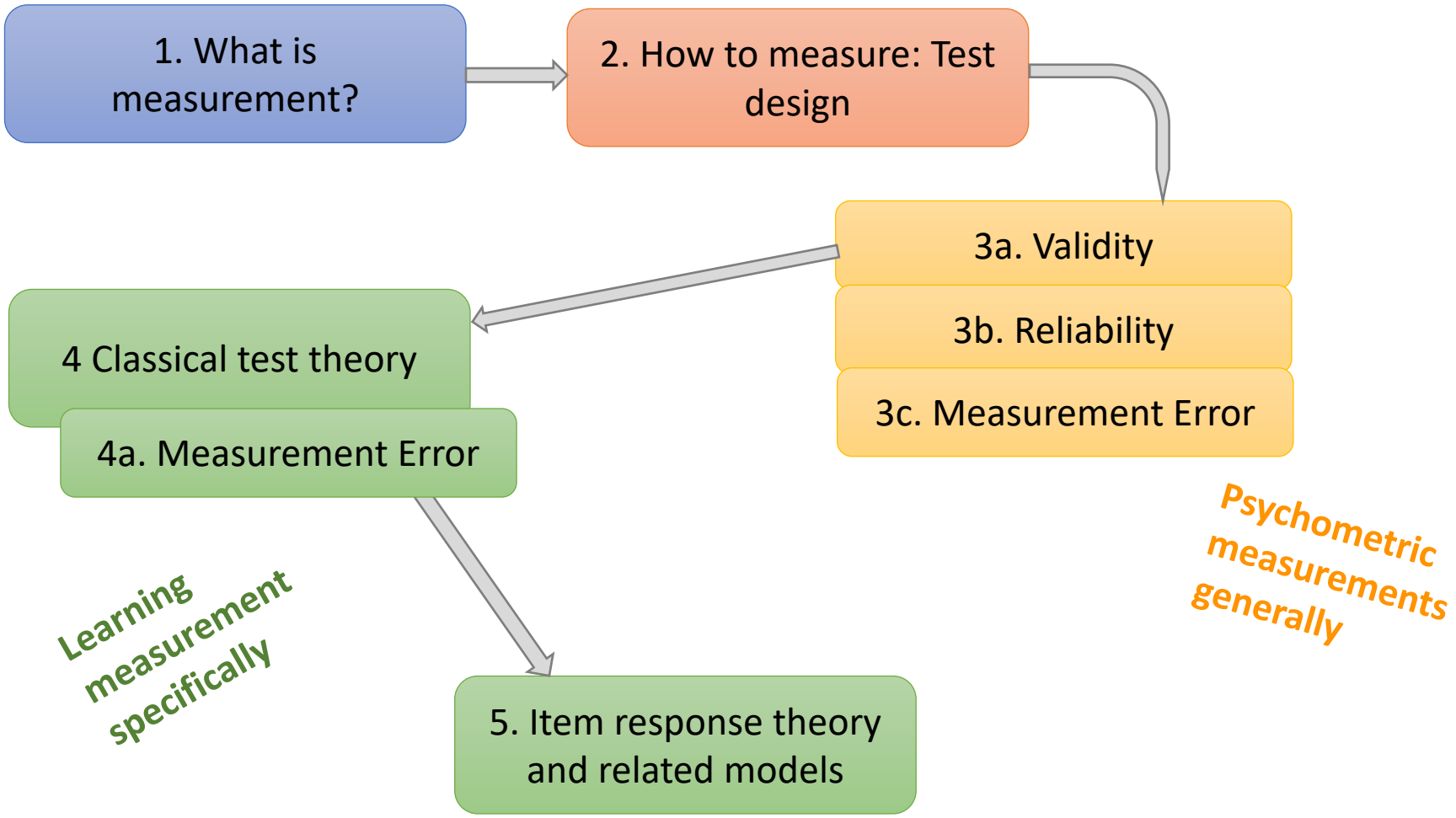
LASI 2018
Workshop

Educational Measurement and the Challenges of Inferring Learning

June 12th and 13th, 2018

geraldine.gray@itb.ie

Overview



Key references:
Bergner, Y. (2017). Measurement and its Uses in Learning Analytics. *Handbook of Learning Analytics, SOLAR*.
Wu, M., Tam, H. P., & Jen, T. H. (2016). *Educational Measurement for Applied Researchers. Theory into Practice*. Singapore: Springer

Overview: looking at statistics to inform . . .

	Item1	Item2	Item3		Item4	Item5	Item6		Item7	Item8	Item9
Person1	1	1	0		1	1	0		1	1	1
Person2	0	0	1		0	0	1		1	0	1
Person3	0	0	1		0	1	0		0	0	0
Person4	1	0	1		0	1	0		1	0	1
Person5	1	0	1		0	1	0		0	1	0
Person6	1	1	1		1	0	1		1	1	1
Person7	1	0	0		1	0	1		0	0	0
Person8	1	0	0		0	1	1		0	0	1
Person9	1	0	1		1	0	1		1	0	1
Person10	1	0	0		0	0	1		1	0	1

What can we say about a person based on their score?

What can we say about the group based on group scores?

What can we say about individual items?

What can we say about constructs measured by groups of items?

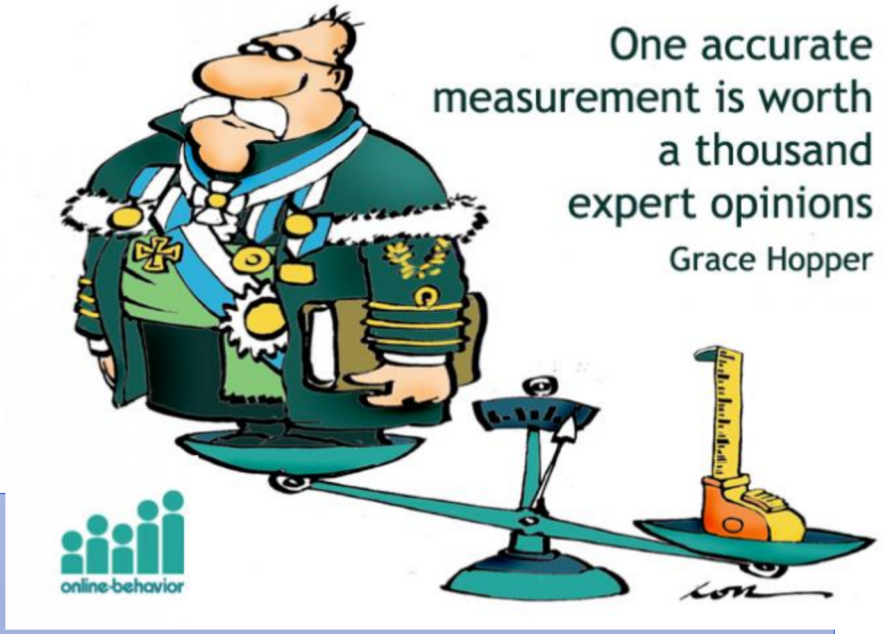
Definitions

- **Construct** is a proposed attribute which often cannot be measured directly, but can be assessed using a number of indicators (variables). e.g. academic achievement.
- **Variable:** It is the measurable expression of the construct, e.g. grade point average
- **Factor**: An independent variable that can be manipulated / controlled for, e.g. gender.
- **Trait**: is stable over time, e.g. conscientiousness
- **Artifact**: a by product of the test procedure itself

- **Effect size**: a way of quantifying the difference between two groups
- **Parallel tests**: two tests that result in the same True score (more formal definition later)
- **Standard error of measurement**: if a student sat a number of parallel tests, SEM is the standard deviation of the resulting scores.

Symbols used

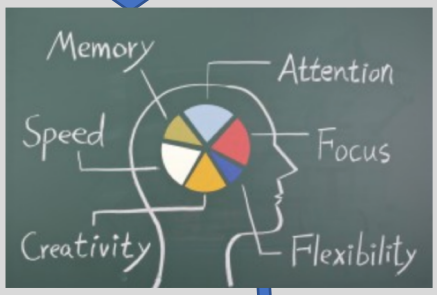
- n, N respondents; i, I test items
- m, μ: the mean
- $s\ and\ s^2$: standard deviation and variance respectively
- δ_{xy}: covariance between x and y
- $\widehat{\ }$ any symbol with a hat is an estimate (e.g. estimate population mean from a sample)
- ρ: correlation & test reliability
- $\mathbb{E}\ or\ E$: both mean expected value
- δ: item difficult
- θ: ability



What is Measurement?

Measurement

Physical measurement



Psycho-social measurement

Aptitude Level	% ile	Interpretation
1	90 - 99	Superior
2	68 - 89	Above Average
3	34 - 65	Average
4	1 - 33	Below Average

Measurement: Assigning numbers to objects to represent quantities of an attribute

Psychometrics: Measurement of a latent trait

- Factor of interest can not be observed
- Factor of interest may lack a clear definition
- It makes sense to ‘measure’ it, i.e. it has various levels from low to high
- Factor is inferred based on indicator variables that can be observed

Levels of Measurement (numeric)

Lowest information content

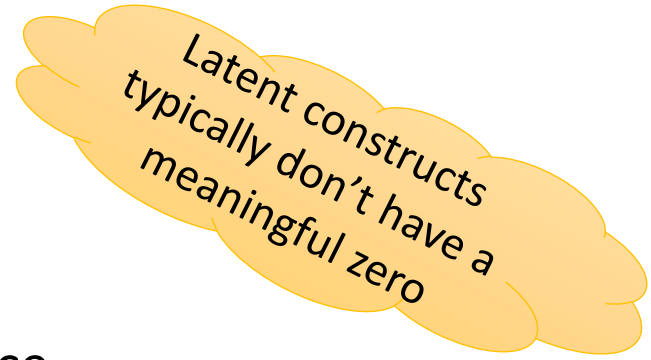
Highest information content

- Nominal
  - Using a number as a label without ordering e.g. student number; gender variables (1=male, 2=female, etc.)
- Ordinal
  - Labels have specific order/rank, but the number does not refer to an amount: 1st in the class, 2nd in the class etc.
- Interval
  - Has an order, and also carries information about the distance between values, but zero does not mean absences of a value e.g. time; ability
- Ratio
  - As for interval but zero means absence of a value, and you can express meaningful ratios based on the numbers in the scale e.g. number of minutes since the start of a test.

Exercise

For each of the following, state if it is nominal, ordinal, interval or ratio

- Grade Point Average
- Academic Alpha Grade (does this change if its mapped to a numeric code?)
- Course code (numeric)
- Likert scale of 1 to 7
- Course credits
- School enrollment number
- Percentage grade
- Time interval viewing an online resource
- Time stamp on an online activity
- Number of clicks on an LMS
- Average clicks per week



Getting started in R

From LASI2018-EducationalMeasurement.html:

1. R Setup

2. A look at some datasets

Mean Absolute Deviation: Median(|X – Median(x)|)

Kurtosis: weight of the tails. 0 for normal distribution.

Trimmed is the median excluding extreme values

Skew: measure of symmetry.
0: normal distribution.
–ive: stretches to the left

Standard error of the mean: $sd/\sqrt{n}$

```
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis
## O1    1 491 5.74 1.60      6    6.03 1.48   1   7     6 -1.31     0.76
## O2    2 491 5.19 1.37      6    5.31 1.48   1   7     6 -0.79     0.04
## O3    3 491 6.18 1.12      6    6.39 1.48   1   7     6 -2.08     5.71
```

```
##       se
## O1 0.07
## O2 0.06
## O3 0.05
```

2. How to measure: Test design

Measuring latent constructs

Latent constructs typically don't have a meaningful zero

Can not be measured using a physical machine

Need to develop an instrument and then prove it measures the factor of interest

1. Define the construct
2. Develop the instrument
3. Deploy the instrument and score
4. Derive a measure
5. Validate the construct

Examples of instruments:

- Written test
- Self reported questionnaire
- Observation checklist

Measuring general ability versus Knowledge of a particular topic(s)

Interpretation in the context of factors influence performance, e.g. Disability (e.g. Visual impairment) Differences in prior knowledge

Multiple choice V imprecision of grading of open ended questions?

Defining a construct: points to ponder

Conflict between content validity (items cover the curriculum) and good properties of measurement (items measure a single construct)

Are subscales different latent traits?

Is the construct measuring one or more traits? . Is their a hierarchical construct (high order & lower order?

Constructs are invented things?

Definition of construct can be fluid depending on context & purpose, and is influenced by how the result will be used

Ideal properties of a test

1. The test is valid: it measures what is says it measures (e.g. geometry, reading ability, etc.). *Result meet the objectives.*

2. The score has a meaning: one can infer a student's ability on a particular topic based on a score (e.g. 7/10)

   - Scores from items measuring a mix of abilities are harder to interpret.

