

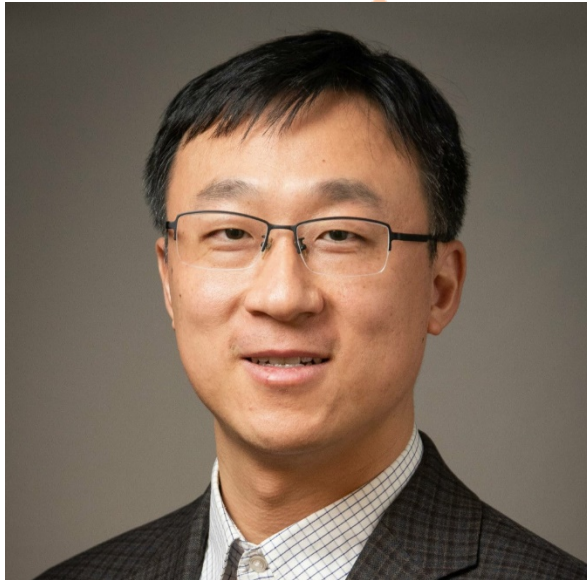
Understanding Item Analysis and Reliability Reporting in Simulation Research

Presented by Kuan Xing, PhD, David Li (Li Li),
MD, PhD, Paul Phrampus, MD, & Yoon Soo Park,
PhD

Pre-recorded video presentation for 2021
IMSH Delivers



WELCOME



Kuan Xing, Ph.D.

Position: Director of Assessment/Research
at the Center for Healthcare Improvement and Patient
Simulation (CHIPS) at the University of Tennessee Health
Science Center (UTHSC), Assistant Professor, Department
of Interprofessional Education at UTHSC;

Education: Ph.D. in Educational Psychology;

Research interests: Simulation assessment & research,
Interprofessional Education, psychometrics

WELCOME



David Li (Li Li), MD, Ph.D.

Work: Physician by training at Guangzhou First People's Hospital, South China University of Technology. Dr. Li oversees the research design and psychometrics perspectives of the national project "Assessment of residents and fellows' clinical competences using simulation" under the National Medical Examination Center.

Education: Doctor of Medicine, Huazhong University of Science and Technology, Tongji Medical College; Doctor of Philosophy, PhD (Biomedical Sciences), The University of Hong Kong, Li Ka Shing Faculty of Medicine.

WELCOME



Paul Phrampus, MD, FACEP, FSSH

Position: Director, Peter M. Winter Institute for Simulation,
Education, and Research (WISER)
Professor, Departments of Emergency Medicine and Anesthesiology
University of Pittsburgh

Past President of the Society for Simulation in Healthcare

Education: MD, Eastern Virginia Medical School

Research interests: Simulation medical education, simulation as a
quality improvement and patient safety tool, simulation as a
competency assessment tool

WELCOME



Yoon Soo Park, PhD

Position: **Associate Professor** at Harvard Medical School and the inaugural **Director** of Health Professions Education Research at the Massachusetts General Hospital; **Chair** of the Research in Medical Education (RIME) committee of the Association of American Medical Colleges (AAMC); **Vice President** for the American Educational Research Association (AERA), serving Division I: Education in the Professions

Education: Ph.D. in Measurement, Evaluation, and Statistics, Columbia University

Research interests including assessment methods in health professions education, advancing the preparation of learners in clinical reasoning and measurement of competencies through validity studies, and psychometric research

Content by sections:

1 Introduction to Basic Psychometrics: Item Analysis & Reliability (Kuan Xing & Yoon Soo Park)

2 Reliability and Real-World Simulation Examples (David Li & Paul Phrampus)

3 More about Reliability – Interrater Reliability: Definition, Misconceptions, and Pitfalls (Kuan Xing & Yoon Soo Park)

4 More about Assessment – Concepts, Reliability, and Item Analysis: Nuts and Bolts (Yoon Soo Park & Kuan Xing)

5 Ten Tips to Improve your Assessment Program (David Li & Paul Phrampus)



Thank You!

SIMULATION:
BRINGING LEARNING TO LIFE

#IMSH2021

IMSH 2021 Preconference Workshop

1. Introduction to basic psychometrics: Item analysis & reliability

Kuan Xing, PhD, & Yoon Soo Park, PhD

Workshop Objectives:

- Describe **key concepts** in item analysis and reliability indices for assessments in simulation;
- Identify and select **appropriate indices** for reporting item analysis and reliability in various simulation/assessment scenarios;
- Understand **best-practice guidelines** for interpreting and improving assessments based on results from item statistics and reliability indices.

Section 1 Outline:

- The basics: Research/scholarship, assessment, & data
- Measurement, reliability, & validity: Concepts
- Item analysis
 - Item difficulty
 - Item discrimination
- Reliability indices
 - Internal consistency

Research/Scholarship:

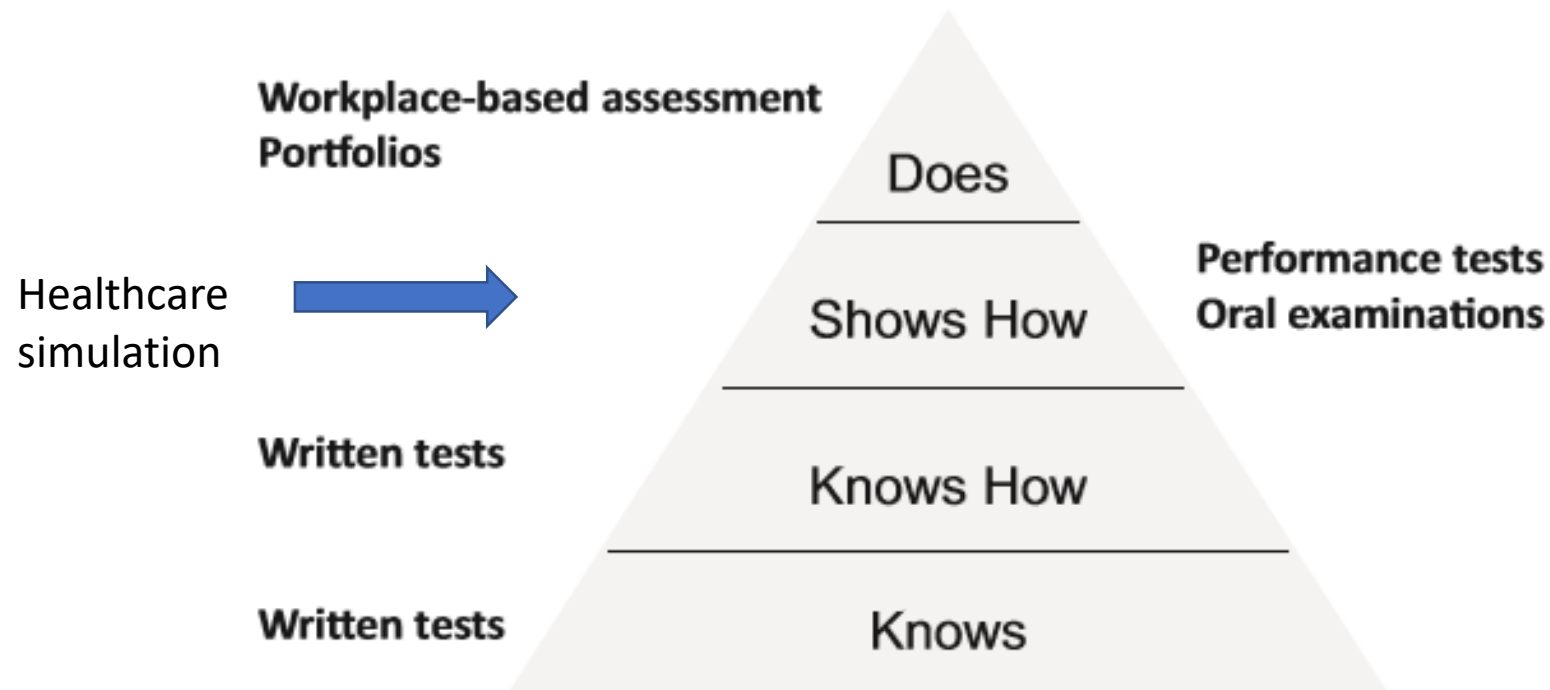
Boyer's definition: 4 types of scholarship –

- Discovery;
 - Integration;
 - Application;
 - Teaching.
-
- Separate yet **overlapping**;
 - A lot of simulation work/research may fall in type **#3** and **#4**

(Glassick, 2000)

Assessment: Miller's framework

Assessment: [systematic process](#) to measure or evaluate the [characteristics or performance](#) of individuals, programs, or other entities, for purposes of drawing [inferences](#) (AERA, APA, & NCME, 2014).




(Miller, 1990; Yudkowsky, Park, & Dawning, 2020)

Data (Level of measurement):

4 types of data:

- Nominal (counts): e.g., Gender, Smoker/Non-smoker;
- Ordinal (ordering): e.g., Ranking;
- Interval (same unit): e.g., (Standardized) test score;
- Ratio (w/ absolute 0): e.g., Length or weight.

- Data type **determines** analysis type;

- Study design/planning: better at earlier stage

Measurement, Reliability, & Validity

- Measurement:

“If something exists, it exists in some amount. If it exists in some amount, then it is capable of being measured.”

-*Rene Descartes*

- Psychometrics:

theory and technique of (quantitative) psychological measurement;



(1596 – 1650)

What is reliability? (1)

- An example: signal vs. noise;
Eating outdoor: how clearly you can hear your friend's talk
- Reliability: The reliability of an assessment is the extent to which can be relied upon to produce 'true' scores
- $\text{Observed Score} = \text{True Score} + \text{Error}$ (Classical Test Theory)

What is reliability? (2)

- Measure of **consistency** across **occasions** or with **different sets of equivalent items**
- What proportion of the data is **useful information** rather than random **noise**?

