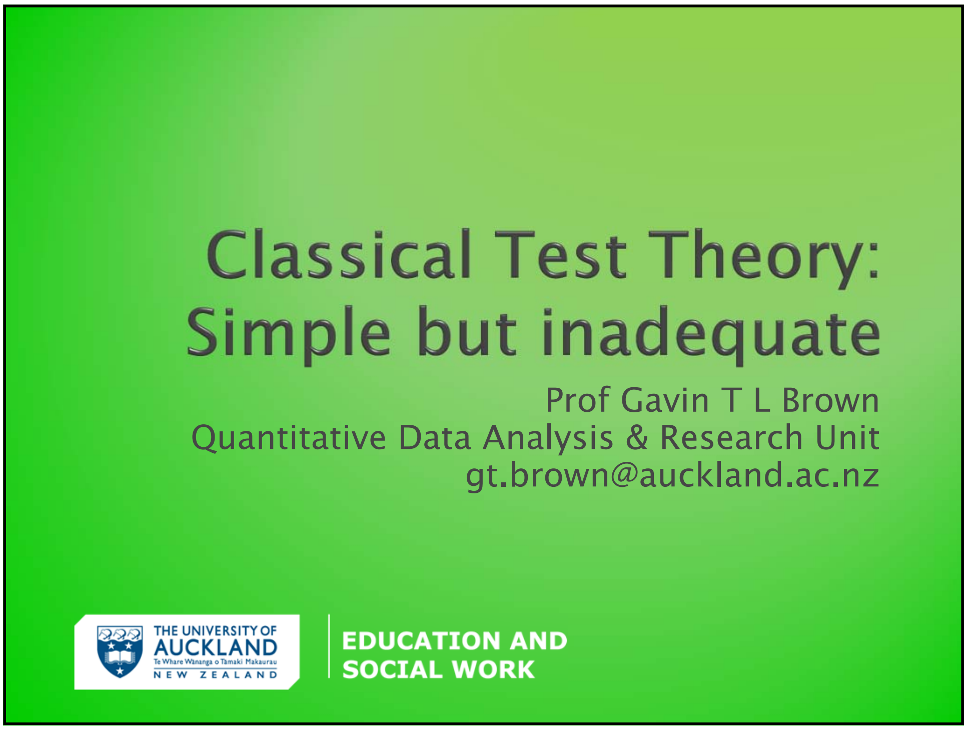
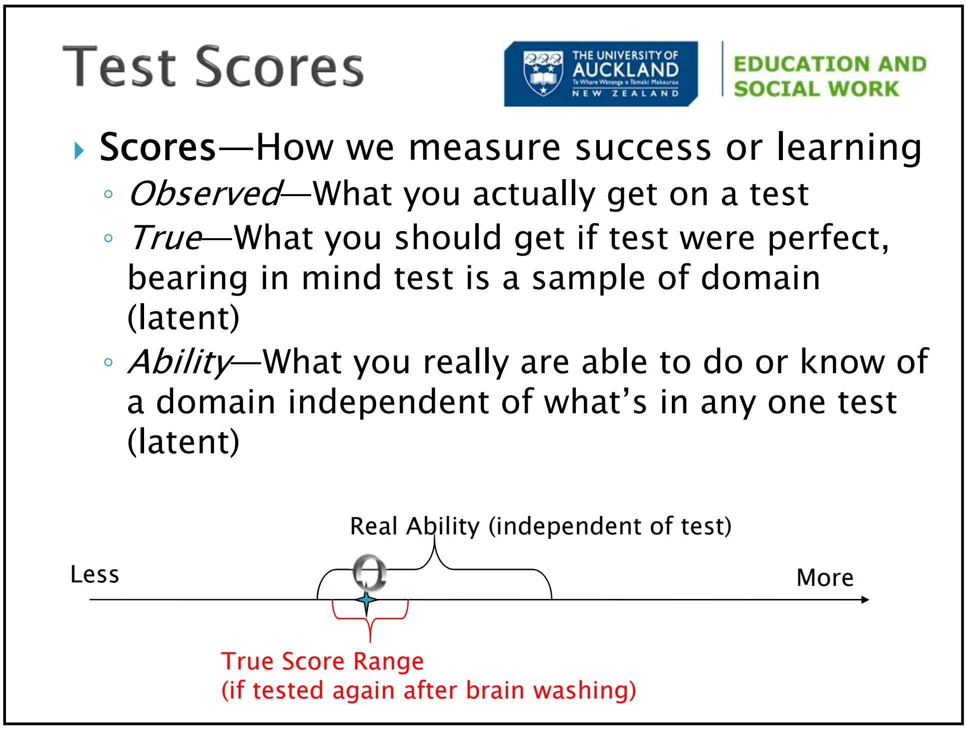


