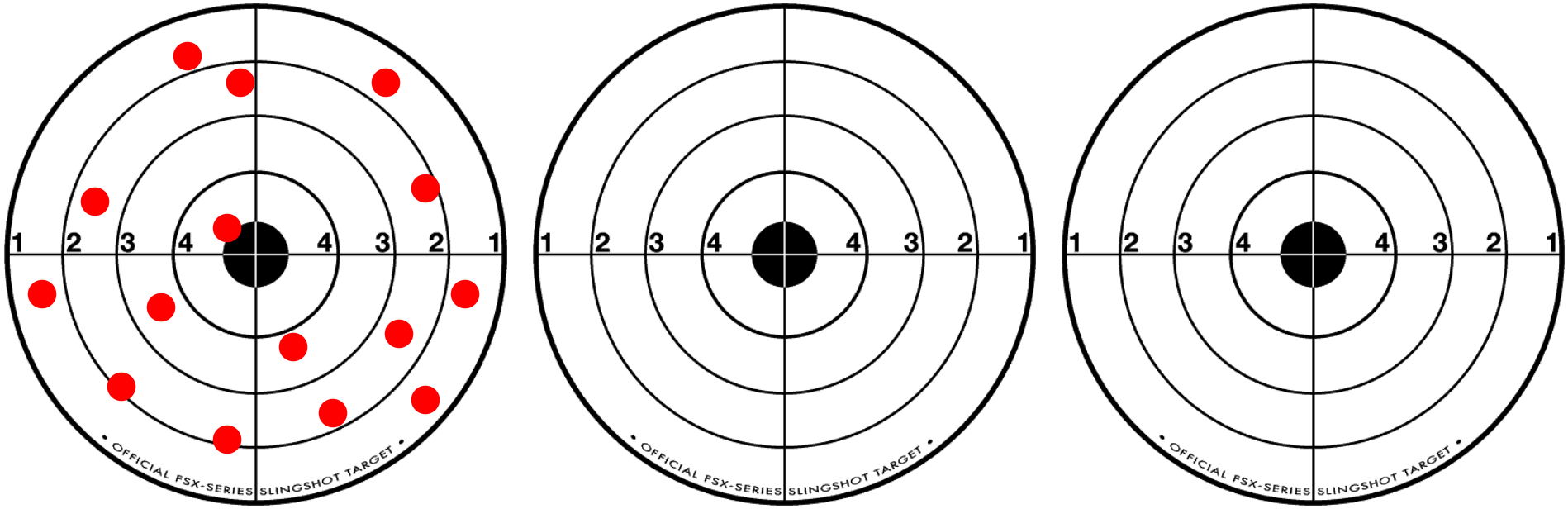
- Consistency
- Minimizing error

Reliability as Consistency



MERC AAMC

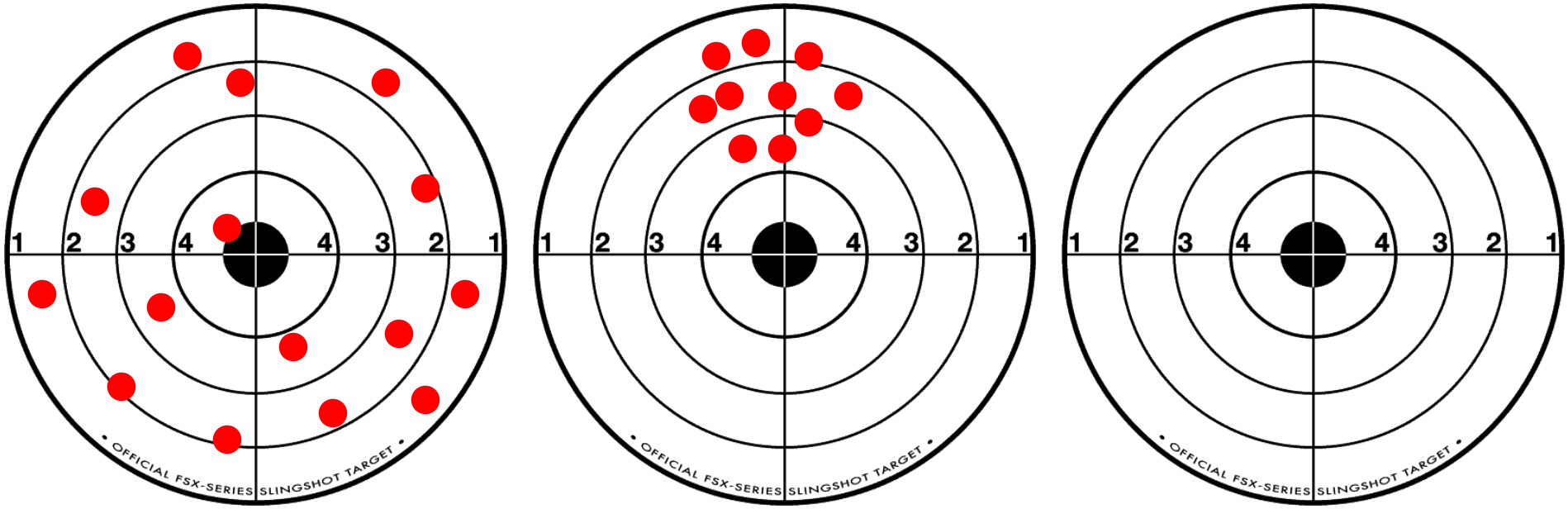
Reliability as Consistency



MERC

AAMC

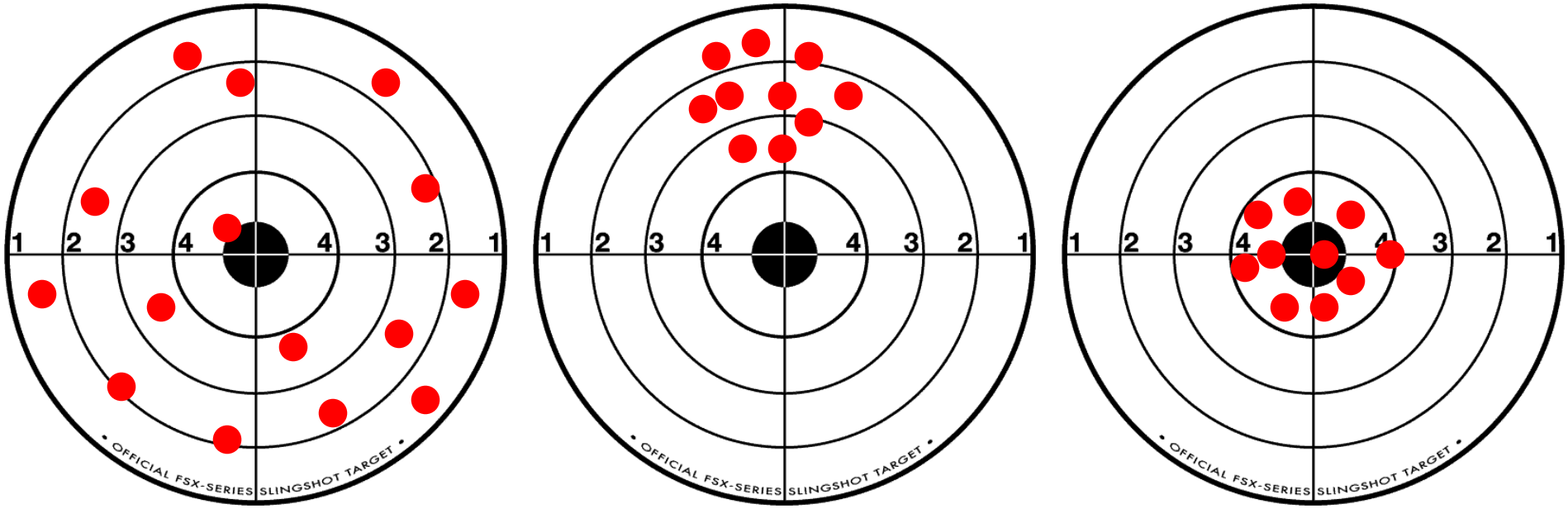
Reliability as Consistency



MERC

AAMC

Reliability as Consistency



Reliability as Reducing Error

Classical test theory

Observed score = true score + error

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:

The issue:

Students need to acquire good oral presentation skills.

Case #1:

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.

What types of reliability should be assessed?

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:

The issue:

Identification and treatment of depression during medical school has important learning and behavioral implications.

Case #2:

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.

What types of reliability should be assessed?

Tjia J, Givens JL, Shea JA. Factors associated with under treatment of medical student depression. Journal of American College Health, 2005 Mar-Apr;53(5):219-24.

Case #3:

The issue:

Residents need to learn to follow clinical guidelines regarding appropriate primary care prevention, screening, and safety issues.

Case #3:

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).

What types of reliability should be assessed?

Kogan JR, Reynolds EE, Shea JA. Effectiveness of report cards based on chart audits of residents' adherence to practice guidelines on practice performance: A randomized controlled trial. Teaching and Learning in Medicine: An International Journal, 2003;15:25-30.

Case #4:

The issue:

Resident work hour regulations have likely had an impact on student learning/educational activities

Case #4:

A time-motion study was done. Random samples of students (pre-reform and post-reform) were given pagers and called randomly, approximately every 90 minutes while they were in the hospital. At the time they responded to a short pocket survey and answered 4 questions: where they were, who they were with, type of activity engaged in, and a rating of the educational usefulness. There were two samples of students: one pre and one post reform

What types of reliability should be assessed?

Kogan JR, Bellini LB, Shea JA. The impact of resident duty hour reform in a medicine core clerkship. Academic Medicine, 2004;79:s58-s61.

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 independently code the same observations (learners or behaviors) 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

Rater A			
Rater B	YES (Occurrence)	NO (Nonoccurrence)	TOTAL
YES (Occurrence)	5 (A)	2 (B)	7 (G)
NO (Nonoccurrence)	1 (C)	2 (D)	3 (H)
TOTAL	6 (E)	4 (F)	10 (I)

Total % Agreement = (A + D) / I
= (5 + 2) / 10
= .70

Percent Agreement

Pros

- Frequently used
- Easy to calculate
- Interpretation is intuitive

Cons

- Does not account for chance agreements
  - This is a 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

Rater A			
Rater B	YES (Occurrence)	NO (Nonoccurrence)	TOTAL
YES (Occurrence)	5	2	7
NO (Nonoccurrence)	1	2	3
TOTAL	6	4	10

Observed agreement = .70

Chance agreement = correction based on observed marginal data – i.e., seeing how unbalanced the observed distributions are – 6 of 10 for Rater A and 7 of 10 for Rater B - the correction for chance is .54

Kappa = (Obs. - Chance) / (1 - Chance)