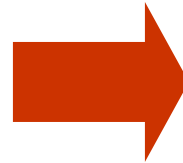
$$\text{reliability} = \frac{\text{Signal}}{\text{Signal} + \text{Noise}} = \frac{\sigma_T^2}{\sigma_X^2}$$

What is validity? (1)

Validity: An argument

- For example: Going to a court; justify for the appropriate use of an assessment tool; you talk about **validity**

Can I defend the use
of the **scores** from
this assessment



To make a **decision**
for a given **purpose?**

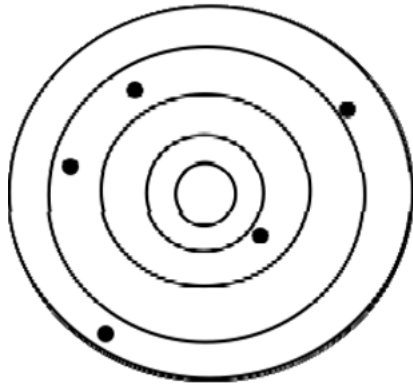
What is validity? (2)

Sources of Validity Evidence

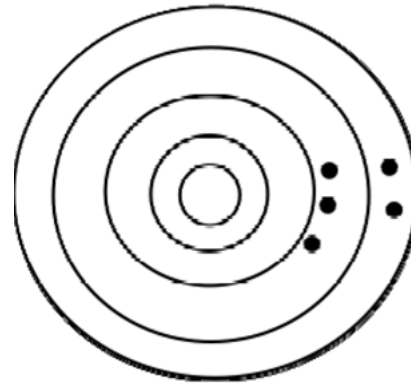
- **Content**
 - Test content reflected in blueprint and relevance of items
- **Response Process**
 - Quality of responses from students, observers, or test administration
- **Internal Structure**
 - Reliability and psychometrics
- **Relations to other variables**
 - Correlation with scores from other relevant assessments
- **Consequences**
 - Impact on students/curriculum, passing standards of students

(Messick, 1990)

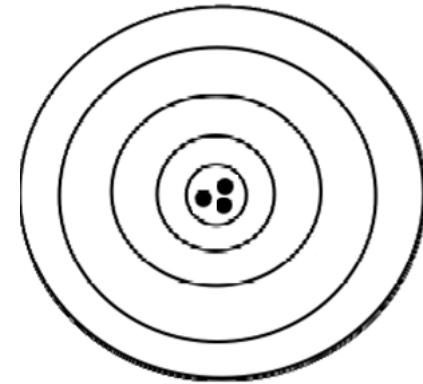
Relationship b/t Reliability and Validity:



Low reliability & validity



High reliability, low validity



High reliability & validity

Item Analysis

- Are items reliable? Do they measure consistently?
- Which items are most difficult to answer correctly?
- What items are easy?
- Are there poor performing items that need to be discarded?

Purpose of Item Analysis

- Evaluate the quality of each item
- Rationale: the quality of items determines the quality of test (i.e., reliability & validity)
- May suggest ways of improving the measurement of a test

Item Difficulty (1)

- The proportion of examinees who get a particular item correct

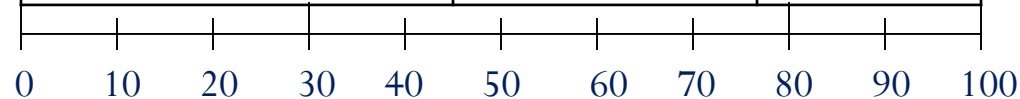
$$p = \frac{\text{number correctly answering the item}}{\text{number taking the test}}$$

Item Difficulty (2)

Item Difficulty Level

- The percentage of students who answered the item correctly

High (Difficult)	Medium (Moderate)	Low (Easy)
< 45%	≥ 45% AND < 75%	≥ 75%



Item Discrimination (1)

- How well an item discriminates between students of high and low performance
- **Item discrimination:** Correlation between each item and the total test score (item-total correlation)



Item Discrimination (2)

- Range: - 1 ~ + 1;
- Rule of thumb: $\geq .20$, good to keep; $0 \sim .20$, revise/drop; < 0 , drop

Item	Item discrimination	Evaluation
1	.33	Keep
2	.40	Keep
3	.10	Revise
4	.00	Drop
5	-.15	Drop

Item Discrimination (3)

- Item-total correlations are directly related to reliability
- Items with higher item-total correlations are more discriminating
- Consider: How to use such information in your own setting?

Reliability Indices

- Types of reliability
 - Test-retest: same person/tool, different occasions
 - Split-half: halves of a test: equivalent
 - Internal Consistency: structure of assessment tool
 - Rater reliability: consistency b/t raters

Internal Consistency Reliability

- If items are measuring the **same construct** they should elicit similar if not identical responses
- *Cronbach's Alpha (α)* is a widely used measure of internal consistency for continuous data
 - .80 to .95 (Excellent)
 - .70 to .80 (Very Good)
 - .60 to .70 (Satisfactory)
 - <.60 (Suspect)

How can we increase reliability?

- Analyze your items
- Increase the number of items

References

- AERA, APA, & NCME. *Standards for educational and psychological testing*. Washington DC: American Educational Research Association; 2014.
- Glassick CE. Boyer's expanded definitions of scholarship, the standards for assessing scholarship, and the elusiveness of the scholarship of teaching. *Acad Med*. 2000 Sep;75(9):877-80.
- Messick S. *Validity of Test Interpretation and Use*. Princeton, NJ: Educational Testing Service;1990.
- Miller G. The assessment of clinical skills/competence/performance. *Acad Med*. 1990 Sep;65(9 Suppl):S63-7.
- Yudkowsky R, Park YS, Downing SM. *Assessment in Health Professions Education*. 2nd ed. New York: Routledge; 2020.

➤ Next: Section #2 – Reporting reliability: Examples from simulations

Section 2

Reliability and Real-World Simulation Examples

David Li (Li Li), MD, Ph.D.

Paul E. Phrampus, MD FSSH

Reliability in Simulation Based Assessment

- Why should we care?
 - Research
 - Assessment of Performance
 - Competence

Reliability – Real World Examples in Simulation

R_x

Rater(s) with Scoring Tool

Sim
X

Simulation with
Student Being Assessed

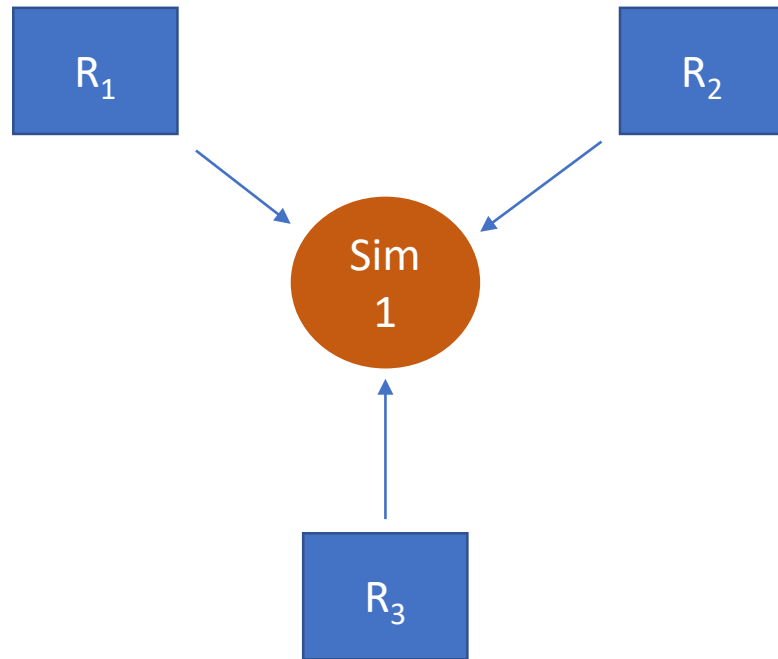
Score Sheet

	Yes	No
Item 1	<input type="checkbox"/>	<input type="checkbox"/>
Item 2	<input type="checkbox"/>	<input type="checkbox"/>
Item 3	<input type="checkbox"/>	<input type="checkbox"/>
Item 4	<input type="checkbox"/>	<input type="checkbox"/>

Score Sheet

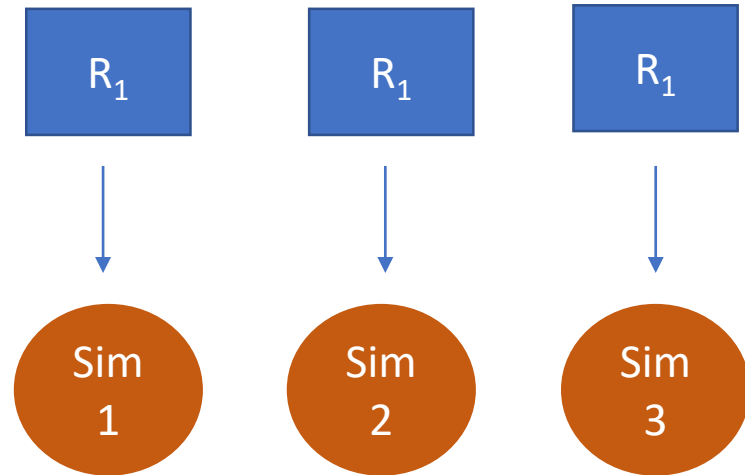
	Not Good		Average		Great
Item 1	1	2	3	4	5
Item 2	1	2	3	4	5
Item 3	1	2	3	4	5
Item 4	1	2	3	4	5

Reliability – Real World Examples in Simulation



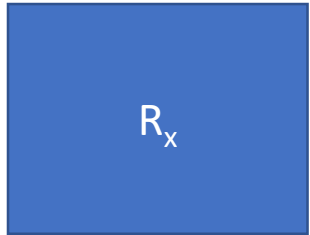
Do all raters come up with a similar score?

