PROBLEMS WITH CLASSICAL TEST THEORY AND RAW SCORES

Bahrul Hayat, Ph.D. (Chicago, 1992)

- Faculty of Psychology, State Islamic University Jakarta
- Faculty of Education, Indonesian International Islamic University

Georg Rasch (1901-1980)

Pioneers of objective and linear
psychological and social measurement

MEASUREMENT

- Is a prerequisite for science development.
- is heuristic to social understanding.
- makes our life easy.

PROBLEMS WITH CLASSICAL TEST THEORY

Sample dependent	Test dependent	Test reliability
Comparability of score	Measurement error	Test design

PROBLEMS WITH CLASSICAL TEST THEORY

Sample dependent

- Item statistics such as item difficulty and item discrimination depend on the particular examinee samples in which they are obtained.
- Test score reliability is directly related to test score variability of the sample.

PROBLEMS WITH CLASSICAL TEST THEORY

Test dependent

- Score obtained is dependent upon the test used
- Interpretation of test score is very test dependent

PROBLEMS WITH CLASSICAL TEST THEORY

Test reliability

- Test reliability is defined in terms of parallel forms which is difficult to achieve in practice.
- The reliability coefficient is affected by homogeneity of items and heterogeneity of sample.

PROBLEMS WITH CLASSICAL TEST THEORY

Comparability of score

- Comparisons of examinees on an ability measured by a test are limited to situations in which examinees are administered the same (or parallel) test items.
- Test equating is problematic under CTT

PROBLEMS WITH CLASSICAL TEST THEORY

Measurement error

- The errors of measurement is the same for all examinees.
- Many achievement and aptitude tests are (typically) designed for middle-ability students and the tests do not provide very precise estimates of ability for either high- or low-ability examinees.

PROBLEMS WITH CLASSICAL TEST THEORY

Test design

- No basis for individual-adaptive testing
- No basis for effective and efficient design of criterion-reference testing

MEASUREMENT

"is the assignment of numbers to objects or events according to rules"

(S.S. Steven, 1946)

Number	Object/Event	Rules

MEASUREMENT INSTRUMENT

We need an instrument (test or non-test)
that gives us:
objective, valid, and reliable measure.

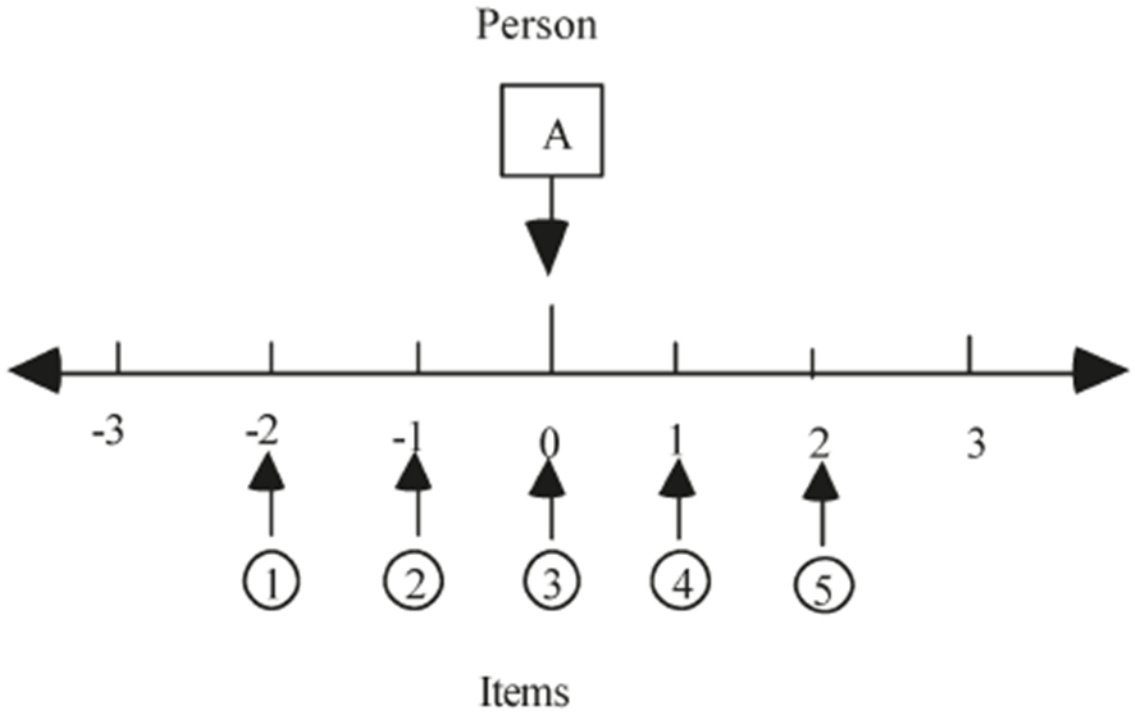
MEASUREMENT INSTRUMENT

We need an instrument (test or non-test)
that gives us:
objective, valid, and reliable measure.