3. The test is reliable: the result of a measurement, calculation, or specification can be depended on to be accurate (e.g. a score of 7/10 is repeatable in similar tests). *Results are consistent.*

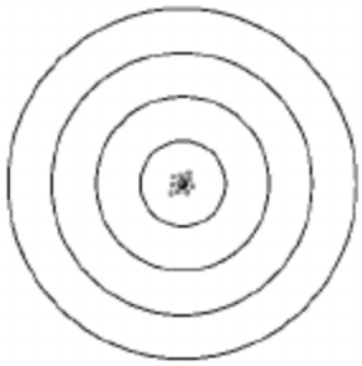
4. Stable frame of references: a comparison of results from different students is meaningful, both in terms of the order of results, and the interval between results.

Results from a single student on multiple items measuring the same construct.

Centre of the circles represents the actual score in the construct of interest.

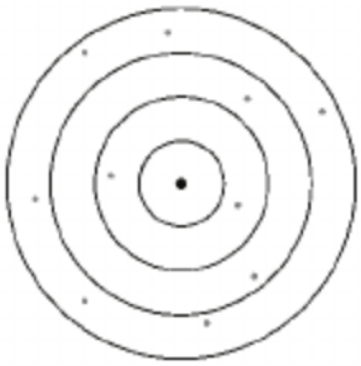
Do items provide a measurement of the construct of interest?

Are results consistent across items?

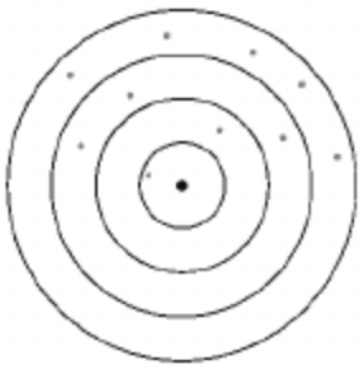


Reliable and Valid

Reliable but not valid



Not reliable, but average score is valid



Not Reliable and Not valid

Test Design covers

Number and arrangement of items

Sample size

and is influenced by PURPOSE

Why are you using educational measurement techniques?

Measuring individual performances?

Measuring cohorts (average / X% meeting a minimum standard)

Evaluating items for a test bank

Preamble

The statistics underpinning educational measurement methods:

Correlation

Covariance

Z-score

Eigen values and Eigen vectors

3a Test Validity

Test validity - scope

Test validity in itself is a huge topic, and can be "a lengthy, even endless process" *Cronbach*.

Technical aspect:

Model fit
Nomological network

Social considerations:
Is the test fair?
What are the unintended consequences?

Philosophical considerations:
What does the score actually mean?

Validate the test → Validate the theoretical construct

Needs to be done over a number of studies, each providing different kinds of evidence from different sources, while systematically ruling out alternative explanations.

History of validity

1. Criterion model: check that test scores correlate with a 'true' criteria score to see how well the test estimates the criterion.
   - Difficult in an educational context, unless its to show that one test (that is cheaper to administer perhaps) is measuring the same thing as another test.
2. Content model: establish a link between procedures to generate the score, and how that score will be interpreted.
   - inferences should be limited to what is externally observable, and not a subjects internal processes as that requires construct validity.
3. Construct model: the test scores match a predefined theory (in the absence of criterion or content).
   - A theory comprises of a network of relationships between constructs and observable variables.
   - Limitations: needs a well defined theory.

Test validation should integrate different kinds of validity evidence (Cronbach)

Some tests of validity

1. Content Validity: items measure the latent construct(s) of interest.
    - Get the opinion of subject matter experts; take items from previous published research
2. Face Validity: items look like they make sense at a cursory glance.

3. Construct Validity confirms that the instrument is appropriate for the **theory** being tested. It generally consists of convergent validity (constructs that should be related are in fact related) and discriminant validity (constructs that should not be related are not in fact related).

- Define the theory behind the constructs.
- Investigate the relationships between the constructs.
- Interpret the construct.

Some tests of validity

4. Internal Validity: constructs only measuring the phenomena under investigation, and are not accidently measuring something that is not part of the research question (cofounders).

- E.g. Does academic achievement only measure learning, or does it also measure ability to handle stress?
- Confirmatory Factor Analysis (CFA) can be used to confirm the internal validity of the constructs.

5. External Validity: can inferences be made about a larger population based on results from a sample, as different populations may answer questions differently.

Multigroup Factor Analysis (MGFA) can determine if the constructs used to measure individual level constructs are equivalent across groups.

Some validation in R

1. Correlations
2. Exploratory Factor Analysis
3. Confirmatory Factor Analysis

Outputs are explained in the following slides.

EFA output:factanal

```
##
## Call:
## factanal(x = myData, factors = 3, rotation = "varimax")
##
## Uniquenesses:
##   O1   O2   O3   O4   O5   C1   C2   C3   C4   A1   A2   A3   A4   A5
## 0.66 0.65 0.60 0.62 0.56 0.21 0.39 0.13 0.22 0.65 0.32 0.61 0.49 0.69
##
## Loadings:
##    Factor1 Factor2 Factor3
## O1                  0.58
## O2                  0.59
## O3                  0.63
## O4                  0.61
## O5                  0.66
## C1   0.87
## C2   0.77
## C3   0.92
## C4   0.87
## A1           0.59
## A2           0.82
## A3          -0.61
## A4           0.71
## A5          -0.54
##
##                Factor1 Factor2 Factor3
## SS loadings       3.02    2.24    1.93
## Proportion Var    0.22    0.16    0.14
## Cumulative Var    0.22    0.38    0.51
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 212.39 on 52 degrees of freedom.
## The p-value is 2.86e-21
```

[0,1]: Want these numbers to be small. A5=0.69 means 31% of variance is contributing to the 3 factors

[-1,1]: Correlations with unobserved factors

Sum of squared loadings

Portion of variance explained

H0: 3 factors is sufficient, this rejects the null hypothesis; is effected by sample size

EFA output: FA

(**summary()** supresses loadings)

Loadings

Variance account for by all factors (MR1^2 here)

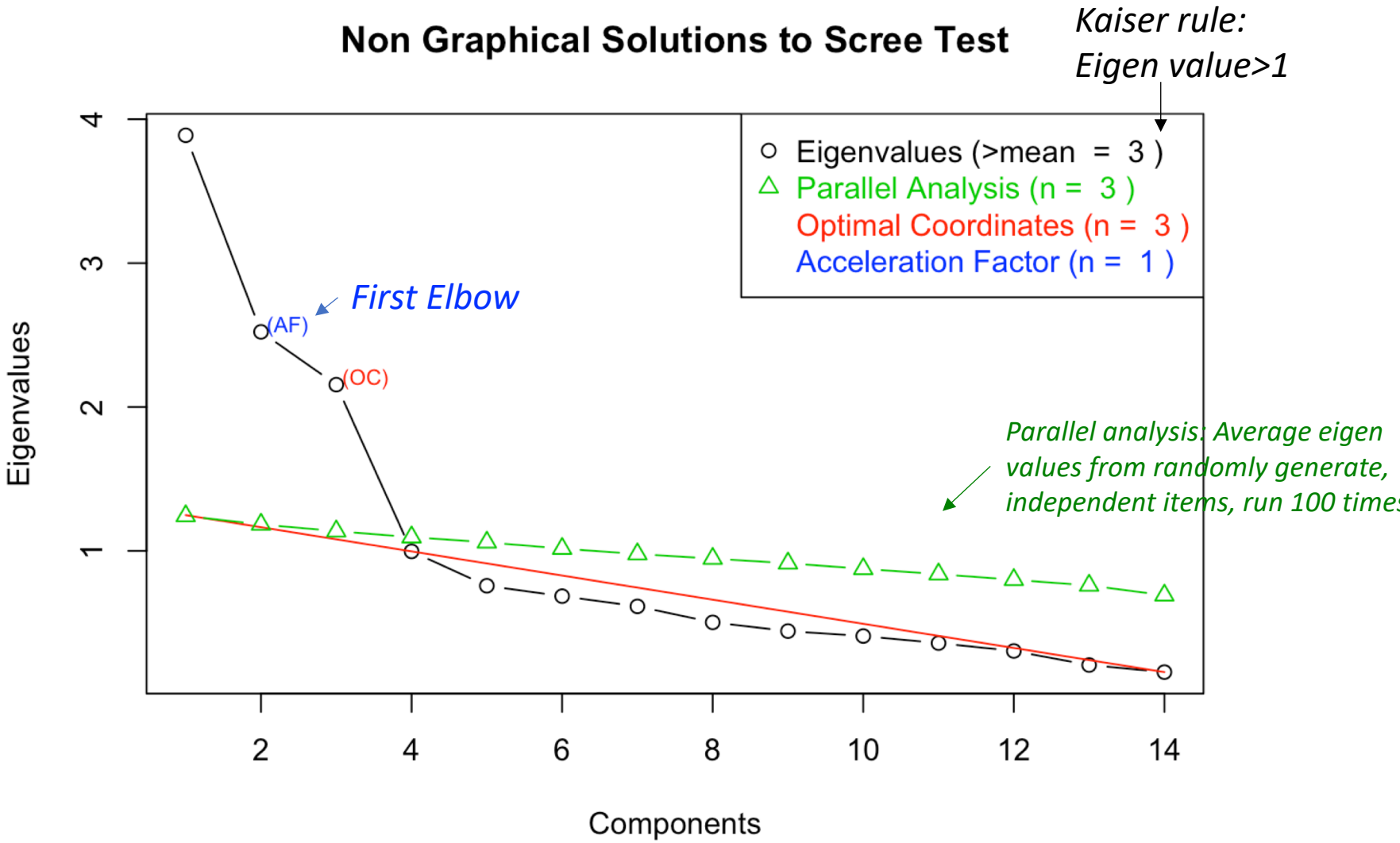
uniqueness

```
##       MR1   h2   u2 com
## I1   0.64 0.41 0.59   1
## I2   0.75 0.57 0.43   1
## I3   0.55 0.30 0.70   1
## I4   0.55 0.31 0.69   1
## I5   0.76 0.58 0.42   1
## I6   0.76 0.57 0.43   1
## I7   0.73 0.53 0.47   1
## I8   0.83 0.69 0.31   1
## I9   0.85 0.72 0.28   1
## I10 0.60 0.36 0.64   1
## I11 0.50 0.25 0.75   1
## I12 0.57 0.33 0.67   1
## I13 0.64 0.41 0.59   1
## I14 0.70 0.48 0.52   1
## I15 0.45 0.21 0.79   1
```

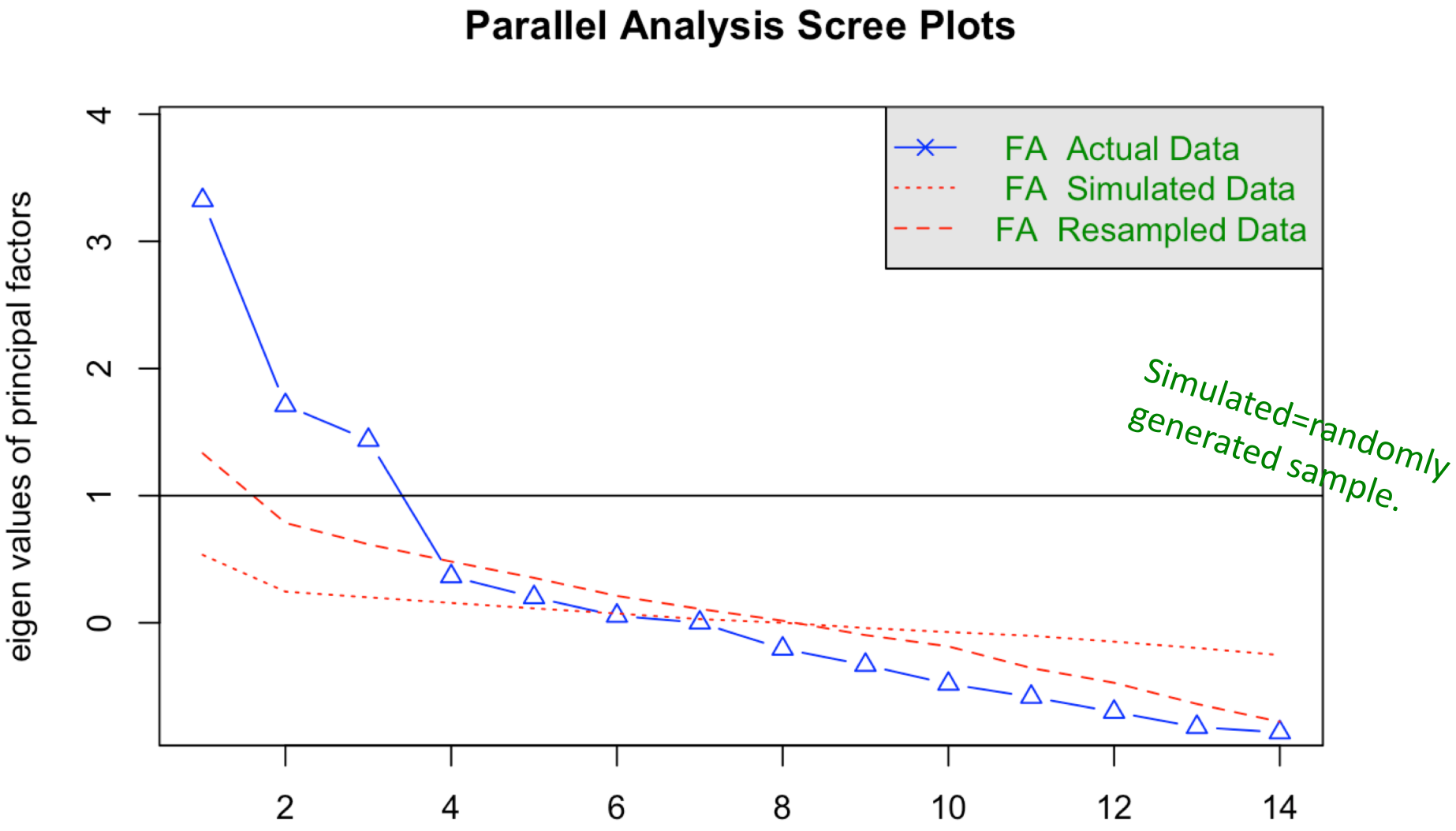
```
                MR1
SS loadings    6.70
Proportion Var 0.45
```

```
The root mean square of the residuals (RMSR) is  0.06
The df corrected root mean square of the residuals is  0.06
 0.3
The harmonic number of observations is  876 with the empirical chi square  642.06  with prob <  3.2e-84
 0.3The total number of observations was  876  with Likelihood Chi Square =  969.57  with prob <  1.7e-147
 0.3
Tucker Lewis Index of factoring reliability =  0.848
RMSEA index =  0.106  and the 90 % confidence intervals are  0.1 0.112 0.3
BIC =  359.79
Fit based upon off diagonal values = 0.98
```

