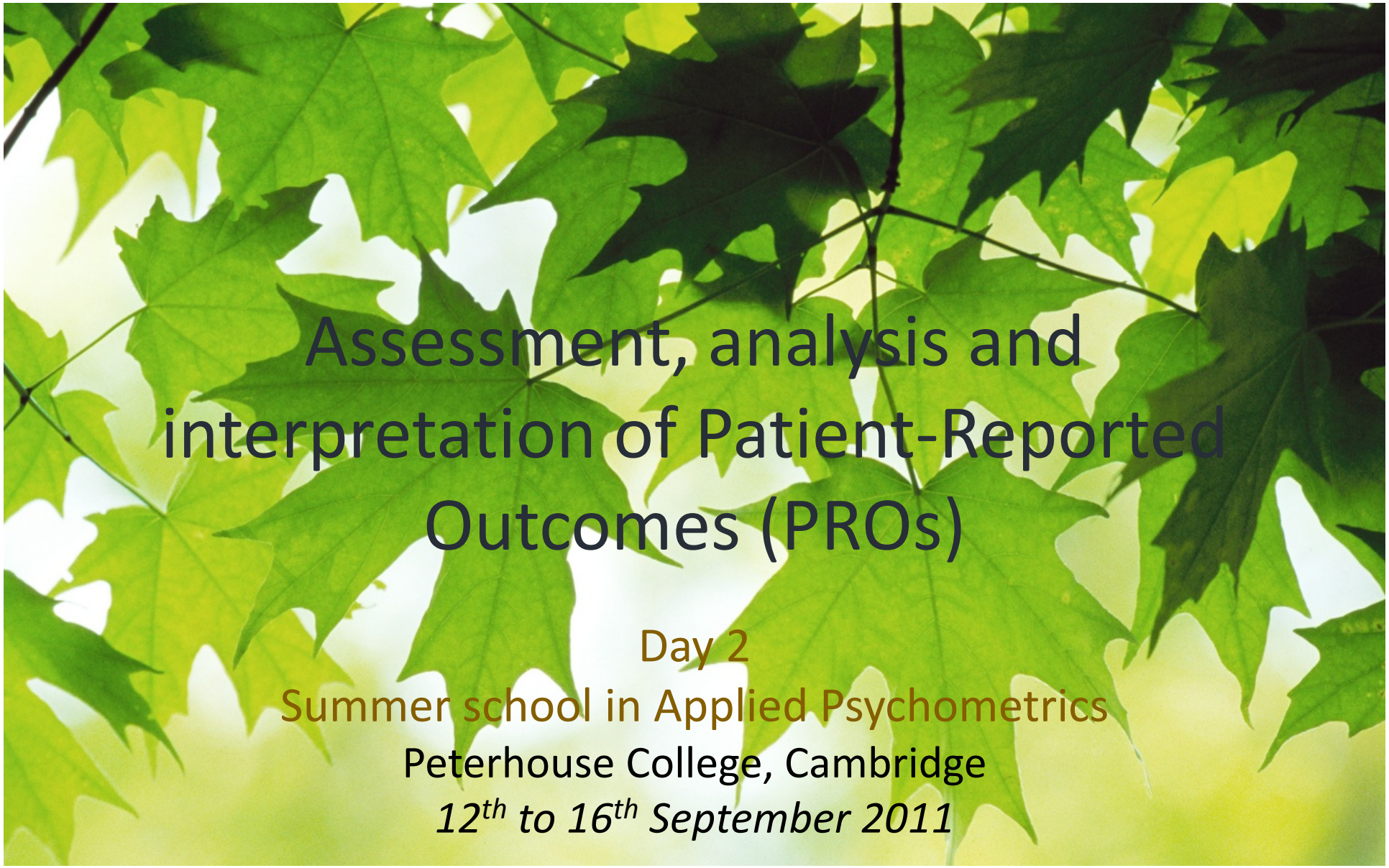


Assessment, analysis and interpretation of Patient-Reported Outcomes (PROs)

Day 2

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

5. STRUCTURE AND QUALITY OF SCALES

The construct validity question

- A central issue in PRO measurement is whether obtained scores represent the measured trait (e.g., severity of depression).
- The empirical question is whether the relationships among scale items can be explained by a single underlying trait (e.g., depression),
  - and are thus unidimensional,
  - or form sub-scales to operationalize the trait's multidimensional structure.
    - (Gibbons, Immekus and Bock, 2007)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Test homogeneity

- Homogeneous test is the test whose items are all of the same kind – they measure something in common
- That "something" is an attribute we want to measure
- Quantity measured in common is a *common factor*, F
- Not all items are equally good indicators of F

