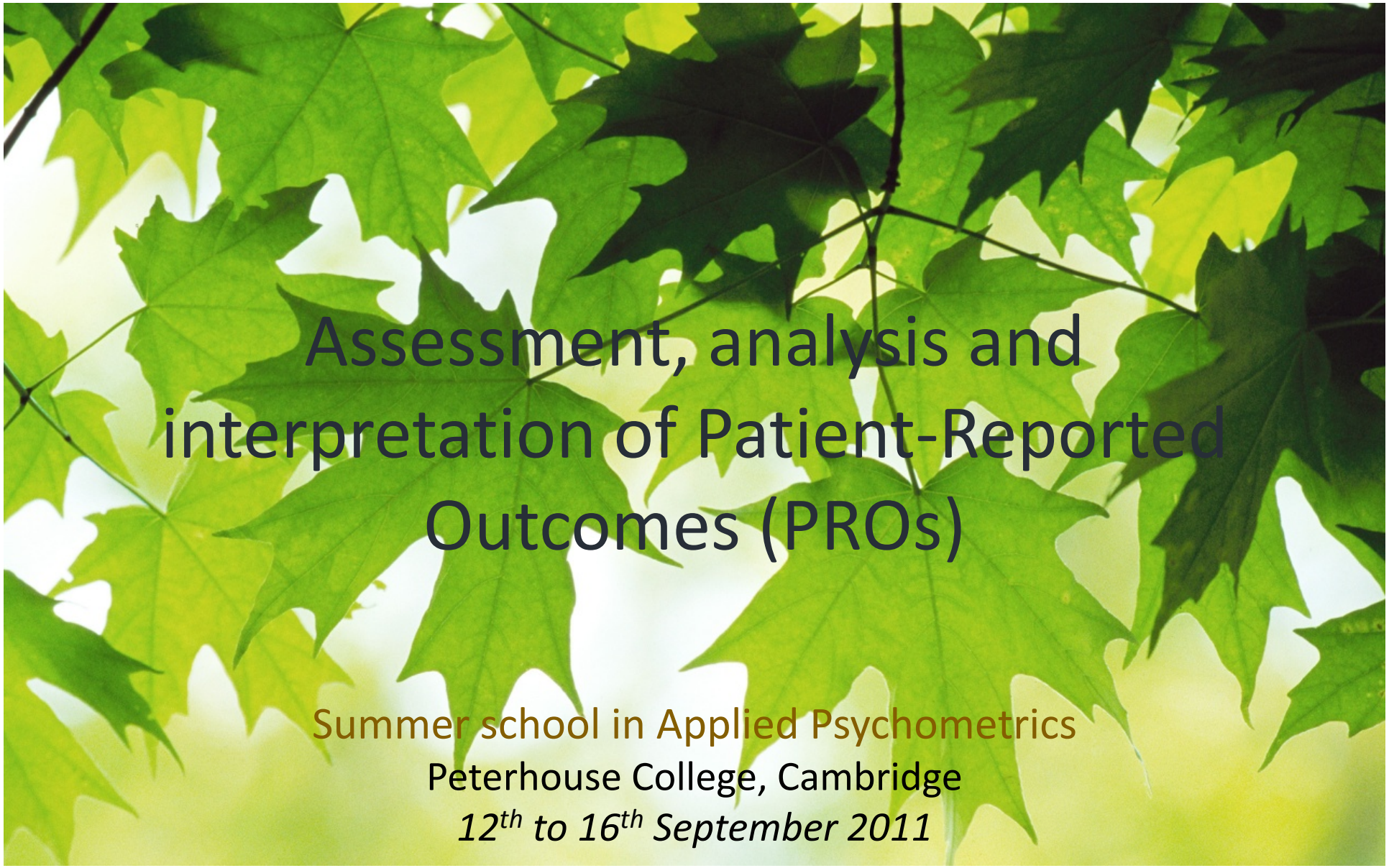

Assessment, analysis and interpretation of Patient-Reported Outcomes (PROs)

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

Jan Stochl

2. VALIDITY, RELIABILITY, SENSITIVITY, GENERALISABILITY

Content

- Validity
- Reliability
- Sensitivity
- Relationships between validity and reliability
- Generalizability
- Practicals:
  - Reliability estimation in R
  - Generalizability with Genova and mGenova

Validity

- “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.” (APA definition)

- Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick, 1989).

Test validity

- Classical model of validity:
  - Content, criterion (concurrent, predictive), and construct validity

- Modern model (Messick, 1995) of validity as single concept with 6 aspects:
  - content, substantive, structural, generalizability, external, and consequential aspects.

Classical model

- Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct.

- Criterion validity compares the test with other measures or outcomes (the criteria) already held to be valid.

- Construct validity evidence involves the empirical and theoretical support for the interpretation of the construct.

Modern model : Aspects of construct validity

- The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality

- The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance, along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks

- The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Modern model : Aspects of construct validity

- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks, including validity generalization of test criterion relationships

- The external aspect includes convergent and discriminant evidence from multitrait - multimethod comparisons, as well as evidence of criterion relevance and applied utility

- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice.

Sources of *in*validity

- *Construct underrepresentation* - the assessment is too narrow and fails to include important dimensions or facets of the construct.

- *Construct-irrelevant variance,* the assessment is too broad, containing excess reliable variance associated with other distinct constructs as well as method variance such as response sets or guessing propensities that affects responses in a manner irrelevant to the interpreted construct.