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

Kappa

Rater A			
Rater B	YES (Occurrence)	NO (Nonoccurrence)	TOTAL
YES (Occurrence)	5 (A)	2 (B)	7 (G)
NO (Nonoccurrence)	1 (C)	2 (D)	3 (H)
TOTAL	6 (E)	4 (F)	10 (I)

Observed agreement = .70
Chance agreement = (E/I x G/I) + (F/I x H/I) = .54
Kappa = (Obs. - Chance) / (1 - Chance)
Kappa = (.70 - .54) / (1 - .54) = .35
= 35% of the improvement possible above chance

Kappa – example #2

Rater A			
Rater B	YES (Occurrence)	NO (Nonoccurrence)	TOTAL
YES (Occurrence)	50	35	85
NO (Nonoccurrence)	15	900	915
TOTAL	65)	935	1000

Observed agreement = (50 + 900) / 1000 = .95

Chance agreement = correction based on observed marginal data – i.e., seeing how unbalanced the observed distributions are – 935 of 1000 for Rater A and 915 of 1000 for Rater B - the correction for chance is .86

Kappa = (Obs. - Chance) / (1 - Chance)

= (.95 - .86) / (1 - .86) = .64

Kappa – example #2

Rater A			
Rater B	YES (Occurrence)	NO (Nonoccurrence)	TOTAL
YES (Occurrence)	50 (A)	35 (B)	85 (G)
NO (Nonoccurrence)	15 (C)	900 (D)	915 (H)
TOTAL	65 (E)	935 (F)	1000 (I)

Observed agreement = (50 + 900) / I = .95
Chance agreement = (E/I x G/I) + (F/I x H/I) = .86
Kappa = (Obs. - Chance) / (1 - Chance)
= (.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

…moving from agreement to association…

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

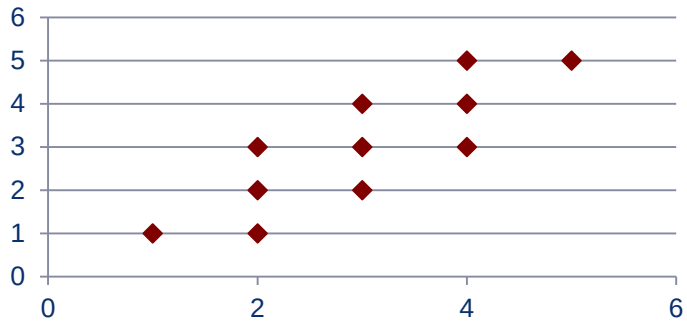
+ sign...as one goes up, the other also goes up