How many factors, based on a plot of the eigen values?



How many factors using poly- / tetr- choric correlation




```
## Parallel analysis suggests that the number of factors =  5
```

CFA model fits

Overall fits are a blunt assessment of a model fit, as a low fit gives no indication of the cause - is it a few rogue items, or a poorly constructed model?

Model fit	Description	Rule of Thumb for good model fit
Chi squared	Similarity between theorized covariance matrix (based on proposed model) and actual covariance matrix	H_0: they are the same. Difference should not be significant
Incremental fit indices: Comparisons with models that assume items are independent (null model):		
CFI	Compare actual model with null model. Penalises complex models (too many degrees of freedom)	Normalised [0,1] >0.9 (good); >0.95:excellent
Tucker Lewis Index	Penalises models for terms added that do not improve the model	Not normalised; good model is a result close to 1 (>0.95)
SRMR	Squared average of the residuals (actual covariance - predicted covariance)	Should be < 0.1, ideally less than <0.06.
RMSEA	Root mean squared error of the model, comparing model to expected outcome.	<0.06
Predictive fit indices: how well a model will replicate on the same sample size randomly drawn from the same population		
Akaike (AIC)	Relative quality of statistical model for comparison. Calculated from the likelihood function, sample size and number of parameters.	Best model = smallest AIC
Bayesian (BIC)	Like AIC, but applies a larger penalty for additional parameters.	Best model = smallest BIC

```
## lavaan (0.5-23.1097) converged normally after  41 iterations
##
##   Number of observations                           491
##
##   Estimator                                         ML
##   Minimum Function Test Statistic              247.388
##   Degrees of freedom                                74
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             3011.174
##   Degrees of freedom                                91
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.941
##   Tucker-Lewis Index (TLI)                       0.927
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -9865.404
##   Loglikelihood unrestricted model (H1)      -9741.710
##
##   Number of free parameters                         31
##   Akaike (AIC)                               19792.807
##   Bayesian (BIC)                             19922.897
##   Sample-size adjusted Bayesian (BIC)        19824.504
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.069
##   90 Percent Confidence Interval          0.060  0.079
##   P-value RMSEA <= 0.05                          0.001
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.048
##
```

CFA output 1: over model fit

```
## Latent Variables:
##                   Estimate  Std.Err  z-value  P(>|z|)
##   open =~
##     O1               1.000
##     O2               0.874    0.094    9.275    0.000
##     O3               0.768    0.080    9.657    0.000
##     O4               0.838    0.088    9.488    0.000
##     O5               0.822    0.083    9.875    0.000
##   conn =~
##     C1               1.000
##     C2               0.838    0.038   22.283    0.000
##     C3               1.009    0.032   31.188    0.000
##     C4               0.956    0.034   28.151    0.000
##   agree =~
##     A1               1.000
##     A2               1.356    0.112   12.075    0.000
##     A3              -1.023    0.096  -10.608    0.000
##     A4               1.040    0.091   11.396    0.000
##     A5              -0.835    0.087   -9.635    0.000
##
## Covariances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##   open ~~
##     conn             0.279    0.080    3.478    0.001
##     agree           -0.015    0.031   -0.476    0.634
##   conn ~~
##     agree           -0.224    0.050   -4.445    0.000
##
## Variances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##    .O1               1.687    0.129   13.063    0.000
##    .O2               1.221    0.095   12.916    0.000
##    .O3               0.758    0.062   12.227    0.000
##    .O4               0.997    0.079   12.564    0.000
##    .O5               0.750    0.064   11.677    0.000
##    .C1               0.630    0.055   11.452    0.000
##    .C2               1.062    0.076   14.002    0.000
```

CFA output 2: individual items

3b Test Reliability

Test reliability

- **Reliability** (ρ) is the degree to which a test produces stable and consistent results. i.e. if a test is administered multiple times, the true scores are the same each time.
- This can be measured as the correlation between the observed score and the actual score:

$$\rho_{XX'} = corr(X, X')$$

- However the actual score is not known.
- One way to estimate reliability is to split test items into two halves, and calculated the correlation between the two halves.
  - This tests the internal consistency of test items
  - Note: This is likely to underestimate reliability, as the two halves have less items.
- Cronbach alpha (α) is an estimate of the average correlation of all possible ways to split test items into two halves.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} s_{x_i}^2}{s_x^2}\right)$$

n=number of test items
s_x^2 is variance of test scores across students,
$s_{x_i}^2$ is the variance of item i across students.

Test reliability

- KR20 is a special case of Cronbach Alpha for dichotomous variables:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_i(1-p_i)}{s_x^2}\right)$$

p_i=proportion of students that got the correct answer for item i;

i.e. test score variance is estimated from probability of success. Variance is lowest when probability of success is high (easy item) or low (difficult item). The largest variance is $\frac{1}{4}$.

Easy and hard items have low variance:

p	p(1-p)
0.00	0.00
0.16	0.13
0.32	0.22
0.50	0.25
0.66	0.22
0.75	0.19
1.00	0.00

Alternatives to Cronbach alpha (α)

- Cronbach alpha (α) has a number of assumptions, including:
  - All items load equally onto the factor (*tau equivalence)*
  - Multivariate normality (Gausian distribution)
  - A single factor is responsible for all shared variance, and so all error is uncorrelated.
- These assumptions are hard to achieve, resulting in α underestimating (or sometimes overestimating) the true score.

- There are a number of other metrics for reliability that address the shortcomings of Cronbach alpha, such as:
  - McDonalds Omega (ω) which is based on hierarchical factor analysis.
  - Guttmans-lambda (G6): the amount of variance in each item that can be accounted for by a linear regression of all of the other items (the squared multiple correlation (smc))
  - Cronbach alpha still remains popular in publications and so is useful for comparison studies

Test reliability in R

- Cronbach Alpha
- Omega & other measures
- Single construct
- Multiple or lower order constructs

```
##  raw_alpha std.alpha G6(smc) average_r S/N   ase mean   sd
##       0.74      0.75    0.73      0.38   3 0.018  5.9 0.92
```

Based on covariance

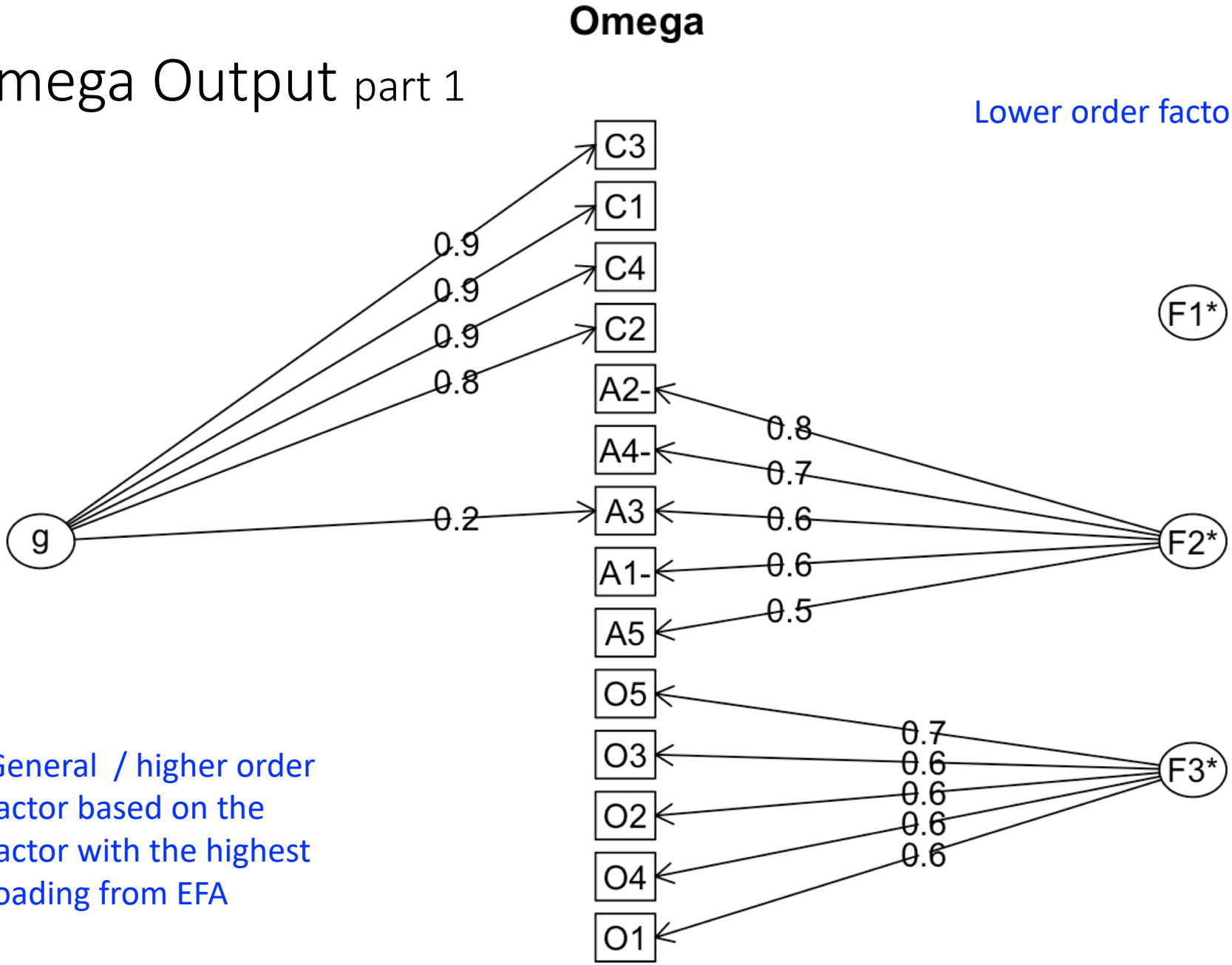
Based on correlation

squared multiple correlation

Average correlation between items

Mean score

Omega Output part 1



General / higher order factor based on the factor with the highest loading from EFA

Omega Output part 2

```
## Omega
## Call: omega(m = myData, nfactors = 3)
## Alpha:                 0.78
## G.6:                   0.85
## Omega Hierarchical:    0.48
## Omega H asymptotic:    0.55
## Omega Total            0.86
##
```

Omega for g only

The rest of the output is as per EFA

3c Measurement Error & SEM

If multiple measurements are taken to estimate a construct, those measurements will vary to some degree. Ideally we want this variability to be low, so we can have confidence in the measurement.

Standard error of measurement (SEM) is a measure of variability.