Classical Test Theory: Simple but inadequate

Prof Gavin T L Brown
Quantitative Data Analysis & Research Unit
gt.brown@auckland.ac.nz

EDUCATION AND SOCIAL WORK



Test Scores

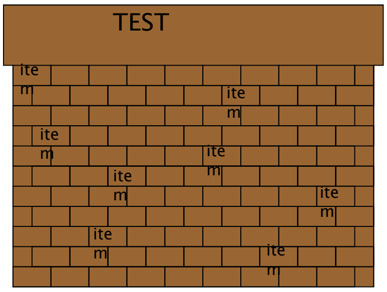
EDUCATION AND SOCIAL WORK

- Scores—How we measure success or learning
  - *Observed*—What you actually get on a test
  - *True*—What you should get if test were perfect, bearing in mind test is a sample of domain (latent)
  - *Ability*—What you really are able to do or know of a domain independent of what's in any one test (latent)

Real Ability (independent of test)
Less
More
True Score Range
(if tested again after brain washing)

Core principle

- Observed score = TRUE score + ERROR
  - O = T + e
- Total Score is simply sum of number of items answered correctly
- All items are equivalent
  - Like another brick in the wall



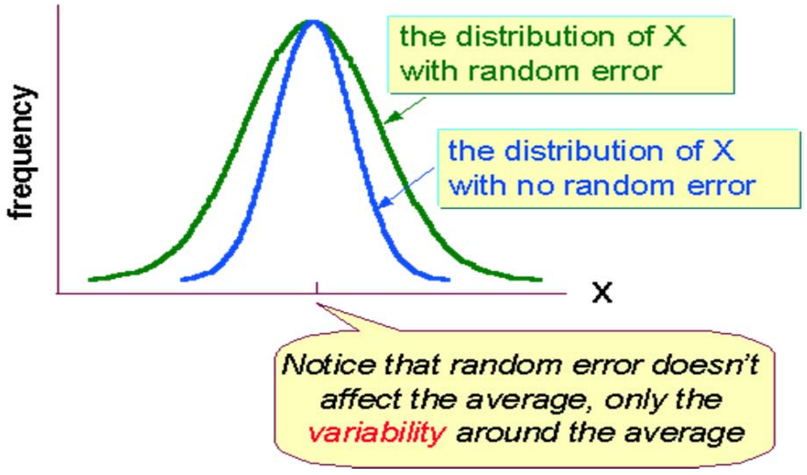
Classical Test Theory

- items only mean something in context of the test they're in
- All items are random sample of domain being tested
- All items have equal weight in making up test statistics
- Error is assumed to be random
  - If not random, then X the measurement is **Biased**
  - $O=T+e_{random}+e_{systematic}$
  - Accept random but try to minimise it
  - but remove systematic

Random Error

- Random error means that
  - Errors will sometimes be positive, sometimes negative
    - tend to cancel out when we add up a person’s score
  - Errors will not be correlated with other things
    - Σ e = 0
    - Thus, test score correlations depend on the true components – not error
    - E(X) = T
  - Thus the higher the proportion of *t* in *X* the higher the correlations will be between items
    - The more items correlate with each other the less disturbance

Effect of random error

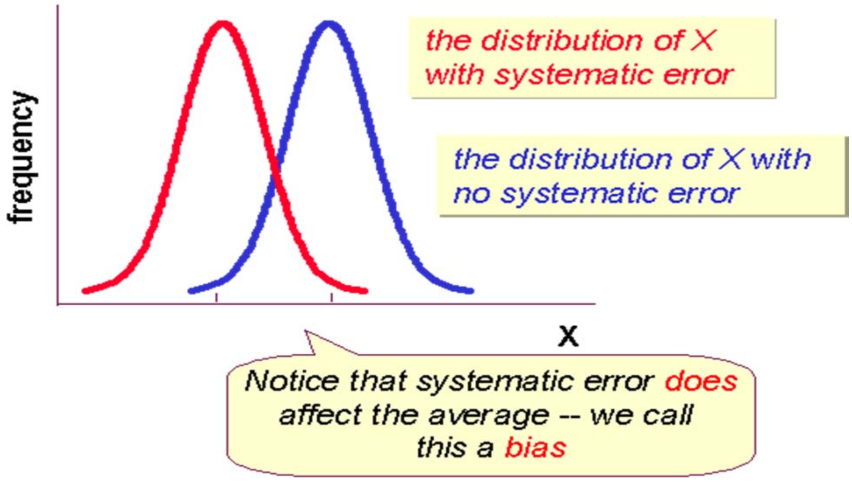


the distribution of X with random error

the distribution of X with no random error

Notice that random error doesn’t affect the average, only the variability around the average

