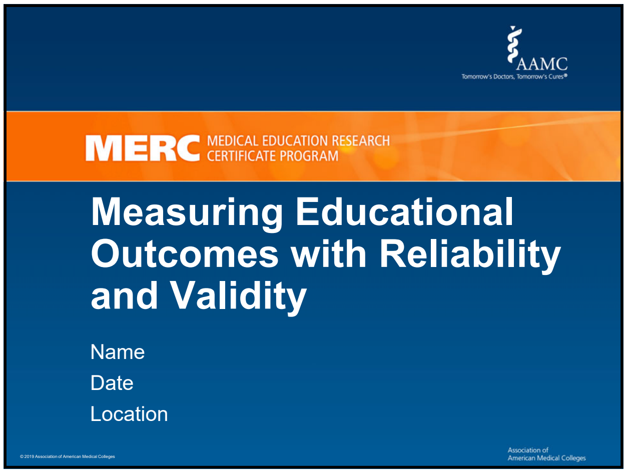


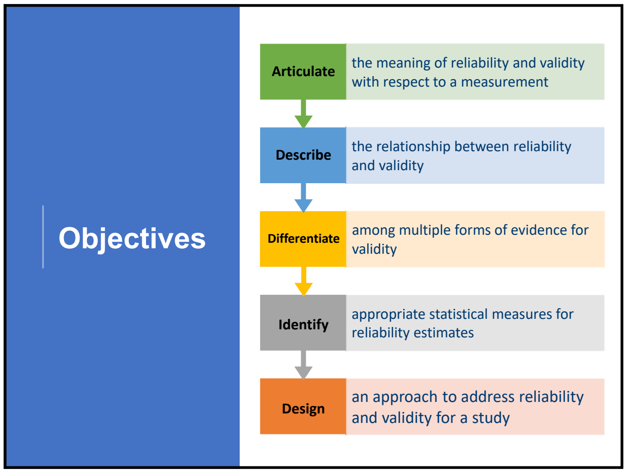
MERC MEDICAL EDUCATION RESEARCH CERTIFICATE PROGRAM

Measuring Educational Outcomes with Reliability and Validity

Name
Date
Location

Association of
American Medical Colleges

1



Objectives

Articulate the meaning of reliability and validity with respect to a measurement

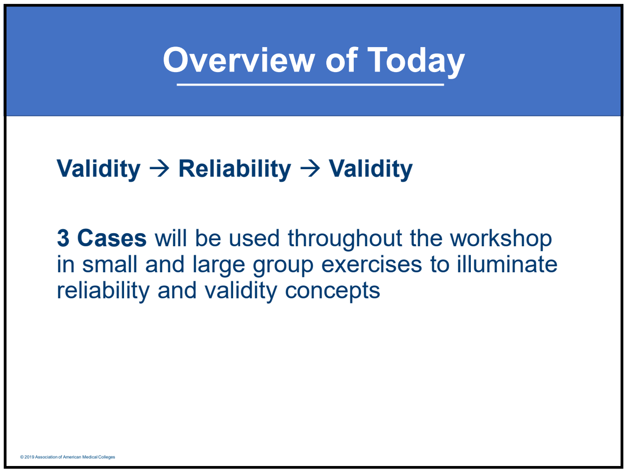
Describe the relationship between reliability and validity

Differentiate among multiple forms of evidence for validity

Identify appropriate statistical measures for reliability estimates

Design an approach to address reliability and validity for a study

2



Overview of Today

Validity → Reliability → Validity

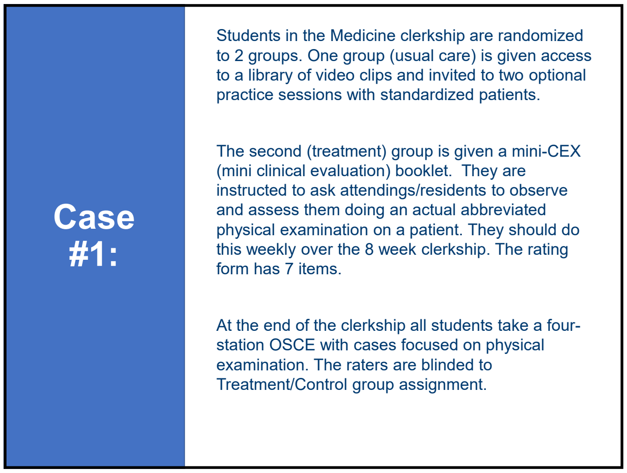
3 Cases will be used throughout the workshop in small and large group exercises to illuminate reliability and validity concepts

3

Case #1:

The issue:
Students need to acquire good physical examination skills.

4



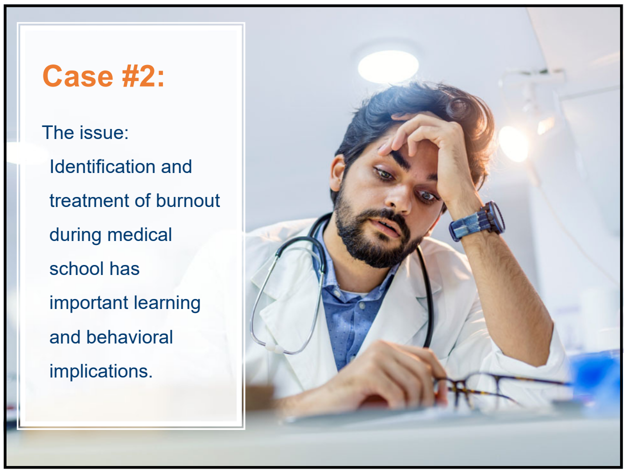
Case #1:

Students in the Medicine clerkship are randomized to 2 groups. One group (usual care) is given access to a library of video clips and invited to two optional practice sessions with standardized patients.

The second (treatment) group is given a mini-CEX (mini clinical evaluation) booklet. They are instructed to ask attendings/residents to observe and assess them doing an actual abbreviated physical examination on a patient. They should do this weekly over the 8 week clerkship. The rating form has 7 items.

At the end of the clerkship all students take a four-station OSCE with cases focused on physical examination. The raters are blinded to Treatment/Control group assignment.

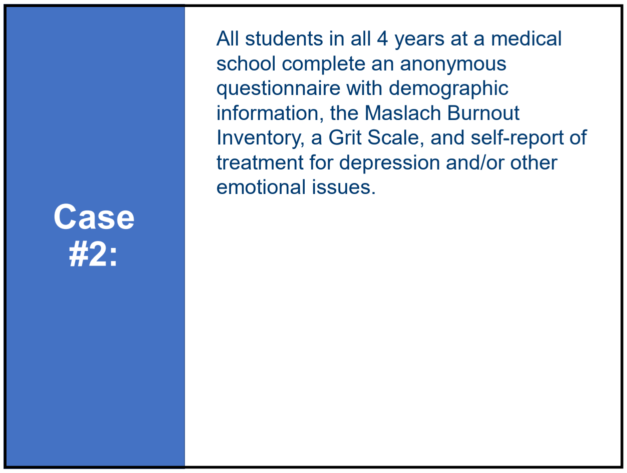
5



Case #2:

The issue:
Identification and treatment of burnout during medical school has important learning and behavioral implications.

6



Case #2:

All students in all 4 years at a medical school complete an anonymous questionnaire with demographic information, the Maslach Burnout Inventory, a Grit Scale, and self-report of treatment for depression and/or other emotional issues.