Potential sources of error

Sampling error:

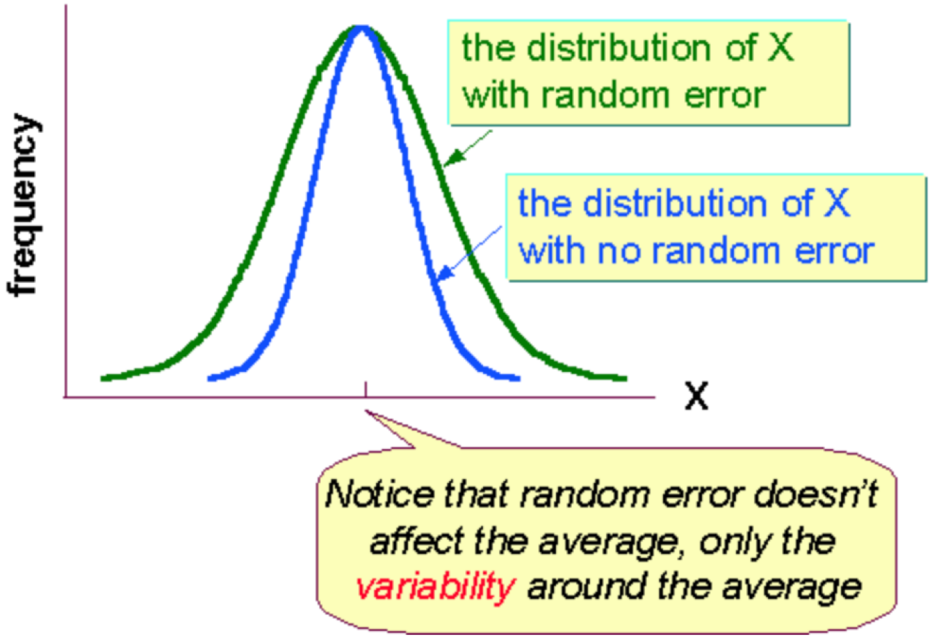
inevitable when working with a sample rather than the full population

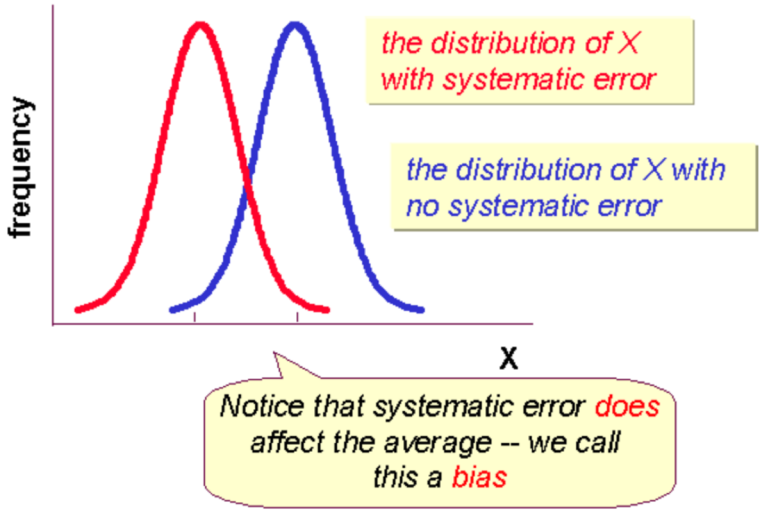
Non-sampling errors, i.e. errors from other sources such as:

- How the data was collected
- Bias in the sample
- Errors in the test items
- Behaviour effects

Measurement error and reliability

- Two types of measurement error:
  - Systematic error / bias: any factors that systematically affects measurement of the variable across the sample
    - E.g. testing maths via English for non-native English speakers; subjective marking that is overly harsh
  - Unsystematic error / random errors occurring by chance:
    - E.g. a mistake on a test; lucky guess; random selection of test items influence score
    - Cause of variability in test scores





Standard Error of Mean – average score in a class

Ignoring test reliability, if there is just one test administered to a number of students (N), that list of results will have a mean value. The standard error of the mean estimates how close the sample mean is to the population mean.

It is estimated as: the standard deviation divided by the square root of the sample size:

Test	Person A	Person B	Person C	Person D	Person E	Mean	Standard deviation (s)
Test 1	36	23	15	23	28	**25**	**7.7**

Standard error of the MEAN: $\frac{s}{\sqrt{N}} = \frac{7.7}{\sqrt{5}} = 3.4$

The bigger the sample size, the lower the error

Educational Measurement specifically

1. Learning/ability can change over time
2. It is difficult to define
3. Scores are influenced by item difficulty and student ability

Sampling in an educational context

Sampling

1. Decide on the population of interest:
   - Age group
   - Region
2. Identify all groups that match the population of interest – the sampling frame
3. Decide on the degree of accuracy needed, to determine inform the sample size.
   - General ranking of students from a low stakes exams; or a more precise measurement, or comparison of change over time when change is likely to be small.
4. Use a probability sampling method, such as random or stratified. Stratified will reduce sampling error if groups are related to target, e.g. group schools into urban and rural if its know performance is lower in rural schools – this ensures a better chance of having a representative sample.

Retuning to Standard Error of the Mean (based on multiple students)

- From Section 3c: $\mathrm{SE} = \frac{s_x}{\sqrt{n}}$ where s_x is the standard deviation of the mean for sample x, and n is sample size.
- This is appropriate of simple random sampling, but its an underestimate if other sampling approaches are used.
- Randomly sampling a population of students is typically not feasible. Instead, cluster sampling is used:
  - One step cluster sample:: where schools are chosen at random, but all students of a particular grade in that school are selected
  - Two step sampling, schools are chosen at random first, and then a sample of students are chosen in each school.
- Clustering sampling requires large sample sizes to achieve similar standard error as random sampling.
  - Design effect: the factor by which the sample size needs to be multiplied by to achieve the same standard error as random sampling (p 60).

Data collection

- Record the RAW data, BEFORE any processing carried out.
- Capture actual responses, and not just whether is was correct / incorrect
  - Can learn about misconceptions from knowing what 'wrong' answer was given.
- Ensure marking guide is as detailed as possible
  - e.g. codes for all predictable incorrect answers
- Prepare a codebook in advance

A gardener mixes 4.45 kilograms of rye grass seed with 2.735 kilograms of clover seed to make a mix for sowing a lawn area. How many kilograms of the lawn mix dies he now have? [TIMSS 2007 grade 8 released mathematics item M022046]

Code	Response	Item: M022046
	Correct Response	
10	7.185	
19	Other responses equivalent to 7.185	
	Incorrect Response	
70	6.780 OR 6.78 [4.045 + 2.735]	
71	Contains one miscalculated digit *(e.g., 7.085, 7.195, 8.185 or similar)*	
72	One of the following: 3.18, 31.8, 318, OR 3180 [misaligns decimals]	
79	Other incorrect (including crossed out/erased, stray marks, illegible, or off task)	
	Nonresponse	
99	Blank	

Variable Values

Value		Label
Rank	1	Freshman
	2	Sophomore
	3	Junior
	4	Senior
Gender	0	Male
	1	Female
Athlete	0	Non-athlete
	1	Athlete
Smoking	0	Nonsmoker
	1	Past smoker
	2	Current smoker
LiveOnCampus	0	Off-campus
	1	On-campus
HowCommute	1	Walk
	2	Bike
	3	Car
	4	Public transit
	5	Other
RankUpperUnder	1.00	Underclassmen
	2.00	Upperclassmen

Data processing

- Avoid manual editing of the data
  - There is no record of changes made, nor can they be reapplied
- All edits should be done using a programming language so they can be easily traced and rerun.
  - R, SAS, SPSS, RapidMiner, MS VBA, Python

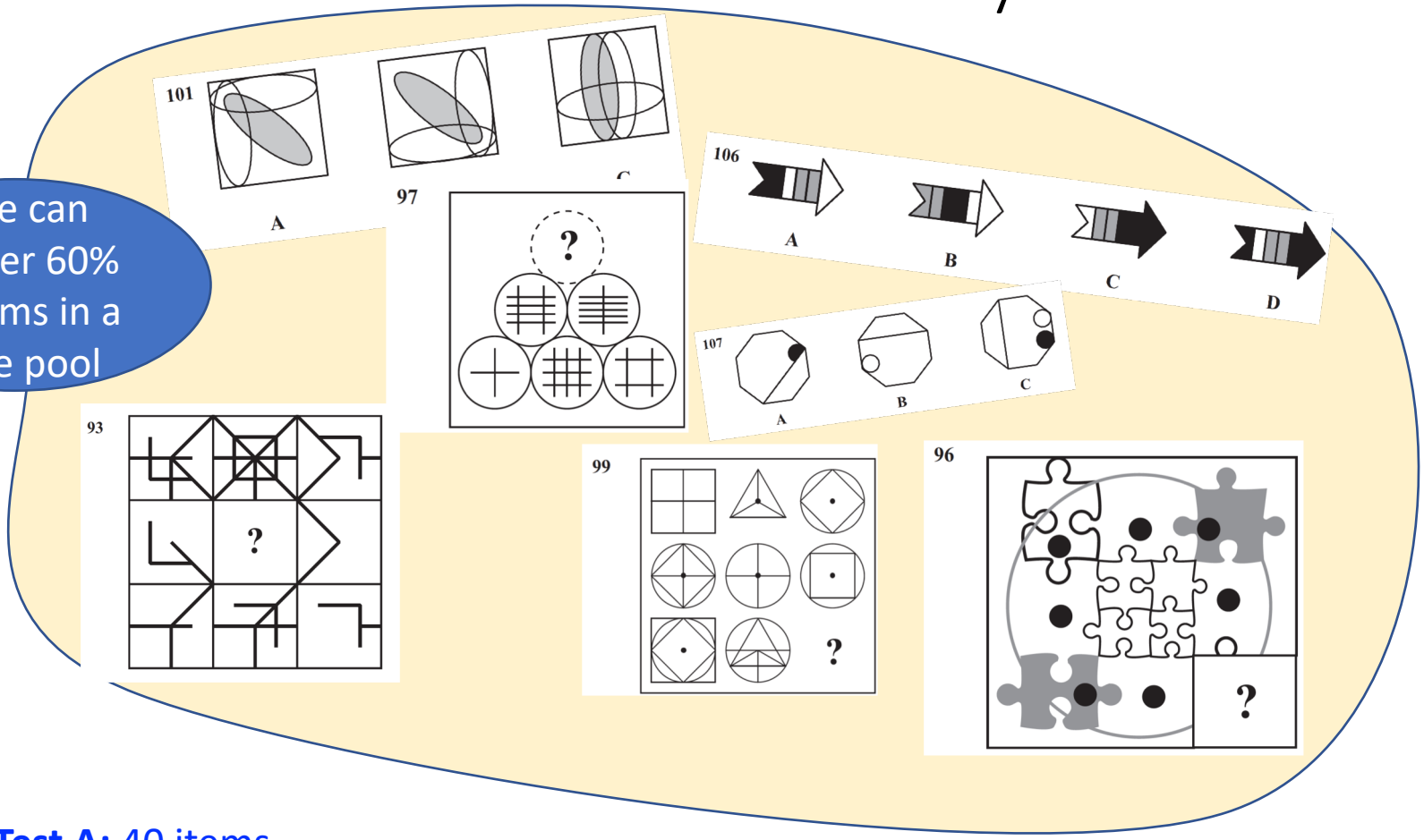
Data cleaning: common issues

- Value range checks
- Missing values
  - Missing by design (student didn't get that test item)
  - An item was skipped
  - An item was not reached
- Duplicate records
- Inconsistencies
  - E.g. giving current year as year of birth
  - Inconsistencies when using double entry for manual entry of test scores.
- Merging data from different courses (e.g. merging a maths test scores with English test scores administered at a different time)

4a. Classical Test Theory (CTT)

Builds on the SEM and reliability tests for generally psychometric data to consider test scores, item difficulty and item discrimination

Measurement error and reliability



Test A: 40 items
Janes score:20/40

Test B: 40 items
Janes score: 28/40

Test C: 40 items
Janes score: 25/40

How do you estimate True Score from a sample test / tests?

Measurement error and reliability

Classical theory assumes

1. Each observed score is the sum of the true score and an error score:

$$X_n = T_n + E_n$$

X_n, T_n and E_n are the observed, true and random error scores for student n.

The true score is the average observed score over repeating administrations of the same test to the same student (expectation of the observed score: $\mathbb{E}(X_n)$).

Other assumptions:

2. $\mathbb{E}(E_n) = 0$, i.e. the average error over repeated administrations of the same test to the same student should be 0.
3. $Corr(T, E) = 0$: there is no correlation between the true score and the error, e.g. a high scoring student will not have consistently higher (or lower) errors than a low scoring student.
4. $Corr(T_A, E_B) = 0$: the true scores of a respondent on one test ($test_A$) does not correlate to the error scores of the same correspondent on another test ($test_B$).
5. $Corr(E_A, E_B) = 0$: similarly, error scores for the same student across two different test are not correlated.

Parallel tests:

Re-administering the same test (to estimate True score for example) would require students to have forgotten the contents of the previous test. Therefore parallel tests can be used.

