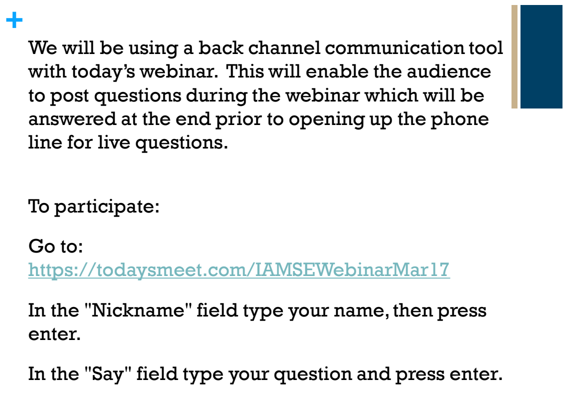


We will be using a back channel communication tool with today's webinar. This will enable the audience to post questions during the webinar which will be answered at the end prior to opening up the phone line for live questions.

To participate:

Go to: https://todaysmeet.com/IAMSEWebinarMar17

In the "Nickname" field type your name, then press enter.

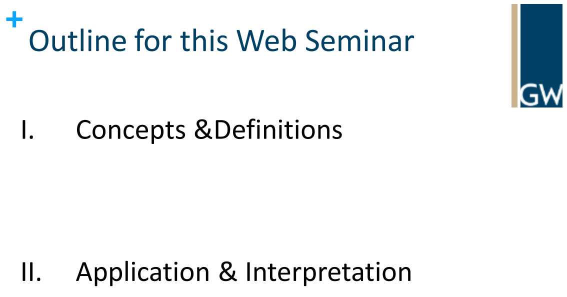
In the "Say" field type your question and press enter.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

IAMSE Web Seminar

Testing your Test: Assessing the Quality of Test Items

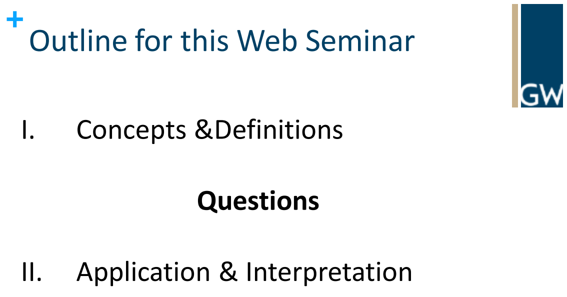
Veronica Michaelsen, MD, PhD
George Washington University
School of Medicine and Health Sciences



Outline for this Web Seminar

I. Concepts &Definitions

II. Application & Interpretation

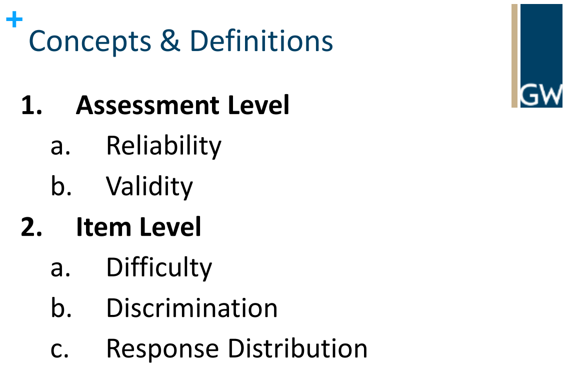


Outline for this Web Seminar

I. Concepts &Definitions

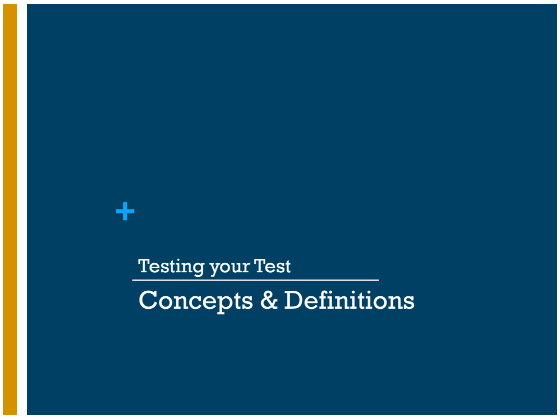
Questions

II. Application & Interpretation



Concepts & Definitions

1. **Assessment Level**
   a. Reliability
   b. Validity
2. **Item Level**
   a. Difficulty
   b. Discrimination
   c. Response Distribution



Testing your Test

Concepts & Definitions

Reliability - Defined

The degree to which an assessment tool produces stable and consistent results.