Homogeneity and symptom clusters

- Homogeneity means that all items are parallel indicators of the latent attribute
- In QoL measures, some symptoms may form clusters related to a common cause
  - E.g. chemotherapy in cancer causes nausea and hair loss. These are highly correlated but might be unrelated to each other in terms of QoL
- Care is needed to identify such items and treat them appropriately from the validity perspective

Validity assessment

- Scale development
  - Expect reasonable correlations between items in the scale, and consequently between the item and its own scale
  - During development, correlation between item and its scale should be 0.3 or greater
  - At the final stages, more stringent criterion of 0.4 could be applied
  - Discriminant validity is supported if the item correlates higher with its scale than with other scales
- Construct validity of multiple scales
  - Convergent and discriminant validity

Correlation coefficients

- Relationships between items, and between items and the scale are usually described by correlations
- Sample size required for classical methods is at least 100, important to check for restriction of range
- Here are some correlation coefficients widely used in *classical* analysis of questionnaire data
  - Pearson's correlation coefficient assumes normally distributed variables
  - Spearman's rank correlation is more suitable for ordered categorical variables
  - Biserial coefficient is suitable for estimating correlation between continuous and binary variable (with underlying normal distribution assumed)

Common factor model

- The appropriate model is the common factor model of Spearman

$$Y_i = \mu_i + \lambda_i F + E_i$$

  - Constants μ allow each item to have unique difficulty
  - Factor loadings λ reflect the difference in response corresponding to difference in attribute – therefore it is a measure of item discrimination
  - Special case is true-score equivalent items

$$(\lambda_l = \lambda_k = \lambda)$$

Dimensionality and Factor analysis

- Unidimensionality is another name for homogeneity
- We can check if items are of homogeneous content – measure just one attribute in common – by checking if responses fit the single factor model
  - This can be done after model parameters are estimated
  - We can compute discrepancies between observed and expected covariances, and summarize them (for example) as ordinary mean of squared differences (ULS function)
  - We can examine overall fit, and any areas of misfit (violating covariances)
- Factor analysis is concerned with structure of correlations (between items or scales)

Reliability of a homogeneous test

- The total summated test score is

$$Y = \Sigma\mu_i + (\Sigma\lambda_i)F + \Sigma E_i$$

- Variance of test score can be partitioned

$$\sigma_Y^2 = (\Sigma\lambda_i)^2 + \Sigma\,\sigma_{Ei}^2$$

- Reliability is defined as
  - *squared correlation between the observed and the true score*
  - or, as proportion of variance due to true score
- Coefficient omega (Browne, 1975)

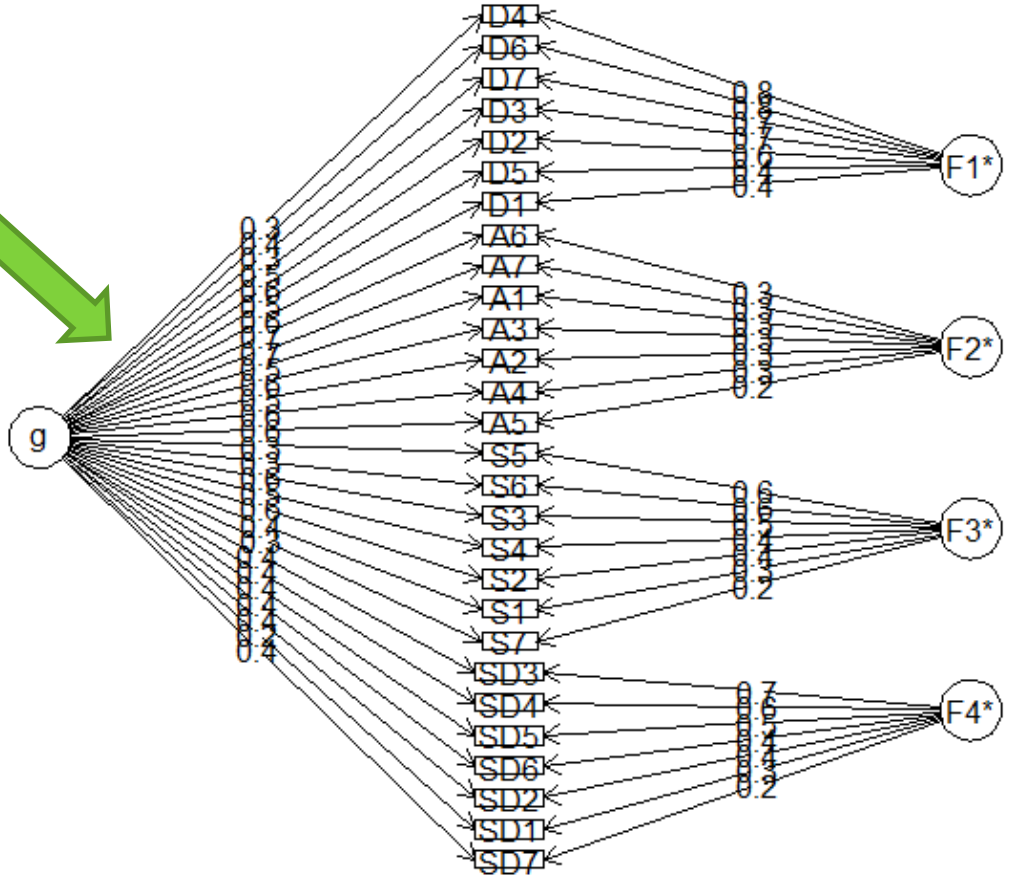
$$\omega = (\Sigma\lambda_i)^2 / \sigma_Y^2$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Coefficient *omega*