Reliability – Real World Examples in Simulation



Does rater score the items the same based on similar performance?

Factors in Simulation Reliability (RST)



Raters



Simulation

Score Sheet

	Yes	No
Item 1	<input type="checkbox"/>	<input type="checkbox"/>
Item 2	<input type="checkbox"/>	<input type="checkbox"/>
Item 3	<input type="checkbox"/>	<input type="checkbox"/>
Item 4	<input type="checkbox"/>	<input type="checkbox"/>

Tool(s)

Factors Involving Raters

- Expertise Level
- Training for Particular Assessment Activity (Rater Calibration)
 - Item Definitions
 - Mock Exams
- Fatigue
- Distractions
- Bias

Factors Involving the Simulation

- Scenario Design Factors
 - Equipment Availability
 - Consistency of Simulation Response to Learner Interventions
 - Computerized Simulators
 - Humans
 - Faculty Involvement
 - Standardized People
 - Built in Features to Enhance Assessment Capabilities
- Equipment Failure / Inconsistencies
- Interpretation of What Is Being Simulated
 - Design Artifact

- Equipment with Hole

Factors Involving Tools

- Overall Length
 - How Many Items Can A Rater Assess
- Anchored Definitions (Closely Related to Training of Raters)
- Technology Assistance
 - Automated Data Collection
 - Electronic Checklist
 - Video Review
- When is Data Being Collected
 - During Simulation
 - During Debriefing
 - Post Simulation Assessment

Score Sheet		
	Yes	No
Item 1	<input type="checkbox"/>	<input type="checkbox"/>
Item 2	<input type="checkbox"/>	<input type="checkbox"/>
Item 3	<input type="checkbox"/>	<input type="checkbox"/>
Item 4	<input type="checkbox"/>	<input type="checkbox"/>

Summary

- Many Factors Can Affect Reliability of Assessment Tools Used in Simulations
- Reliability can be affected through three main factors
 - Raters
 - Simulation
 - Tool (Assessment)

IMSH 2021 Preconference Workshop

3. More about reliability - Interrater reliability: Definition, misconceptions, and pitfalls

Kuan Xing, Ph.D., & Yoon Soo Park, Ph.D.

Section 3 Outline

- Interrater reliability: Can you measure?
- How to measure?
- Misconceptions and pitfalls
- Quality monitoring

Rater Challenges



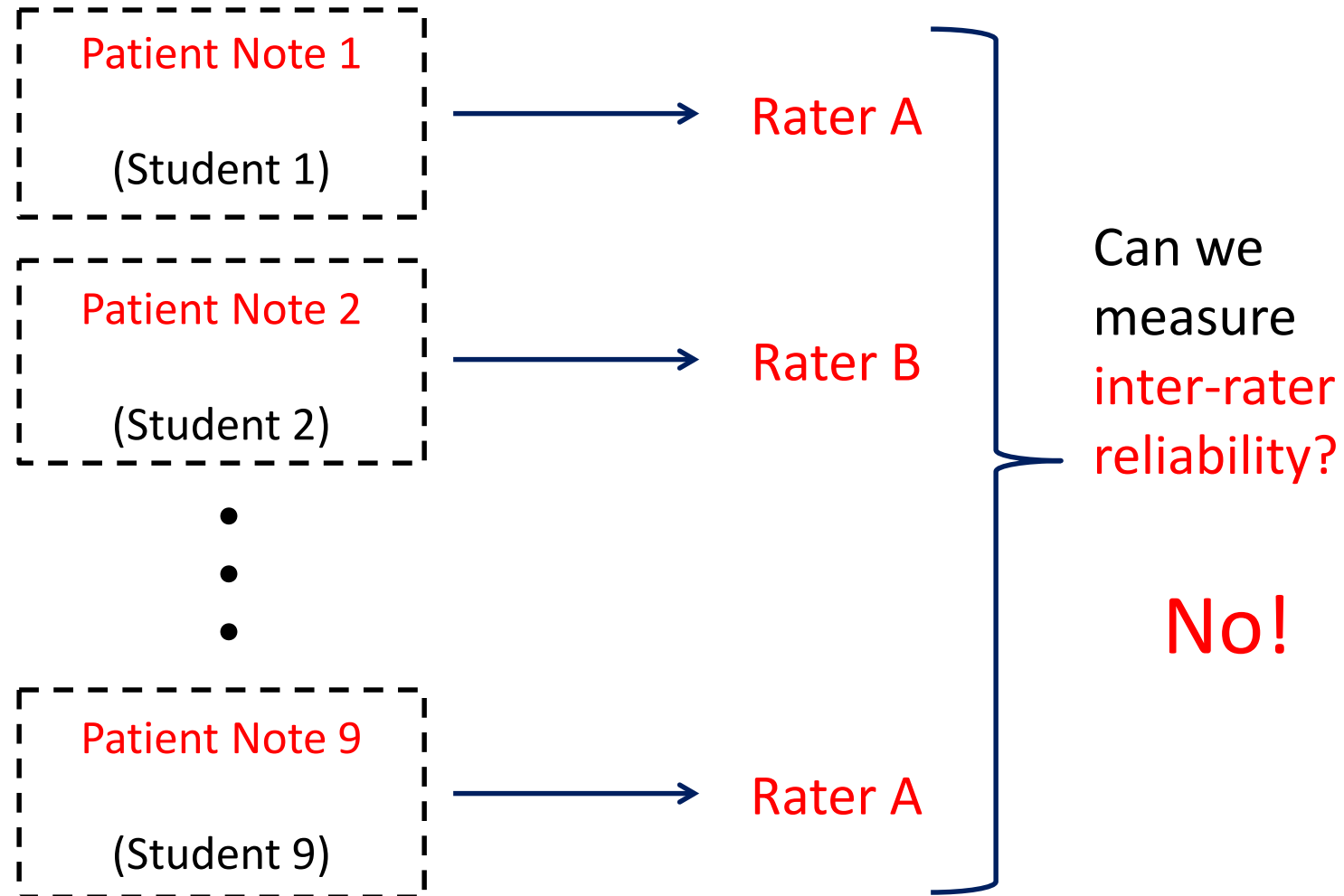
*A person with one watch knows what time it is;
a person with **two watches is never quite sure.***

– Robert Brennan

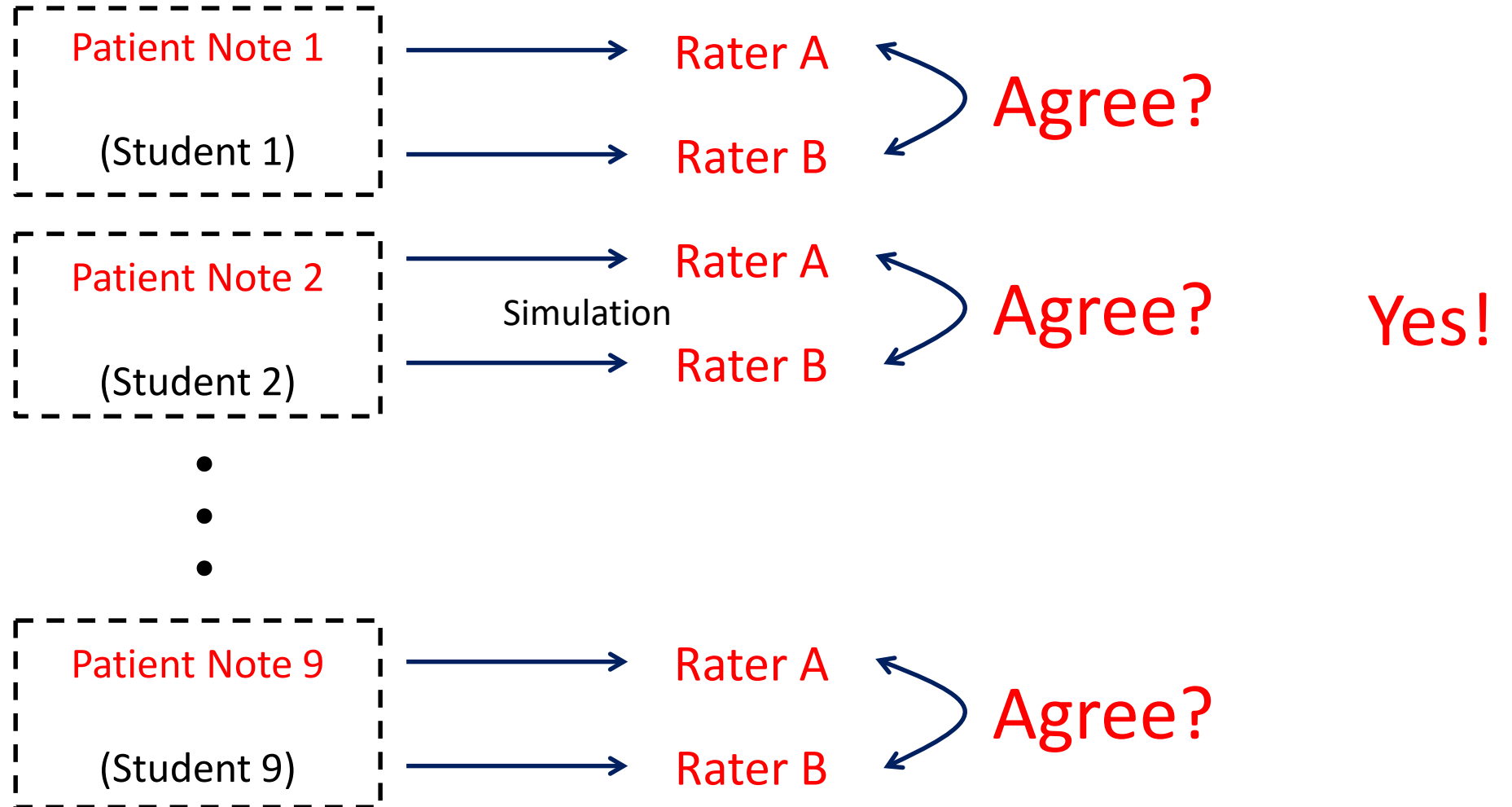
Questions to consider...