Effect of systematic error



the distribution of X with systematic error

the distribution of X with no systematic error

Notice that systematic error **does** affect the average -- we call this a **bias**

Basic Properties of Items

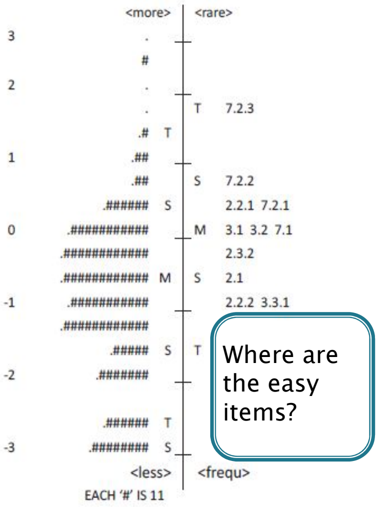
- Core total test statistics are:
  - **DIFFICULTY**: the average test score (mean) **DISCRIMINATION**: Who gets the items correct? The spread of scores (standard deviation)
  - **RELIABILITY**: how small is the error?
- All statistics for persons and items are sample dependent
  - Requires robust representative sampling (expensive, time consuming, difficult)
  - Classrooms are not large or representative; schools might be

Item Difficulty

- Not about the complexity or obscurity of the item
- Nor does it relate to an individual's subjective reaction
- Derived from the responses to an item
- Item Difficulty: % answer correct or wrong
  - How hard is the item?
  - Mean correct across people is *p*
  - Usually delete items too easy ($p>.9$) or too hard ($p<.1$) for generalised ability test

Optimal Item Difficulty

- Don't want all items to have a p = .50
- Need to spread items out to measure the full range of the trait
- Accuracy in score determination requires enough information for each person's ability



Where are the easy items?

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

Item Discrimination r_{pb}

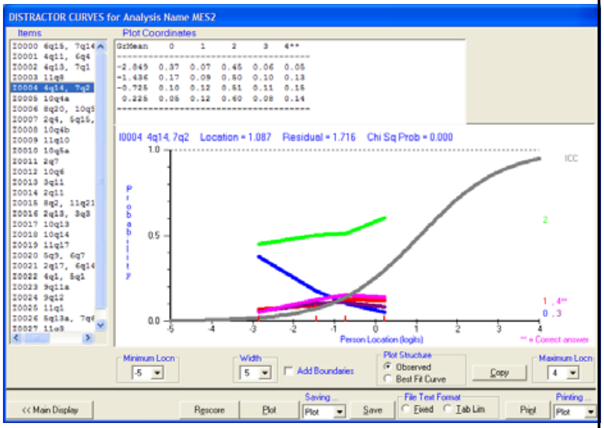
- Who gets the item right?
  - Correlation between item and total score, person by person – expect best students to get items correct, and least able to get it wrong
  - Are the distractors working properly?
  - Look for values > .20
  - Beware negative or zero discrimination items

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

Low discrimination

- Almost everyone chooses the wrong answer



DISTRACTOR CURVES for Analysis Name MES2

Items

```
I0000 6q15, 7q14
I0001 4q11, 6q4
I0002 4q13, 7q1
I0003 11q8
I0004 4q14, 7q2
I0005 10q4a
I0006 8q20, 10q5
I0007 2q4, 5q15,
I0008 10q4b
I0009 11q10
I0010 10q5a
I0011 2q7
I0012 10q6
I0013 3q11
I0014 2q11
I0015 8q2, 11q21
I0016 2q13, 3q8
I0017 10q13
I0018 10q14
I0019 11q17
I0020 5q9, 6q7
I0021 2q17, 6q14
I0022 4q1, 5q1
I0023 9q11a
I0024 9q12
I0025 11q1
I0026 5q13a, 7q6
I0027 11o3
```

Plot Coordinates

```
GrMean    0     1     2     3     4**
-----------------------------------------
-2.849  0.37  0.07  0.45  0.06  0.05
-1.436  0.17  0.09  0.50  0.10  0.13
-0.725  0.10  0.12  0.51  0.11  0.15
 0.225  0.05  0.12  0.60  0.08  0.14
-----------------------------------------
```

I0004 4q14, 7q2 Location = 1.087 Residual = 1.716 Chi Sq Prob = 0.000

Minimum Locn: -5
Width: 5
□ Add Boundaries
Plot Structure: ● Observed ○ Best Fit Curve
Maximum Locn: 4
Saving ...: Plot
File Text Format: ○ Fixed ○ Tab Lim
Printing ...: Plot

