Test Theory: Basic Notions

Norman Verhelst
Eurometrics, The Netherlands

EALTA Webinar
October 28 2016

Overview

- Part 1: Classical Test Theory (CTT)
- Answering questions
- Part 2: Item Response Theory (IRT)
- Answering questions

Overview
Classical Test Theory

- Questions, Items, Answers and Scores
- Where is the theory?
- Difficulty and Discrimination
- Reliability
  - What is it?
  - How to measure it?
- Two important formulae

Items and Item scores

- A test consists of a series of
  - Stimulus texts (together with a rubric)
  - One or more **questions** (explicit or implicit) associated with the stimulus
- An item is not the same as a question
  - In principle an item can be defined arbitrarily
- An item **score** is a **number** that can be interpreted as ‘the number of points earned for the given answer’.
- Test score is the sum of the item scores

Examples of items and item scores

- A multiple choice question with alternatives A, B, C and D
  - Possible answers are ‘A’,…,’D’, no choice, multiple choices, one or more choices accompanied by written comments,…
  - Usual scores: ‘1’ for a (unique) correct choice, ‘0’ otherwise
  - Other possibility (1): ‘2’ for a correct answer, ‘1’ for a partial correct answer, ‘0’ otherwise.
  - Other possibility (2): ‘2’ for correct, ‘0’ for incorrect (because…?)

Examples (continued)

- Matching task: ‘Match (4) capitals to (5) countries’
  - How many items represents this task?
  - How would you score the answers?
- Match 4 gaps to 5 [words | sentences | paragraphs]

Iceland	Tallin
Portugal	Bucharest
Roumania	Sofia
Bulgaria	Lisbon
Estonia	

Binary and Partial Credit Items

- Binary: score can take **two** values, usually 0 and 1.
  - Synonym: dichotomous
- Partial Credit: score can take **more than two** values, usually 0, 1,…,*m*
  - *m* can differ from item to item
  - Synonym: polytomous

Examples of partial credit scoring

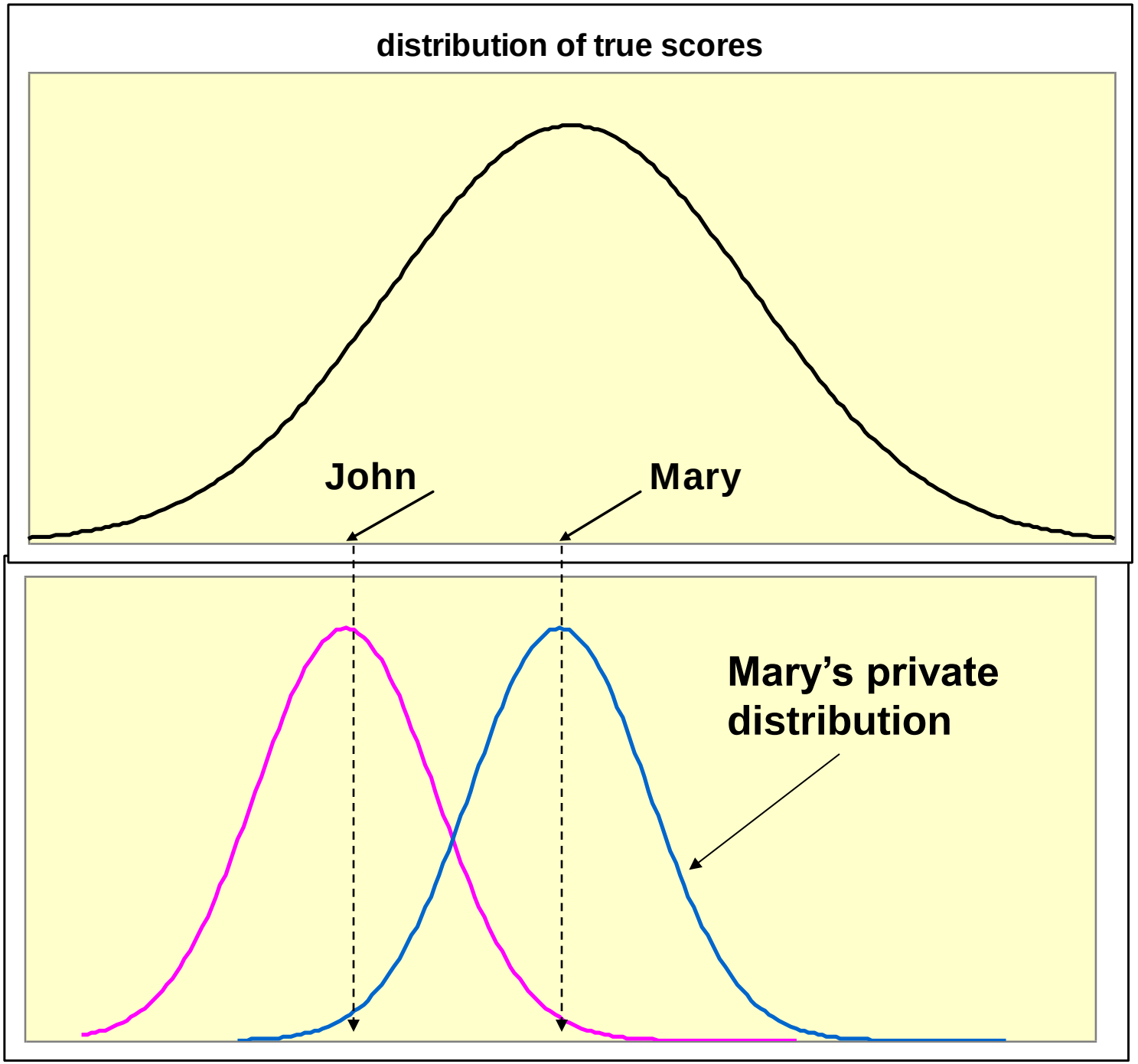
- Matching tasks: score is number of correct matchings
- Text with 5 true-false questions
  - No credits if below chance level
  - 0, 1 or 2 correct → score = 0
  - All correct → score = 2
  - Other cases (3 or 4 correct) → score = 1
- A good scoring rule is a matter of validity

