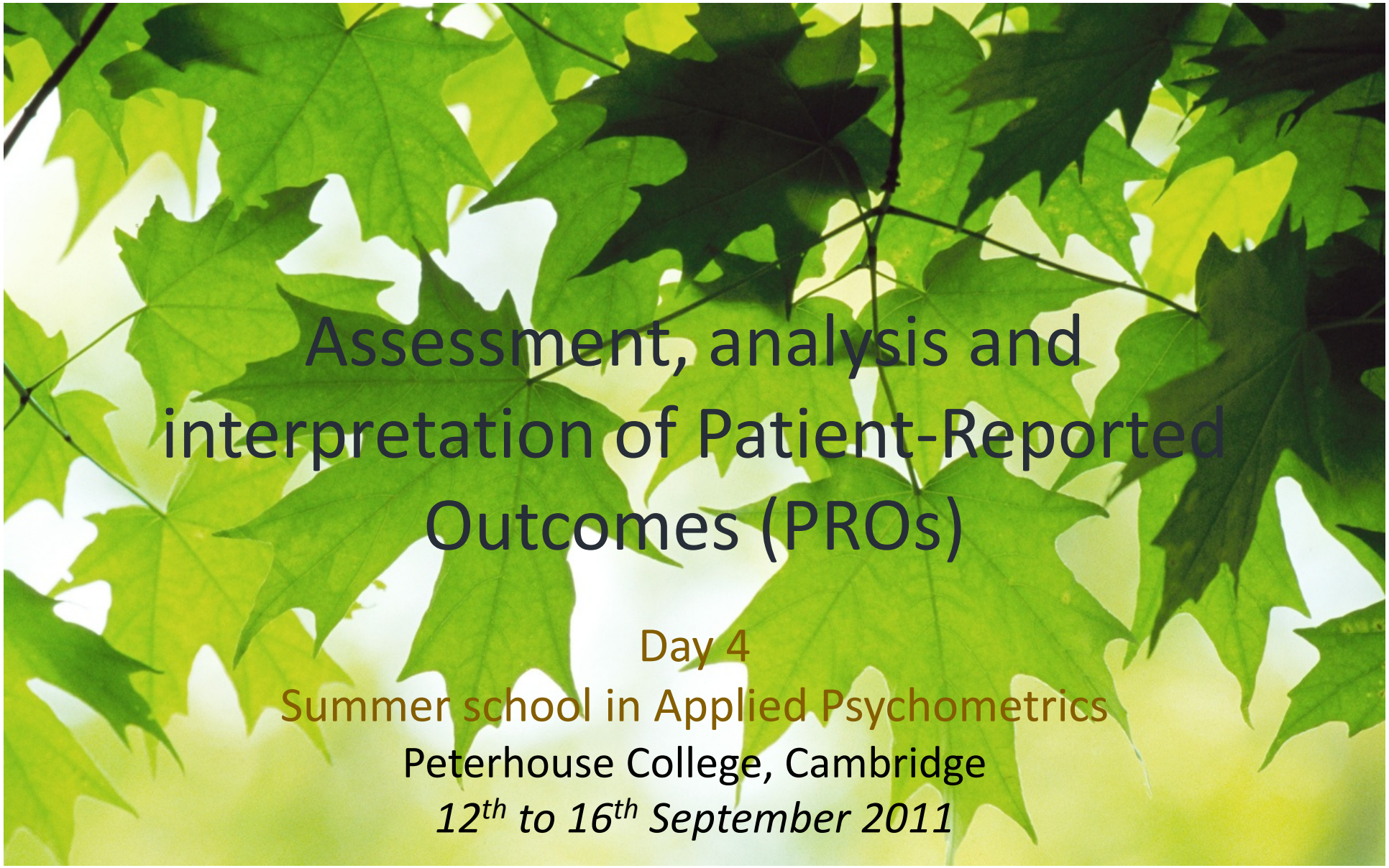


Assessment, analysis and interpretation of Patient-Reported Outcomes (PROs)

Day 4

Summer school in Applied Psychometrics

Peterhouse College, Cambridge

12th to 16th September 2011

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

This course is prepared by

Anna Brown, PhD ab936@medschl.cam.ac.uk
Jan Stochl, PhD js883@cam.ac.uk
Tim Croudace, PhD tjc39@cam.ac.uk
(University of Cambridge, department of Psychiatry)
Jan Boehnke, PhD boehnke@uni-trier.de
(University of Trier, Department of Clinical Psychology and Psychotherapy)