Indicate the size of the association

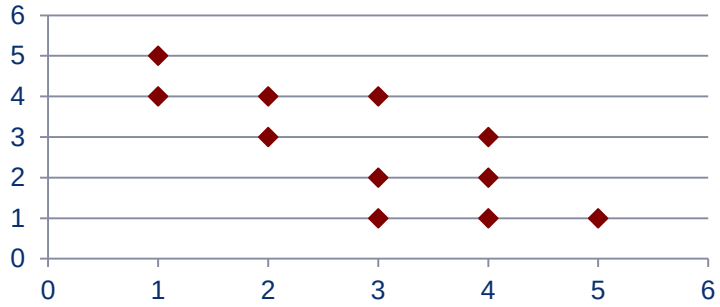
–1 = perfect negative relationship

+1 = perfect positive relationship

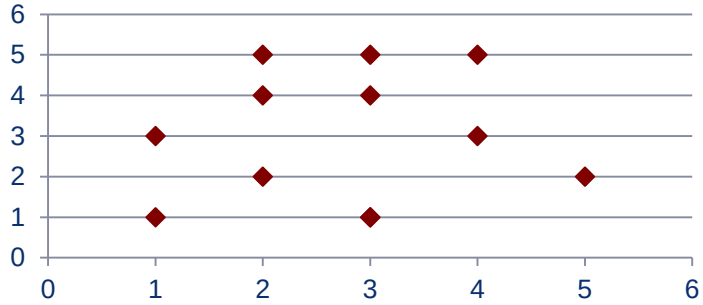
Correlations



r = .84
positive
strong correlation

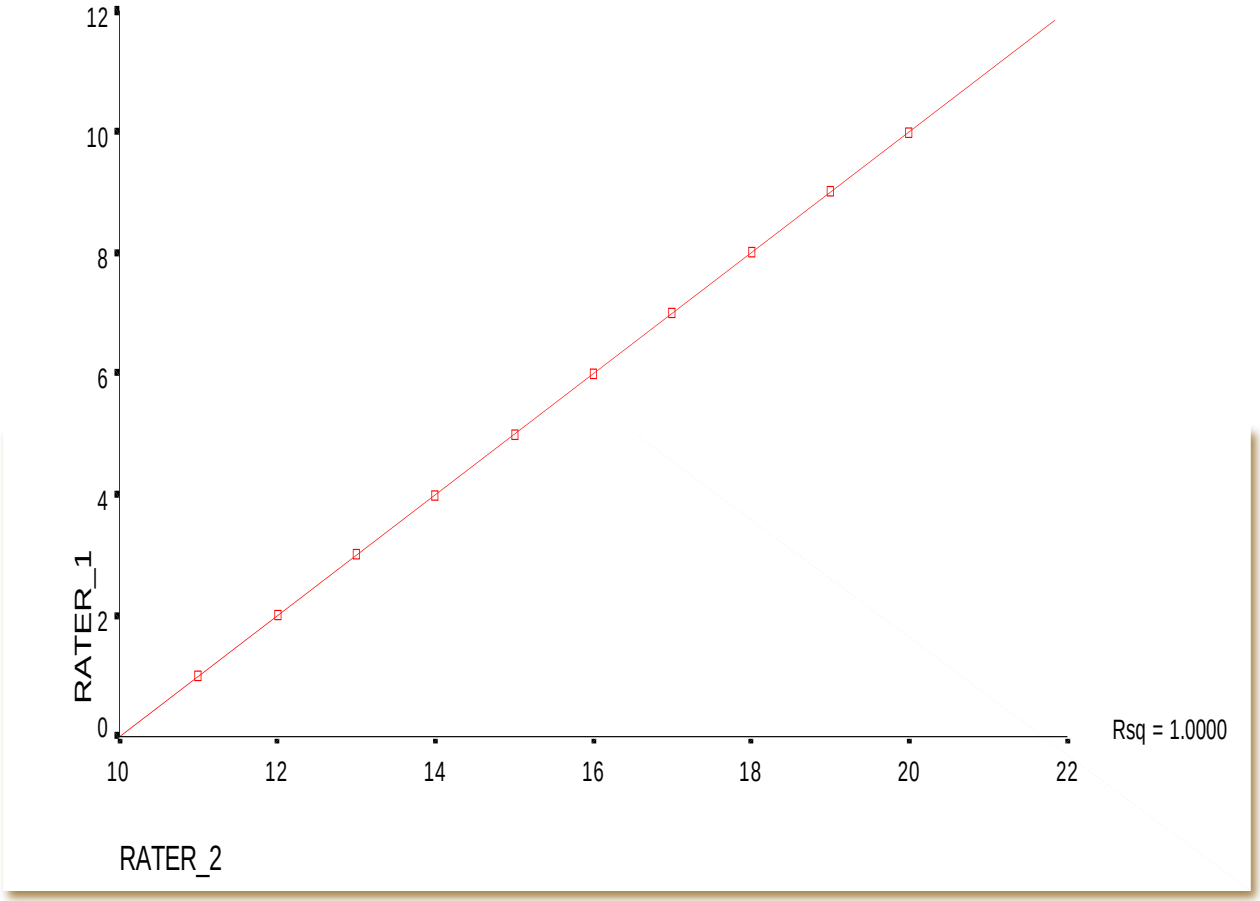


r = - .77
negative
moderate correlation



r = -.04
neither positive nor negative
no correlation

Correlation Coefficient



pt	*rater_1*	*rater_2*
1	1	11
2	2	12
3	3	13
4	4	14
5	5	15
6	6	16
7	7	17
8	8	18
9	9	19
10	10	20

Note that a perfect 1.0 correlation coefficient does <u>not</u> 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

Attractive features:

Can handle multiple raters and multiple stimuli (e.g., charts, SPs, notes) simultaneously

Can deal with multiple designs – e.g., all raters rate all cases (crossed design) versus subsets of cases assigned to subsets of raters (nested)

Can look at both consistency and absolute agreement

Intraclass Correlation Coefficient (ICC) – Example 1

Context: use of the Learning Goal Scoring Rubric to assess the quality of written learning goals and residents' goal writing skills

Data: five raters independently scores 48 goals written by 48 residents

two raters independently scored 48 goals written by 12 residents

Reliability:

- among raters
- among items in rubric
- among goals within a resident

Lockspeiser TM, Schmitter PA, Lane JL, Hanson JL, Rosenberg AA, Park YS. Assessing Residents' Written Learning Goals and Goal Writing Skill: Validity Evidence for the Learning Goal Scoring Rubric. Acad Med, 2013:88;1558-1563.

ICC – Example 2

Context: rater training for mini-cex

Data: Mini-CEX ratings at baseline (just before workshop for workshop group), and four weeks later using videotaped resident–patient encounters; mini-CEX ratings of live resident–patient encounters one year preceding and one year following the workshop

Reliability:

among raters for cases (absolute agreement)
among raters for live encounters
within a rater? (if did test-retest with same stimulus)

Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS, Effect of Rater Training on Reliability and Accuracy of Mini-CEXScores: A Randomized, Controlled Trial. JGIM, 2008, 24:74-9.

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

Most common metric:

Cronbach's alpha (α)
assesses homogeneity of ***continuous*** items

Cronbach's Alpha (α)