Classical Test Theory

- Fundamental equation: X = T+ E
- X is the observed score
- T is the true score: average or expected score (computed from ‘similar’ replications).
- E = X-T is the measurement error
- But also, and almost never told in textbooks…

Important principle: Independence

- Make sure that a certain answer on one (or more) item(s) **does not imply** a correct or an incorrect answer on another item. (deterministic)
- Make sure that a certain answer on one (or more) item(s) **does not make** a correct or an incorrect answer on another item **more or less likely** (probabilistic).
- Make sure that every item is a **new opportunity** to show one's ability, knowledge, competence…


distribution of true scores
John
Mary
Mary's private distribution

Difficulty and Discrimination

- Index of difficulty: *p*-value (of an item)
  - Proportion correct in the sample
  - Population dependent
  - Sampling dependent
  - High stakes – Low stakes (pretest effect)
- Discrimination
  - There are several indices used: be specific
  - Population and sampling dependent
  - High stakes – Low stakes (???)
  - My favorite: item-test correlation

The reliability of a test score *X*

- Two sources of variation determine the scores: between subjects (Var(T)) and within subjects (Var(E))
- Var(X) = Var(T) + Var(E)

$$\mathrm{Rel}(X) = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} = \frac{\mathrm{Var}(T)}{\mathrm{Var}(T) + \mathrm{Var}(E)}$$

More on the reliability

- Reliability is a proportion ($0 \leq \mathrm{Rel}(X) \leq 1$) of the variance of the true scores relative to the variance of the observed scores.
- The formula is a definition formula
- It cannot be used to compute the reliability

Computing the reliability of the scores

- New concept: parallel test (X')

$$\mathrm{Rel}(X) = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} = \rho(X, X') = \rho^2(X, T)$$

(ρ is the Greek letter rho, and denotes the correlation)

A parallel test X' measures the same thing as X equally accurately, but with independent measurement errors (and therefore the correlation is less than one)

But how do we know that two tests are parallel?

$$\text{Rel}(X) = \frac{\text{Var}(T)}{\text{Var}(X)} = \rho(X, X') = \rho^2(X, T)$$

Example: Suppose $\text{Rel}(X) = \rho(X, X') = 0.81$

Then, $\rho^2(X, T) = 0.81$ or $\rho(X, T) = \sqrt{0.81} = 0.90$

Our vision on the true score (how well will the candidate perform on the average?) is blurred by measurement error (we don't observe T, but only X)

To compute the reliability, we need a parallel measurement

- This means we need at least two test administrations on the same sample (expensive) $\alpha \leq \mathrm{Rel}(X)$
- We need a parallel measurement.
  - Parallel form
  - Retesting
- What about Cronbach's alpha (α)?