The course is funded by the ESRC RDI and hosted by
The Psychometrics Centre

Anna Brown

10. TEST SCORING - CLASSICAL AND IRT METHODS

Test scoring

- The ultimate goal of measurement is to produce a score by which individuals can be assessed and differentiated
- Such a score has to approximate the latent trait well
- Its precision must be known in order to make correct judgements (significance of clinical change, significance of difference etc.)
- *Thissen, D. & Wainer, H. (2001). Test scoring*. - is a very good book

Sum score

- Test scoring is always about estimating true score
- Classical Test Theory tells us that $X = T + E$
- Summated test score
  - Infinite number of repetitions (parallel tests) will give expected true score
  - Unbiased estimate
  - Any administration of the test, however, inevitably has an error component
    - A very high score will probably not be so high next time, and a very low score will not be so low
    - Regression to the mean is often observed
    - Scores regress more if they are unreliable

Kelley's equation

- Can we do better in estimating the true score using the simple item scores?
  - Regression between scores on two parallel forms of the test

$$\frac{\hat{x}-\mu}{\sigma}=\rho_{xx'}\left(\frac{x'-\mu}{\sigma}\right) \qquad \text{or,} \qquad \hat{x}=\rho_{xx'}x'+\left(1-\rho_{xx'}\right)\mu$$

  - And because the observed score is true plus error (and error is unrelated to x')

$$\hat{\tau}=\hat{\rho}_{xx'}x'+\left(1-\hat{\rho}_{xx'}\right)\hat{\mu}$$

  - Due to Truman Kelley (1927); this is a more precise but biased estimate of true score.
  - Makes no distributional assumptions

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Bayesian approach

$$\mathrm{P}\left(\tau \mid x'\right) = \frac{\mathrm{P}\left(x' \mid \tau\right)\mathrm{P}\left(\tau\right)}{\mathrm{P}\left(x'\right)}$$

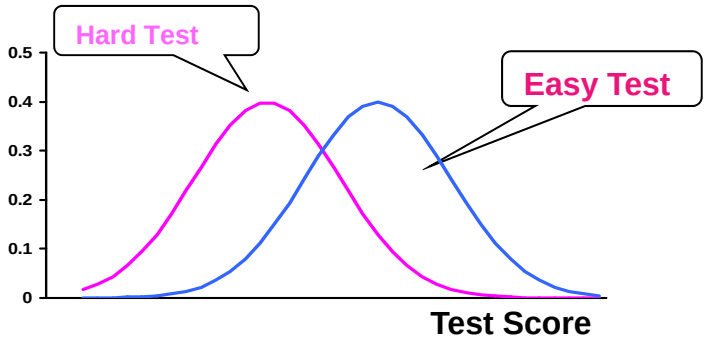
- Making use of known distributions of scores
  - We have beliefs about the distribution of true score in the population (prior distribution)
  - Gaussian with mean μ and variance $\sigma_T^2 = \rho_{xx'}\sigma_x^2$
- We also have a theory about the observed score *given the true score* τ
  - Error variance component only
  - Gaussian with mean τ and variance $\sigma_E^2 = \left(1 - \rho_{xx'}\right)\sigma_x^2$
- We know that the observed score is distributed (posterior distribution) as Gaussian with mean $\overline{x}$ and variance σ_x^2
- Using simple rule for computing the mean of resulting distribution as mean of component distributions weighted by inverses of variances, we obtain Kelley's equation again

$$\overline{x} = \rho_{xx'}x' + \left(1 - \rho_{xx'}\right)\mu$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Limitations of CTT scoring

- In CTT, respondents' scores are item dependent.
  - Properties of test items are nuisance variables that escape standardisation
- True score is fully determined by the test as designed.
  - True score only has meaning conditional on standardised error variables
- Errors of measurement are absorbed by scores.
  - Many spurious effects are due to error variance affecting results in non-random way