- What **rater-related issues** have you encountered?
- How did you **overcome these issues**?
- When can you measure **inter-rater agreement/reliability**?

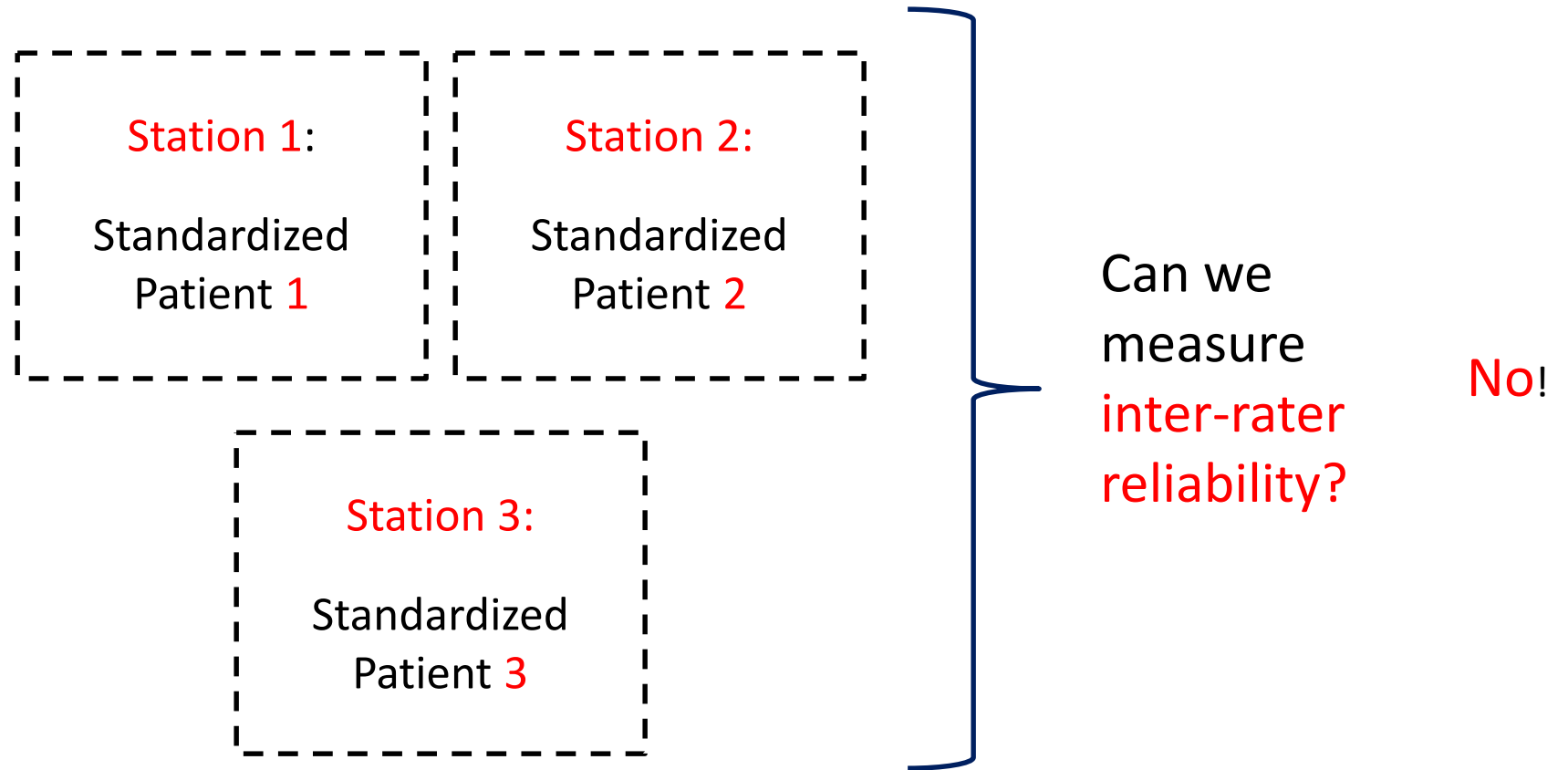
Example 1: Scoring Patient Notes



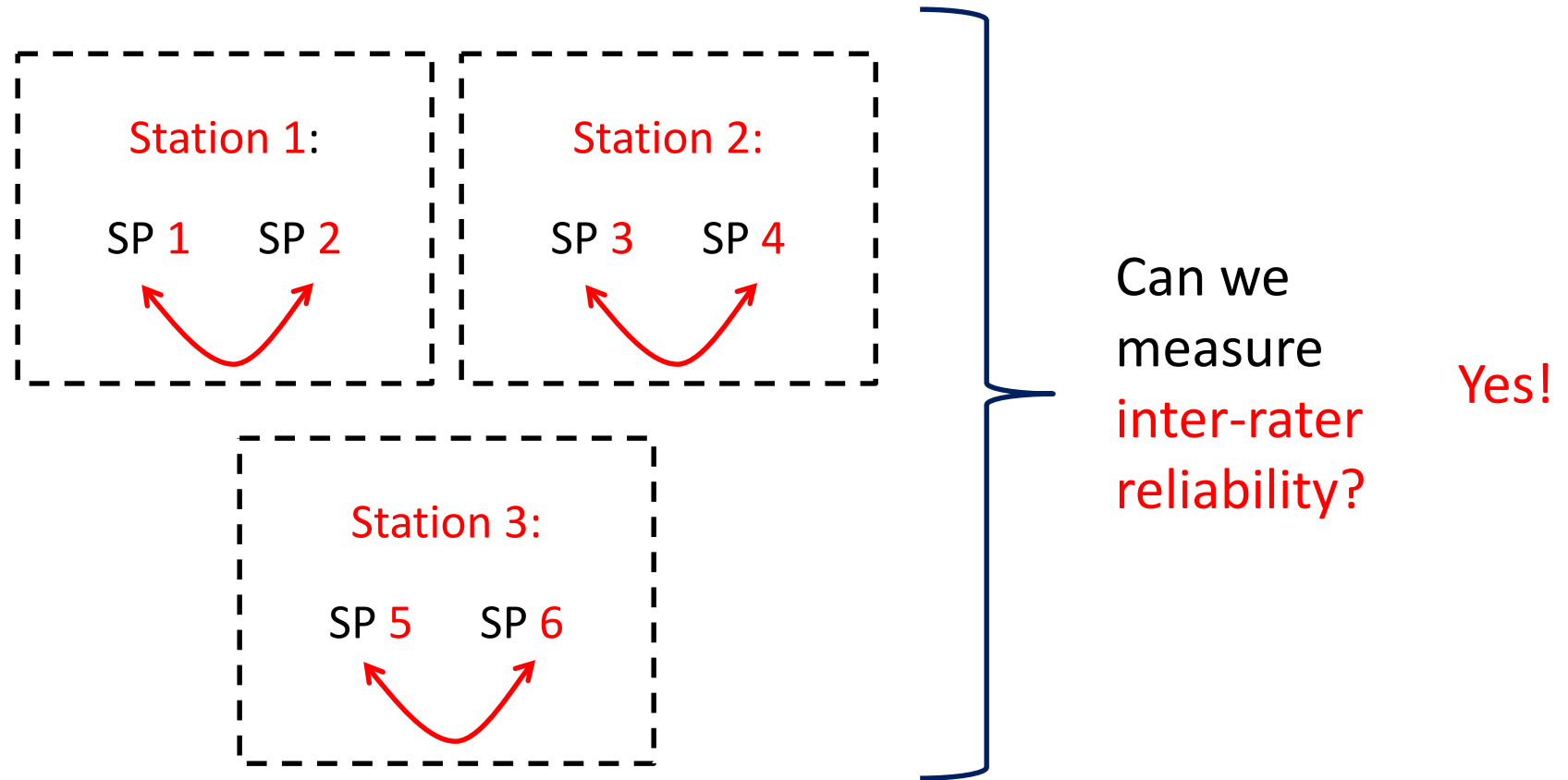
Example 1: Scoring Patient Notes (continued)



Example 2: OSCE with 3 stations and 3 different standardized patients



Example 2: OSCE with 3 stations and 6 different SPs (2 SPs per station)



Overcome these issues?

- Single rater per observation?
- No
- Multiple raters for multiple observations, no double-scoring?
- No
- For a single observation, at least 2 raters (double-scoring)
- Rater design (e.g., fully-crossed, nested)

Interrater Reliability

- Definition: the extent to which **independent** evaluators produce **similar** ratings in judging the **same** abilities or characteristics in the **same** target person or object.
- Agreement (between raters)
- Consistency vs. accuracy?

(APA Dictionary of Psychology)

Measures of interrater reliability: Example

Table 2 Reliability of marking

July	Station: Agreement Kappa	Psoriasis 72% 0.39	Hyperthyroid 79% 0.58	TIA 84% 0.62	DVT 74% 0.45	Back pain 77% 0.32
September	Station: Agreement Kappa	Psoriasis 94% 0.88	Hyperthyroid 91% 0.82	TIA 83% 0.67	Prostate 87% 0.69	Back pain 93% 0.85
October	Station: Agreement Kappa	Asthma 67% 0.26	Contraception 88% 0.73	TIA 89% 0.76	Prostate 89% 0.74	Back pain 88% 0.76
November	Station: Agreement Kappa	Asthma 74% 0.31	CVD 92% 0.71	Diabetes 82% 0.61	Smear 83% 0.04	Earache 88% 0.73

DVT = Deep Vein Thrombosis; TIA = Transient Ischaemic Attack; CVD = Cardiovascular Disease.