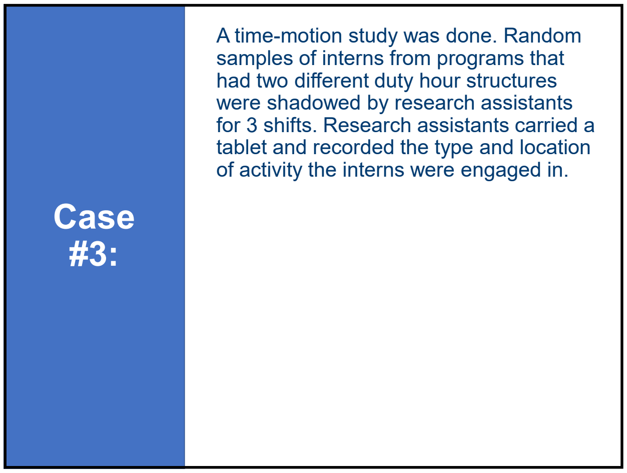
7

Case #3:

The issue:

Duty hours limitations have likely impacted how and where residents spend their time

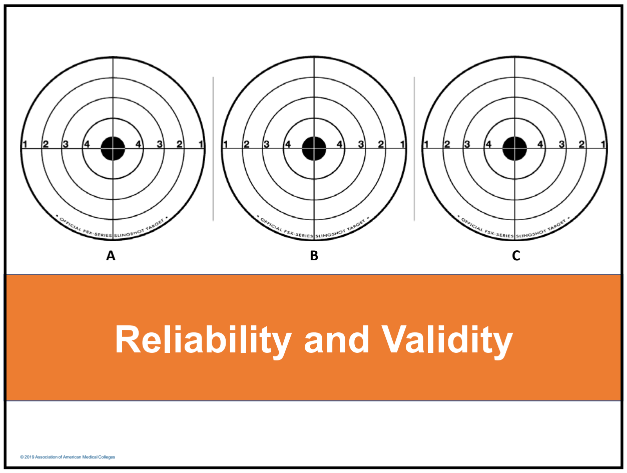
8



Case #3:

A time-motion study was done. Random samples of interns from programs that had two different duty hour structures were shadowed by research assistants for 3 shifts. Research assistants carried a tablet and recorded the type and location of activity the interns were engaged in.

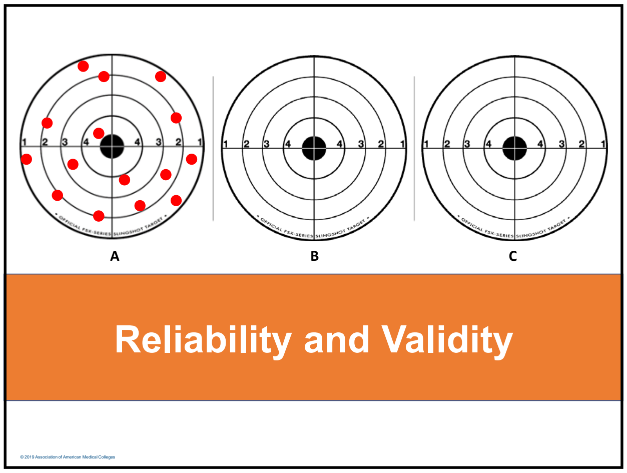
9



Reliability and Validity

© 2019 Association of American Medical Colleges

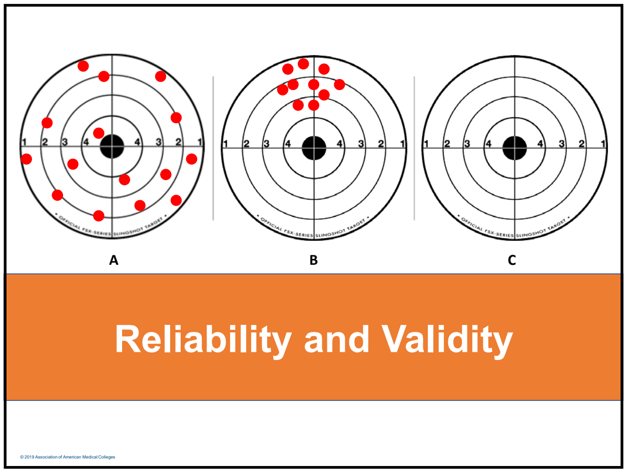
10



Reliability and Validity

© 2019 Association of American Medical Colleges

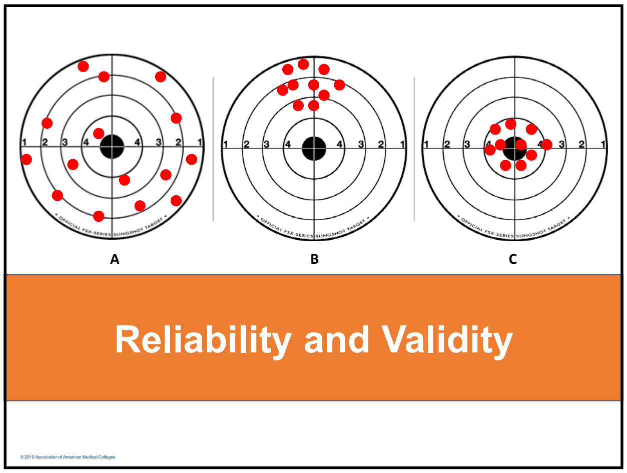
11



Reliability and Validity

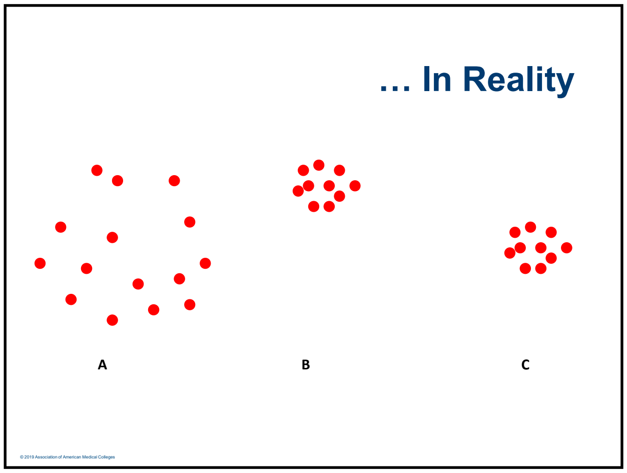
© 2019 Association of American Medical Colleges

12



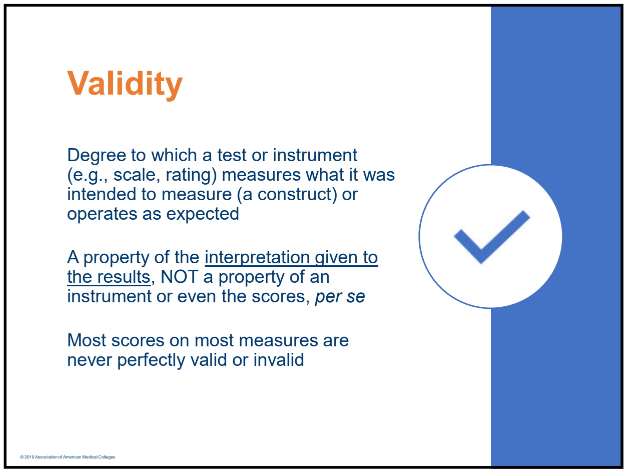
Reliability and Validity

13



… In Reality

14



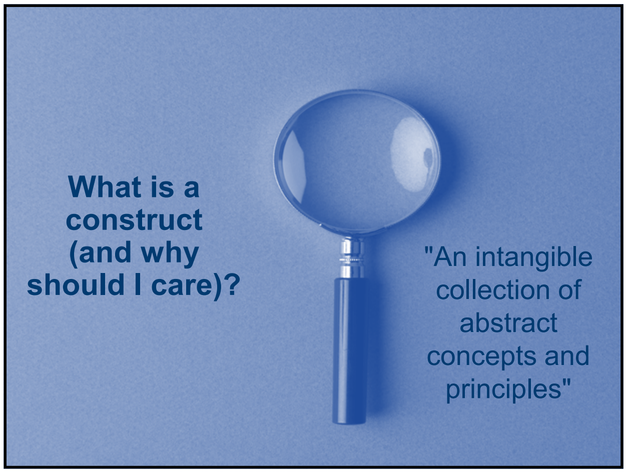
Validity

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

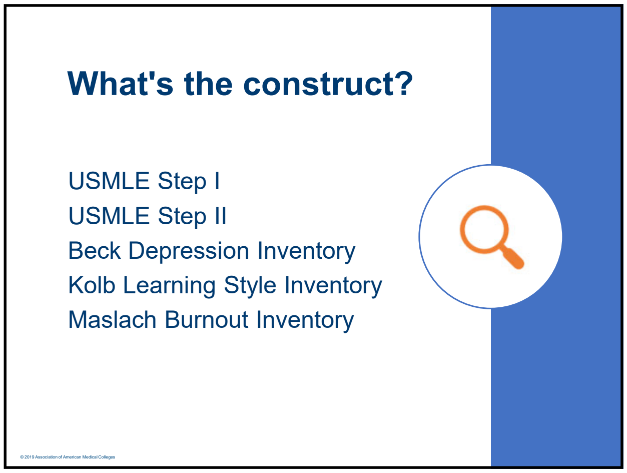
15



What is a construct (and why should I care)?