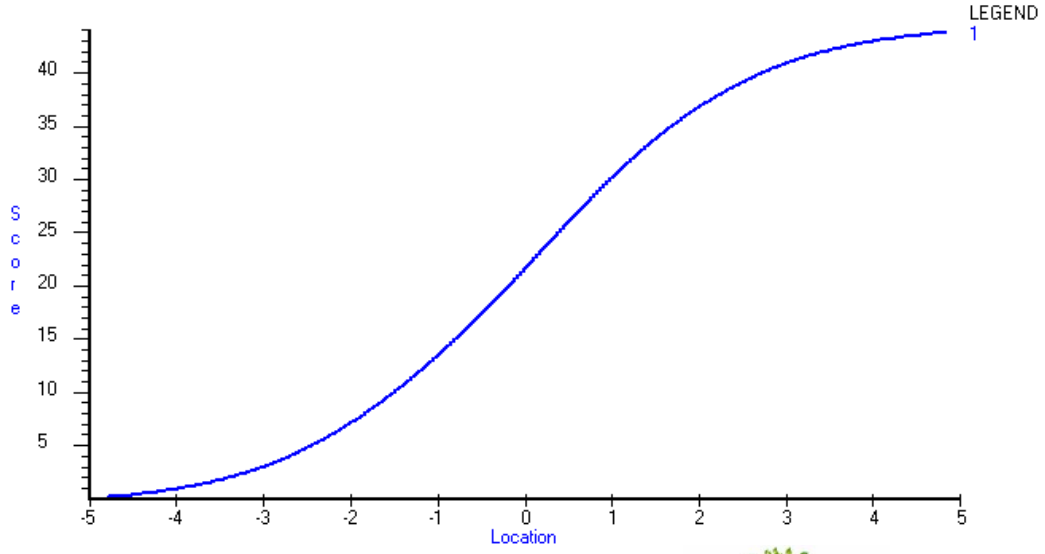
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Limitations of CTT scoring - scaling

- Measurement requires interval scale, so that scores can be added, subtracted etc.
- Strictly speaking, sum scores of ordinal items are ordinal
- When measuring change is needed, simple sum scores are not recommended:
  - reliability of the change measure is inversely related to the correlation between the tests,
  - change may be not measured on the same scale for persons with different scores on the original measure,
  - spurious negative correlations between the baseline and the change score.

True and summated score

- Probability of keyed item response is also an item score expectation over infinite number of occasions, given a trait score
- Test characteristic curve gives the expected relationship between the summated score and the latent trait score
  - This relationship is not linear



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Limitations of CTT scoring – other

- Item statistics are sample dependent.
- The common estimate of measurement error (SEm) is group-based.
- Modeling of data is at the test score level (X=T+E) but item level modeling is needed for flexibility of use
  - item banks
  - computer-adaptive tests
  - improved score reporting, and more…

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

IRT scores have desirable features:

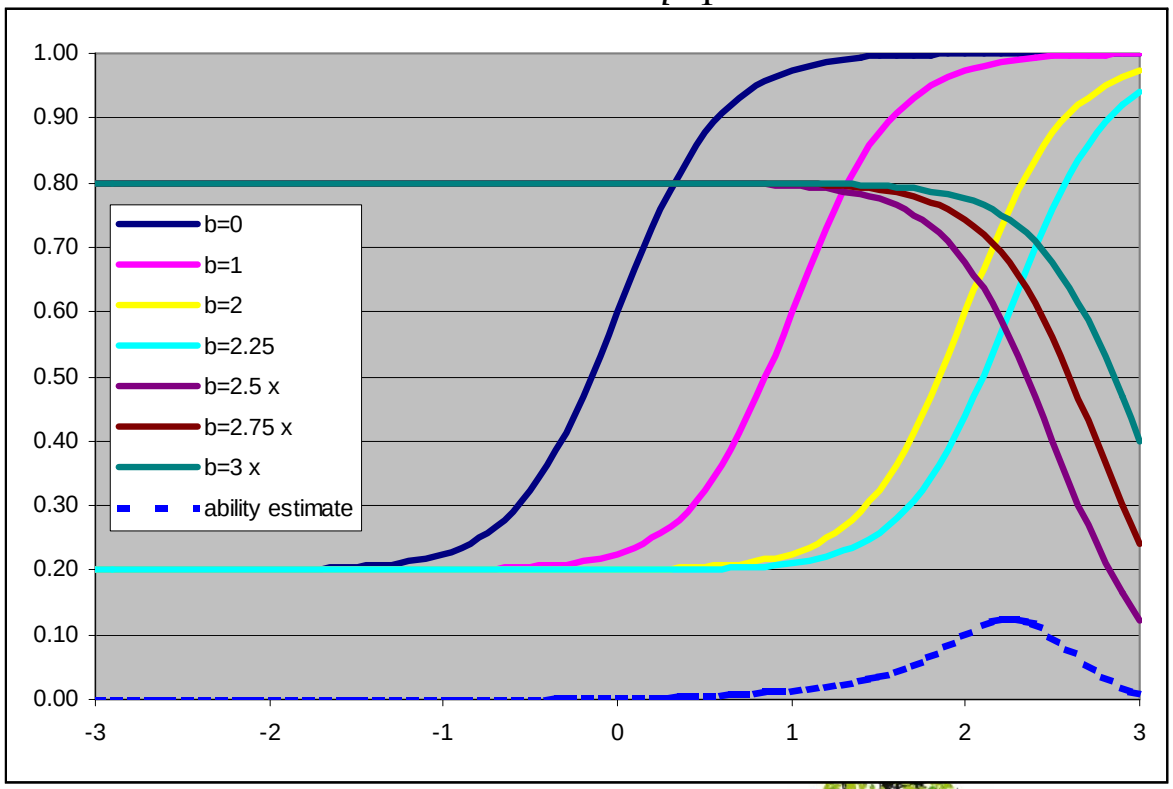
- Items and Persons are on the same scale
  - Particularly helpful in test design and score reporting
- Person parameter invariance
  - *Person scores are independent of the particular test items* - critical in computer adaptive testing and randomised testing
- Item parameter invariance
  - *Item Parameters are independent of the persons used to calibrate them [within a linear transformation]* - useful in field-testing and item banking