For two tests to be parallel:

1. Their respective observed scores (X and X') must satisfy assumptions 1 to 5 on the last slide
2. Their True scores must be equal ($T = T'$).
3. The variance in their errors must be equal ($Var(E) = Var(E')$)

Item Difficulty

In CTT, item difficulty is measured as the percentage of respondents who got the item correct, i.e. the average score.

This ignores the abilities in the group taking the test.

Dichotomous items:
Percentage of correct answers

Score	Percentage the got this score
0	0.2
1	0.8

Proportion correct / Difficulty:
0.8

Scoring items out of two (0, 1, 2):

1. Calculate an average score
2. Express it as a percentage

Score	Percentage the got this score
0	0.2
1	0.45
2	0.35

Average: 0*0.2 + 1*0.45 + 2* 0.35 = 1.15
So average score of 1.15 out of 2
So proportion correct is 1.15/2 = 0.575

Item Discrimination

- Item discrimination refers to an items ability to distinguish between strong and weak students w.r.t. to the latent variable of interest.
  - For dichotomous items, it refers to the degree to which high achieving students will get the score correct, and low achieving students will get the score incorrect.
- Low discrimination can be caused by:
  - An item is not related to the latent trait (poor reliability)
  - A item is poorly worded / confusing
  - The item is too easy (mostly answered correctly) or too difficult (most answered incorrectly)
    - Note: Variety in item difficulty may be desirable to facilitate discrimination between a wide range of abilities.

Item discrimination and item difficulty

- An items ability to discriminate between strong and weak students in a particular trait will be effected by item difficulty.
  - Many students will get a high score on easy items
  - Many students will get a low score on difficult items

- Therefore very hard and very easy questions may have lower correlation with overall test score, but may still be useful to include:
  - Starting with a few easy questions can help relax a student.
  - A very hard question can be useful in identifying high ability students

Item discrimination

- Assuming an item is measuring the latent trait of interest (valid and reliable), CTT estimates item discrimination as

$$\frac{(\textit{number correct in the top third - number correct in the bottom third})}{\textit{size of each group}}$$

Ordered by overall score:	Easy item	Average difficulty	Difficult item
Person 1	0	0	0
Person 2	1	0	0
Person 3	1	0	0
Person 4	1	0	0
Person 5	1	0	0
Person 6	1	1	0
Person 7	1	1	0
Person 8	1	1	1
Person 9	1	1	1
Item Discrimination	0.33	1	0.67

4b Measurement Error (again)

Standard error of measurement (SEM)

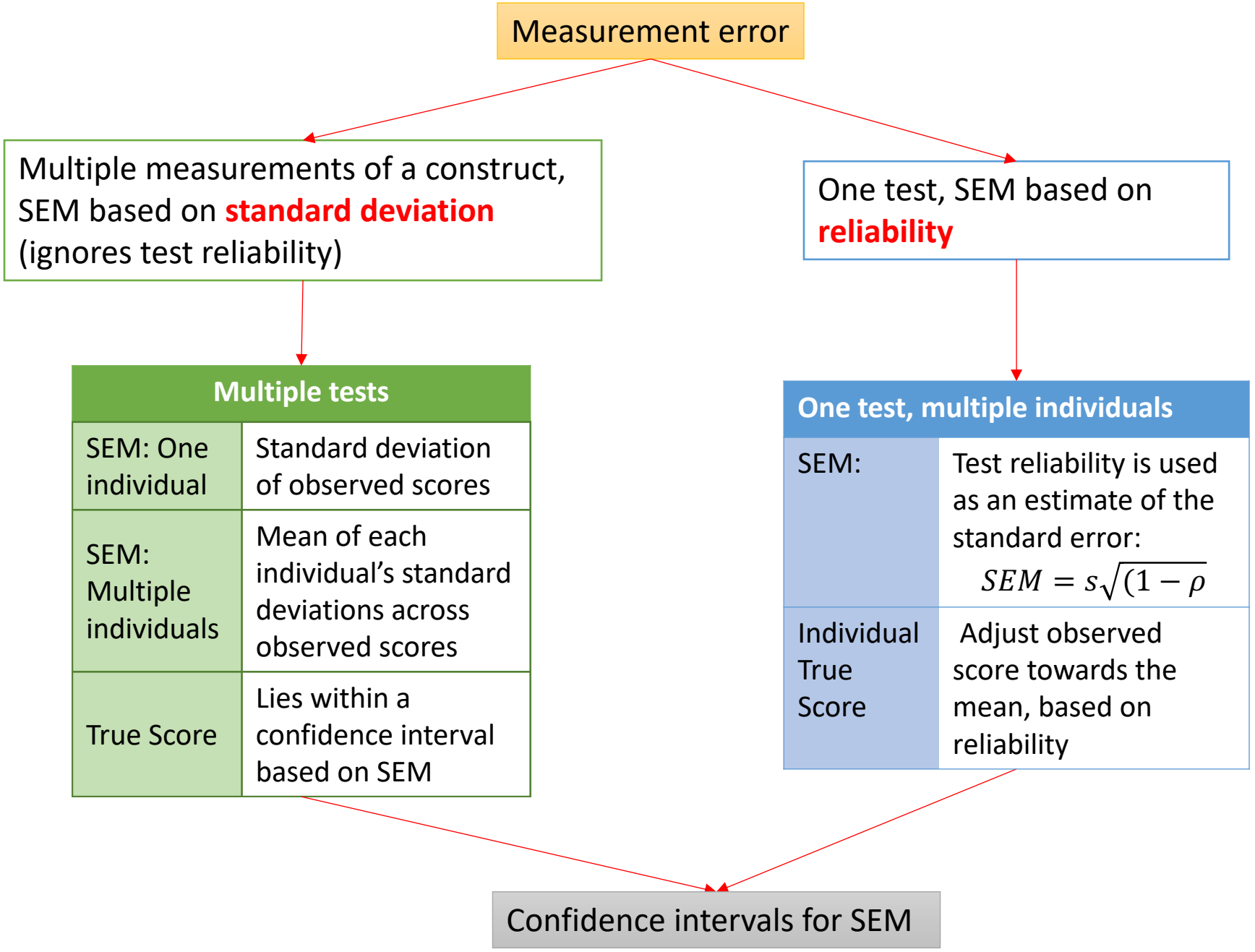
Previously looked at error of the mean score in a test.

Will now look at errors in measurement related to a student's score.

Measurement error / effect size

1. What is the possible variation in a student's test scores if similar tests are administered?
   - A test is a limited window on a students ability; different tests measuring the same construct can results in different scores.
   - The score is an estimate of the true score
   - Error measure is the likely gap between actual and true score

2. Is this level of variation appropriate?


Measurement error
Multiple measurements of a construct, SEM based on standard deviation (ignores test reliability)
One test, SEM based on reliability
Multiple tests
SEM: One individual
Standard deviation of observed scores
SEM: Multiple individuals
Mean of each individual's standard deviations across observed scores
True Score
Lies within a confidence interval based on SEM
One test, multiple individuals
SEM:
Test reliability is used as an estimate of the standard error:
$SEM = s\sqrt{(1-\rho}$
Individual True Score
Adjust observed score towards the mean, based on reliability
Confidence intervals for SEM

Multiple tests:
SEM based on **standard deviation**

Ignoring test reliability, the **variability of one persons score** over multiple measurements (SEM) is the standard deviation of the scores:

- Example: conscientiousness instrument, repeated 5 times, range [0,40]

Test	Result for Person A
Test 1	36
Test 2	37
Test 3	33
Test 4	37
Test 5	35
SEM = SD =	1.7

If the measurement covers **multiple tests for multiple people**: SEM across people and tests is the root mean squared average of their standard deviations:

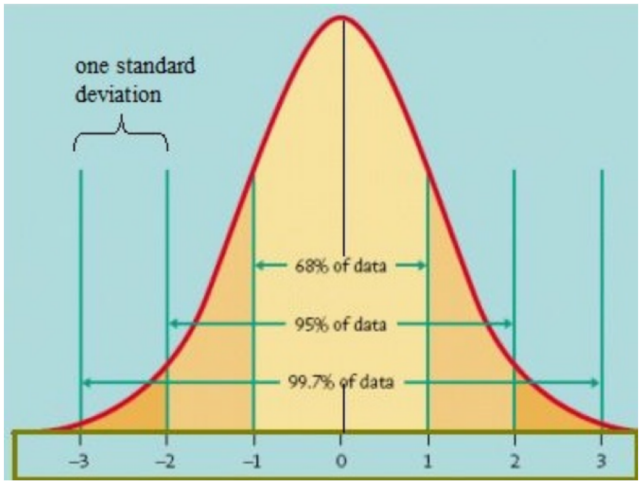
Note: the **mean** of the standard deviations could also be used

Test	Person A	Person B	Person C	Person D
Test 1	36	23	15	23
Test 2	37	25	19	24
SD =	0.7	1.4	2.8	0.7
SEM	$\sqrt{\frac{0.7^2+1.4^2+2.8^2+0.7^2}{4}} = 1.66$			

Estimating **True Score** from SEM & confidence intervals

SEM is used to calculate the interval within which the true score is likely to lie. The size of the interval will depend on how confident we want to be:

- If we want to be 68% confident, the interval is one SEM from the score
- If we want to be 95% confident, the interval is two SEMs from the score.



Test	Person A
Test 1	36

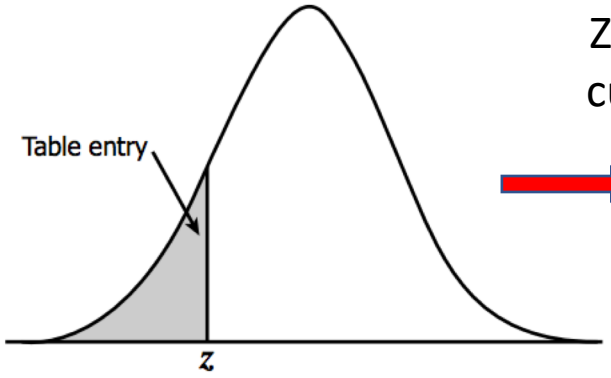
The 95% confidence interval for Person A is:

$$= Score \pm (2 * SEM)$$

$$= [36 - (2 * 0.77), 36 + (2 * 0.77)]$$

$$= 95\%\ CI\ [34.5, 37.5]$$

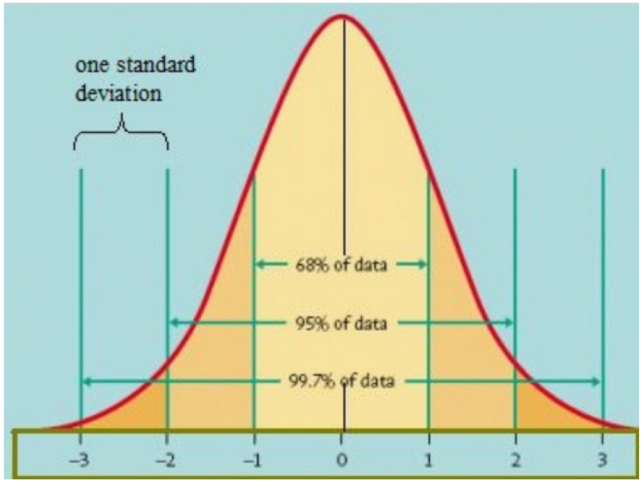
Aside: explaining the 68–95–99.7 rule



Z-scores for standard normal curve = AUC to left of z-score

–1.1	.1357
–1.0	.1587
–0.9	.1841

One standard deviation from the mean has 15.8% of values to the left ∴ area for <u>both tails</u> is 15.8*2=31.6%. ∴ Area <u>inside</u> one standard deviation from the mean is:
1-31.6 ≅ 68%



68–95–99.7 rule

One test: SEM based on reliability

- Returning to a **single test**, the standard error of measurement (observed score) will be influenced by test reliability.
  - The more reliable the test, the lower the standard error and vice versa. Therefore, standard error can alternatively be calculated as:

$$SEM = s\sqrt{1-\rho}$$

s:standard deviation of the test;
ρ: test reliability.

Test	Person A	Person B	Person C	Person D	Person E	Mean	Standard deviation (s)
Test 1	36	23	15	23	28	**25**	**7.7**

If we assume reliability is 0.9,

standard error of MEASUREMENT is: $s\sqrt{1-\rho} = 7.7 * \sqrt{1-0.9} =$ **2.43**

SEM and True scores / Correction of Attenuation

- The variance in True Scores is less than the Variance of Observed score, because observed scores are attenuated by measurement error.

$$Var(T) = \rho \times Var(X)$$

- So from the example in the previous slide:
- If the variance in scores across all students is 59 (s^2); and reliability is 0.9, the variance in True Scores is (59 x 0.9 = 53).