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_{gi}\right)^2}{\mathrm{var}(y)}$$

- Based on the common factor model
- Works whether one or more factors are present
- λ_{gi} are factor loadings on the general factor



Coefficient *alpha*

- For true-score equivalent items the expression for coefficient omega reduces to

$$\omega_\tau = m^2 \lambda^2 / \sigma_Y^2 = m^2 \sigma_T^2 / \sigma_Y^2$$

  - where σ_T^2 is covariance between any two items
- Sample-based estimates of population values

$$\hat{}\sigma_T^2 = (\Sigma\Sigma s_{ik} - \Sigma s_{ii})/[m(m-1)]$$

$$\hat{}\sigma_Y^2 = \Sigma\Sigma s_{ik}$$

- we obtain coefficient alpha

$$\omega_\tau = [m/(m-1)]*(1 - \Sigma s_{ii}/ \Sigma\Sigma s_{ik}) = \alpha$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Do not misuse alpha

- In a unidimensional test, adding items usually increases alpha
- Alpha has been used as measure of "internal consistency" (homogeneity)
- Alpha itself provides no evidence of homogeneity
  - Applying alpha itself requires a lot of assumptions
  - Such as that the single-factor model holds
  - It is an estimate of reliability for true-score equivalent items
  - Computing alpha is no substitute for dimensionality assessment
  - Dimensionality should be always assessed first