For continuous items

Preferred method of calculating internal consistency

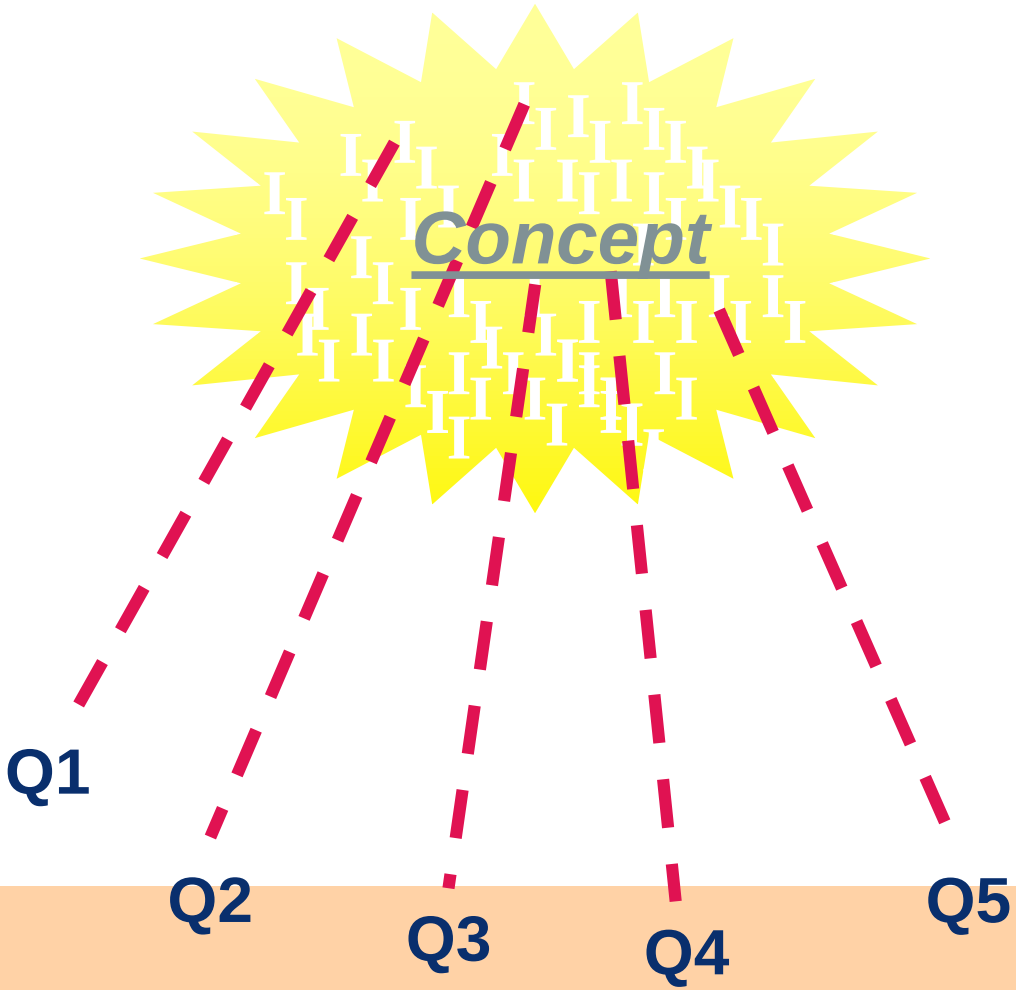
Easy to interpret

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

Reliability Example

Domain Sampling Theory Review



measures are composed of a random sample of items from a hypothetical domain of items

purpose of any measure is to estimate the measurement that would be obtained if one could employ ALL items in the domain

Cronbach's Alpha (α)

Formula:

is based on inter-item correlations- *the idea is that if the items all measure the same construct, then the items should all be highly correlated*

$$\alpha = \frac{k\,\bar{r}}{1 + (k-1)\,\bar{r}}$$

k = number of items

$\bar{r}$ = average inter-item correlation

Sample Questions

1. The course goals, objectives and expectations were clear.
2. The course was organized and coherent.
3. The course director(s) were committed.
4. The course achieved its stated goals.
5. The course had educational value – I learned a lot.
6. I understood how I would be evaluated.

*on a scale where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree

Inter-Item Correlation Matrix: For Computing Alpha

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1.00					
Q2	.54	1.00				
Q3	.24	.38	1.00			
Q4	.39	.39	.37	1.00		
Q5	.44	.47	.26	.22	1.00	
Q6	.37	.46	.27	.37	.55	1.00

Calculate alpha

Calculation

$$\alpha = \frac{k\,\bar{r}}{1 + (k-1)\,\bar{r}}$$

$$\alpha = \frac{6\ (.38)}{1 + (6-1)\ .38}$$

$$\alpha = .79$$

Sample Questions

1. The course goals, objectives and expectations were clear.
2. The course was disorganized and incoherent.
3. The course director(s) were committed.
4. The course achieved its stated goals.
5. The course had little educational value.
6. I understood how I would be evaluated

*on a scale where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree

Inter-Item Correlation Matrix: For Computing Alpha

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1.00					
Q2	-.54	1.00				
Q3	.24	-.38	1.00			
Q4	.39	-.39	.37	1.00		
Q5	-.45	.47	-.26	-.22	1.00	
Q6	.38	-.46	.27	.37	-.54	1.00

Calculate alpha

Calculation

$$\alpha = \frac{k\,\bar{r}}{1 + (k-1)\,\bar{r}}$$

$$\alpha = \frac{6\ (.05)}{1 + (6-1)\ .05}$$

$$\alpha = \text{junk}$$

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

Factors Influencing Reliability

Test length

- Longer tests give more reliable scores

Group heterogeneity

- The more heterogenous the group, the higher the reliability

Objectivity of scoring

- The more “objective” (i.e., clear) the scoring, the higher the reliability

Summary of Reliability

This reliability…	assesses this error…	and estimates…
1. Inter-rater	•rater/scorer	•rater reliability
2. Test-retest & intra-rater	•individual changes over time or administration	•stability
3. Cronbach's alpha	•sampling	•internal consistency

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

Know that reliability is a characteristic of the scores rather than the test

Be familiar with 3 types of reliability

Match reliability types with statistical measures

Select the best type for particular study

MERC

Exercise

AAMC

Tomorrow's Doctors, Tomorrow's Cures®

Case #1:

The issue:

Students need to acquire good oral presentation skills.

Case #1:

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.

What types of reliability should be assessed?

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.

Case #2:

The issue:

Identification and treatment of depression during medical school has important learning and behavioral implications.

Case #2:

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.

What types of reliability should be assessed?

Tjia J, Givens JL, Shea JA. Factors associated with under treatment of medical student depression. Journal of American College Health, 2005 Mar-Apr;53(5):219-24.