"An intangible collection of abstract concepts and principles"

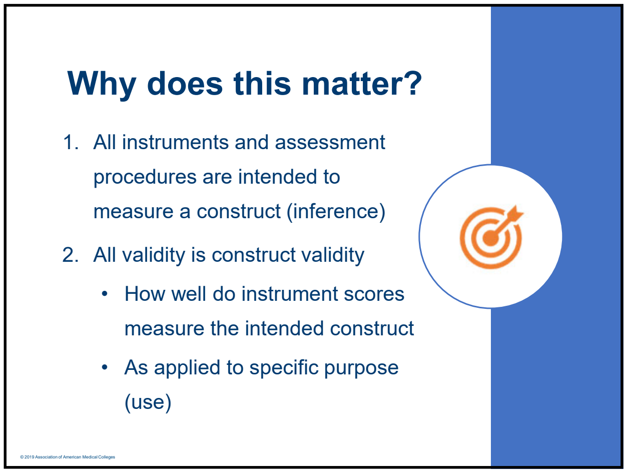
16



What's the construct?

USMLE Step I
USMLE Step II
Beck Depression Inventory
Kolb Learning Style Inventory
Maslach Burnout Inventory

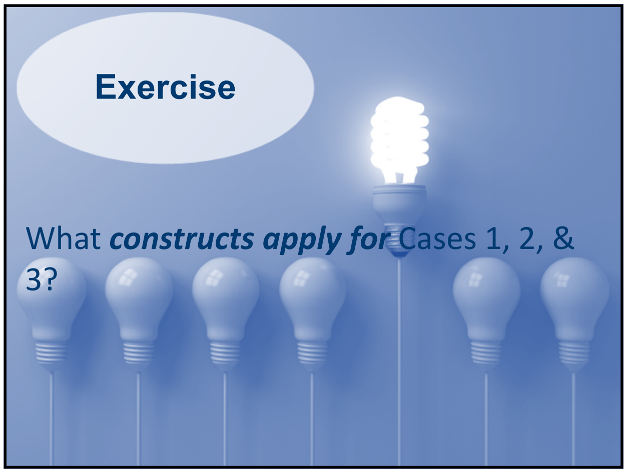
17



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)

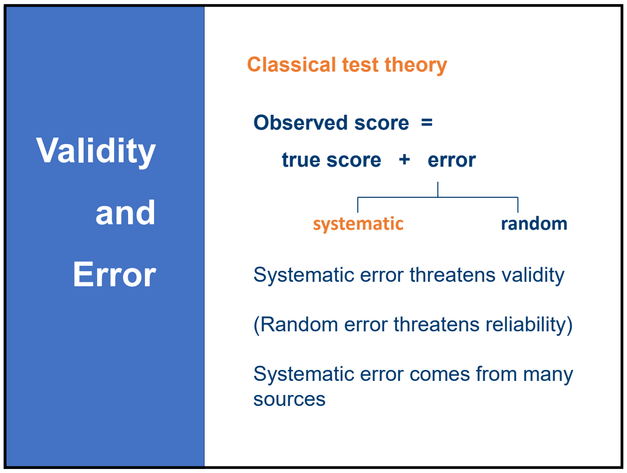
18



Exercise

What ***constructs apply for*** Cases 1, 2, & 3?

19



Validity and Error

Classical test theory

Observed score =

true score + error

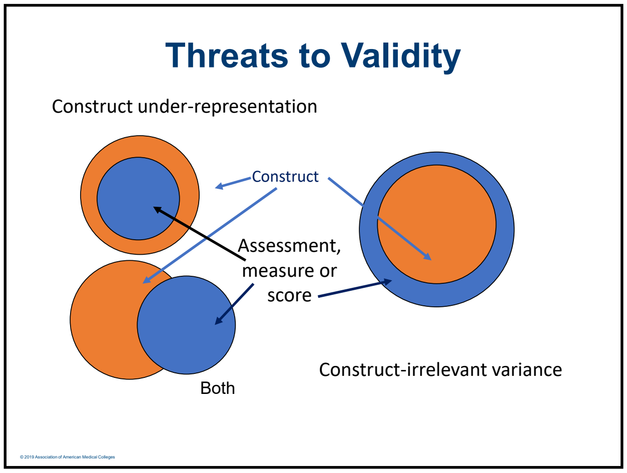
systematic random

Systematic error threatens validity

(Random error threatens reliability)

Systematic error comes from many sources

20



Threats to Validity

Construct under-representation

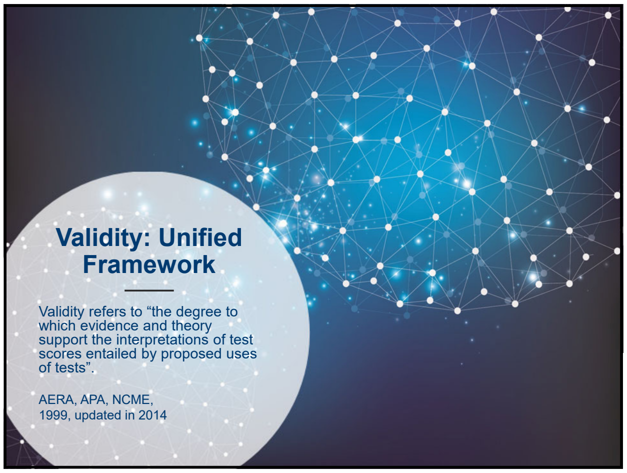
Construct

Assessment, measure or score

Both

Construct-irrelevant variance

21

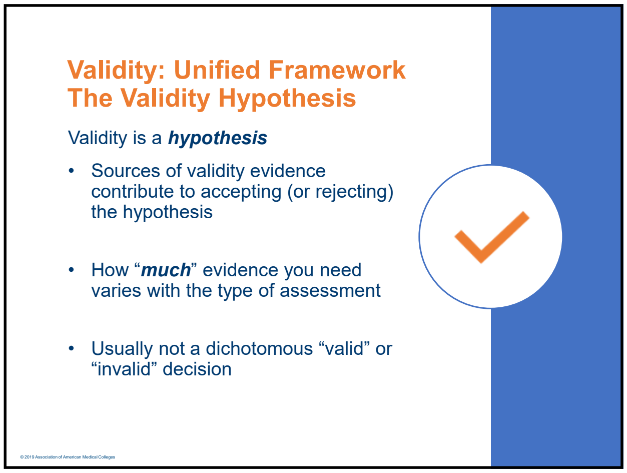


Validity: Unified Framework

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, updated in 2014

22

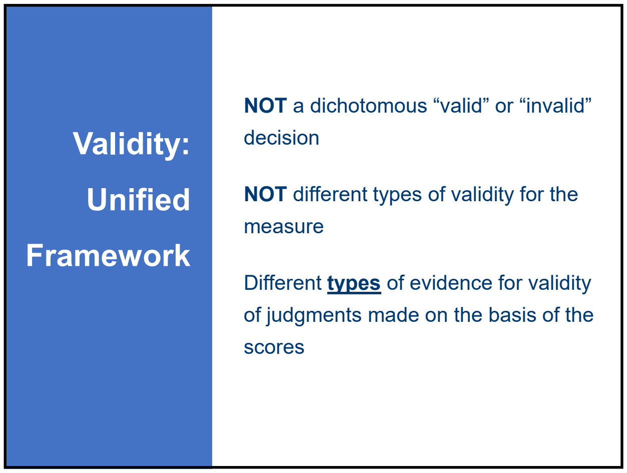


Validity: Unified Framework
The Validity Hypothesis

Validity is a ***hypothesis***

- Sources of validity evidence contribute to accepting (or rejecting) the hypothesis
- How “***much***” evidence you need varies with the type of assessment
- Usually not a dichotomous “valid” or “invalid” decision