Cronbach's alpha

- Alpha is not the reliability
  - Except under very special conditions (which are difficult to verify)
  - In general it holds that **$\alpha \leq \mathrm{Rel}(X)$**
- Examples
  - α = 0.95: the reliability is **at least** 0.95 (hurrah!)
  - α = 0.55: the reliability is at least 0.55, it may be 0.80 or 0.95, but we cannot know this from a single test administration.
- Heterogeneous tests in general yield lower α

Determinants of the reliability

- Homogeneity of the population
  - The more homogeneous, the lower the reliability.
  - Beware of zero scores
- The quality of the items
- The test length: Spearman-Brown formula

Spearman-Brown formula

$$\rho(f \times k) = \frac{f \times \rho(k)}{1 + (f - 1) \times \rho(k)}$$

$\rho(k)$ is the reliability with k items (given)

f is the factor of test lengthening or shortening

$f = 2$: double the number of items

$f = 0.5$: halve the number of items

Example of S-B $\rho(f \times k) = \dfrac{f \times \rho(k)}{1 + (f - 1) \times \rho(k)}$

My test has $k = 20$ items and a reliability of 0.67

What will the reliability be if $k = 40$?

Answer: $f = \dfrac{40}{20} = 2$

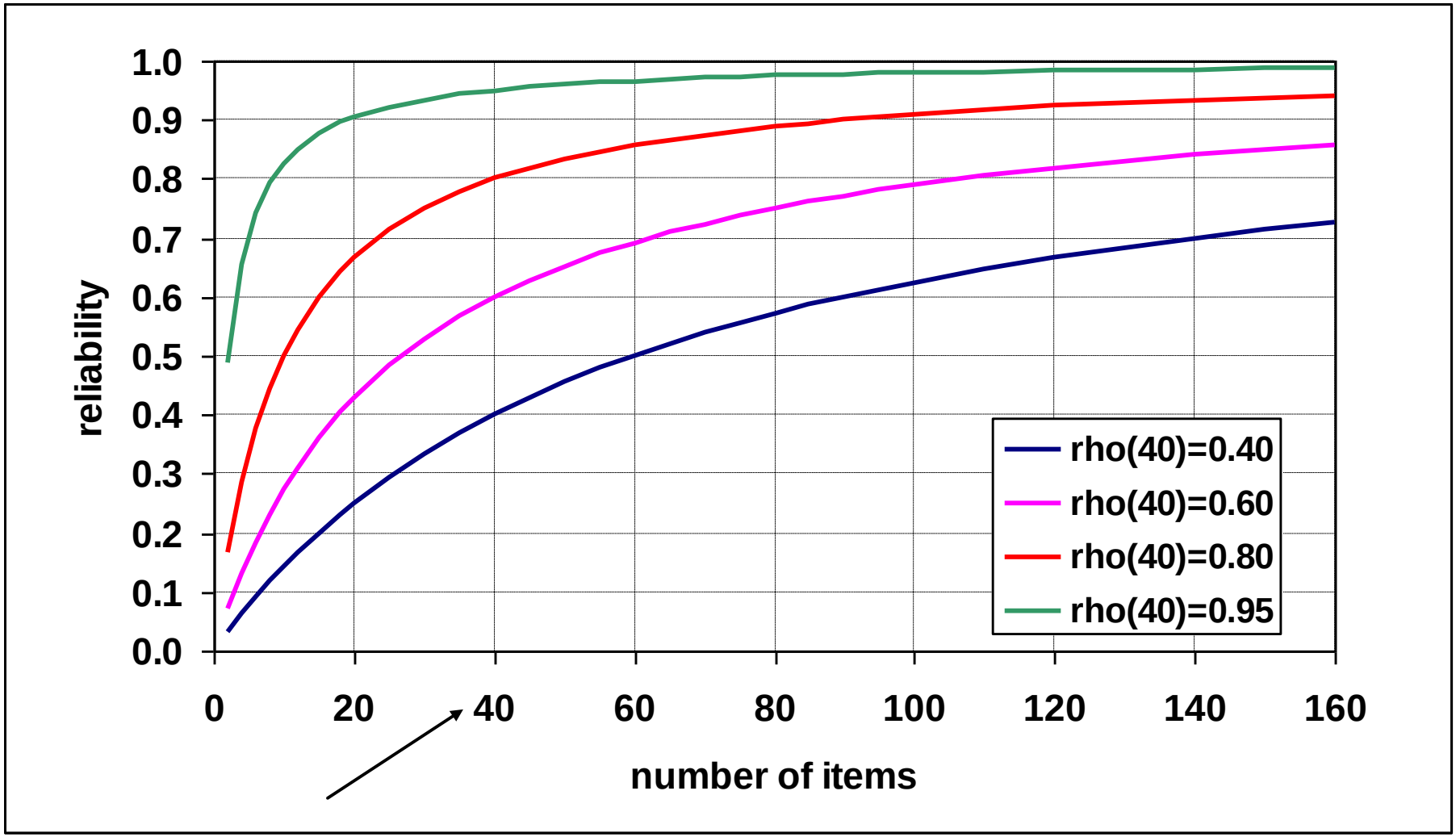
$$\rho(2 \times 20) = \frac{2 \times 0.67}{1 + (2 - 1) \times 0.67} = \frac{1.34}{1.67} = 0.8024$$