IRT scoring basics

- In routine applications of tests item parameters will be known (calibrated during standardisation)
- Given individual pattern of item responses, probabilities of responses will depend only on the latent trait
- Assuming responses are independent after controlling for the latent trait, the joint probability of the response pattern equals the product of probabilities of responses to individual items

Probabilities of responses to several items

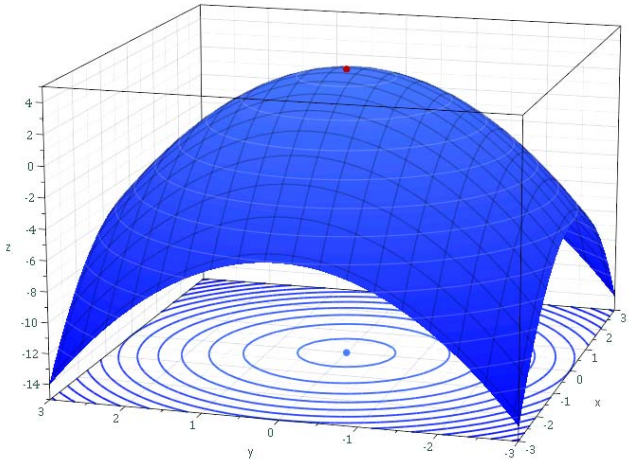
$$L(u_1,u_2,\ldots,u_p\,|\,\theta)=\prod_{i=1}^{p}P_i^{\,u_i}Q_i^{\,1-u_i}$$

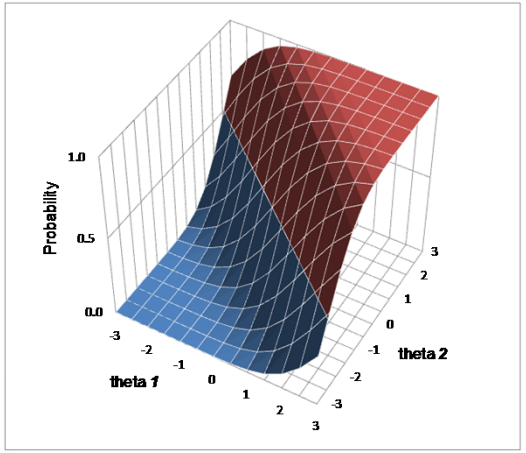


UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Maximum likelihood estimation

- Maximising the likelihood function (iterative process)
- Finding a score (or a set of theta scores for multidimensional model) that maximises the likelihood of observed response
  - ML estimator is unbiased, and its errors are normally distributed
  - Problems with ML is that convergence is not guaranteed with aberrant responses, and no estimator exists for “perfect” response patterns





UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Interval measurement

- IRT-estimated scores can be considered interval scale*
  - the constrained iterative estimation process will yield something close to interval measurement when the structures underlying the data are consistent with the model
- IRT-estimated scores can be used to compute difference scores
  - Between measurement occasions
  - Difference scores are free of problems associated with classical scoring
  - Meaning of difference scores between persons can be tricky to interpret

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Bayesian estimation

- Based on the Bayes theorem

$$\mathrm{P}(\boldsymbol{\theta} \mid \mathbf{u}) = \frac{\mathrm{P}(\mathbf{u} \mid \boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{u})} = \mathrm{P}(\mathbf{u} \mid \boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta})$$

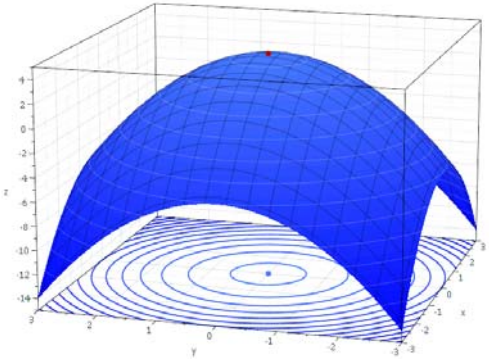
- Likelihood of a specific latent score given the observed response pattern (posterior distribution) is computed from
  - the likelihood of the observed response pattern given the true score
  - known proportion or density of the score in the population (prior distribution, usually standard normal).
- Make use of information about the score distribution

Expected A Posteriori (EAP)

- Maximises the mean of the posterior distribution

$$\mathrm{EAP}\left(\mathbf{u}\middle|\theta\right) \approx \frac{\sum_q \theta_q \left[\prod_{i=1}^{m} P_i\left(u_i\right)\right] \phi\left(\theta_q\right) d\theta_q}{\sum_q \left[\prod_{i=1}^{m} P_i\left(u_i\right)\right] \phi\left(\theta_q\right) d\theta_q}$$

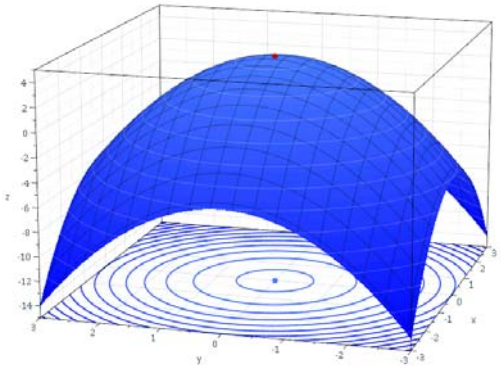
- Non-iterative
- Excellent choice for one-dimensional models
  - Requires numerical integration therefore is computationally demanding with more than one dimension
- Properties
  - Exists for all response patterns
  - Better precision than ML
  - Biased towards the population mean



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Maximum A Posteriori (MAP)

- Maximises the mode of the posterior distribution
- Iterative process
- Excellent choice for multidimensional models
  - Searching for all traits simultaneously
  - Number of iterations is unaffected by the number of dimensions
- Properties
  - Exists for all response patterns
  - Better precision than ML
  - Biased towards the population mode



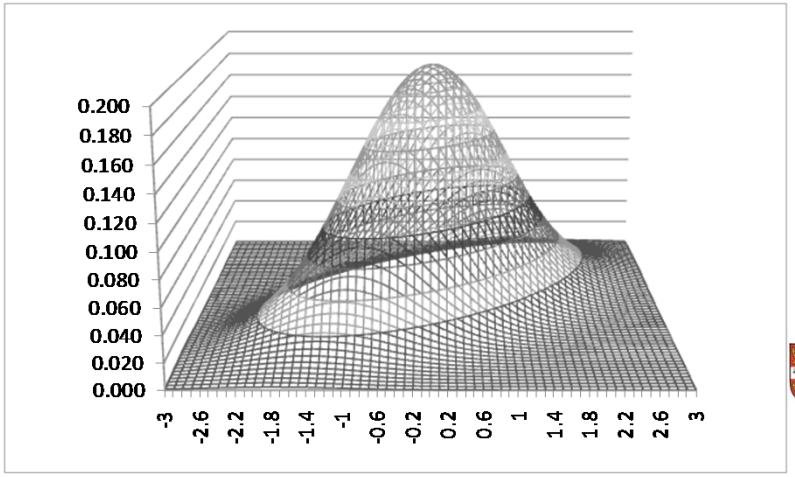
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Practical – IRT univariate score estimation