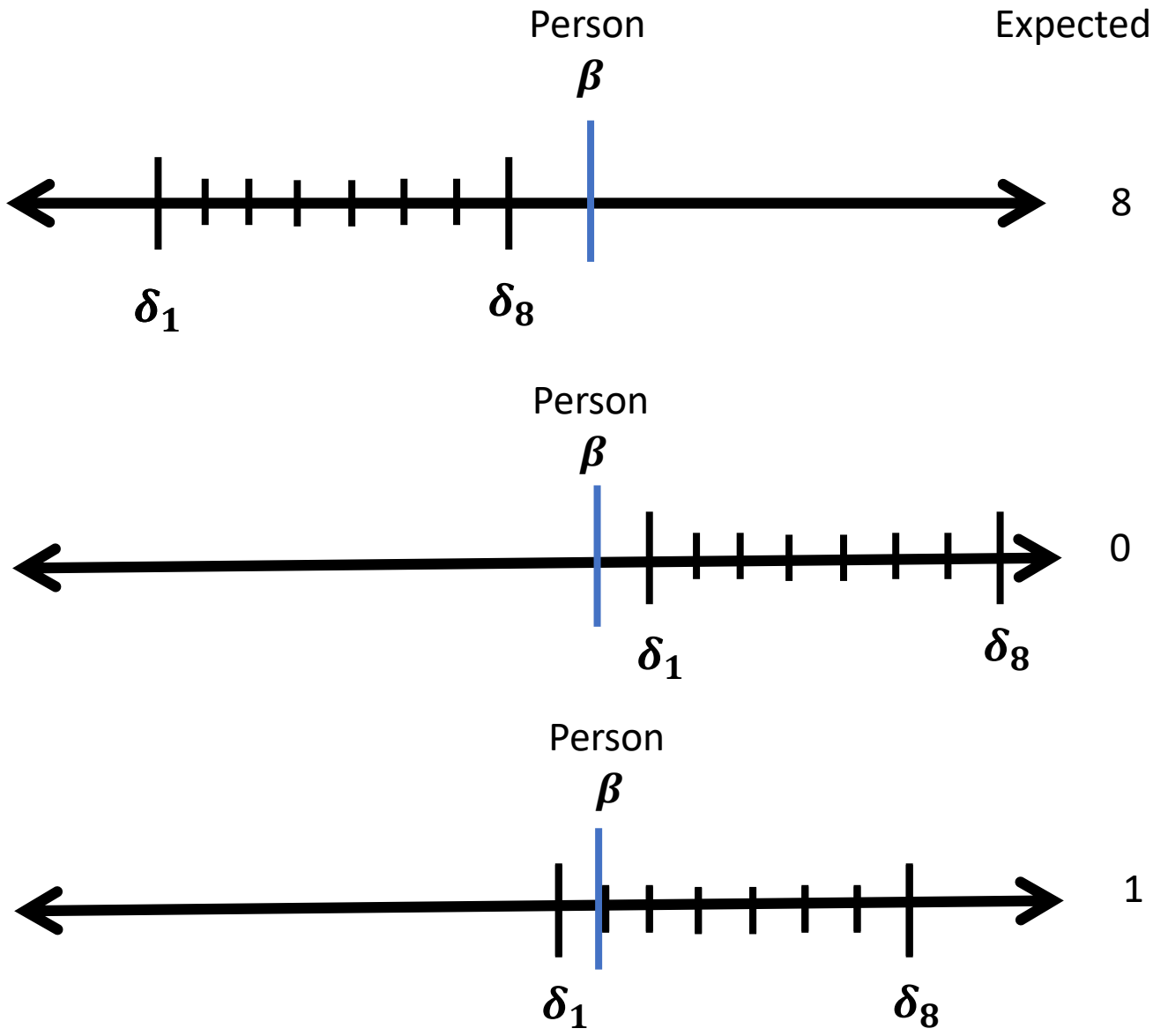
RAW SCORES

- Most standardized tests are initially scored with raw scores.
- Raw scores are simply the number questions or problems the student answered or solved correctly.