What will the reliability be if the test is shortened to 15 items?

Answer: 0.6036

The next three slides were not presented at the webinar due to a lack of time

S-B graphically



The relationship between constructs

- Two constructs, measured by X and Y
- First thought: the correlation $\rho(X,Y)$
- But: our sight on the construct is blurred
- We want the correlation between T_X and T_Y
  - i.e., we want $\rho(T_X,T_Y)$
  - But it is **attenuated** (pushed down) by measurement error

Correction for attenuation

$$\rho(T_X, T_Y) = \frac{\rho(X, Y)}{\sqrt{\rho(X, X') \times \rho(Y, Y')}}$$

Example: $\rho(X, Y) = 0.70$

but $\rho(X, X') = 0.85$ and $\rho(Y, Y') = 0.60$

So, we find: $\rho(\mathrm{T}_X, T_Y) = \frac{0.70}{\sqrt{0.85 \times 0.60}} = 0.98$

That was it for now,
Thank you for listening

I need a short break to look at your questions…

The remaining slides were prepared in anticipation of questions (which were not asked)

Cronbach's alpha: a formula

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_i Var(X_i)}{Var(X)}\right]$$

k is the number of items

X_i denotes the score on item i

X is the test score

Negative scores

- Example: write a summary of a text in maximally 500 words
  - Positive (or zero) scores for several aspects of the summary
  - Penalty points for excess to 500 words (e.g. minus one point for every 20 words in excess of 500)
- Why (not)?
  - Usually not socially acceptable
  - Usual formulae (implemented in the software) will not work properly as most of them assume that the minimal score is zero.

Test score

- Usually the test score is the **sum** of the item scores
  - 0/1 scoring: test score is number of items correct
  - Sometimes called ‘raw’ score or ‘unweighted’ score
- Sometimes a weight is given to a correct item answer and the test score is the sum of the weights one has collected
  - weighted score
  - what are good weights?
  - often problems in reporting

p-values as average relative scores (1)

score	relative score	proportion
2	1	0.75
0	0	0.25

$$\text{Relative score} = \frac{\text{score}}{\text{maximum score}}$$

$$\text{Average relative score} = 1 \times 0.75 + 0 \times 0.25 = 0.75$$

p-values as average relative scores (2)

score	relative score	proportion
2	1	0.29
1	0.5	0.48
0	0	0.23

$$\text{Av. rel. score} = 1 \times 0.29 + 0.5 \times 0.48 + 0 \times 0.23 = 0.53$$

Some statistical concepts

- The variance is a measure of variability
  - It cannot be negative
  - If it is zero, then there is no variability: 'everybody' has the same value
- The standard deviation is the square root of the variance
- Notation: $SD^2(X)$ or SD_X^2 or $\mathrm{Var}(X)$

Eğitim ve Bilim
2014, Cilt 39, Sayı 172
Geniş Ölçekli Test Uygulamaları Özel Sayısı

Education and Science
2014, Vol. 39, No 172
Large-Scale Assessment Special Issue

Test Theory: Some Basic Notions

Test Teorisi: Bazı Temel Kavramlar

Norman VERHELST[1]
Eurometrics

Abstract

This article discusses basic concepts in Classical Test Theory and Item Response Theory. In the context of Classical Test Theory the concepts of observed and true score, reliability of observed scores, and item indices are discussed. Some common rules of thumb to interpret numeric values of these indices are also presented in line

When looking for the article on the internet,
use the Turkish name of the journal