(Singleton et al., 1999)

Interrater reliability indices (1)

- Exact agreement (EA):

$$EA = \frac{\text{Number of concordant responses} * 100\%}{\text{Total number of responses}}$$

- Kappa:

$$\text{kappa} = \frac{\text{Proportion observed agreement} - \text{Proportion expected chance agreement}}{1 - \text{Proportion expected chance agreement}}$$

Correction for chance? (1)

Inter-rater agreement		Rater 1	
		No	Yes
Rater 2	No	20	2
	Yes	0	5

% agreement = $25 / 27 = 92.6\%$

Kappa = **0.787**

Kappa takes into account **chance agreement**

(Hasnain et al., 2004)

Correction for chance? (2)

Inter-rater agreement		Rater 1	
		No	Yes
Rater 2	No	25	2
	Yes	0	0

% agreement = $25 / 27 = 92.6\%$

Kappa = 0.000

Both Raters 1 and 2 have no agreement for "YES"!

Interrater reliability indices (2)

ICC: Intraclass correlation coefficient

$$\text{ICC} = \frac{\textit{Between subjects variance}}{\textit{Between subjects variance} + \textit{Within subjects variance}}$$

- was originally applied to the evaluation of differences between interval or ratio variables;
- Applicable for multiple raters' scenario (≥ 3);
- Mathematically **equivalent** to **weighted kappa** under certain circumstances

Scoring Design and Rater Training (Misconceptions/Pitfalls)

- Carefully craft meaningful and clear rubric **before** scoring
- Most rater training programs focus on **rater severity**
- Rather, focus should be placed on how well **raters discriminate differences between scoring categories!**
 - Focusing on discrimination can increase classification by up to 20%

Quality Monitoring (1)

- % Exact
- Kappa
- ICC
- Guidelines for % Exact
 - 7 pt. scale: ~50% or better
 - 5 pt. scale: 70% or better
 - 4 pt. scale: 80% or better
- Depending on number of scoring categories, guidelines can vary

Quality Monitoring (2)

- Guidelines for kappa
 - > 0.75 : Excellent agreement
 - $0.40 - 0.75$: Intermediate to Good agreement
 - < 0.40 : Poor agreement

(Landis & Koch 1977)

Quality Monitoring (3)

- Guidelines for ICC:
 - > 0.75 : Excellent agreement
 - $0.6 - 0.75$: Good agreement
 - $0.4 - 0.6$: Fair agreement
 - < 0.4 : Poor agreement

(Cicchetti, 1994)

Scoring accuracy vs. Scoring consistency

- Raters can be **consistent, but not accurate**
- Two inaccurate raters can have high agreement and two accurate raters can disagree
 - Hard to know who is right and who is wrong
- **Maintain standards** for score quality even in the face of challenging score reporting demands

References

American Psychological Association (APA) Dictionary of Psychology. *Interrater reliability*.

<https://dictionary.apa.org/interrater-reliability>

Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994;**6**:284-90.

Hasnain M, Onishi H, Elstein AS. Inter-rater agreement in judging errors in diagnostic reasoning. *Med Educ* 2004;**38**:609-16.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
Singleton A, Smith F, Harris T, Ross-Harper R, Hilton S. An Evaluation of the Team Objective Structured Clinical Examination (TOSCE). *Med Educ* 1999;**33**:34-41.

- Next: Section #4 More about Assessment –
Concepts, Reliability, and Item Analysis: Nuts and bolts



4. More About Assessment Concepts, Reliability, and Item Analysis: **Nuts and Bolts**

Yoon Soo Park, PhD

Kuan Xing, PhD

Example: How is item analysis conducted?

Statistic	Value
-----------	-------

Number of Examinees	100
---------------------	-----

Number of Items	50
-----------------	----

Reliability	0.70
--------------------	-------------

Mean Item Difficulty	0.56
----------------------	------

Mean Item Discrimination	0.25
--------------------------	------



Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
1	0.60	0.20	26	0.28	0.12
2	0.46	0.34	27	0.67	0.35
3	0.84	0.27	28	0.32	0.28
4	0.40	0.36	29	0.56	0.17
5	0.83	0.27	30	0.39	0.03
6	0.68	0.32	31	0.62	0.49
7	0.85	-0.01	32	0.47	0.26
8	0.80	0.38	33	0.55	0.38
9	0.35	0.25	34	0.37	0.04
10	0.46	0.39	35	0.57	0.38
11	0.48	0.34	36	0.37	0.30
12	0.48	0.26	37	0.86	0.24
13	0.46	0.34	38	0.61	0.32
14	0.59	0.30	39	0.38	0.01
15	0.59	0.39	40	0.51	0.24
16	0.53	0.39	41	0.44	0.19
17	0.49	0.05	42	0.83	0.37
18	0.41	0.19	43	0.74	0.30
19	0.75	0.06	44	0.33	0.24
20	0.96	0.23	45	0.33	-0.05
21	0.52	0.35	46	0.49	0.29
22	0.49	0.13	47	0.49	0.23
23	0.69	0.41	48	0.35	0.07
24	0.83	0.18	49	0.45	0.44
25	0.54	0.35	50	0.83	0.24

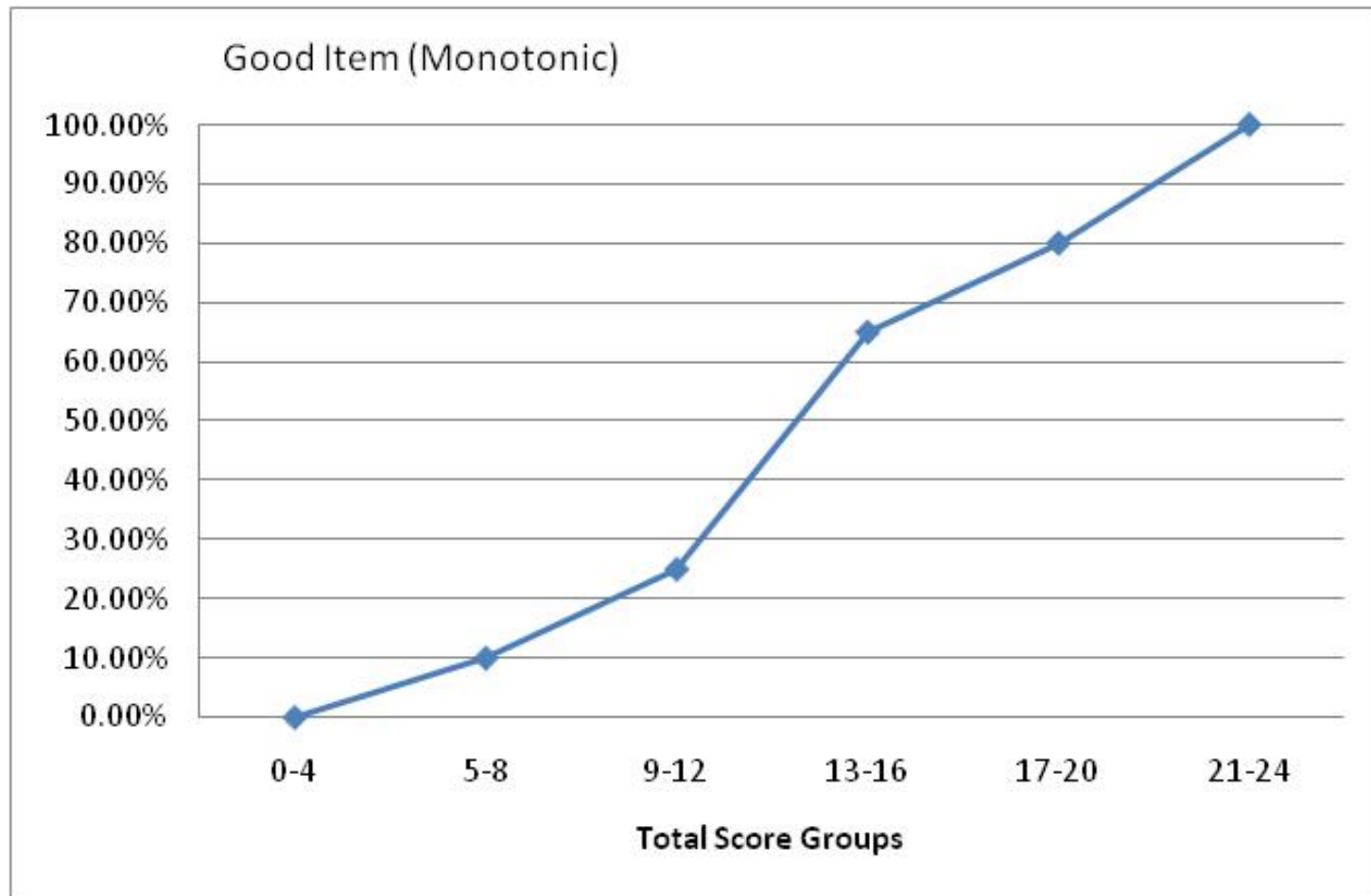
Results of Item Analysis

Statistic	Original Data	Best Items Only	Worst Items removed
Number of Items	50	19	35
Reliability	0.70	0.75	0.78
Mean Item Difficulty	0.56	0.58	0.58
Mean Item Discrimination	0.25	0.43	0.34