23



Validity: Unified Framework

NOT a dichotomous “valid” or “invalid” decision

NOT different types of validity for the measure

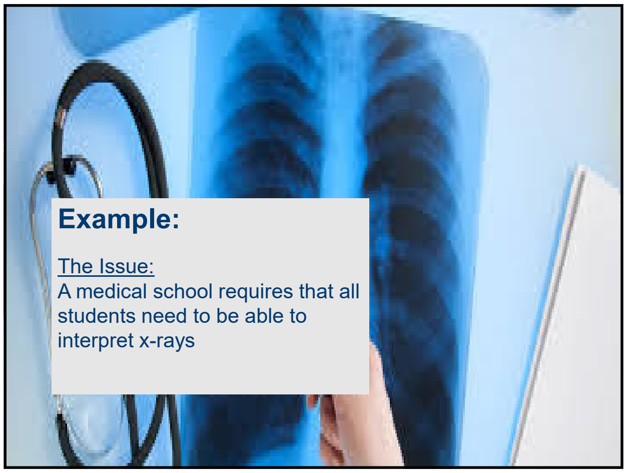
Different **types** of evidence for validity of judgments made on the basis of the scores

24

Types of Evidence

1. Content
2. Internal Structure
3. Relations to Other Variables
4. Response Processes
5. Consequences

25



Example:

The Issue:
A medical school requires that all students need to be able to interpret x-rays

26

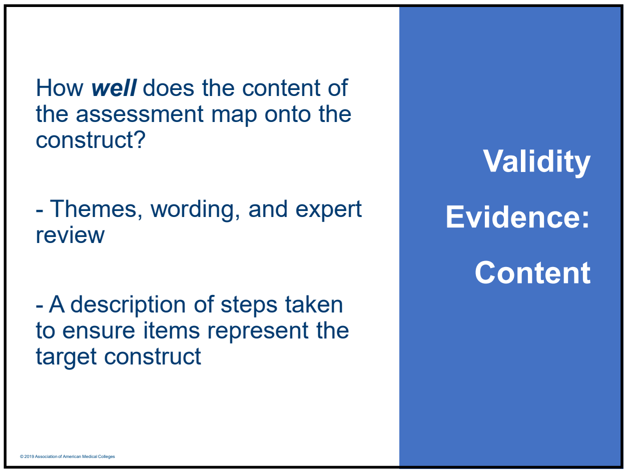
Example

Fourth year medical students complete an online quiz with 10 x-rays.

For each x-ray quiz item, the student selects the preferred diagnosis from an extended matching list of 15-20 options.

Students have 15 minutes to complete the quiz.

27



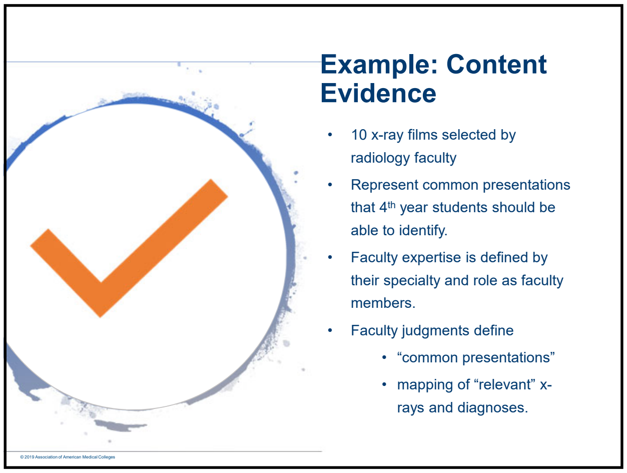
Validity Evidence: Content

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

28



Example: Content Evidence

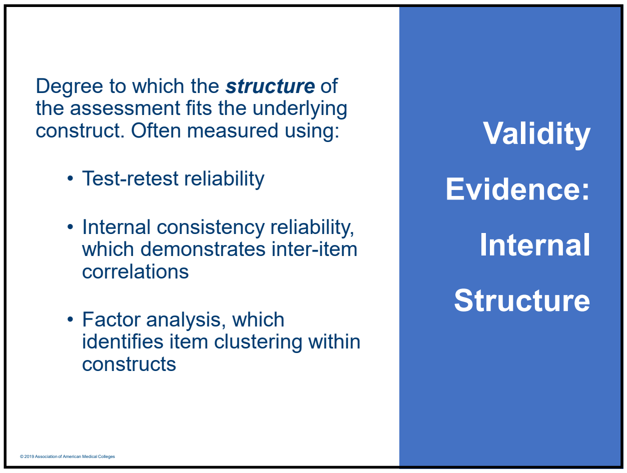
- 10 x-ray films selected by radiology faculty
- Represent common presentations that 4th year students should be able to identify.
- Faculty expertise is defined by their specialty and role as faculty members.
- Faculty judgments define
  - "common presentations"
  - mapping of "relevant" x-rays and diagnoses.

29

Exercise

What might ***content validity evidence*** look like for Cases 1, 2, & 3?

30

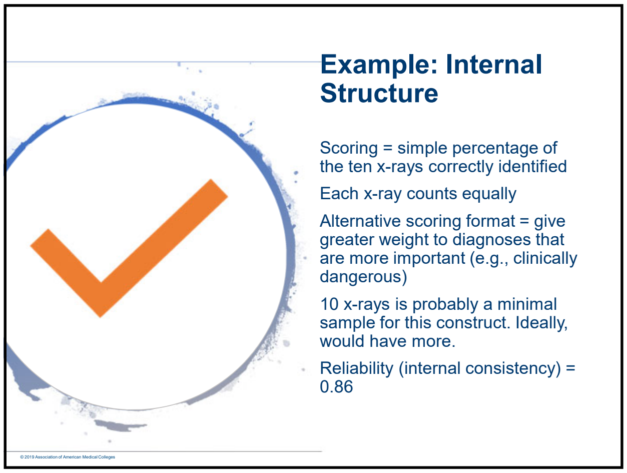


Validity Evidence: Internal Structure

Degree to which the ***structure*** of the assessment fits the underlying construct. Often measured using:

- Test-retest reliability
- Internal consistency reliability, which demonstrates inter-item correlations
- Factor analysis, which identifies item clustering within constructs

31



Example: Internal Structure

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

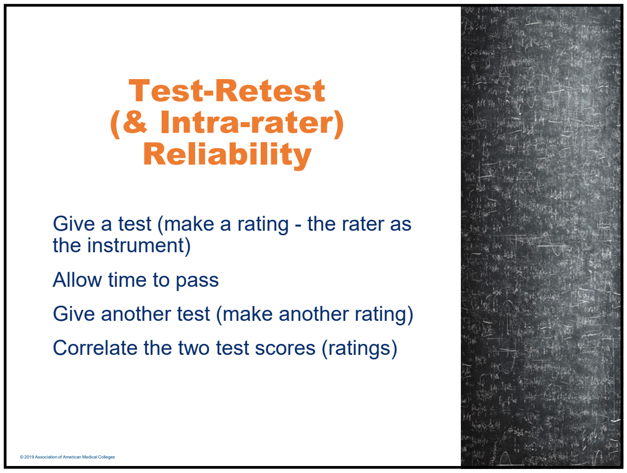
Each x-ray counts equally

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

32



Test-Retest (& Intra-rater) Reliability

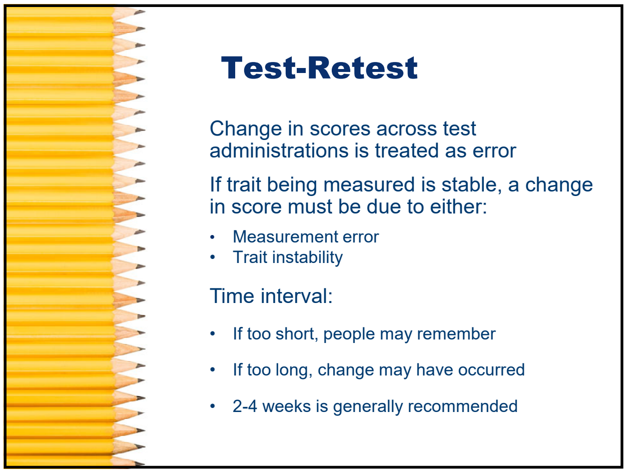
Give a test (make a rating - the rater as the instrument)

Allow time to pass

Give another test (make another rating)

Correlate the two test scores (ratings)