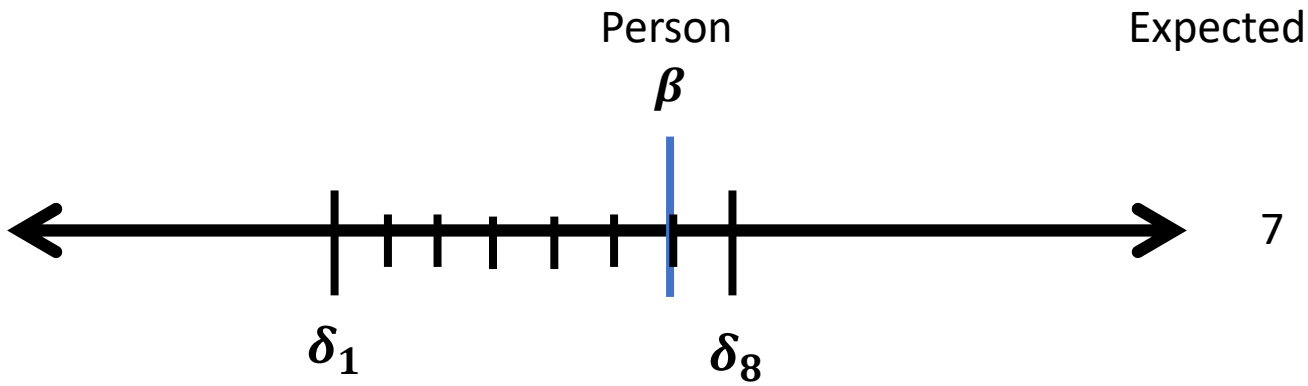
How Tests Are Used To Measure

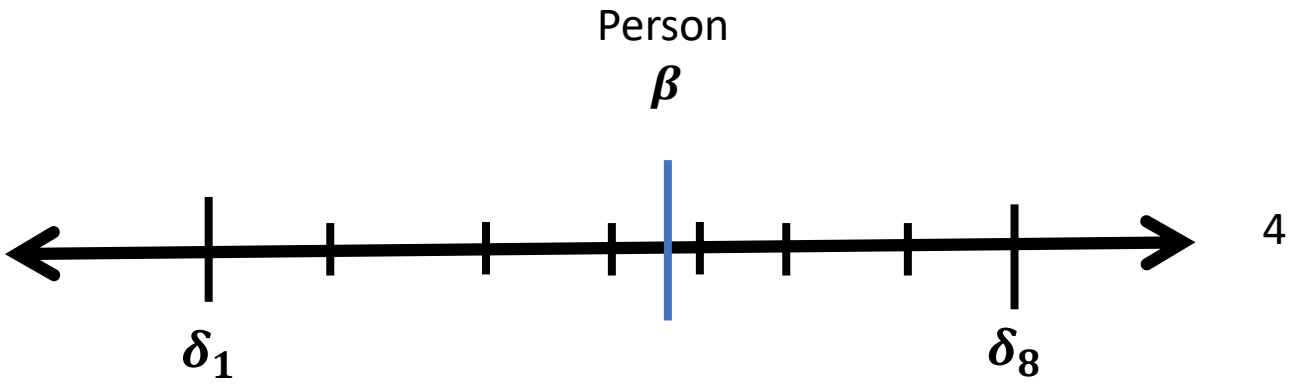


Problems with raw scores