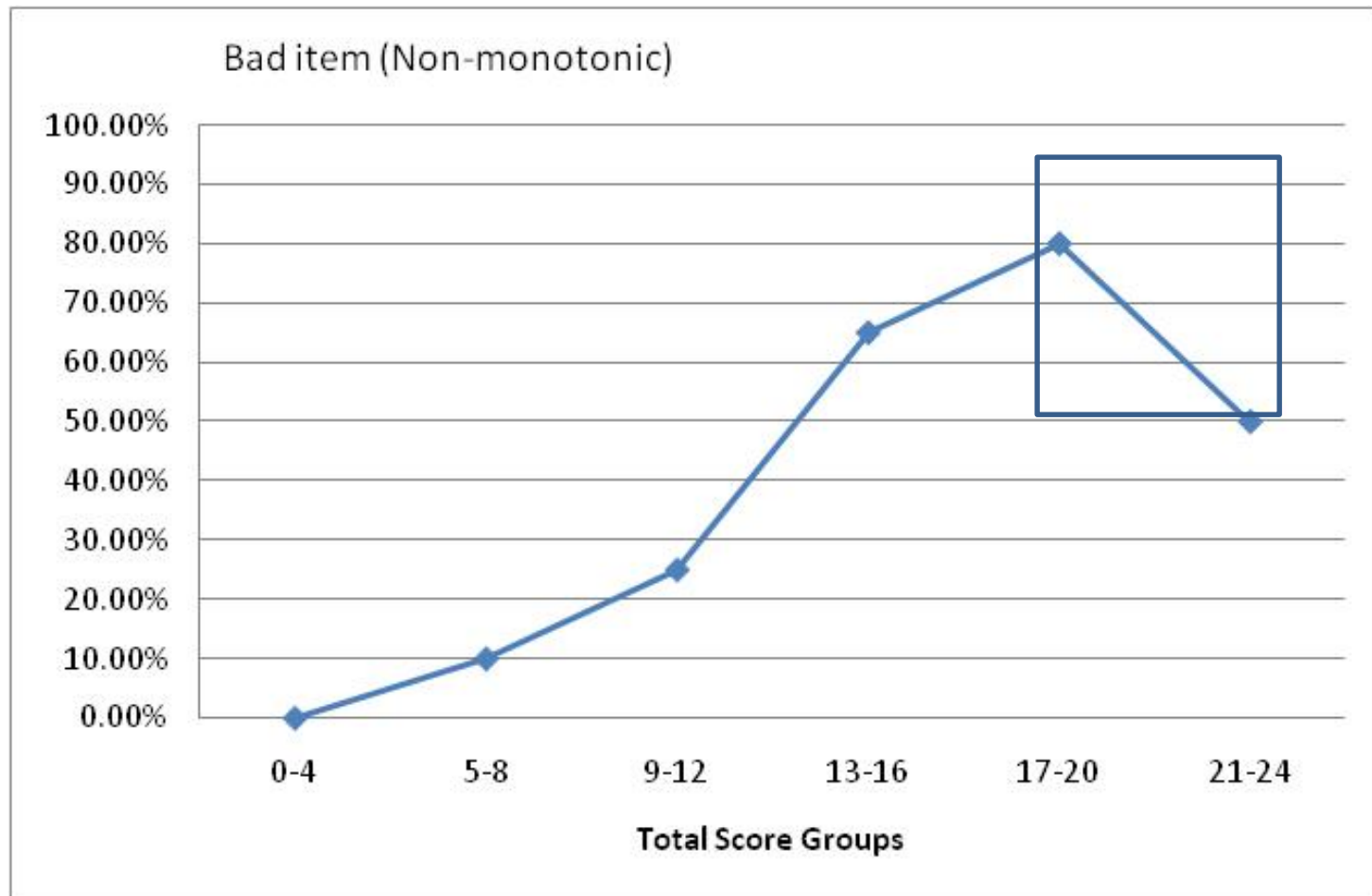
Item Characteristic Curves

- A graph of the proportion of examinees getting each item correct, compared to total scores on the test
- Ideally, lower test scores → lower proportions of examinees getting a particular item correct
- Ideally, higher test scores → higher proportions of examinees getting a particular item correct

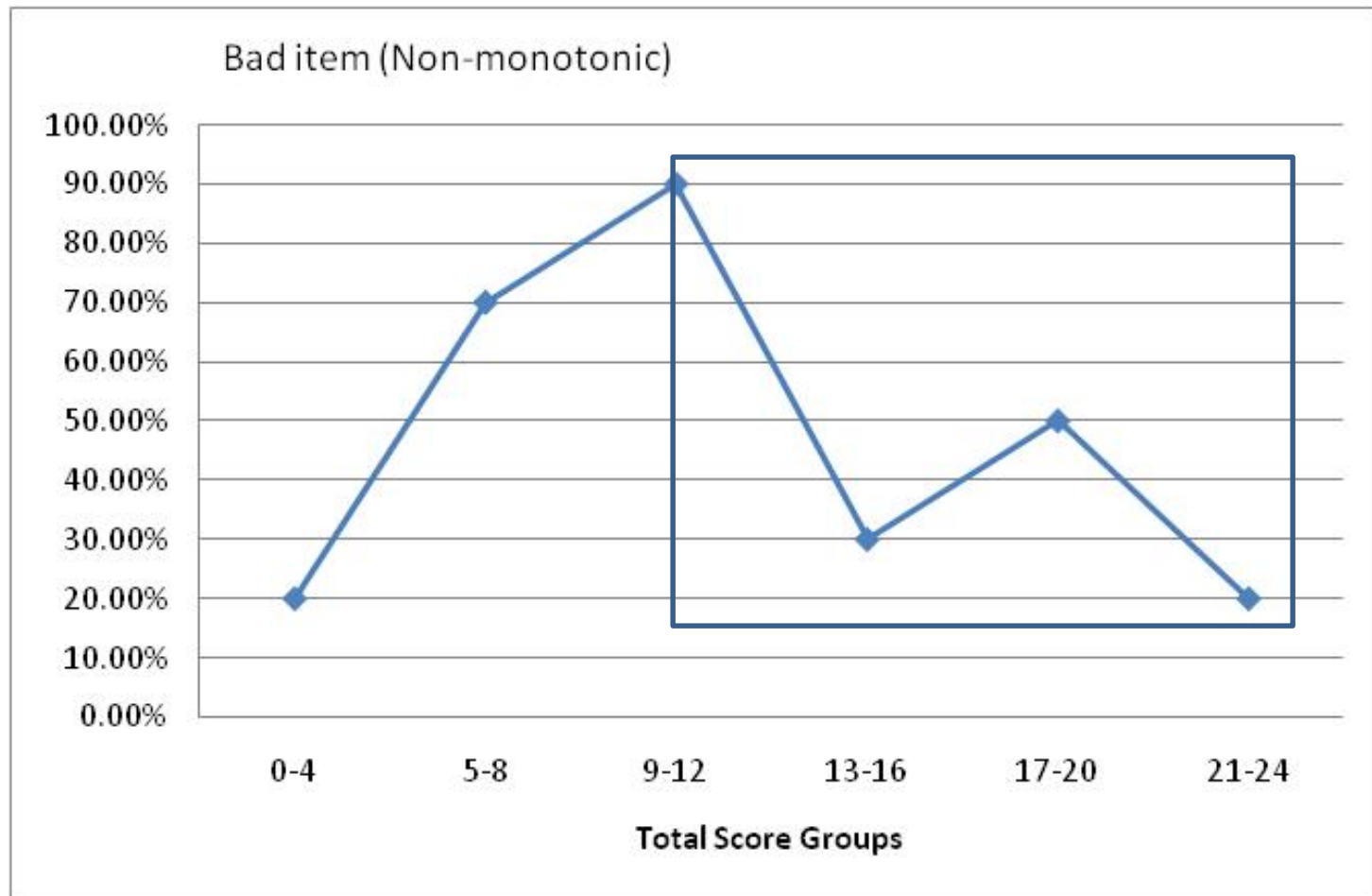
Item Characteristic Curve: Good item



Item Characteristic Curve: Bad item (1)

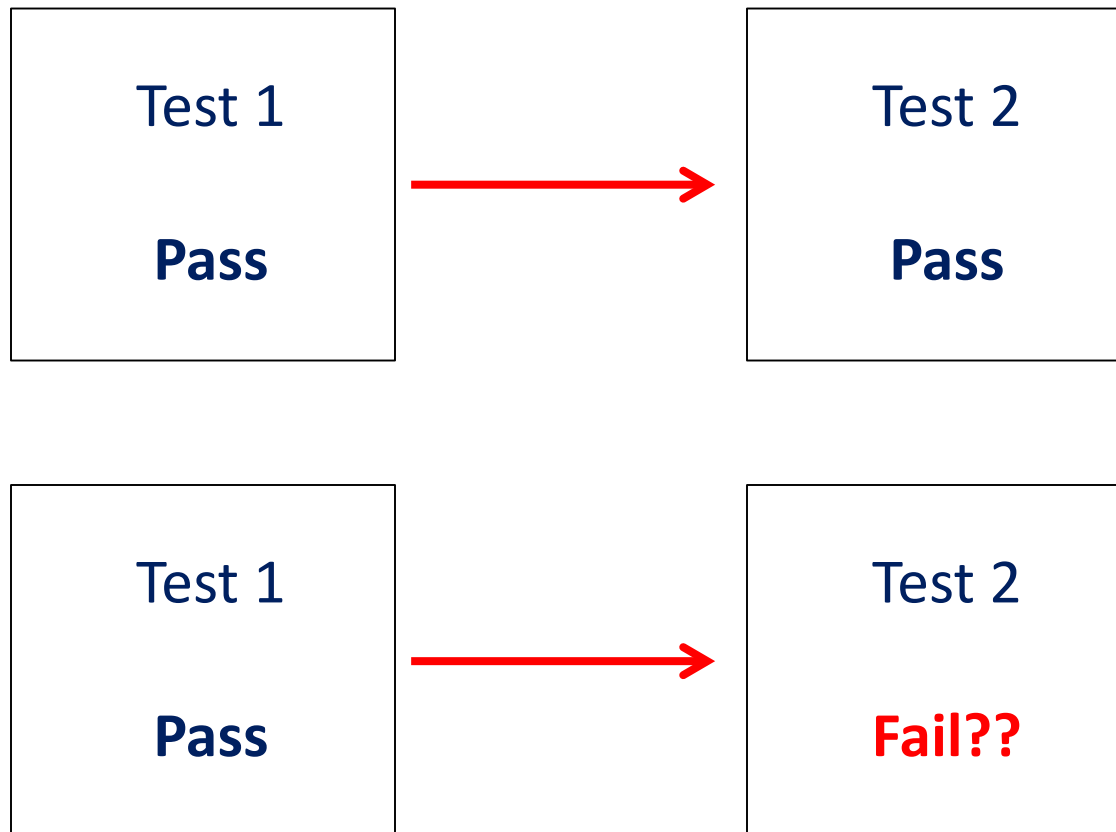


Item Characteristic Curve: Bad item (2)