Case #3:

The issue:

Residents need to learn to follow clinical guidelines regarding appropriate primary care prevention, screening, and safety issues.

Case #3:

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).

What types of reliability should be assessed?

Kogan JR, Reynolds EE, Shea JA. Effectiveness of report cards based on chart audits of residents' adherence to practice guidelines on practice performance: A randomized controlled trial. Teaching and Learning in Medicine: An International Journal, 2003;15:25-30.

Case #4:

The issue:

Resident work hour regulations have likely had an impact on student learning/educational activities

Case #4:

A time-motion study was done. Random samples of students (pre-reform and post-reform) were given pagers and called randomly, approximately every 90 minutes while they were in the hospital. At the time they responded to a short pocket survey and answered 4 questions: where they were, who they were with, type of activity engaged in, and a rating of the educational usefulness. There were two samples of students: one pre and one post reform

What types of reliability should be assessed?

Kogan JR, Bellini LB, Shea JA. The impact of resident duty hour reform in a medicine core clerkship. Academic Medicine, 2004;79:s58-s61.

MERC

Validity

AAMC

Tomorrow's Doctors, Tomorrow's Cures®

Objectives - Validity

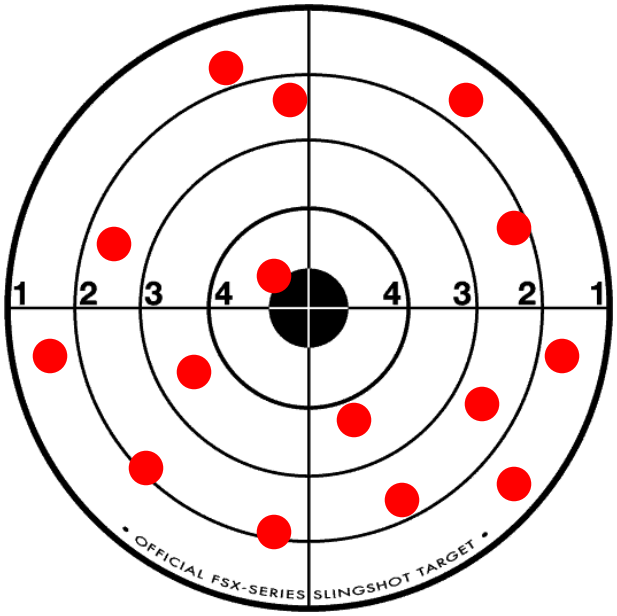
Explain that validity is a characteristic of the interpretation of the scores rather than of the test

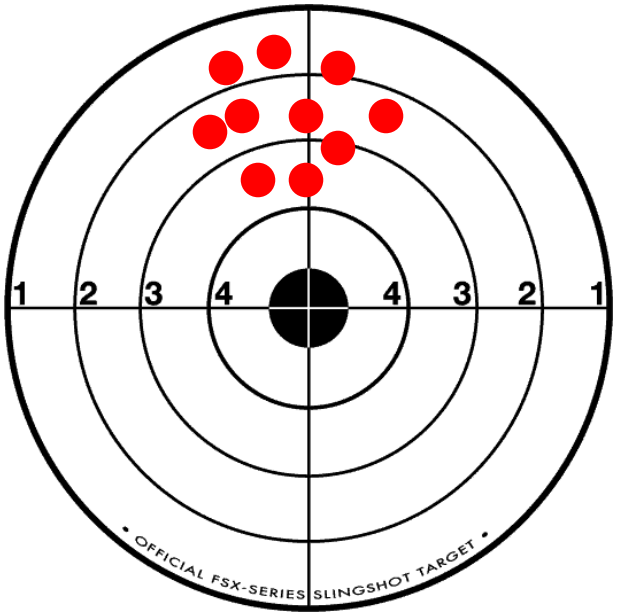
Describe the relationship between reliability and validity

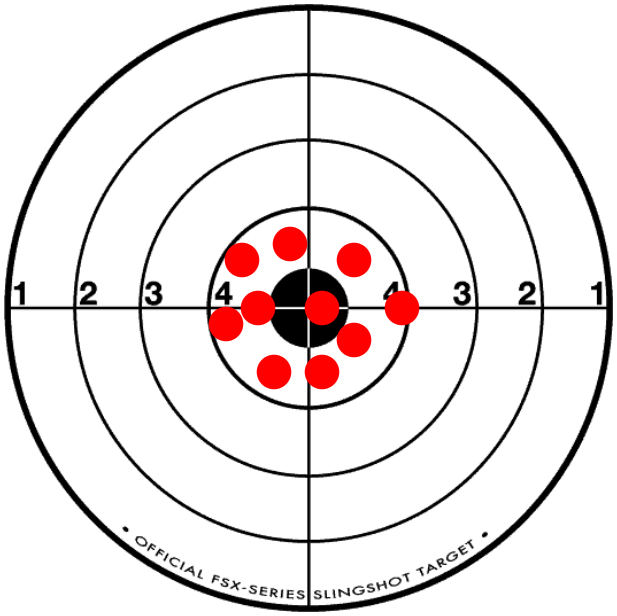
Describe five sources of validity evidence