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

Correlational Indexes of Discrimination

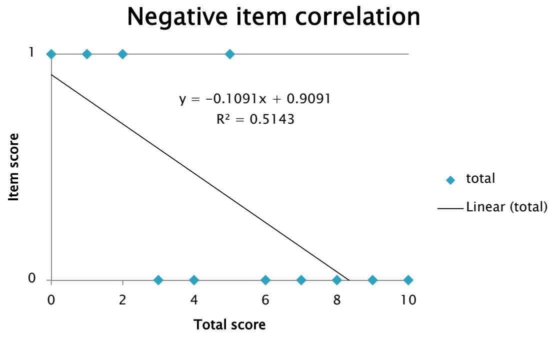
- Item to total correlations
- Point-biserial – dichotomous and continuous variable
  - The correlation of the item to the total without the item in the total

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

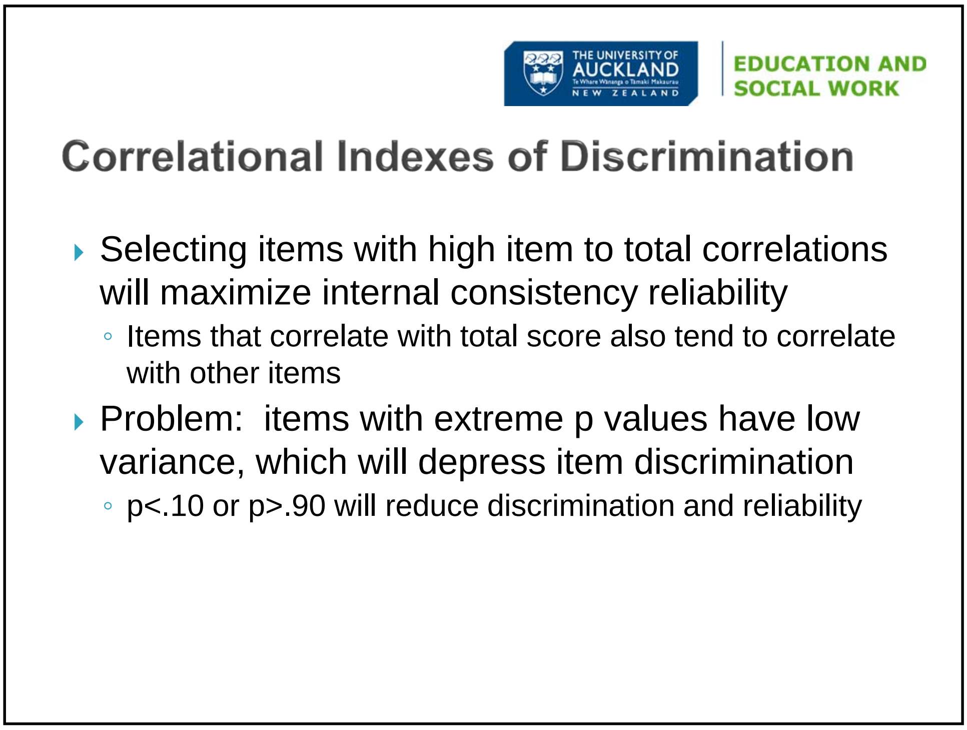
EDUCATION AND SOCIAL WORK

Negative item discrimination

item	total
1	0
1	1
1	2
0	3
0	4
1	5
0	6
0	7
0	8
0	9
0	10

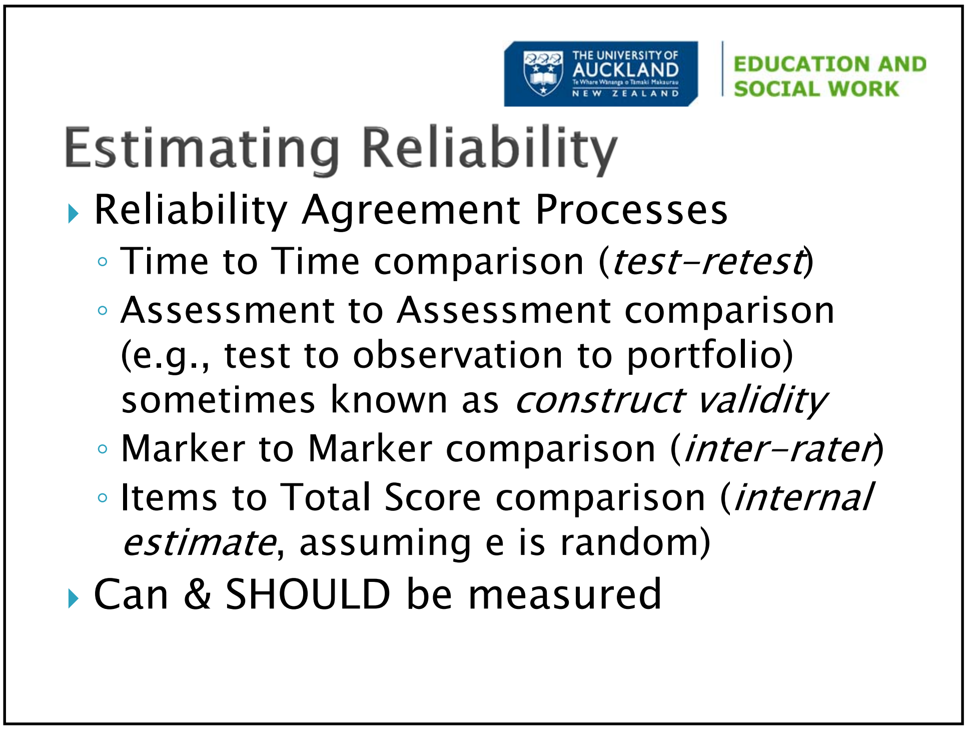


What does it mean if low scoring students do better on an item than high scoring students?



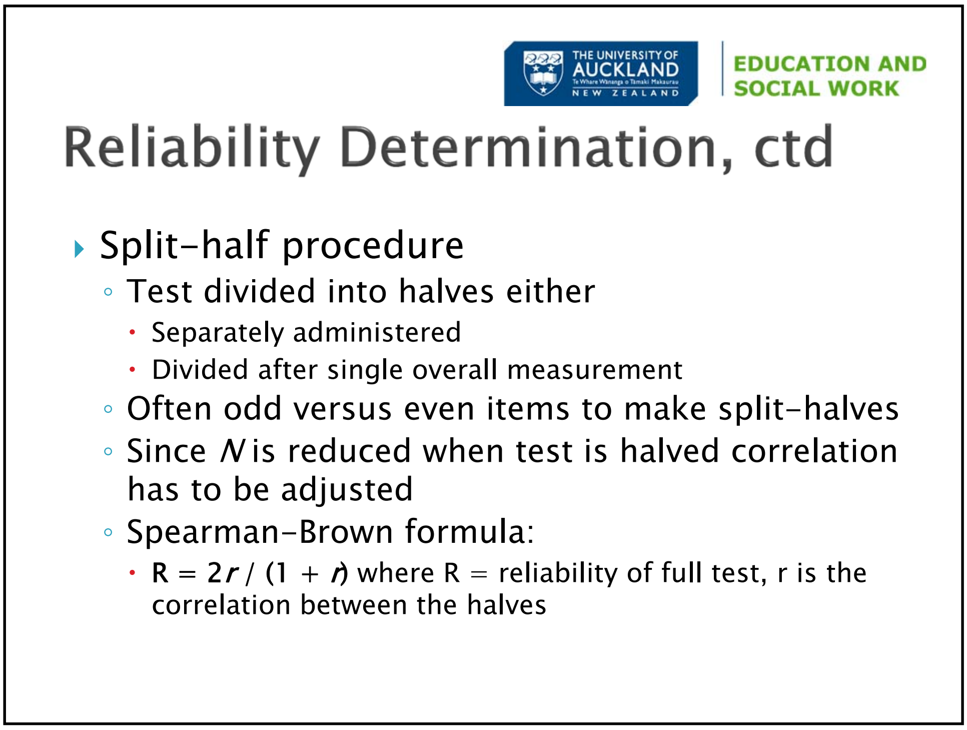
Correlational Indexes of Discrimination

- Selecting items with high item to total correlations will maximize internal consistency reliability
  - Items that correlate with total score also tend to correlate with other items
- Problem: items with extreme p values have low variance, which will depress item discrimination
  - p<.10 or p>.90 will reduce discrimination and reliability



Estimating Reliability

- Reliability Agreement Processes
  - Time to Time comparison (*test-retest*)
  - Assessment to Assessment comparison (e.g., test to observation to portfolio) sometimes known as *construct validity*
  - Marker to Marker comparison (*inter-rater*)
  - Items to Total Score comparison (*internal estimate*, assuming e is random)
- Can & SHOULD be measured