Aside: True correlation between two parallel tests is the correlation in observed scores divided by reliability of each test: $corr(T_A, T_B) = \frac{corr(X_A, X_B)}{\sqrt{\rho_A \rho_B}}$

SEM and True scores / Correction of Attenuation:

- A previous slide had the standard error for observed score as:

$$SEM = s\sqrt{1-\rho}$$

- Using the estimated variance from the last slide, the standard error for TRUE scores can be estimated as:

$$SEM = s\sqrt{(1-\rho)*p}$$

s=standard deviation of the test;
ρ is test reliability.

Test	Person A	Person B	Person C	Person D	Person E	Mean	Standard deviation (s)
Test 1	36	23	15	23	28	**25**	**7.7**

If we assume reliability is 0.9,

Estimated standard error of TRUE SCORE is: $s\sqrt{\rho*(1-\rho)}$=$7.7*\sqrt{0.9*(1-0.9)}$ =**2.31**

Estimate True Score from Observed score

- Estimating true score from observed score accounts for regression to the mean, so it will be closer to the mean than the observed score.
- It can be calculated as: $Tn = m * (1 - \rho) + (\rho * Xn)$
- Example:

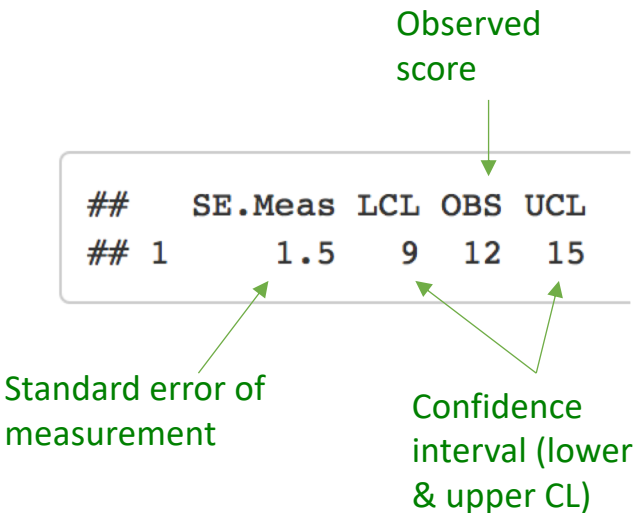
Test	Person A	Person B	Person C	Person D	Person E	Mean	Reliability
Observed Score	36	23	15	23	28	25	0.9
True Score:	**34.9**	**23.2**	**16**	**23.2**	**27.7**		

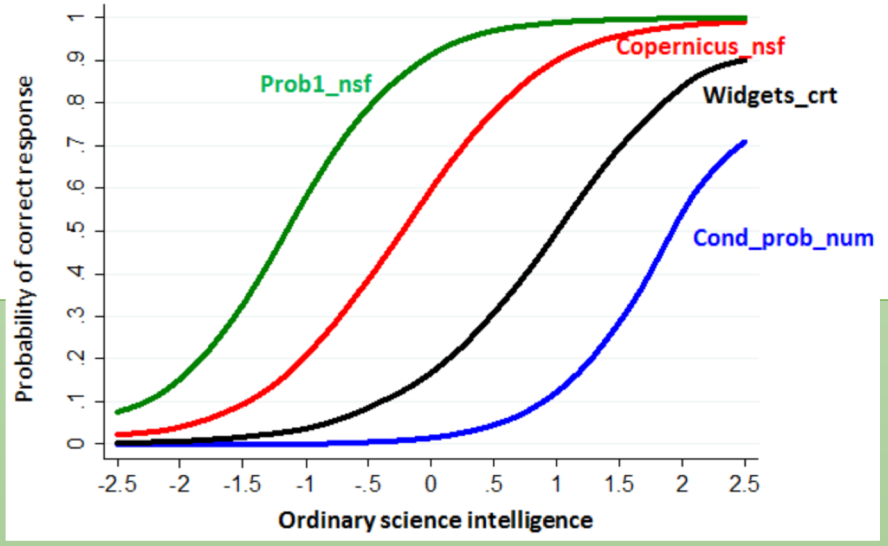
Exercise in R

- Calculate item difficulty and item discrimination

- From the data of 1 test item:
  - Mean
  - Standard deviation
  - Sample error of measurement
  - Confidence intervals for observed and true scores

R output explained:





Item Response Theory (IRT)

More detailed interrogation of the meaning of the scores

In addition to CTT's measure if test score is accurate (reliability) and useful (validity), IRT considers if test scores allow us to compare individuals based on scores achieved.

Retuning back to what does a raw test score mean . . .

Difficulty

Consider a test is made up of questions testing understanding of:

Averages: (correctly answered by 90% of students)

Standard deviation: (correctly answered by 75% of students)

Error measurement: (correctly answered by 45% of students)

Consider a set of students with the following results:

Student A: 60%

Student B: 50%

Student C: 25%

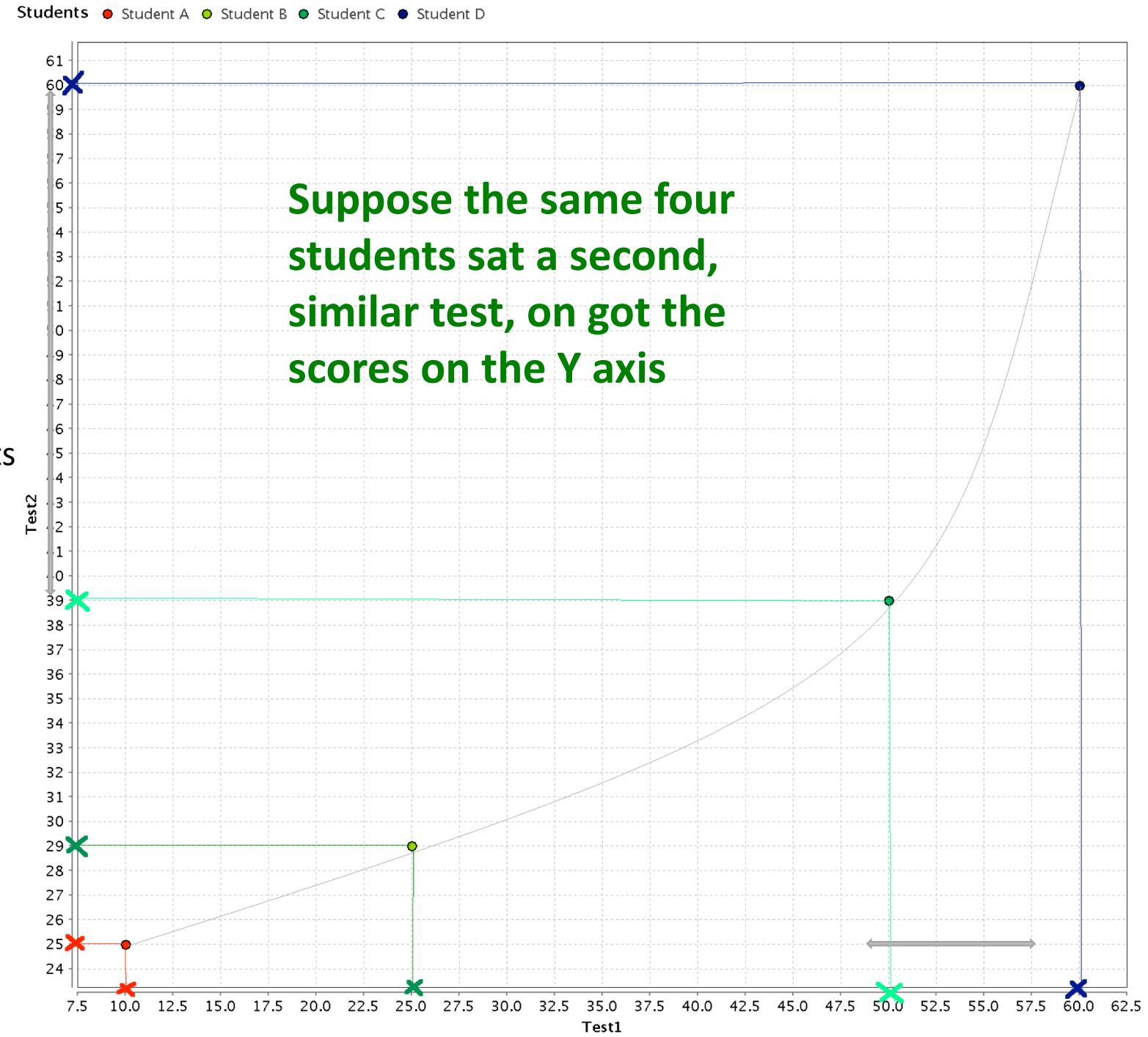
Student D: 10%

Ability

What can be said about the abilities of student B in terms of their understanding of the concepts tested?

What can be said about student B's abilities relative to Student A?

Invariance across tests (are raw scores interval or ordinal?)



IRT models

- The following slides will cover:
- 1PL IRT model: 1 parameter logistic model (Rasch model)
  - Determining ability from item difficulty
- 2PL IRT model: 2 parameter logistic model
  - Determining ability from item difficulty and item discrimination

And briefly look at other related topics including 3PL

Symbols used:

δ : Item difficulty

θ: Persons ability

α: discrimination power of an item

1PL: Item Response Theory (IRT)

A respondent's test score for an item is based on:

1. The difficulty of the item itself (δ)
2. The ability of the respondent (θ)

IRT mathematically models the probability of getting an item correct given item difficulty and student ability

IRT response function must have the following properties:

- Take an input (ability) in the range $[-\infty, \infty]$ i.e. there is no lower or upper bound on ability.
- Output a number in the interval [0,1] probability of getting an item correct
- Output increases with ability (higher abilities should result in a higher probability of success)

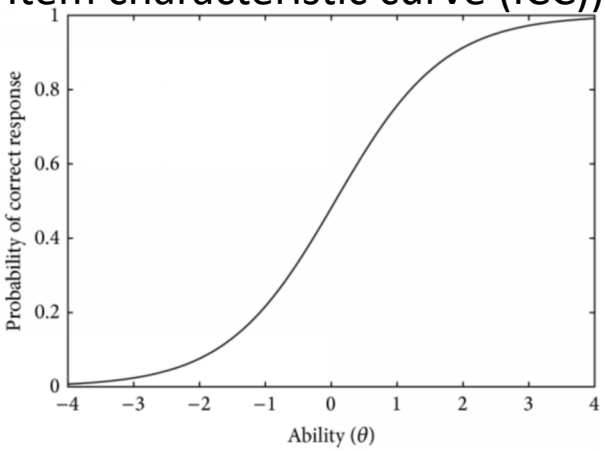
Rasch Dichotomous Model - used when a score is either correct or incorrect (1 or 0)

- The simplest IRT model is a Rasch Model, where the IRT response function is logistic, e.g. of the form:

$$p = P(X = 1) = \frac{e^{(\theta-\delta)}}{1 + e^{(\theta-\delta)}}$$

θ is a "person" parameter: **ability** on the latent variable
δ is an "item" parameter: generally **item difficulty**

Item characteristic curve (ICC))



The resulting curve (ICC) represents ONE test item. It plots the probability of getting the answer correct against student ability.

Y axis: the probability of getting that item correct

X axis: student ability

A single point on X (point at which Y=0.5) is defined as the item difficulty.

Interpreting probability of success

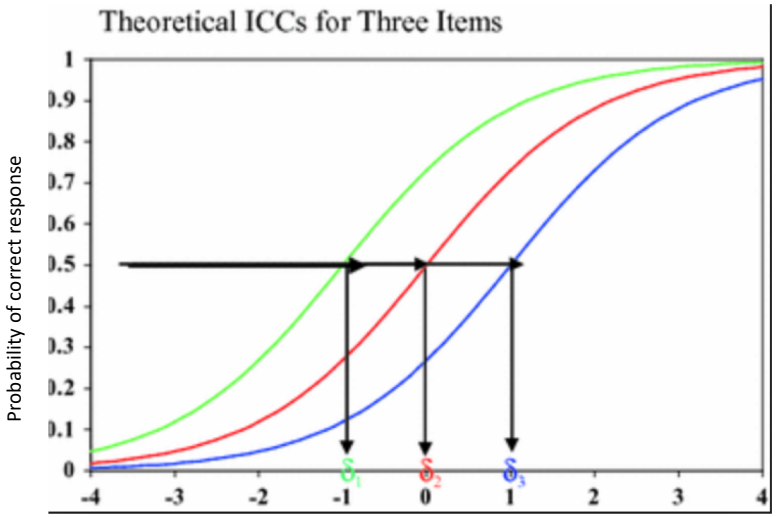
Suppose Jane has a 0.7 probability of success on a particular item. This can be interpreted in the following two ways:

- 70% of students at Jane's ability level will answer the question correctly
- On a set of items at this level of difficulty, Jane will answer 70% of them correctly.

Note: It does NOT say Jane will get a score of 70% on a that item. On dichotomous items, the answer will either be correct (100%) or incorrect (0%), regardless of the probability of success.

Item difficulty (δ)

Item difficulty (δ) is defined as the level of ability needed to have a 50% chance of getting the item correct.