Select the best type of validity evidence for your particular study

Reliability and Validity



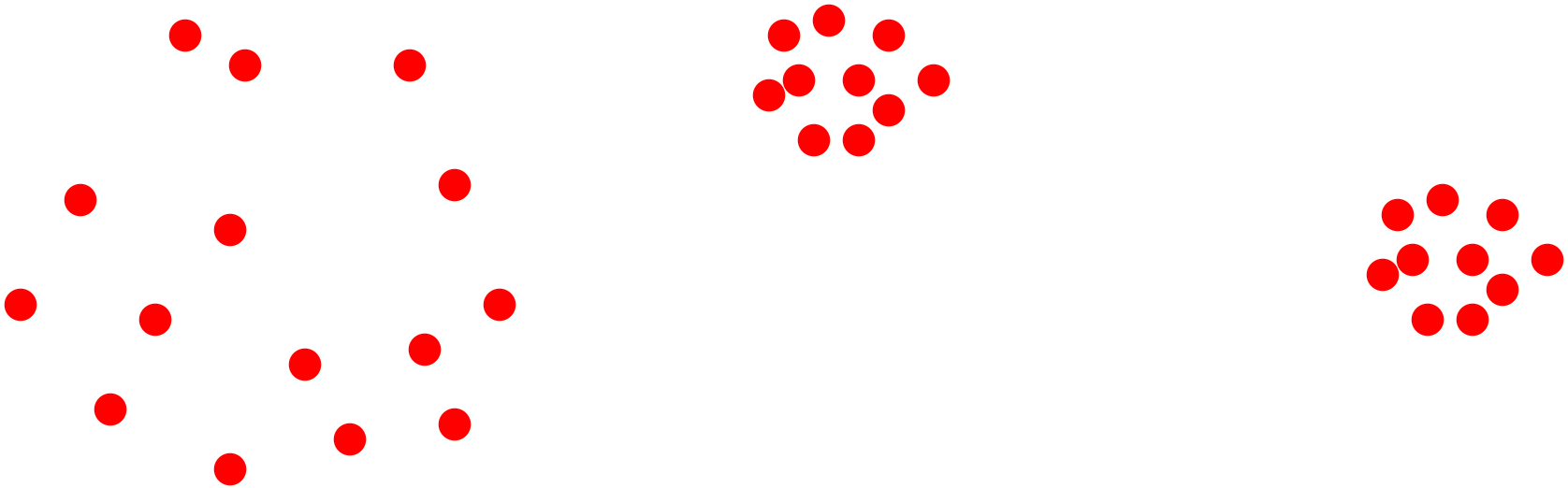




MERC

AAMC

...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

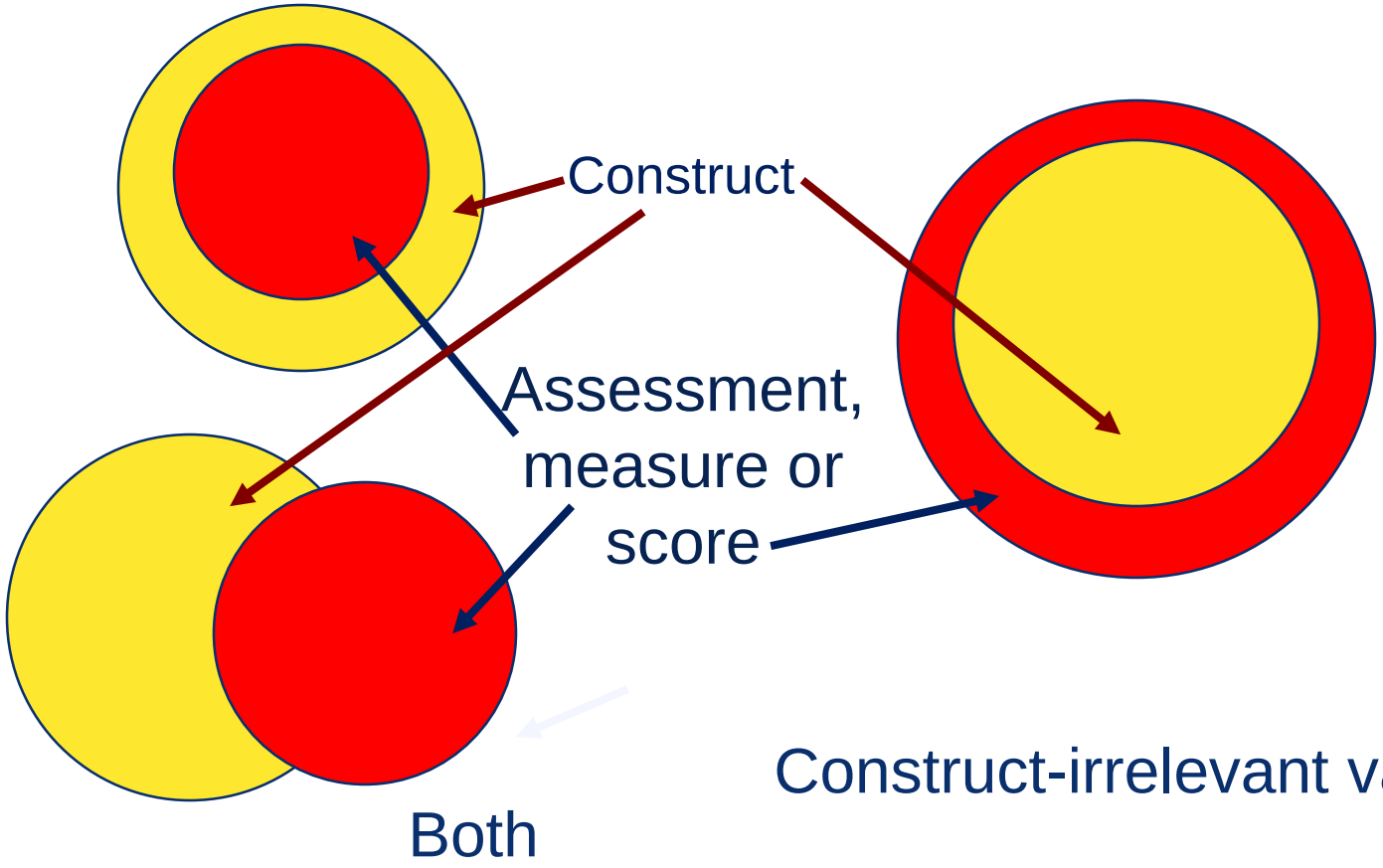
Systematic error threatens validity

(Recall that reliability was concerned with random error)