Reliability - Types

1. Test-Retest
2. Split Half
3. Alternate (Parallel) Forms
4. Internal Consistency

Validity - Defined

How well an assessment measures what it is purported to measure.

Validity - Types

1. Face
2. Construct
3. Predictive
4. Concurrent
5. Convergent

Item Difficulty

The percentage of students who answered an item correctly.

Item Discrimination

The ability to which an item differentiates between high and low performing test-takers.

Response Distribution

The distribution of students selecting each response option for a given item.

Testing your Test

Questions

Testing your Test

Applications & Interpretation

Assumptions

1. Summative Assessments
2. Individual Assessments
3. Selected Response Items

Reliability - Types

1. Test-Retest
2. Split Half
3. Alternate (Parallel) Forms
4. Internal Consistency

Reliability - Types

1. ~~Test-Retest~~
2. ~~Split Half~~
3. ~~Alternate (Parallel) Forms~~
4. Internal Consistency

Reliability Measures

1. KR-20
   a. Dichotomous Variables only
2. Chronbach's Alpha
   a. Continuous or Dichotomous

Reliability Measures

If each item on an assessment has **only one correct answer** <u>and</u> each item is worth the **same number of points**, Chronbach's alpha and KR-20 will be identical.

Reliability Measures

1. Can be impacted by:
   a. Score variance
   b. Length of assessment
   c. Overall difficulty
2. Range from 0 – 1.00

Reliability Measures

> 0.90	Level of standardized tests
0.80-0.90	Very Good
0.70-0.80	Good for instructor designed
0.60-0.70	Somewhat Low, needs revision
0.50-0.60	Significant Revisions Needed
<0.50	Questionable

Reliability Measures

> 0.90	Level of standardized tests
0.80-0.90	Very Good
0.70-0.80	Good for instructor designed
0.60-0.70	Somewhat Low, needs revision
0.50-0.60	Significant Revisions Needed
<0.50	Questionable

REMEMBER:

1. Reliability can be impacted by:
   a. Score variance
   b. Length of assessment
   c. Overall difficulty

+ REMEMBER:

Homogeneity of Learners

1. Reliability can be impacted by:
   a. Score variance
   b. Length of assessment
   c. Overall difficulty

+ REMEMBER:

Homogeneity of Learners

1. Reliability can be impacted by:
   a. Score variance
   b. Length of assessment
   c. Overall difficulty

Quiz vs. Exam

+ REMEMBER:

Homogeneity of Learners

1. Reliability can be impacted by:
   a. Score variance
   b. Length of assessment
   c. Overall difficulty

Quiz vs. Exam

Formative vs. Summative

+ Difficulty

1. Most instructor-designed exams will see mean difficulty of .75-.85
2. Too high risks inadequate preparation for qualifying examinations

+ Difficulty

Should fall between .3-.9
Ideal is ~.63

Exception is Mastery Items!

+ Discrimination

The ability to which an item differentiates between high and low performing test-takers.

Discrimination

The ability to which an item differentiates between high and low performing test-takers.

High performers are top 27%
Low performers are bottom 27%

Discrimination - Measures

1. Discrimination Index (DI)

   $\mathrm{DI} + \%C_h\text{-}\%C_l$

2. Point Biserial Correlation Coefficient (PBCC)

   Considers variance across all students.

Discrimination (DI and PBCC)

Range: -1.00 - +1.00

Generally:

<0.20 needs to be reviewed
>0.40 is good discrimination

Keep goals of assessment in mind!

Response Distribution

1. Review distribution of responses selected for each item.
2. Also note if distribution is different for high and low performing students.

Response Distribution

1. Remove distractors with <5%
2. Choose Quality over Quantity
3. All alternatives should be plausible.

Handling poor-performing items

1. Double-Key
2. Delete ("throw out")
3. Nullify

Cheat Sheet

KR-20	>0.70
Difficulty	0.3-0.9
DI	>0.25
PBCC	>0.20

Cheat Sheet

KR-20	**>0.70**
Difficulty	**0.3-0.9**
DI	**>0.25**
PBCC	**>0.20**

Testing your Test

Questions