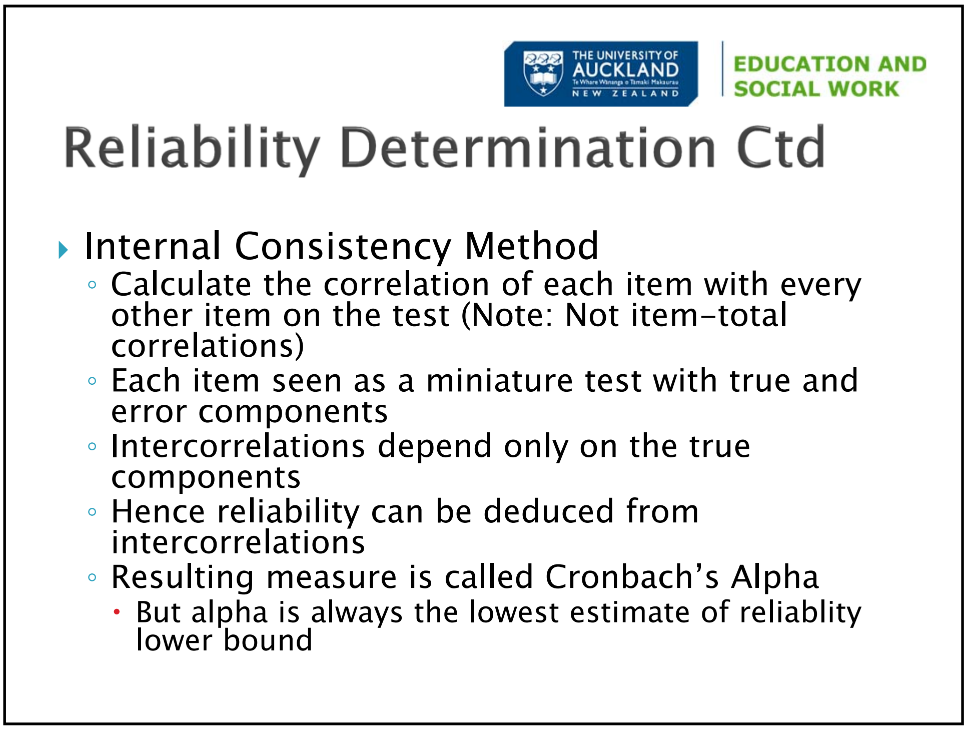


THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

Reliability Determination, ctd

- Split-half procedure
  - Test divided into halves either
    - Separately administered
    - Divided after single overall measurement
  - Often odd versus even items to make split-halves
  - Since *N* is reduced when test is halved correlation has to be adjusted
  - Spearman-Brown formula:
    - $R = 2r / (1 + r)$ where R = reliability of full test, r is the correlation between the halves