- We are going to estimate IRT scores for one GHQ subscale using R software
  - Compute summated scores for one subscale that we will use for comparison with IRT scores
  - Estimate item parameters using Graded Response Model (grm function) for one subscale
  - Then estimate persons' scores using Bayesian approaches (with standard normal as prior)
    - Both EAP and MAP
  - Plot the obtained IRT scores against summated scores and against each other

Benefits of multivariate prior

- Even with an independent-cluster structure (technically one-dimensional IRT), one can use multivariate Bayesian score estimation
  - it uses prior information about scale correlations
  - helps obtain more precise scores when scales share some variance
  - "borrowing strength"
- We will be able to quantify how much more precise this estimation is when discuss measurement precision in IRT



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Approximating an IRT score

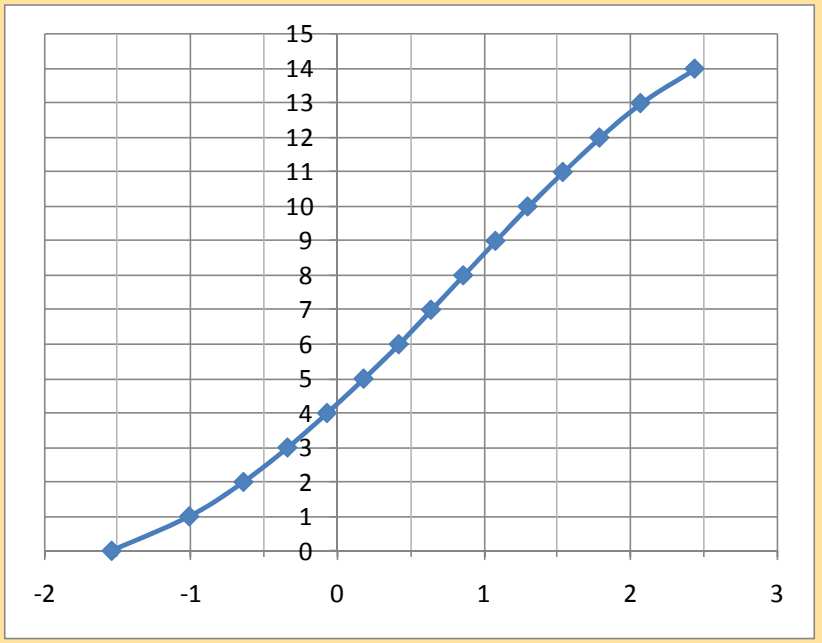
- What if the full IRT estimation is impractical in some applications?
  - Score is needed there and then
  - Scoring is done by people unskilled in IRT
- Summated score could be used to approximate the IRT score
  - described by Thissen, Nelson and Swygert (2001)
  - We know the relationship between the expected summated and the true score

Approximating an IRT score – cont.

- Each response pattern can be associated with:
  1. corresponding summed score (number of '**points**')
  2. corresponding IRT theta score, estimated with, say, the EAP estimator.
- Every unique response pattern will have only one sum score and only one EAP score associated with it.
- A particular sum score, on the other hand, might be a result of several response patterns.
- Likelihood of observing a particular sum score is sum of likelihoods of observing all possible response patterns leading to that sum score
- Conversion tables from the sum scores to the expected IRT scores together with their standard errors can be rapidly built using the freely available software IRT*Score* (Flora & Thissen, 2002).

Practical – approximated IRT scores

- We are going to use IRT*Score* software to produce approximated theta scores from sum scores for the Anxiety facet of GHQ
- Will need input files of item parameters (take from our previous practical using R software)
- Will obtain simple-to-use conversion tables



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Conclusions

- There are multiple options for scoring
- Summated score is a reasonable proxy for true score
  - preserves the order of respondents quite well, but distorts the interval properties of the scale
- IRT approach produces a better proxy for true score
  - attempts to produce the interval measurement scale
  - takes to account item-dependent nuisance factors
- Both approaches have Bayesian alternatives
- IRT scores have another advantage - conditional estimates of standard error
  - can be only understood after considering how precision of measurement is addressed in IRT