33



Test-Retest

Change in scores across test administrations is treated as error

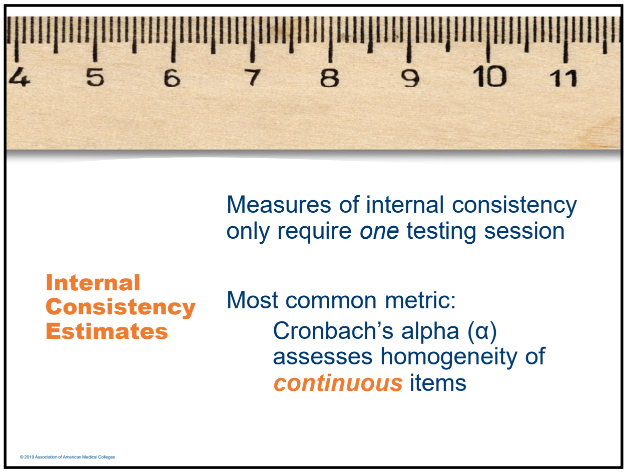
If trait being measured is stable, a change in score must be due to either:

- Measurement error
- Trait instability

Time interval:

- If too short, people may remember
- If too long, change may have occurred
- 2-4 weeks is generally recommended

34


4
5
6
7
8
9
10
11

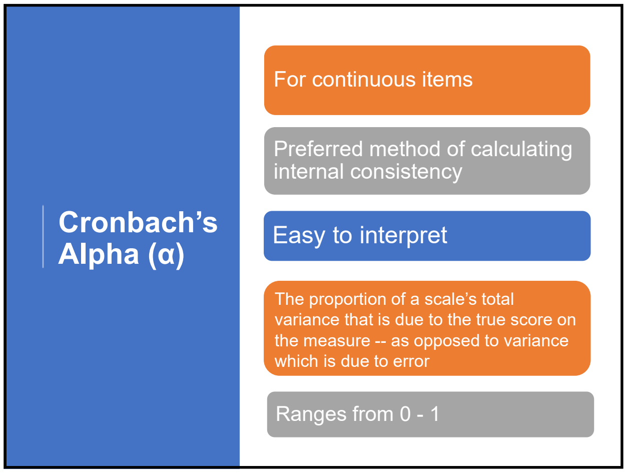
Internal Consistency Estimates

Measures of internal consistency only require *one* testing session

Most common metric:

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

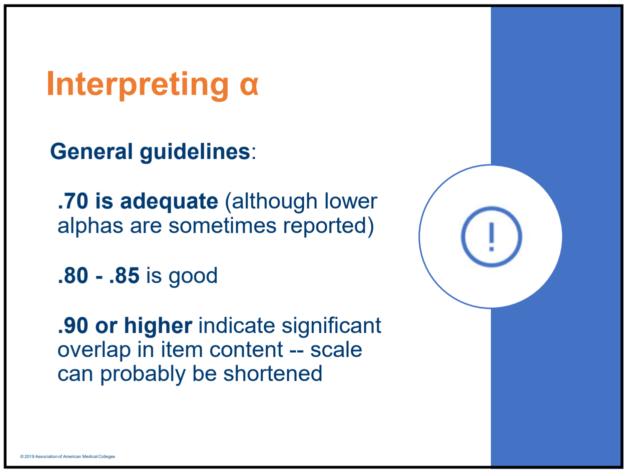
35



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

36



Interpreting α

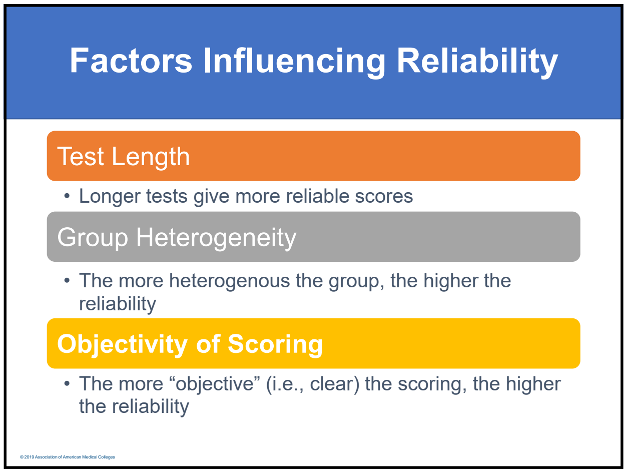
General guidelines:

.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

37



Factors Influencing Reliability

Test Length

- Longer tests give more reliable scores

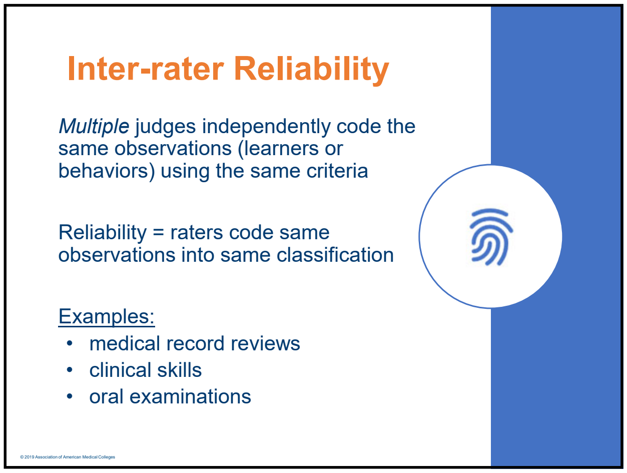
Group Heterogeneity

- The more heterogenous the group, the higher the reliability

Objectivity of Scoring

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

38



Inter-rater Reliability

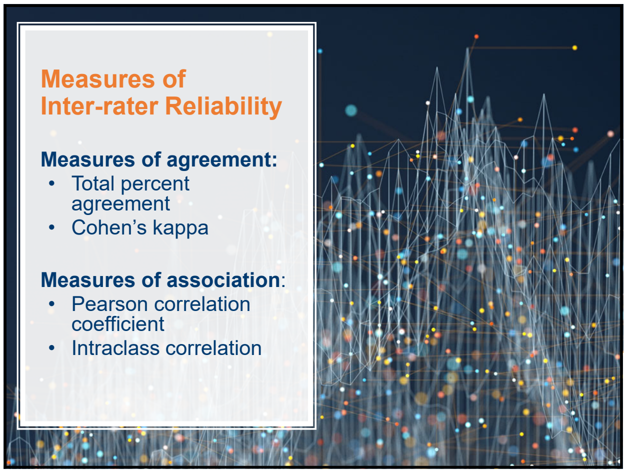
Multiple judges independently code the same observations (learners or behaviors) using the same criteria

Reliability = raters code same observations into same classification

Examples:

- medical record reviews
- clinical skills
- oral examinations

39

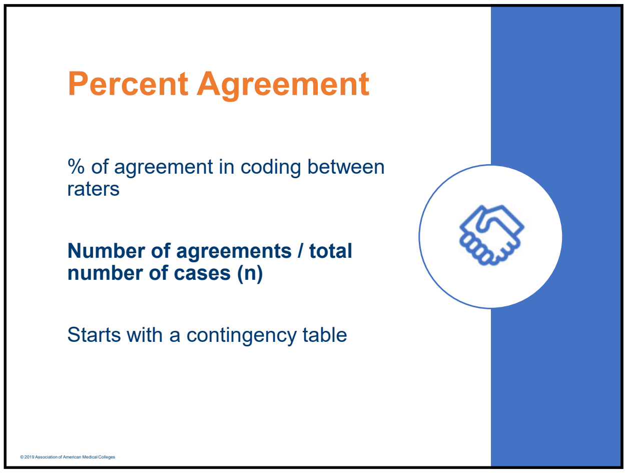


Measures of Inter-rater Reliability

Measures of agreement:
- Total percent agreement
- Cohen's kappa

Measures of association:
- Pearson correlation coefficient
- Intraclass correlation

40



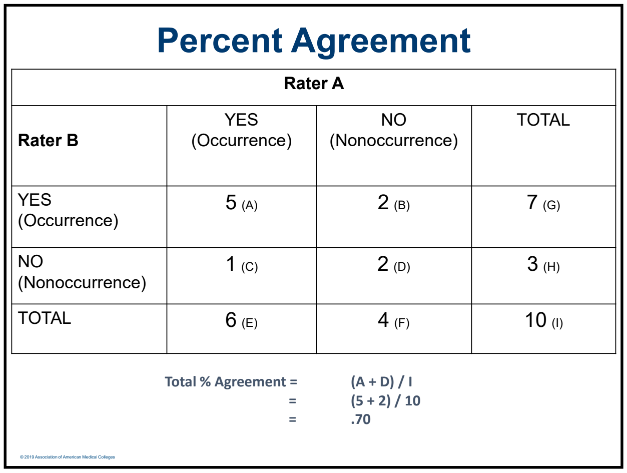
Percent Agreement

% of agreement in coding between raters

Number of agreements / total number of cases (n)