Problems with raw scores





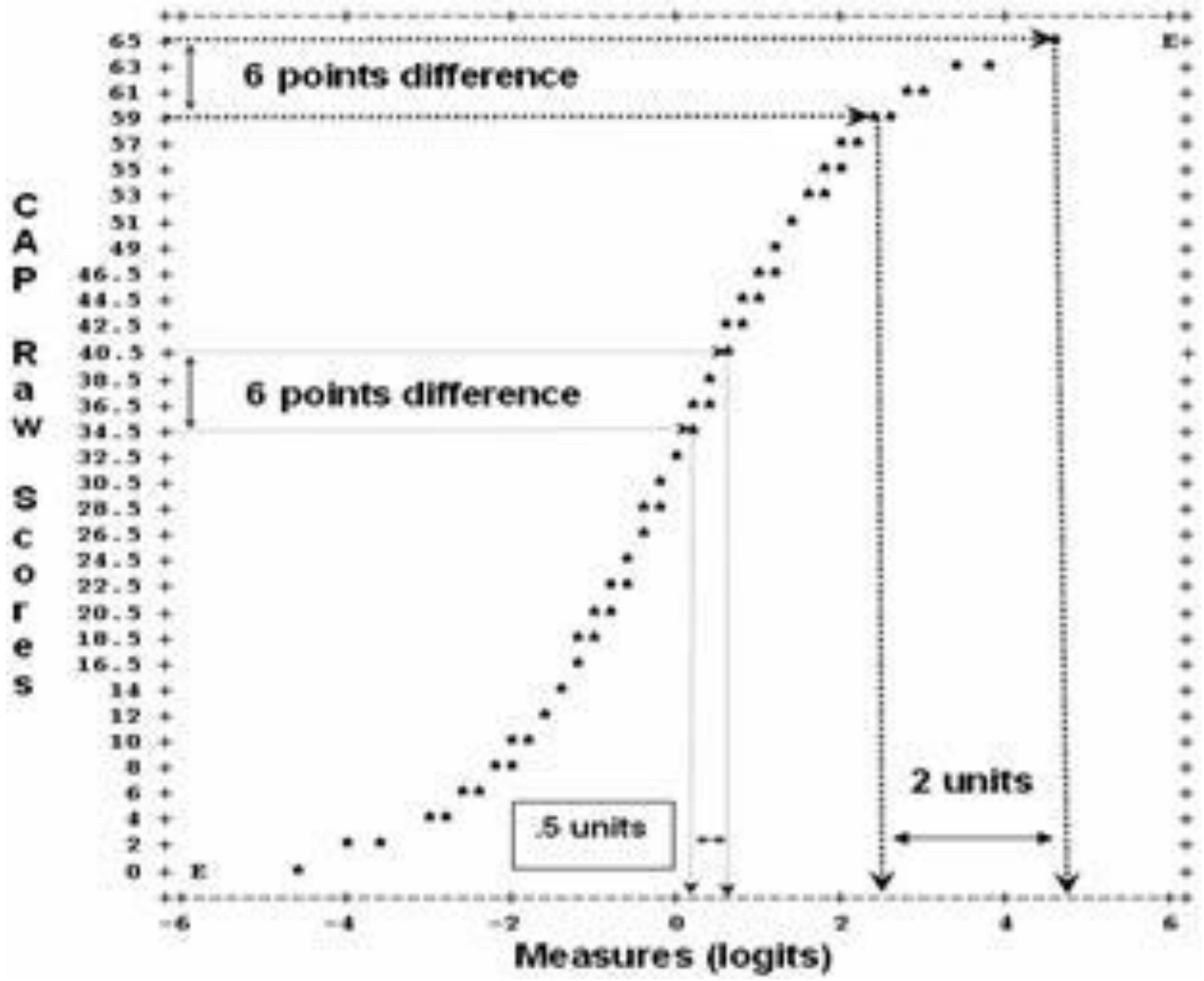
PROBLEMS WITH RAW SCORES

Test raw scores depend upon:

- *level*
- *spacing*
- *distribution*

of item difficulty of the test

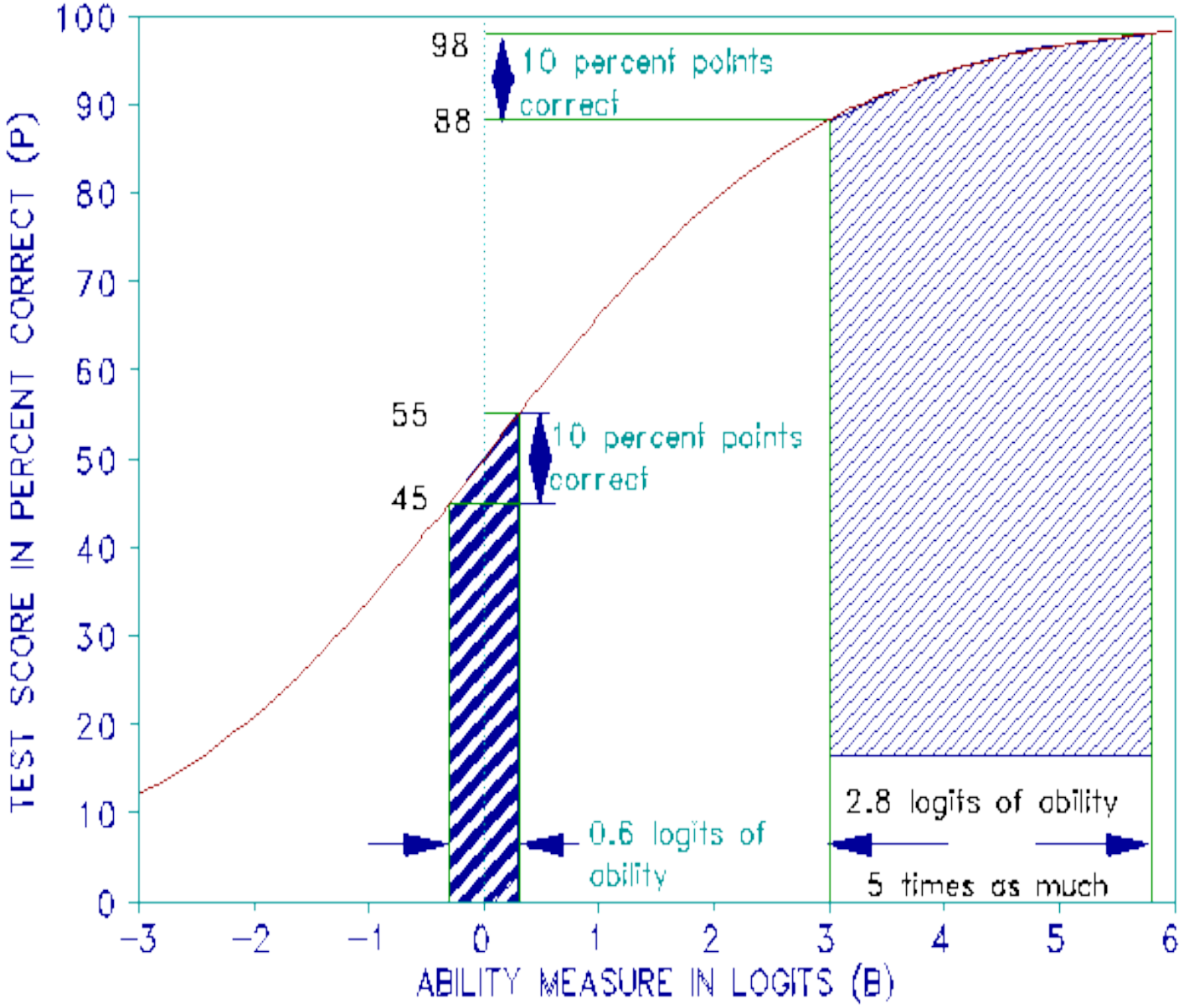
RAW SCORES ARE NOT LINEAR MEASURES



RAW SCORES ARE NOT LINEAR MEASURES

- The figure shows initial transformation of Child Assessment Profile (CAP) raw scores to linear logit measures, and, as expected, upper and lower tails show substantial raw score distortions. A six point raw score difference in upper tail is four times greater when represented with logits.
- The data are 25,000 CAP Chicago Public Schools, 1993-2002) raw score records that were collected by Chicago preschool teachers.