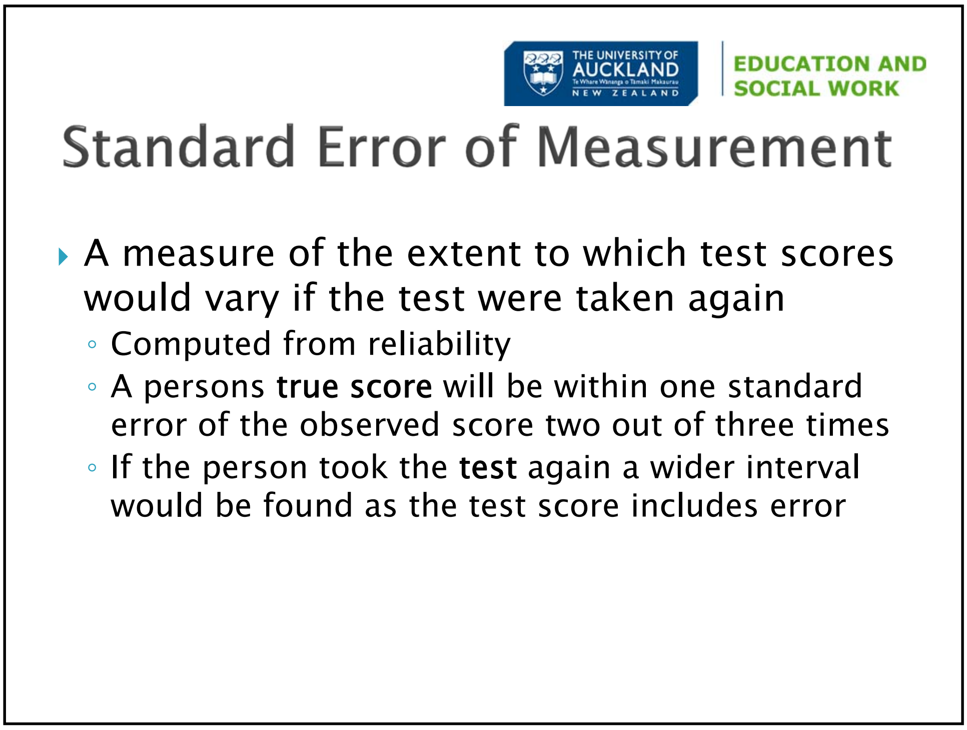


THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

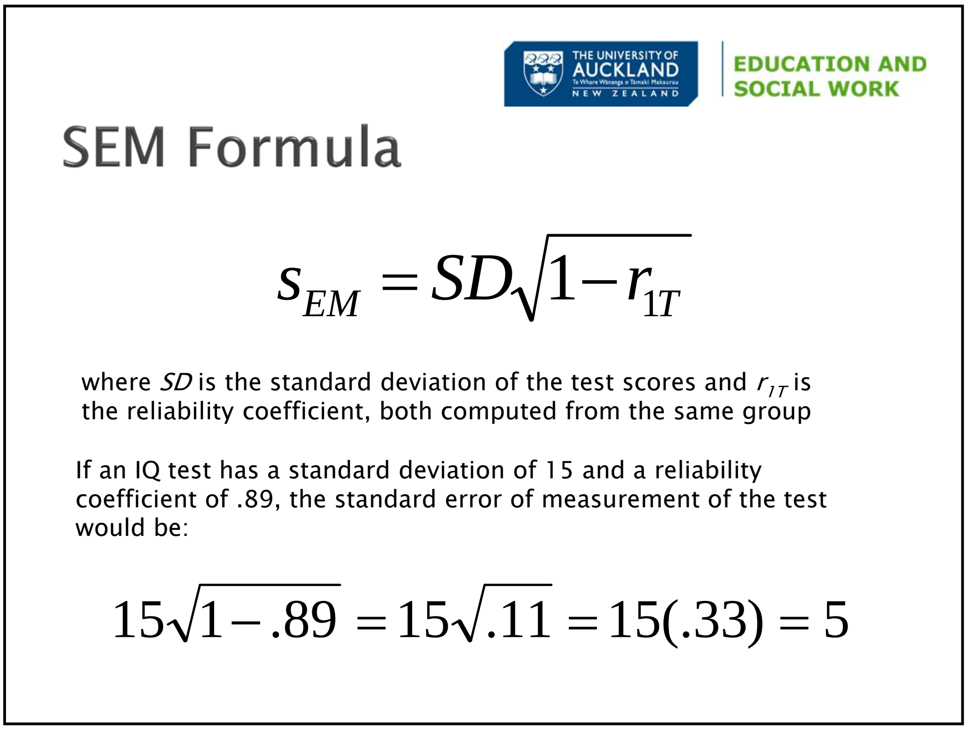
Reliability Determination Ctd

- Internal Consistency Method
  - Calculate the correlation of each item with every other item on the test (Note: Not item-total correlations)
  - Each item seen as a miniature test with true and error components
  - Intercorrelations depend only on the true components
  - Hence reliability can be deduced from intercorrelations
  - Resulting measure is called Cronbach's Alpha
    - But alpha is always the lowest estimate of reliablity lower bound



Standard Error of Measurement

- A measure of the extent to which test scores would vary if the test were taken again
  - Computed from reliability
  - A persons **true score** will be within one standard error of the observed score two out of three times
  - If the person took the **test** again a wider interval would be found as the test score includes error