Note: ability and item difficulty are on the same scale

Measurement Invariance: a student's ability (grade) is the item difficulty of the group of items at which the students gets about 50% correct; its not a percentage, or number of items correct (raw score). It will be the same group of items regardless of whether or not a greater percentage of easier items (or a greater percentage of harder items) were included in the test.

Rasch Dichotomous Model, some more formulas:

From the Rasch model below . . . it can be derived that:

$$p = P(X = 1) = \frac{e^{(\theta-\delta)}}{1 + e^{(\theta-\delta)}} \quad \Rightarrow \quad ln\left(\frac{p}{1-p}\right) = \theta - \delta$$

Derived in Appendix 1;
p is probability of success
ln is log to base e (≅ 2.7)
θ : person **ability**
δ : **item difficulty**

Therefore the distance between a person's ability and item difficult is the log of the odds unit that person on that item, so the measurement scale for item difficulty and ability is generally referred to ***logit***.

Odds is the probability of success over the probability of failure: $\left(\frac{p}{1-p}\right)$.

Called a 1 parameter model (1PL), because ability is determined by one parameter: item difficulty.

Note: probabilty of failure is $p = P(X = 0) = \frac{1}{1 + e^{(\theta-\delta)}}$

Rasch Dichotomous Model

$$ln\left(\frac{p}{1-p}\right) = \theta - \delta$$

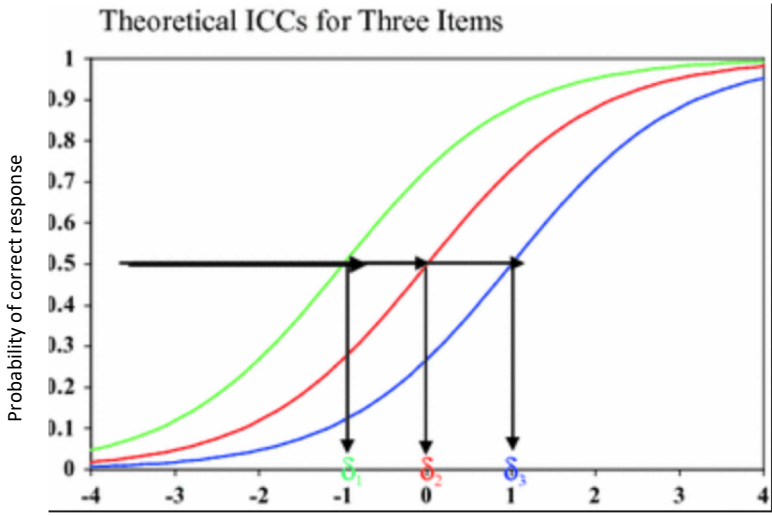
- The left hand side is essentially a transformation of p (percentage correct for a specific $\theta \; and \;\; \delta$) into the logit scale
- Initial estimates of item difficulty can be calculated as $-ln\left(\frac{p}{1-p}\right)$, i.e. a transformation of percentage of correct scores.

Recap: in CTT, p was an estimate of item difficulty calculated as percentage of correct answers. If θ is 0 (mid point), then item difficulty here can be estimated as p transformed to a log scale.

Deriving θ, δ *and* p from raw scores

$$p = P(X = 1) = \frac{e^{(\theta-\delta)}}{1 + e^{(\theta-\delta)}}$$

$$ln\left(\frac{p}{1-p}\right) = \theta - \delta$$



- From the formulas above, once two parameters are known, the other can be calculated.
- However, initially none are known, so estimates are required that match the raw scores with the answers.

Step 1: assume all respondents have the same ability, and that ability is mid point in the range (set θ = 0 for all respondents).

Step 2: set probability p as the percentage of correct answers for an item

Step 3: estimate item difficult δ as $-ln\left(\frac{p}{1-p}\right)$

Step 4: for each respondent, revise their ability estimate θ based on p and δ above, based on the answers given by that respondent.

Step 5: for each item, revise the difficulty estimate δ based on p , $\theta's$, and the answers given for that item

Repeat steps 5 and 4 until estimates converge

Worked example: first iteration in joint maximum likelihood (JMP) to derive θ *and* δ from raw scores

Raw scores

	Item1	Item2	Item3
Person1	**1**	**0**	**0**
Person2	0	0	1
Person3	0	0	1
Person4	1	0	1
Person5	1	0	1
Person6	1	1	1
Person7	1	0	0
Person8	1	1	0
Person9	1	0	1
Person10	1	0	0

p	0.8	0.2	0.6
1-p	0.2	0.8	0.4

Step 3: assume θ = 0 to estimate δ:

$$ln\left(\frac{p}{1-p}\right) = \theta - \delta$$

	Item1	Item2	Item3
Difficulty	-1.39	1.39	-0.41

Step 4: For each person, which ability level maximises the likelihood of their responses/scores?

For Person 1, Using the formula below:

Ability	Item1	Item2	Item3	Likelihood (product)
-3	0.17	0.99	0.93	0.15
-2	0.35	0.97	0.83	0.28
-1	0.60	0.92	0.64	0.35
0	0.80	0.80	0.40	0.26
1	0.92	0.60	0.20	0.11
2	0.97	0.35	0.08	0.03
3	0.99	0.17	0.03	0.01

$$p = P(X = 1) = \frac{e^{(\theta-\delta)}}{1 + e^{(\theta-\delta)}} \qquad p = P(X = 0) = \frac{1}{1 + e^{(\theta-\delta)}}$$

Combining gives one formula (X=response):

$$p = \frac{e^{(X*(\theta-\delta))}}{1 + e^{(\theta-\delta)}}$$

Based on the responses given by person 1, and the estimated δ's, the most likely ability level is -1.

Do the same for the remaining respondents, and return to δ estimates using revised θ

Worked example: first iteration in joint maximum likelihood (JMP) to derive θ and δ from raw scores

Step 5: For each item, calculate the probability if getting that list of responses based on each persons ability

$$p = \frac{e^{(X*(\theta-\delta))}}{1+e^{(\theta-\delta)}}$$

Difficulty	-3	-2	-1	0	1	2	3	Item response (raw score)	Revised abilities (from Step 4)
Person1	0.97	0.91	0.79	0.59	0.34	0.16	0.07	1	0.35
Person2	0.04	0.10	0.23	0.44	0.68	0.85	0.94	0	0.23
Person3	0.04	0.10	0.23	0.44	0.68	0.85	0.94	0	0.23
Person4	0.97	0.92	0.81	0.62	0.37	0.18	0.07	1	0.47
Person5	0.97	0.92	0.81	0.62	0.37	0.18	0.07	1	0.47
Person6	0.98	0.95	0.87	0.70	0.47	0.24	0.11	1	0.87
Person7	0.97	0.91	0.79	0.59	0.34	0.16	0.07	1	0.35
Person8	0.96	0.89	0.75	0.52	0.29	0.13	0.05	1	0.1
Person9	0.97	0.92	0.81	0.62	0.37	0.18	0.07	1	0.47
Person10	0.97	0.91	0.79	0.59	0.34	0.16	0.07	1	0.35
Likelihood:	0.0011	0.0047	0.0090	0.0034	0.0001	0.0000	0.0000		

The most likely difficulty item difficulty for item1 is -1 (highest probability)

Repeat Steps 4 and 5 . . .

Properties of the Rasch Model

Differences between ability and item difficulty on the logit scale are relative, not absolute.

1. Specific objectivity (one ICC):

- The ability of two students can be compared without reference to the test item being used.
- Item difficulties can be compared without reference to the person who completed those items.

$$ln\left(\frac{p_1}{1-p_1}\right) - ln\left(\frac{p_2}{1-p_2}\right)$$
$$= \theta_1 - \delta - \theta_2 - \delta$$
$$= \theta_1 - \theta_2$$

2. Indeterminacy (multiple ICCs):

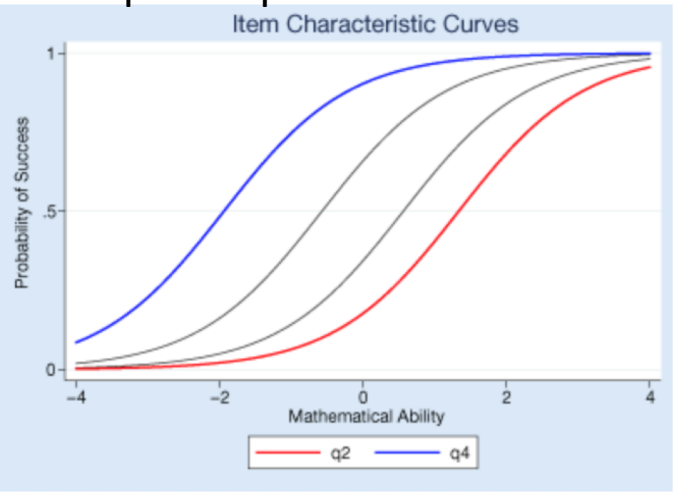
- However, absolute values across two ICC's can not be compared unless they use the same origin
- When scaling items to estimate ability or difficulty, an arbitrary origin can be selected (average of item difficulties). A class achieving high test scores can not be interpreted correctly without reference to class ability and item difficulty.
- Item difficulty itself is evaluated with reference to the group completing the item.

Properties of the Rasch Model

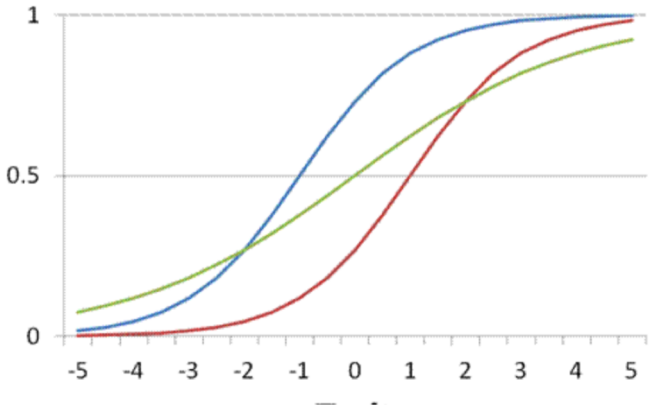
3. Equal Discrimination: The ICC curves for sets of items do not cross (equal slopes at $p = 0.5$).

- A Rasch model assumes all items are equally good as discriminating between weak and strong students.
- The discriminating power of an item is its slope at $p = 0.5$. Flatter ICC's have lower discriminating power, and in turn lower reliability.

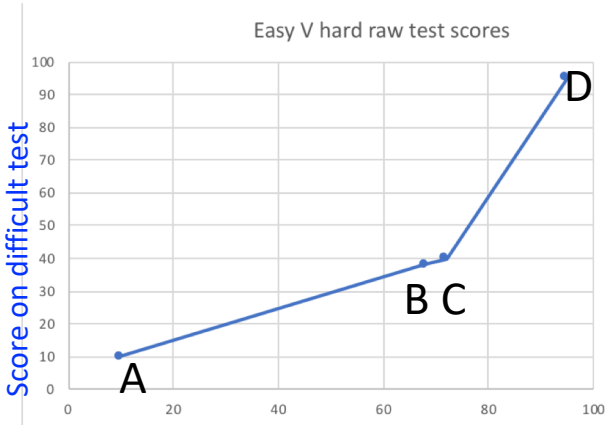
Equal slopes

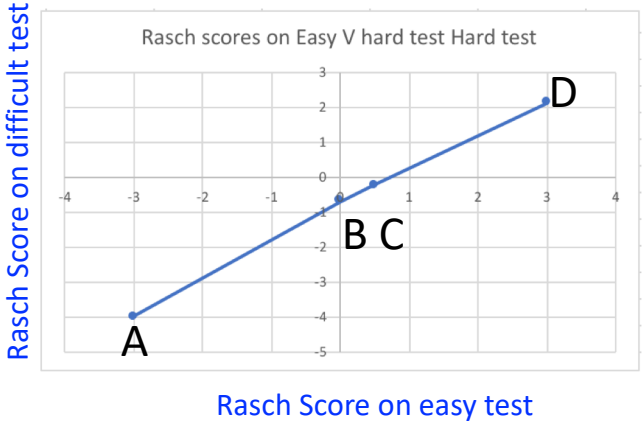


Unequal slopes



CTT V IRT : ordinal V interval





In individual raw scores, the ordering of students is the same above (ordinal), but the interval between students differs with item ability. On the easy questions, Person C's score is closer to Person D's score compared to a difficult item.

A Rasch score can be considered a transformation that preserves distances between respondents regardless of the difficulty of test items.