EXTREME RAW SCORES ARE BIASED AGAINST MEASURES



SCORE COMPARABILITY ACROSS TEST FORMS

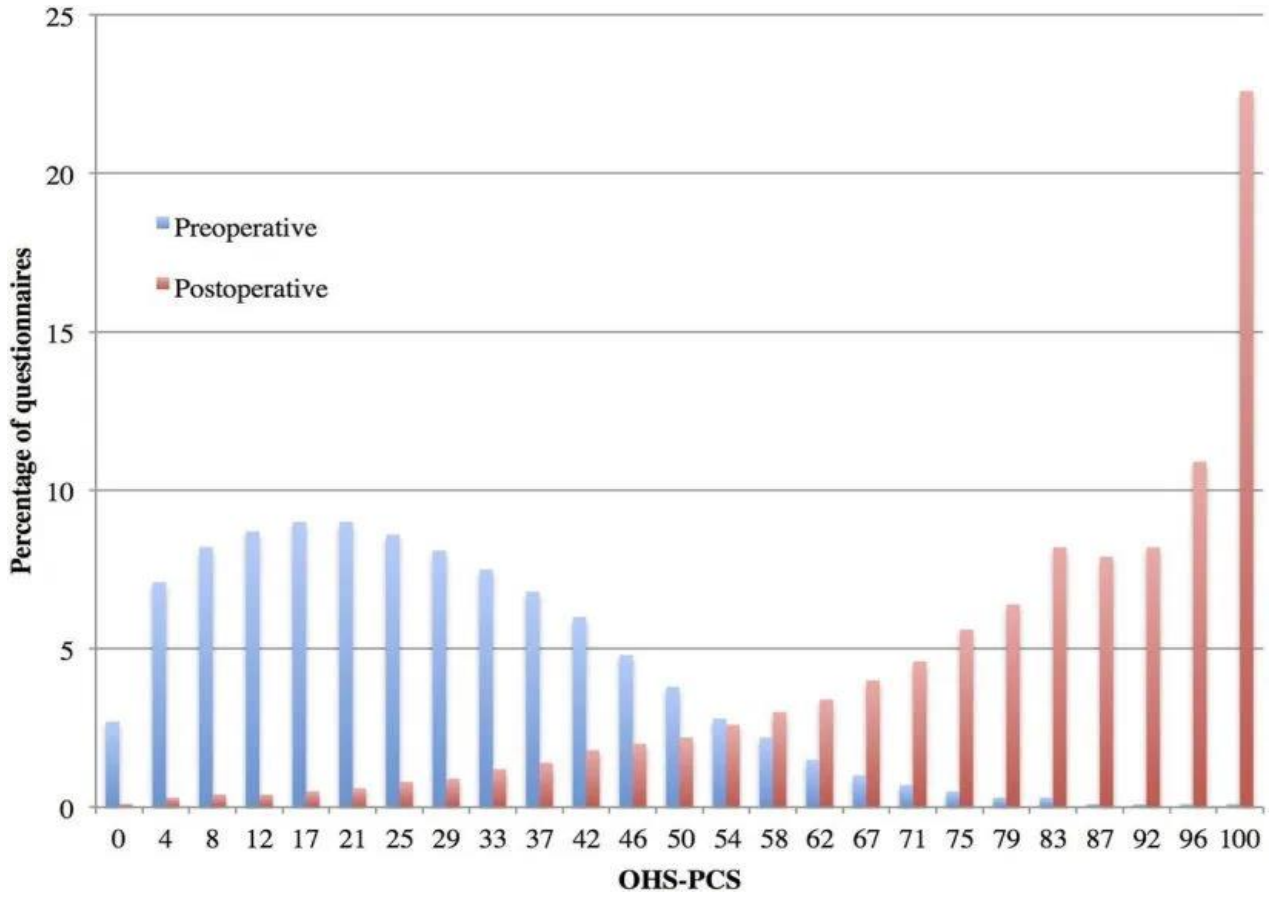
"For an easier form, a test taker needs to answer slightly more questions correctly to get a particular scaled score. For a more difficult form, a test taker can get the same scaled score answering slightly fewer questions correctly."

Table 1. Scaled Scores for Form A and Form B

Raw Score	Scaled Score	
	Form A	Form B
100	200	200
99	200	199
98	199	197
97	197	195
96	195	194
95	194	192
Etc.	Etc.	Etc.

Table 1 shows an example of scaled scores associated with different raw scores for two different forms. In this hypothetical example, Form A is the more difficult form. To achieve the same scaled score of 195, a test taker needs to answer 96 out of the 100 questions correctly on Form A, but needs to answer 97 questions correctly on Form B.

RAW SCORES ARE CONSTRAINED BY FLOOR AND CEILING EFFECTS



Ceiling effects and floor effects both limit the range of obtained scores from a test.

RAW SCORES LACK INTERPRETIVE MEANING

A student receives a raw score of 52. Without more information, the raw score has no meaning. If the test consisted of 55 questions, a raw score of 52 would be a superior score. Alternatively, if the test had 112 questions, a raw score of 52 would be a below-average score.

RAW SCORES LACK INTERPRETIVE MEANING

- No link between obtained score from a test with the trait measured by the test.
- No basis to measure growth or progress of the construct of interest

MEASURING IS NOT COUNTING

A

B

The end