Starts with a contingency table

41

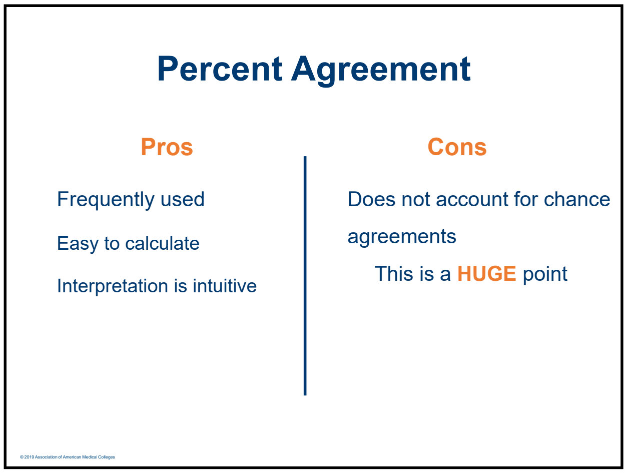


Percent Agreement

	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

42



Percent Agreement

Pros

Frequently used

Easy to calculate

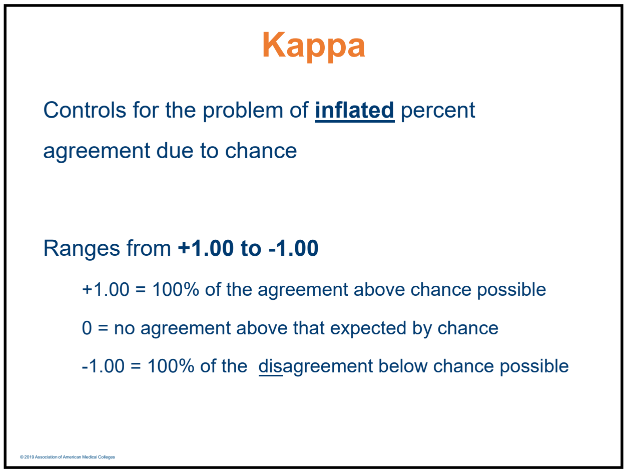
Interpretation is intuitive

Cons

Does not account for chance agreements

This is a **HUGE** point

43



Kappa

Controls for the problem of **inflated** percent agreement due to chance

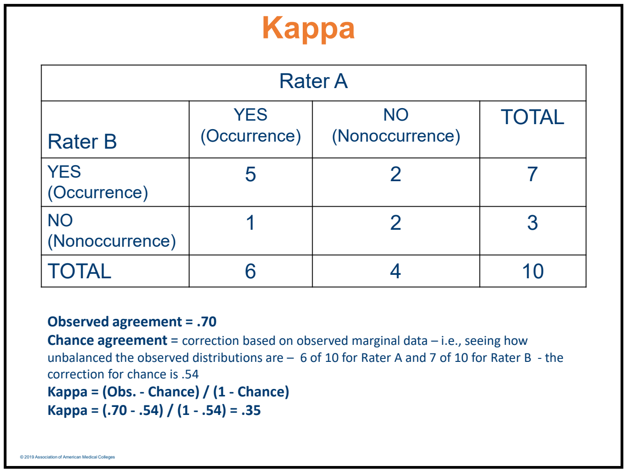
Ranges from **+1.00 to -1.00**

+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

44

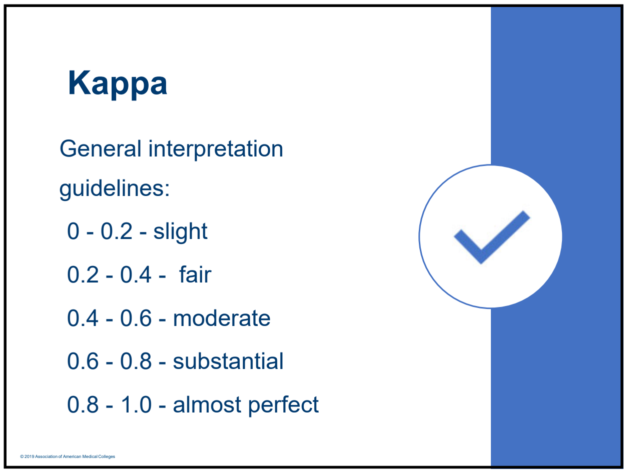


Kappa

	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

45



Kappa

General interpretation guidelines:

0 - 0.2 - slight

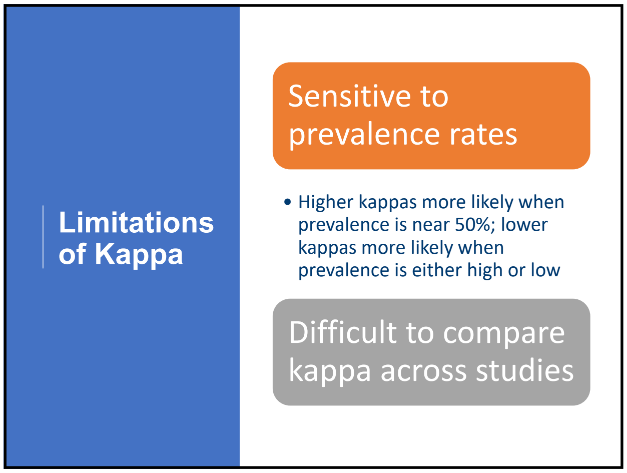
0.2 - 0.4 - fair

0.4 - 0.6 - moderate

0.6 - 0.8 - substantial

0.8 - 1.0 - almost perfect

46



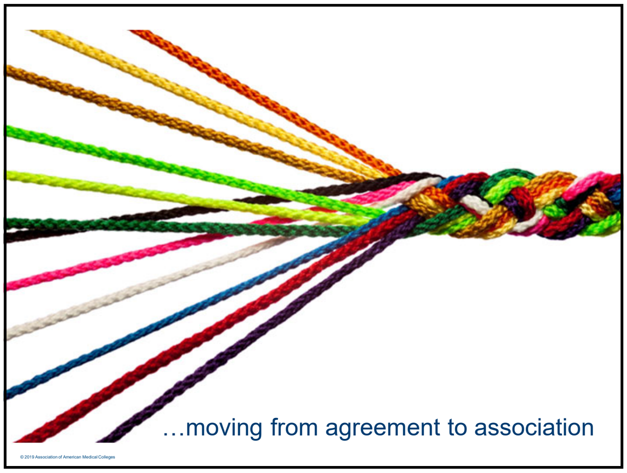
Limitations of Kappa

Sensitive to prevalence rates

- Higher kappas more likely when prevalence is near 50%; lower kappas more likely when prevalence is either high or low

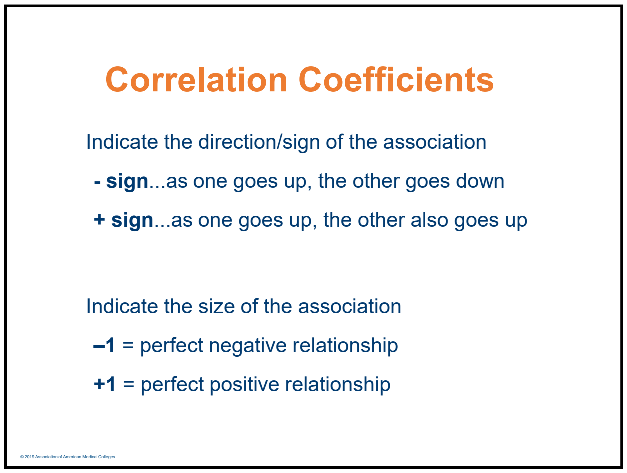
Difficult to compare kappa across studies