Measures of Model Fit

- Parameters estimates will never fit the data perfectly, so there is a need to determine how good the fit is.
- There are many fit metrics, none are perfect, so need to look at a few:
  - Residual based fit : how many standard deviations is each score from the expected score/mean (zscore of the residuals)

$$z_{ni} = \frac{(x_{ni} - \mathrm{E}(X_{ni}))}{\sqrt{Var(X_{ni})}}$$

x_{ni}: observed score for person n on item i

$E(X_{ni})$: expected score = p_{ni} for dichomomous items

$\sqrt{Var(X_{ni})}$:standard deviation where = $Var(X_{ni}) = p_{ni}$*(1- p_{ni}).

Variance is highest when item difficulty = person ability (point at which 50% of responses will be correct)

Easy and hard items have low variance:

p	p(1-p)
0.00	0.00
0.16	0.13
0.32	0.22
0.50	0.25
0.66	0.22
0.75	0.19
1.00	0.00

Measures of Model Fit

z_{ni} is a standardised metric for the residuals of the model fit. (The sum will be zero as some are +ive and some are –ive).

How well each person fits the model can be calculated from the sum of their residuals on each item squared, and scaled (mean).Simple scaling uses N (outfit – more dispersed from the mean); a weight mean weights each residual by the variance, meaning residuals of simple and hard items have less of an influence (infit – converges towards the mean).

Squaring z_{ni} over summing of all items is a fit index for person n:

$\frac{\sum_i z_{ni}^2}{I}$ or $\frac{\sum_i z_{ni}^2 Var(X_{ni})}{\sum_i Var(X_{ni})}$

Outfit Infit

How well an item fits the model can be calculated from the sum of the item residuals for each response, and scaled (mean).

Squaring z_{ni} over summing of all persons (n) is a fit index for item i:

$\frac{\sum_n z_{ni}^2}{N}$ or $\frac{\sum_n z_{ni}^2 Var(X_{ni})}{\sum_n Var(X_{ni})}$

Outfit Infit

N = number of respondents; I =number of items

R

- Train an IRT model on the maths dataset
- Plot the ICC
- Review item difficulty and ability estimates

- Fit statistics

2 parameter logistic IRT model (2PL)

- The basic RASCH model can be adapted to include the slope of the ICC curve as follows:
  - (The Rasch model assumes all ICC's have the same slope, convention is to assume a =1, but a can be any constant)

$$p = P(X = 1) = \frac{e^{(a(\theta-\delta))}}{1 + e^{(a(\theta-\delta))}}$$

θ is a "person" parameter: ability on the latent variable

δ is an "item" parameter: generally item difficulty

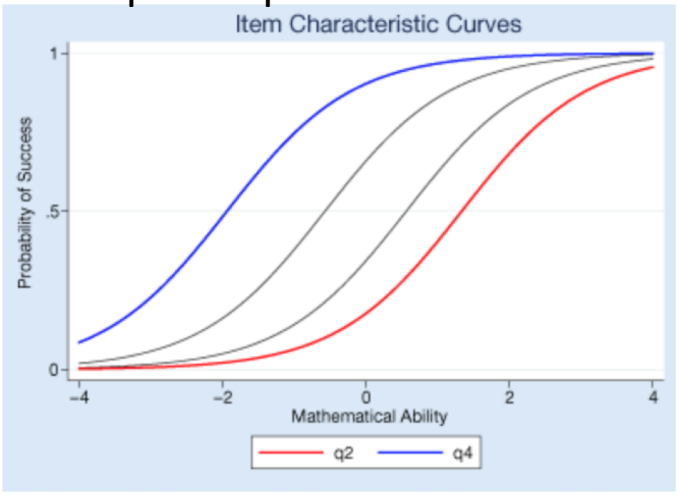
a determines the slope of the ICC

Note:
Small variations in slope can be tolerated in a Rasch model; considered as noise in the discrimination power of items.

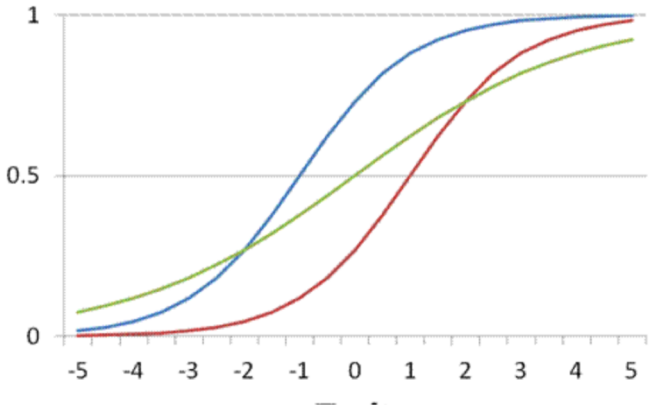
Ranking item difficulty with 1PL and 2PL

- When ICC curves do not cross (Rasch model), ordering items by difficulty is straight forward (blue item is easiest for all students; red item is hardest).
- However if slopes are not equal, ranking items by difficulty can be problematic. For example, the green ICC below is easiest for weaker students, but hardest for stronger students.

Equal slopes



Unequal slopes



R

- Train an 2PL IRT model on the maths dataset
- Plot the ICCs
- Review item difficulty and ability estimates

- Review fit statistics.
  - How has the fit of Item8 changed?

Other topics to explore

3PL IRT model

- In multiple choice questions, of say pick one of four answers, even the weakest student has a 25% chance of getting the answer correct.
- Therefore the lowest probability of a correct answer is 0.25 rather than 0.
- Pseudo guessing (c) is a third parameter in a 3PL model, it sets a minimum probability of a correct answer at a value > 0.
- A 3PL model allows

Item difficulty (mid point between min and max probability) is now:

$$\frac{1+c}{2}$$

$$p = P(X = 1) = c + (1 - c)\ \frac{e^{(a(\theta-\delta))}}{1 + e^{(a(\theta-\delta))}}$$

Incorporating other facets

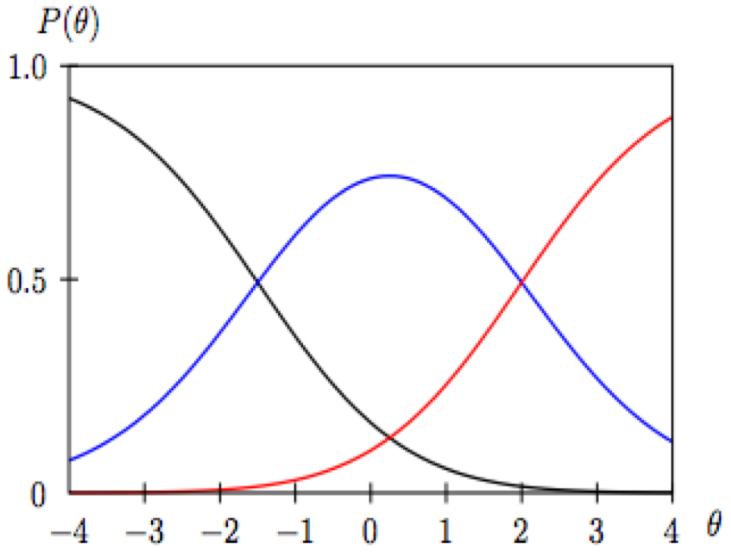
- Factors other than ability and item difficult may impact on a persons score, such as the leniency of the marker. Essentially such factors adjust the item difficulty, resulting in a IRT model that incorporates other facets (f):

$$p = P(X = 1) = \frac{e^{(\theta-(\delta+f))}}{1 + e^{(\theta-(\delta+f))}}$$

The more parameters you add, the greater the chance of overfitting the data.

Partial credit model

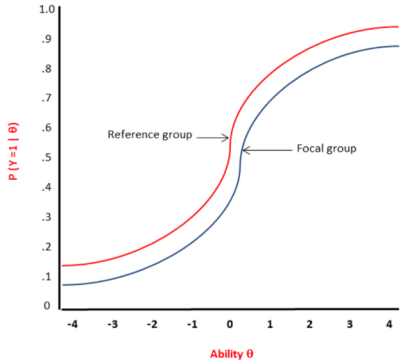
- Examples in this workshop were based on dichotomous scoring. To extend this to a scenario of multiple answers. The diagram below represents a question with three possible scores: 0, 1 or 2.



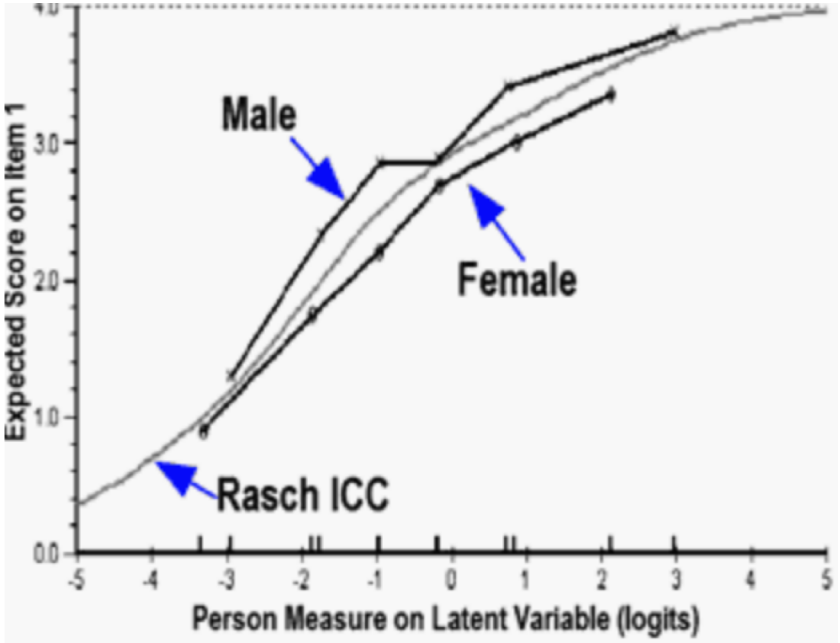
The black ICC represents the probability of getting a score of 0 (most likely scenario if ability < -2).
The blue ICC represents the probability of getting a score of 1. (most likely scenario if ability [-2,2]).
The red ICC represents the probability of getting a score of 2. (most likely scenario if ability > 2)

As each are mutually exclusive, the sum of each probability at any ability level of is 1.

Differential Item function



- Student responses may be influenced by context.
  - For example boys may click to a maths question framed around baseball scores than girls
- Differential Item function looks at how consistent responses to an item are across groups of the same ability.
- This can be looked at visually, or objectively as a chi-squared statistic



Scores at ability k for item i:	1 (correct)	0 (incorrect)
Group 1	A_{ik}	B_{ik}
Group 2	C_{ik}	D_{ik}

$$\chi^2_{MH} = \frac{\sum_k A_{ik} D_{ik} / T_{ik}}{\sum_k B_{ik} C_{ik} / T_{ik}}$$

Where T_{ik} is total responses

Information function of a test item

- The information of a test item is the product of ability and difficulty.
  - Hard items given little information; easy items give little information; best items have a difficulty level similar to the person ability.

- I(item(θ, δ)) = $\theta * \delta$.
  - Max value is 0.5*05 = 0.25.
- The information value of a test or a particular ability level is the sum of the values of the individual items information at that ability level.

To ponder:

"The use of standardised tests has also been strongly criticised. First, most tests do not provide information on what a student has learned, only how he/she stands relative to other students. Secondly, tests put pressure on teachers to teach to the test, leading to a narrowing of the curriculum. Thirdly, tests encourage a competitive atmosphere in the classroom. Fourthly, when standardised test results are used to select and classify students, they lead to labelling, which, in turn may be associated with the perpetuation of distinctions based on race, gender, or socioeconomic status. Even when not consciously used to classify students, it has been argued that test information can influence teachers' expectations"

- http://www.erc.ie/documents/standardised_testing_lowersecondary_education.pdf

CTT V IRT

- A study based on a large sample size found CTT and IRT item discrimination and student ability estimates were comparable.
- So when dealing with a single test, item difficulty estimates and item discrimination index from CTT is adequate for selection of test items.
- IRT is better than CTT when multiple tests need to be compared, e.g. monitoring trends over time.
- IRT is better if there is a need to link ability levels with item difficulty

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, *58*(3), 357-381.

Is it appropriate to conduct experiments in an educational setting – and with what level of transparency?

- http://blogs.edweek.org/edweek/edtechresearcher/2018/04/ben_harold_an_ed_week.html, April 2018

Appendix 1-Rasch model: logit derivation

$$p = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}}$$

$$1-p = 1-\frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} = \frac{1+e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} - \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} = \frac{1}{1+e^{(\theta-\delta)}}$$

So:

$$\frac{p}{1-p} = \frac{\frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}}}{\frac{1}{1+e^{(\theta-\delta)}}} = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} * 1+e^{(\theta-\delta)} = e^{(\theta-\delta)}$$

$$odds = \frac{P(sucess)}{P(failure)} = \frac{p}{1-p}$$

Therefore:

$$\ln\left(\frac{p}{1-p}\right) = \theta - \delta$$

Changing ln to log of anther base just changes the scale

This is referred to as the *"log of odds unit"*, shortened to logit