Systematic error comes from many sources

Threats to Validity

Construct under-representation



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

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

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

Messick S, Educational Measurement, 3rd ed. 1993

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

Downing SM. Med Educ. 2003; 37: 830

Illustrative Assessment

Target construct: x-ray interpretation skill

Assessment: 10 x-rays presented on a computer screen.

After viewing an x-ray, the examinee selects their preferred diagnosis from an extended matching list of 15-20 options.

They then go on to the next x-ray.

Examinees are allowed 15 minutes for this assessment

Validity Evidence: Content

How well does the content of the assessment map onto the construct

Themes, wording, and expert review

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

Example of Content Evidence

10 x-ray films were selected by radiology faculty

Represent common presentations that 4th year medical students should be able to identify.

Faculty expertise is defined by their specialty and role as faculty members.

Faculty judgments define

- "common presentations"
- mapping of "relevant" x-rays and diagnoses.

Validity Evidence: Internal Structure

Degree to which the structure of the assessment fits 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

Scoring = simple percentage of the ten x-rays correctly identified

Each x-ray counts equally

Alternative scoring format = give greater weight to diagnoses that are more important (e.g., clinically dangerous)

10 x-rays is probably a minimal sample for this construct. Ideally, would have more

Reliability (internal consistency) = 0.86

Validity Evidence: Relations to Other Variables

The relationships between scores on the assessment 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

Predict that x-ray interpretation should correlate positively with other visual interpretation skills, like reading EKGs and CT

Should not correlate with interviewing or communication skills

This assessment focuses on common diagnoses - may not generalize to unusual diagnoses.

Validity Evidence: Response Process

How well the cognitive processes required by the assessment map onto the processes of the underlying construct

Examining the reasoning and thought processes of learners

Does cognitive processes required by assessment map onto those required in ‘real life’?

Systems that reduce the likelihood of response error

Example of Response Process Evidence

Students view the x-ray films and select a diagnosis from an extended list of alternatives

Viewing the x-ray on screen is identical to actual practice of this construct

Selecting a diagnosis from a list is not the same and could be a evidence against validity

Validity Evidence: Consequences

Do the decisions made on the basis of the assessment “work”

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

Passing score is set at 60%

Students who fail must remediate and retake the station.

Up to two retakes are allowed before other interventions take place, such as repeating a rotation or the whole third year.

What are the pros and cons of raising or lowering the pass/fail cut-point and the consequences on examinees.

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

MERC

Exercise

AAMC

Tomorrow's Doctors, Tomorrow's Cures®

Case #1:

The issue:

Identification and treatment of depression during medical school has important learning and behavioral implications.

Case #1:

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.

What types of validity should be assessed?

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.

Case #2:

The issue:

Students need to acquire strong clinical evaluation skills, including history-taking and physical examination.

Case #2:

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.

What types of validity should be assessed?

Tjia J, Givens JL, Shea JA. Factors associated with under treatment of medical student depression. Journal of American College Health, 2005 Mar-Apr;53(5):219-24.

Case #3:

The issue:

Residents need to learn to follow clinical guidelines regarding appropriate primary care prevention, screening, and safety issues.

Case #3:

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).

What types of validity should be assessed?

Kogan JR, Reynolds EE, Shea JA. Effectiveness of report cards based on chart audits of residents' adherence to practice guidelines on practice performance: A randomized controlled trial. Teaching and Learning in Medicine: An International Journal, 2003;15:25-30.

Case #4:

The issue:

Resident work hour regulations have likely had an impact on student learning/educational activities

Case #4:

A time-motion study was done. Random samples of students (pre-reform and post-reform) were given pagers and called randomly, approximately every 90 minutes while they were in the hospital. At the time they responded to a short pocket survey and answered 4 questions: where they were, who they were with, type of activity engaged in, and a rating of the educational usefulness. There were two samples of students: one pre and one post reform

What types of validity should be assessed?

Kogan JR, Bellini LB, Shea JA. The impact of resident duty hour reform in a medicine core clerkship. Academic Medicine, 2004;79:s58-s61.

References

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

Downing SM. Validity: on the meaningful interpretation of assessment data. Medical Education. 2003;37:830-837.

References

Fraenkel JR, Wallen NE. How to design and evaluate research in education.(4th Ed). Boston: McGraw Hill, 2000.

Gall MD, Borg WR, Gall JP. Educational research: an introduction (6th edition). White Plains NY: Longman Publishers, 1996.

Linn RL, Gronlund NE. Measurement and assessment in teaching (8th Ed.), Upper Saddle River NY: Merrill, Prentice Hall, 2000.

References

Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. 1995;50:741-749.

Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use (3rd ed.) Oxford: Oxford University Press, 2003.

Questions?

MERC Evaluation Link

Please go to the link below and complete the evaluation:

https://goo.gl/QaViXn