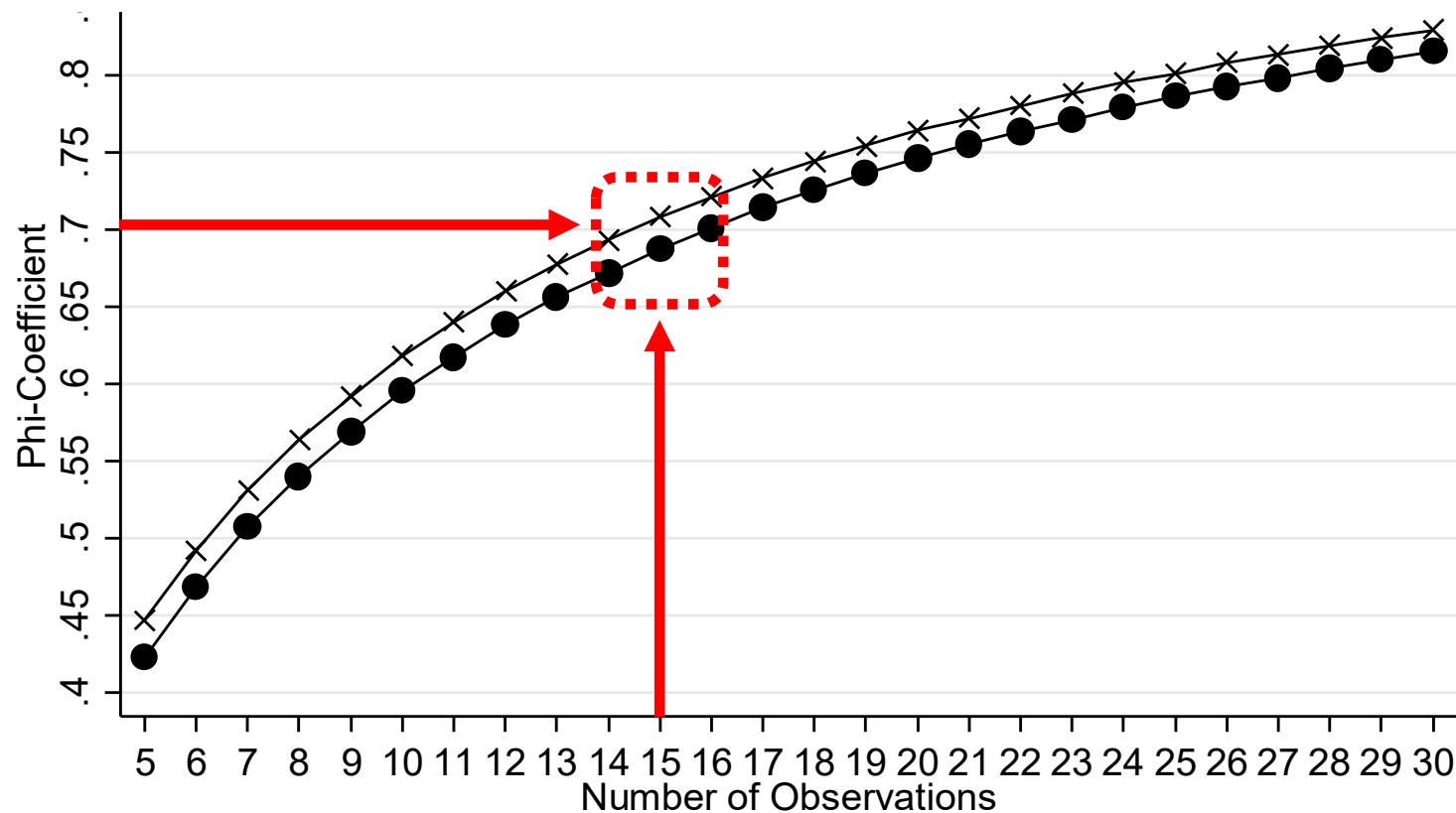


1. Check the **Reliability** of your assessment

Reliability = Consistency

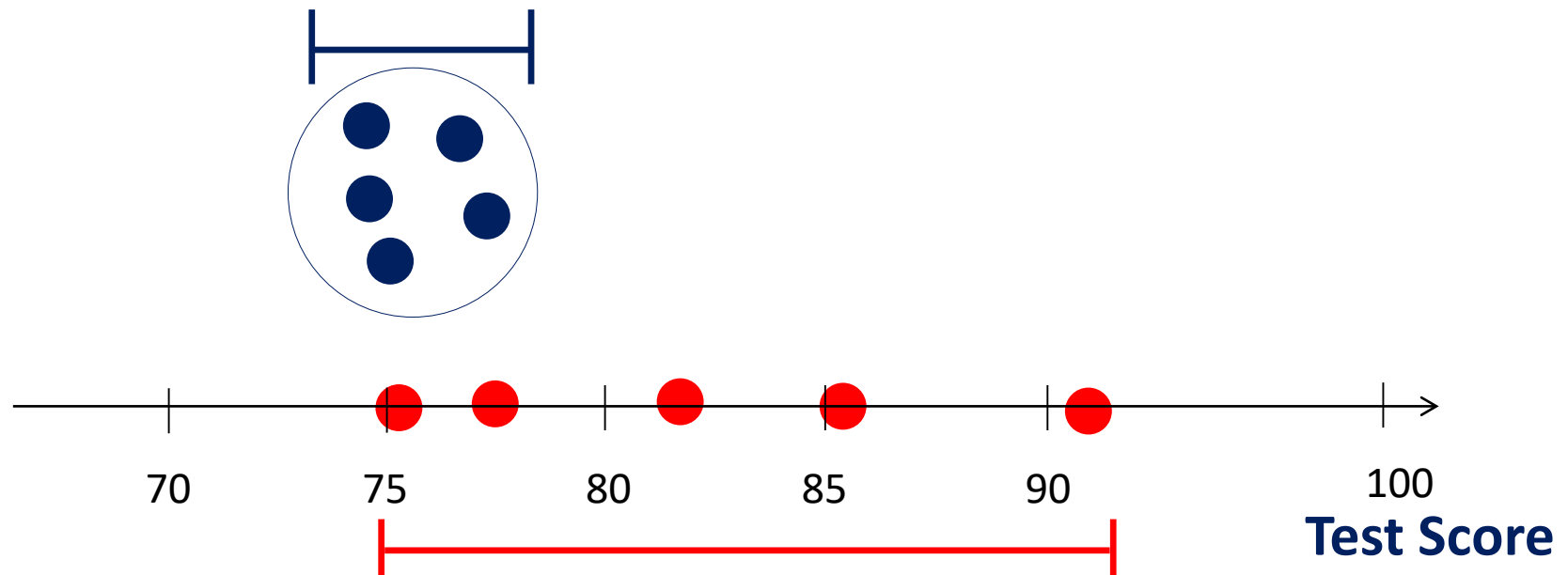


2. ↑ Items and Observations → Higher Reliability



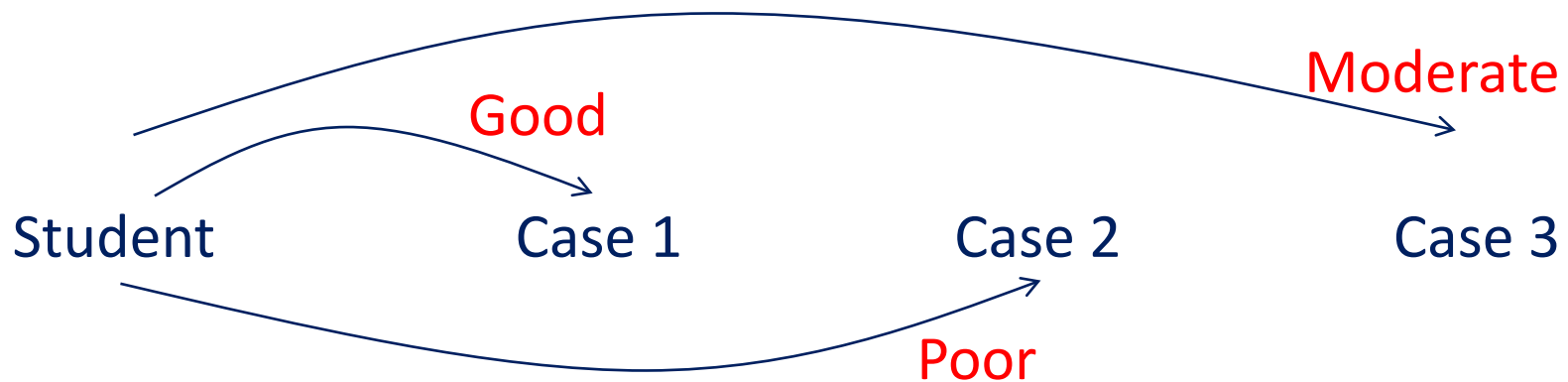
3. Can the Assessment Identify High and Low Performing Students?