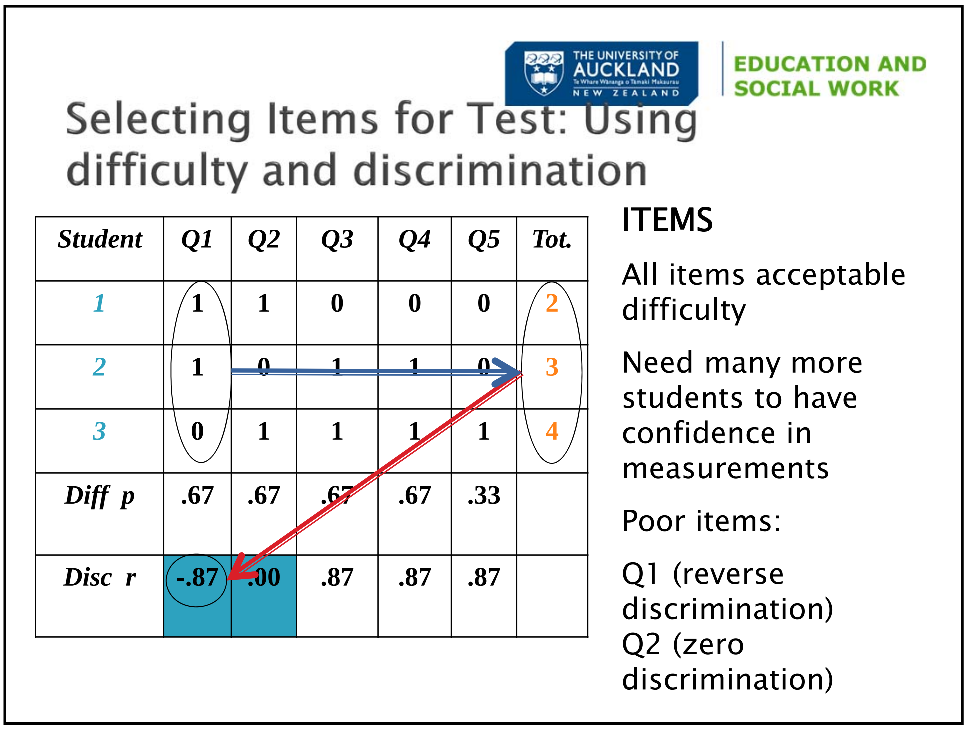
SEM Formula

$$s_{EM} = SD\sqrt{1-r_{1T}}$$

where *SD* is the standard deviation of the test scores and r_{1T} is the reliability coefficient, both computed from the same group

If an IQ test has a standard deviation of 15 and a reliability coefficient of .89, the standard error of measurement of the test would be:

$$15\sqrt{1-.89} = 15\sqrt{.11} = 15(.33) = 5$$



EDUCATION AND SOCIAL WORK

Selecting Items for Test: Using difficulty and discrimination

Student	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Tot.*
1	**1**	**1**	**0**	**0**	**0**	2
2	**1**	**0**	**1**	**1**	**0**	3
3	**0**	**1**	**1**	**1**	**1**	4
Diff p	**.67**	**.67**	**.67**	**.67**	**.33**	
Disc r	**-.87**	**.00**	**.87**	**.87**	**.87**	

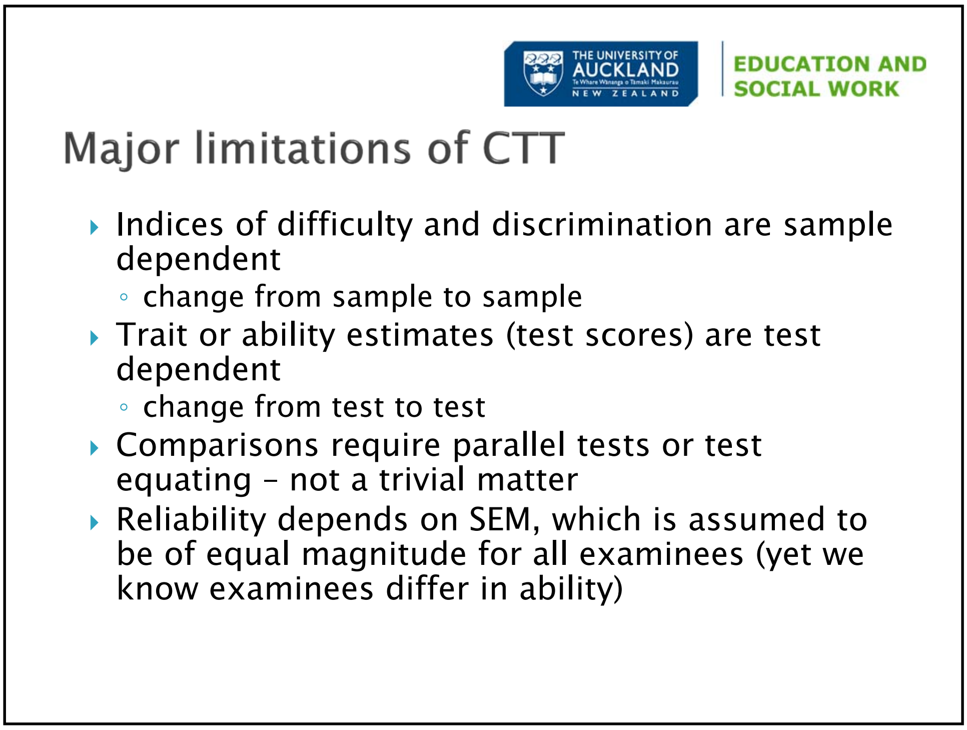
ITEMS

All items acceptable difficulty

Need many more students to have confidence in measurements

Poor items:

Q1 (reverse discrimination)
Q2 (zero discrimination)



EDUCATION AND SOCIAL WORK

Major limitations of CTT

- Indices of difficulty and discrimination are sample dependent
  - change from sample to sample
- Trait or ability estimates (test scores) are test dependent
  - change from test to test
- Comparisons require parallel tests or test equating – not a trivial matter
- Reliability depends on SEM, which is assumed to be of equal magnitude for all examinees (yet we know examinees differ in ability)