Applying alpha and omega

- It can be shown that alpha is a lower bound to omega

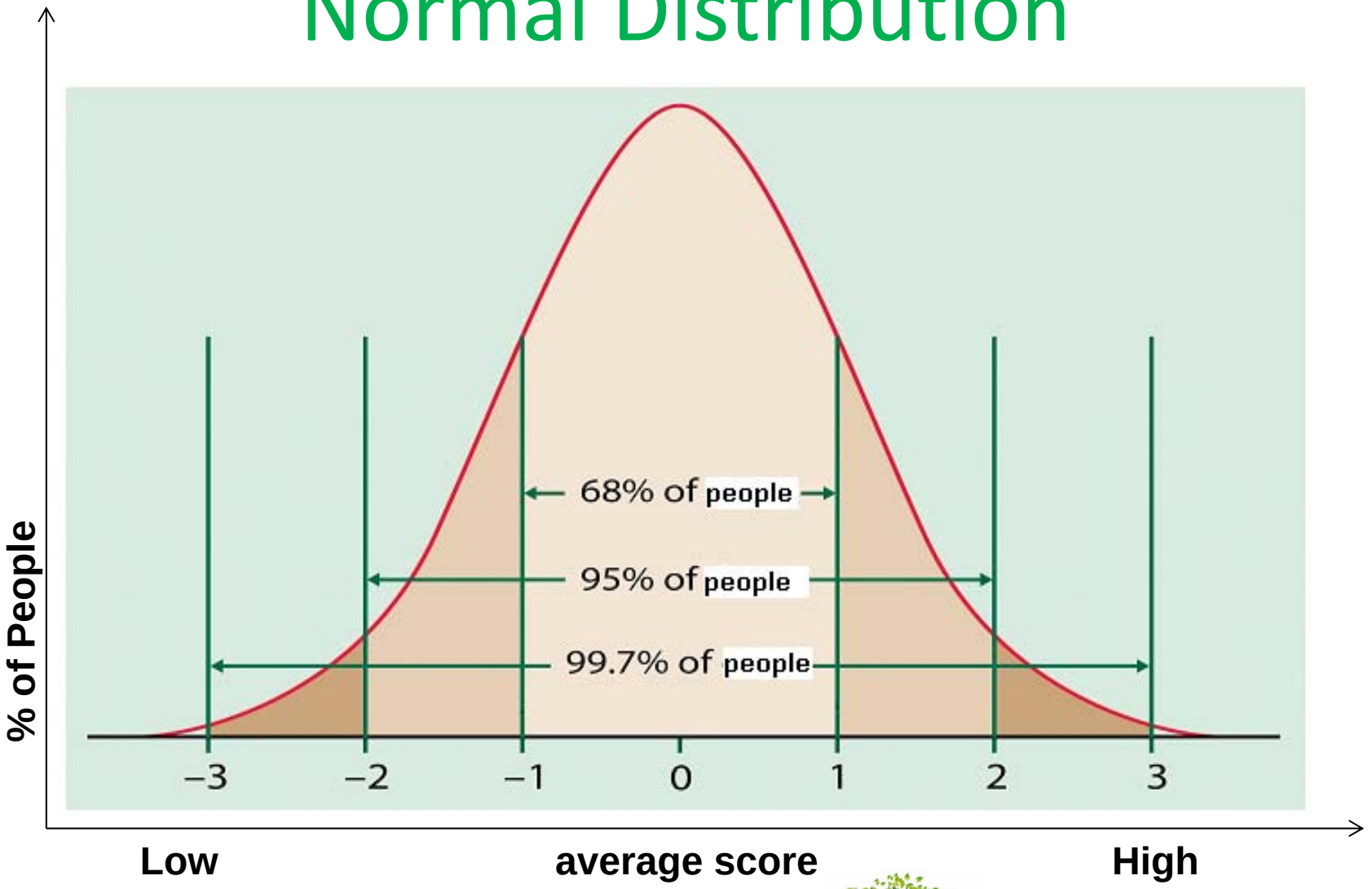
$$\omega_\tau \geq \alpha$$

- If unidimensionality holds in sum-scored test
  - alpha can be used as a crude tool to estimate the lower bound to reliability
  - If the factor model is fitted, omega is a better measure of reliability
- Guidelines to "sufficient" alpha or omega are meaningless without understanding what precision of measurement is required in a particular application
  - SE of measurement is more informative

Binary (and ordinal) items

- All the above calculations hold for the linear factor model
  - That is, relationships between expected item and test score are assumed linear
  - These relationships are not linear for binary items (and for items with small number of response categories)
- To overcome this problem, it is assumed that item scores are observed manifestations of latent response tendencies
  - Continuous and assumed normally distributed
  - Item responses relate to response tendencies through threshold process
- Correlations between items
  - Tetrachoric correlation is suitable for two binary variables (with underlying normal distribution assumed)
  - Polychoric correlation is suitable for two ordinal variables (with underlying normal distribution assumed)

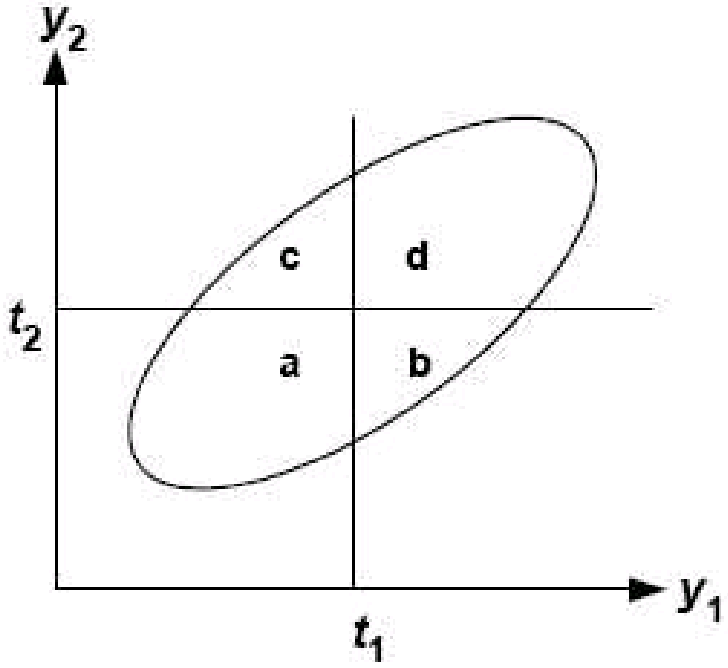
Normal Distribution



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Tetrachoric correlation

	-	+	y1
-	a	b	a + b
+	c	d	c + d
y2	a + c	b + d	1



- Assumes that two normally distributed variables have been dichotomised using threshold process
- Appropriate assumption for item responses, where response tendency is unobserved, only its dichotomisation is observed

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Limitations of classical scale assessment

- Classical test theory assumes
  1. linear relationships between items and attributes
  - Relationships between binary (and ordinal) items and attributes are not linear
  - Linear model is an approximation only (still a useful summary, as summation of items tends to cancel the effects of non-linearity)
  2. independence of true and error variance
  - In general, the error variance of the total test score from a set of binary items cannot be independent of the test score