- How well does the assessment discriminate **differences of learners?**



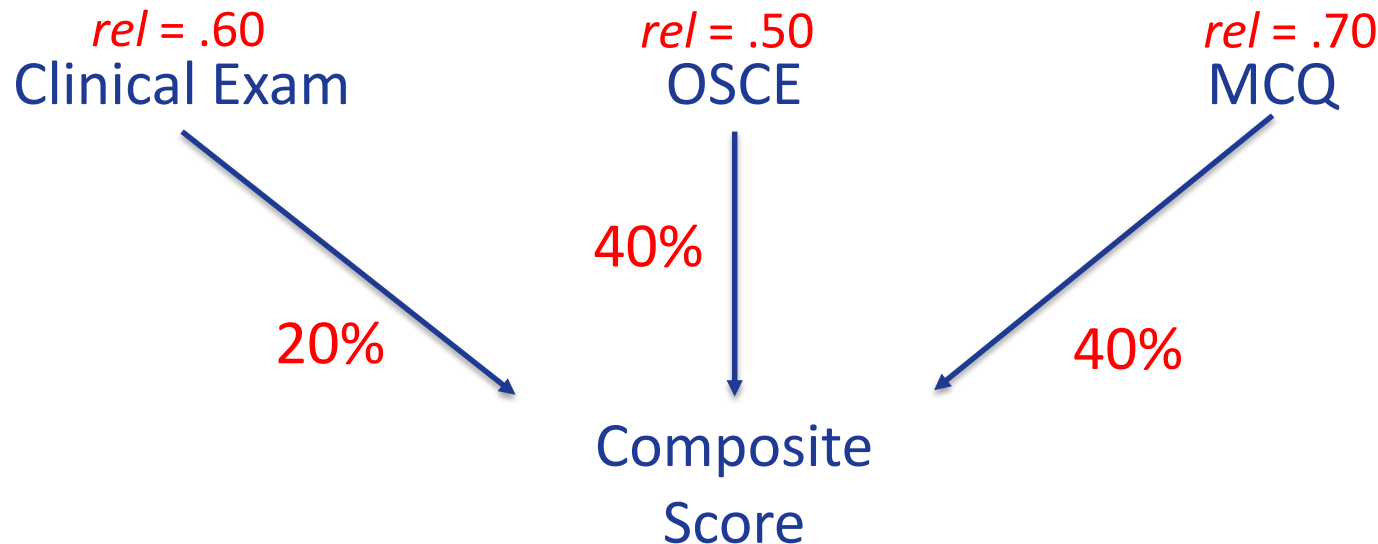
4. Case Specificity

- Difference in student performance by case
 - *generally large* in medical education studies



5. Gathering information from multiple assessments

Example



- Composite score reliability = .75

Implications

- Assessments based on **sampling** and **structure**
 - Consider **case specificity**
 - Identify **clinically discriminating** items
- To increase reliability
 - Increase **number** of items
 - Conduct **item analysis**
- Maximize **variability** of learners
- Developing an **assessment system**

Top Ten Tips to Improve Your Assessment Program

David Li (Li Li), MD, Ph.D.

Paul E. Phrampus, MD FSSH

Tip 1

- Engage Psychometricians Early



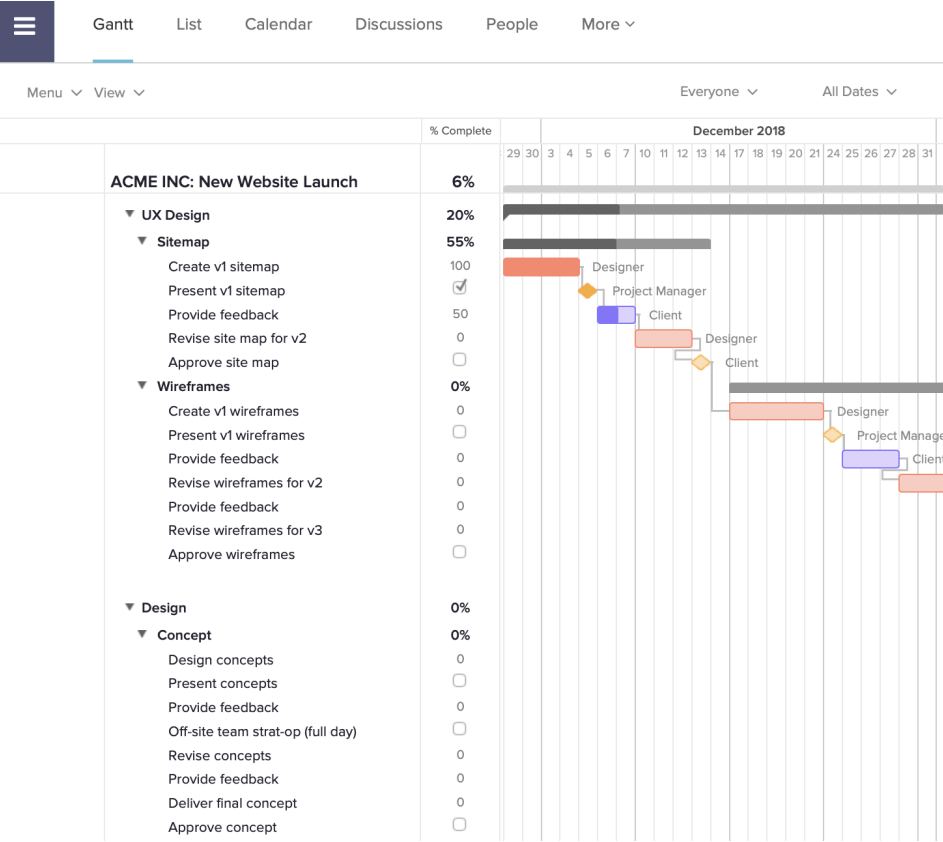
Tip 2

- Develop a Team Interested in Assessment



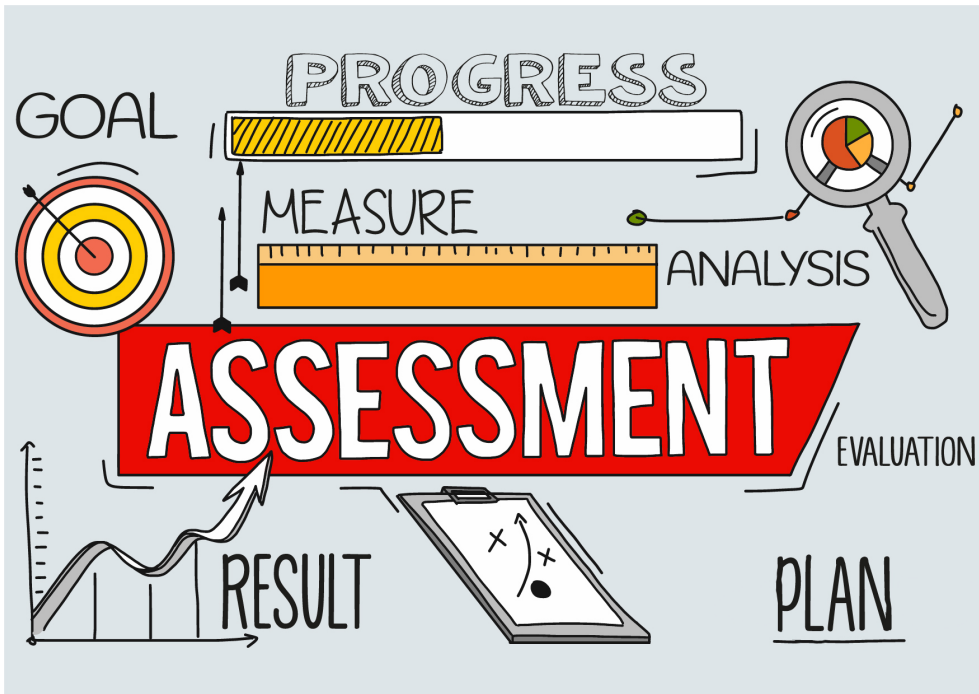
Tip 3

- Attention to Detail Planning



Tip 4

- Carefully Plan Assessment Objectives
 - Be careful of Scope Creep



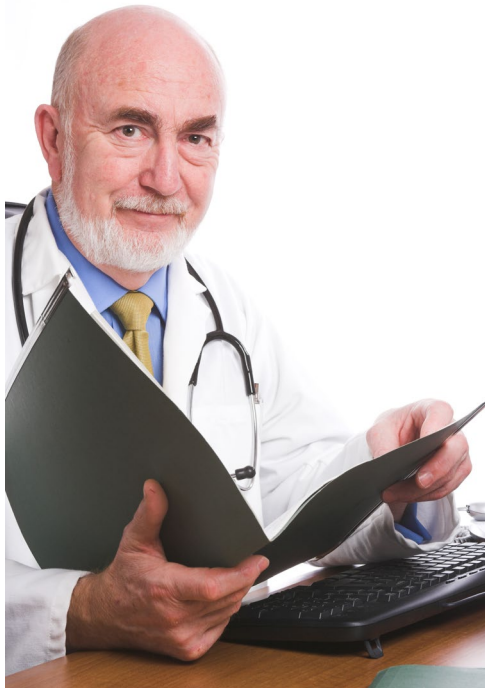
Tip 5

- Budget Time for Rater Training
 - Mock Rating Exams



Tip 6

- Don't Rely Soley on Rater Expertise
 - Difficuilt to Control Bias



Tip 7

- Plan for Equipment/Data Collection Failures
 - Have Back Up Plans



Tip 8

- Acquire Appropriate Technology



Tip 9

- Test Your Rating Tool(s)



Tip 10

- Evaluate Feasibility
 - Overall Investment
 - Costs
 - The Assessment Stakes



THANK YOU!



SIMULATION:
BRINGING LEARNING TO LIFE

#IMSH2021