47



…moving from agreement to association

48



Correlation Coefficients

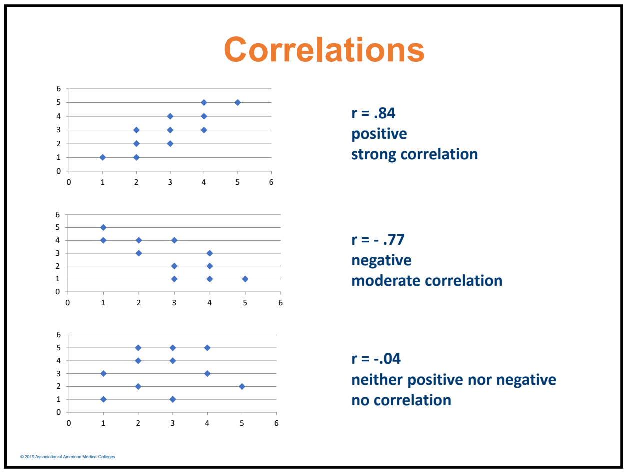
Indicate the direction/sign of the association

- **- 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

49



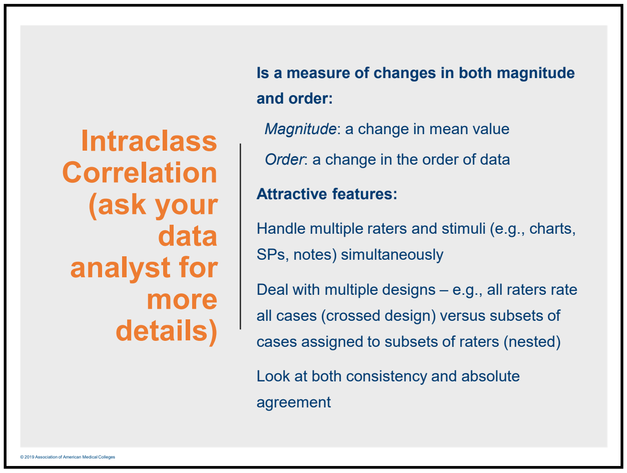
Correlations

r = .84
positive
strong correlation

r = - .77
negative
moderate correlation

r = -.04
neither positive nor negative
no correlation

50



Intraclass Correlation (ask your data analyst for more details)

Is a measure of changes in both magnitude and order:

Magnitude: a change in mean value

Order: a change in the order of data

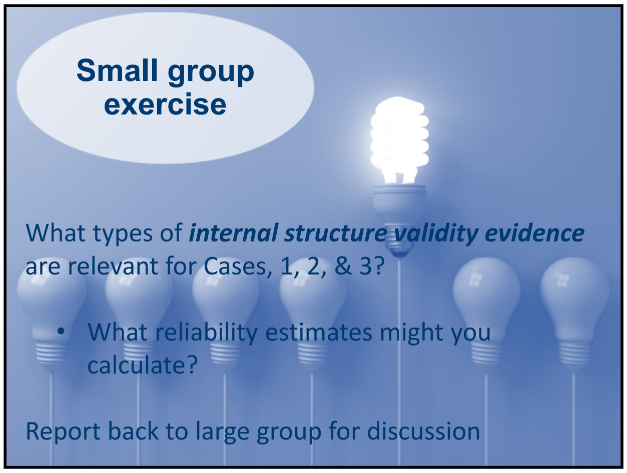
Attractive features:

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

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

Look at both consistency and absolute agreement

51



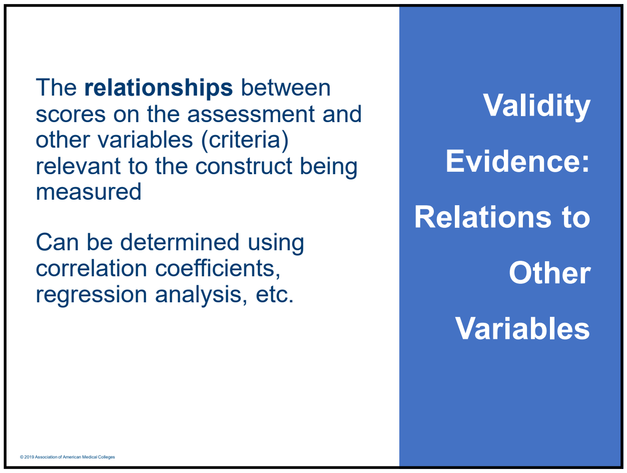
Small group exercise

What types of ***internal structure validity evidence*** are relevant for Cases, 1, 2, & 3?

- What reliability estimates might you calculate?

Report back to large group for discussion

52

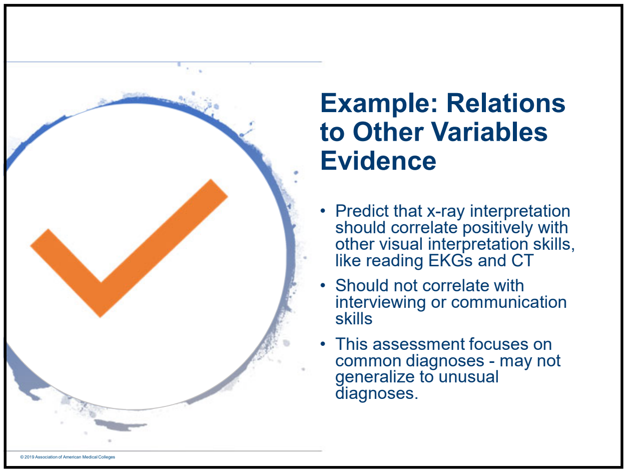


Validity Evidence: Relations to Other Variables

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.

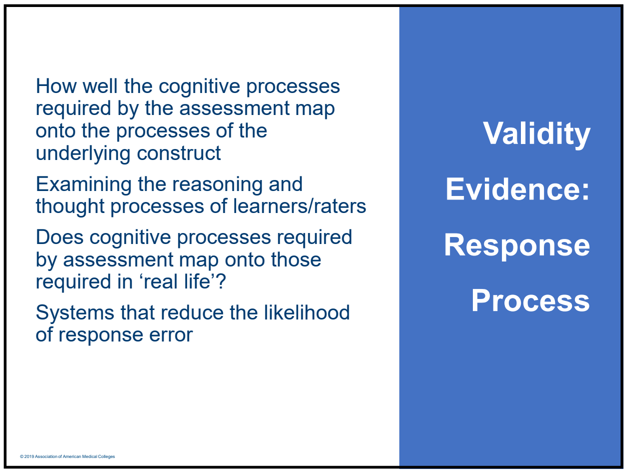
53



Example: 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.

54



Validity Evidence: Response Process

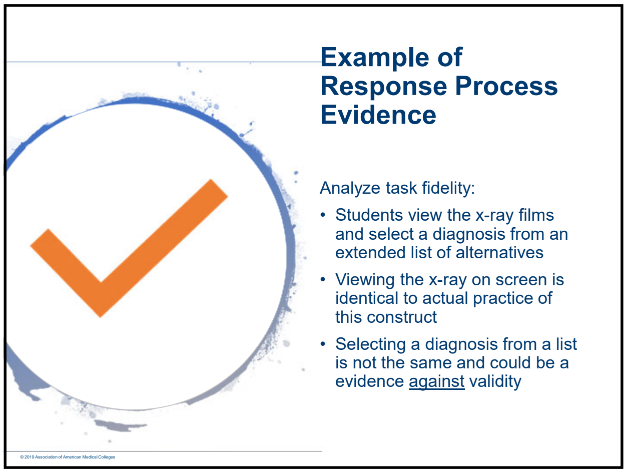
How well the cognitive processes required by the assessment map onto the processes of the underlying construct

Examining the reasoning and thought processes of learners/raters

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

Systems that reduce the likelihood of response error

55

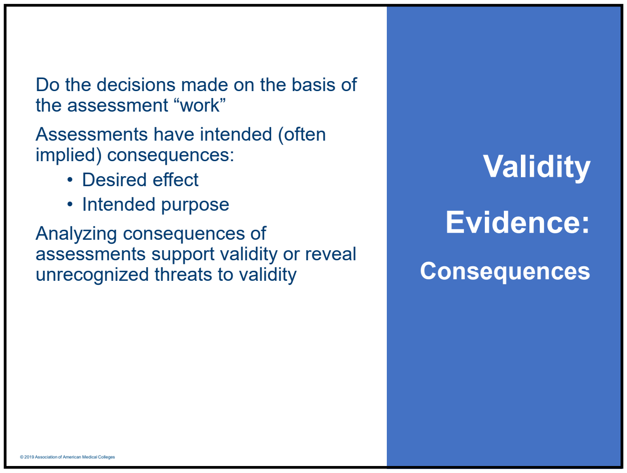


Example of Response Process Evidence

Analyze task fidelity:

- 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 <u>against</u> validity

56



Validity Evidence: Consequences

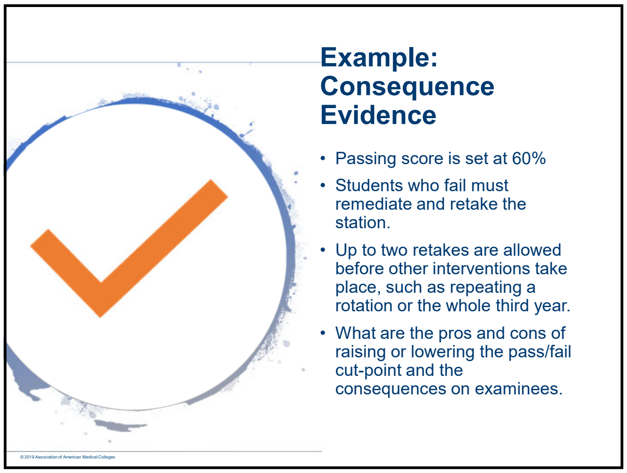
Do the decisions made on the basis of the assessment "work"

Assessments have intended (often implied) consequences:

- Desired effect
- Intended purpose

Analyzing consequences of assessments support validity or reveal unrecognized threats to validity

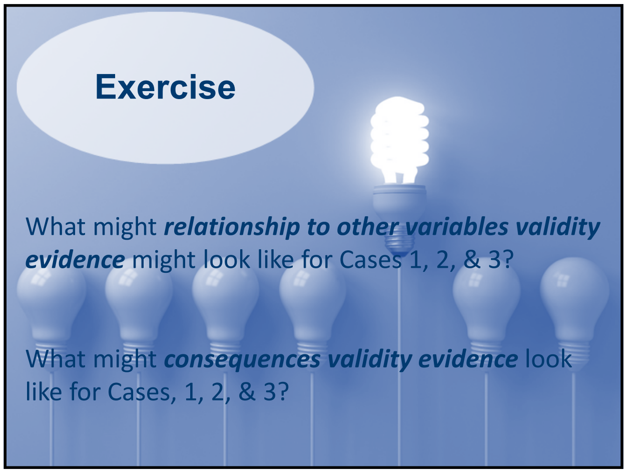
57



Example: Consequence 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.

58



Exercise

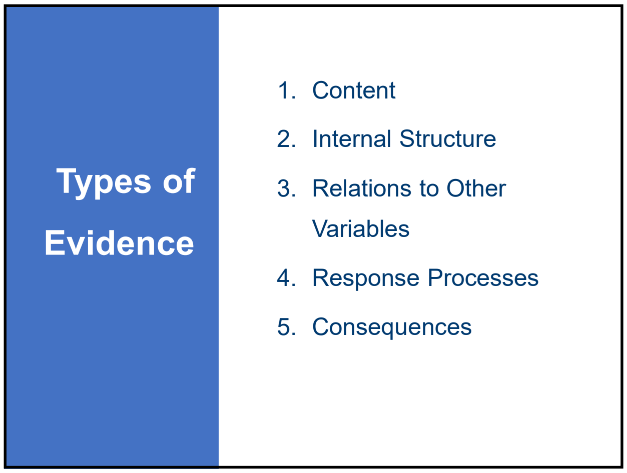
What might ***relationship to other variables validity evidence*** might look like for Cases 1, 2, & 3?

What might ***consequences validity evidence*** look like for Cases, 1, 2, & 3?

59

Let's Review

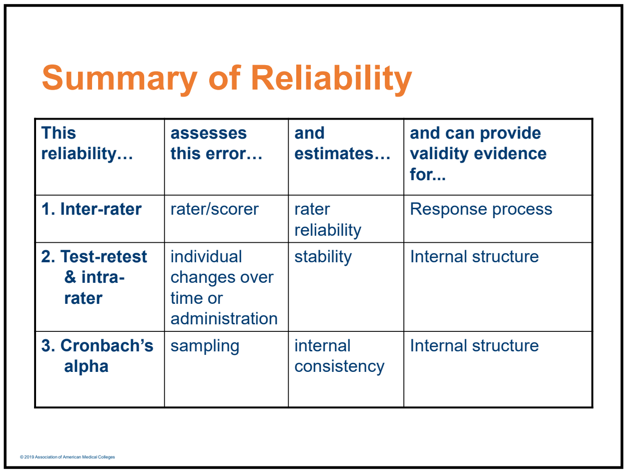
60



Types of Evidence

1. Content
2. Internal Structure
3. Relations to Other Variables
4. Response Processes
5. Consequences

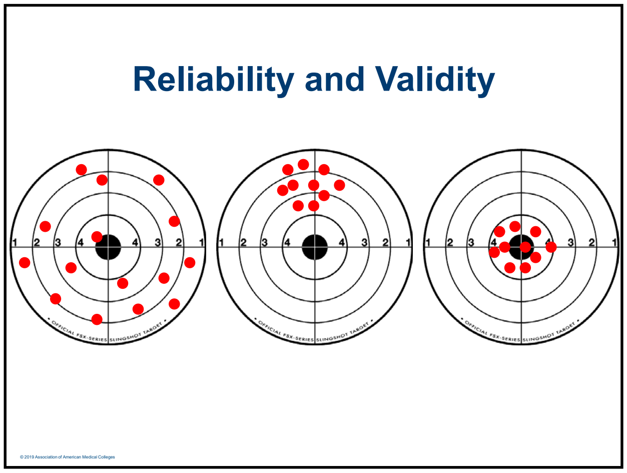
61



Summary of Reliability

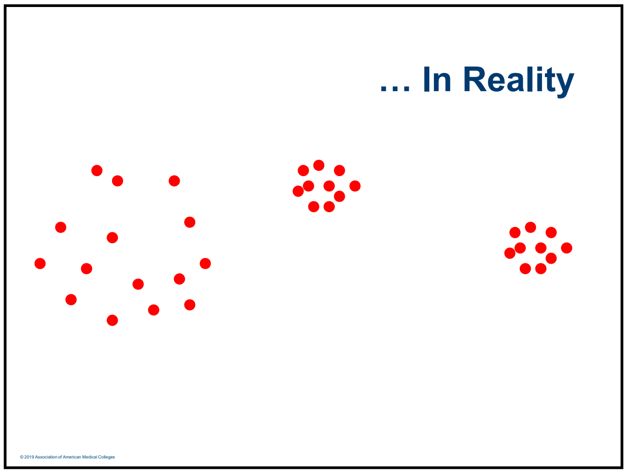
This reliability…	assesses this error…	and estimates…	and can provide validity evidence for…
1. Inter-rater	rater/scorer	rater reliability	Response process
2. Test-retest & intra-rater	individual changes over time or administration	stability	Internal structure
3. Cronbach's alpha	sampling	internal consistency	Internal structure

62



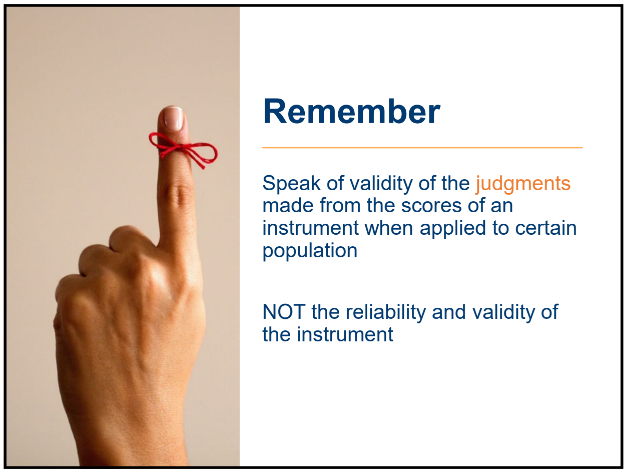
Reliability and Validity

63



… In Reality

64



Remember

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

NOT the reliability and validity of the instrument

65

Questions?

66

MERC Evaluation Link

Please go to the link below and complete the evaluation

http://goo.gl/mYQ3Dn

67