Sources of evidence in construct validity

- Historically - internal and external test structures
- Studies of expected performance differences over time, across groups and settings, and in response to experimental treatments and manipulations.
- Modelling of the processes underlying test responses
- Content relevance and representativeness as well as criterion-relatedness.
- Social consequences of test interpretation and use

Psychometric assessment of validity

- Modelling approach

- Construct validity:
  - when items are continuous - factor analysis (FA) – focus is on factor loadings

Measurement bias

- *Selection bias* exists if some potential subjects are more likely than others to be selected for the study sample.
  - telephone surveys, volunteering, nonresponse bias, or dropout in longitudinal studies

- *Information bias* may enter the study through the methods used to collect and record data
  - Interviewer bias, social desirability

Reliability

- A reliable test is one that measures something in a consistent, repeatable and reproducible manner.
  - Example: if a patient's QoL were to remain stable over time, a reliable test would be one that would give very similar scores on each measurement occasion

- Reliability is a property of the *scores of a measure* rather than the measure itself

- Two approaches to reliability: classical test theory view (we cover this today) and Item Response Theory view (covered later)

Reliability – Classical test theory view

- Reliability of a test is a single index
- observed score ***x*** = **true** score τ + error **ε**
- Definition:

$$\text{reliability} = \frac{\text{true score variance}}{\text{observed score variance}} = \frac{\text{var}(\tau)}{\text{var}(\tau) + \text{var}(\varepsilon)}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Reliability classifications

- Multiple occasions reliability, parallel forms reliability, Internal-consistency, inter-rater reliability
  - Multiple occasion: test-retest
  - Parallel forms: Split-half
  - Inter-rater reliability: Cohen´s kappa, Intraclass correlation coefficient
  - Internal consistency: Cronbach´s alpha, McDonald´s omega

- Specific (for items) reliability and generic reliability (for constructs)
  - Specific - methods like test-retest, inter-rater reliability
  - Generic – methods like internal consistency , KR-20, split-half reliability

Multiple occasion reliability (test-retest)

- Estimates reliability as correlation between the same items administered (at least) twice
- Usually computed as correlation between two administrations of the same item (or test)
- It is necessary to be aware of assumptions of this method: measured trait has not changed over occasions

Parallel forms reliability

- This reliability is estimated as correlation between 2 parallel sets of items measuring the same construct and
  - Parallel means that $\tau_{ip} = \tau_{jp}$ for all *i* and *j* and $Var(\varepsilon_i) = Var(\varepsilon_j)$ for all *i* and *j*
- Parallel forms can be created from pool of items that are believed to be parallel.
- Also, parallel forms can be created for example by random splitting existing (unidimensional) test into 2 halves. In this case we talk about split-half reliability.
- Problematic feature of split-half method is that many different split-halves of single test can be created.

Inter-rater reliability

- Relevant if there are raters or observers involved in testing. Assessed usually using *Intraclass correlation coefficient* among raters. Usually not focus in PROs.

- *Percent agreement*: it is calculated by dividing the number of cases in which the raters agreed by the total number of ratings.

- *Cohen's kappa:* is preferable to percent agreement because it is corrected for agreement due to chance.

- *Intraclass correlation coefficient* is preferred over product-moment correlation

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Internal consistency reliability

- *Internal consistency* is typically a measure based on the correlations between different items on the same test (or the same subscale on a larger test). It measures whether several items that propose to measure the same general construct produce similar scores.
  - Average inter-item correlation – not recommended
  - Average item-total correlation – not recommended
  - Cronbach´s alpha – somewhat useful
  - McDonald´s omega –recommended

  - Revelle´s beta – we will not cover here

Cronbach´s alpha

$$\alpha = \frac{k\bar{c}}{\bar{\upsilon} + (k-1)\bar{c}} \qquad \alpha_{\text{standardized}} = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

k = number of items

$\bar{c}$ = average of all covariances

$\bar{\upsilon}$ = average variance

$\bar{r}$ = average of non-redundant correlations

- measure of “internal consistency” (homogeneity)
- It can be shown that alpha equals to mean of all possible split-half estimates.
- In a unidimensional test, adding items usually increases alpha

Alpha is problematic because

- Alpha is problematic because it assumes any covariance among items is due to *true-score* variance
  - i.e. even spurious covariance

- Consequence is that alpha can take on quite high values even when the set of items measures several unrelated latent constructs

- Provides *no evidence* of homogeneity
  - Applying alpha itself requires a lot of assumptions
  - Computing alpha is no substitute for dimensionality assessment
  - Dimensionality should be always assessed first

Coefficient Omega

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\text{var}(y)}$$

Where λ_i are factor loadings on the general factor and $y = \sum_{i=1}^{k} x_i$ (i.e. sum of items)

- Based on the common factor model
- Advantage comparing to alpha is that it addresses the true score variance.
- Omega can be also considered as generalizability coefficient (we will cover this later today)
- It is necessary to test unidimensionality prior the computation of omega.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

How high should reliability be?

- I believe it is wrong question. It rather should be – how large is the error that I can accept? Or what is my desired measurement precision?

- Error associated with reliability also depends on variability of observed scores.

$$SEM = \sigma_x \sqrt{1 - rel}$$

- Confidence intervals: 68% ……(score-SEM, score+SEM)
  95% ……(score-2xSEM, score+2xSEM)
  99% ……(score-3xSEM, score+3xSEM)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Spearman-Brown prophecy formula

- Predicted reliability (rel^*) if we make the test N times longer (values over 1) or shorter (values below 1):

$$rel^* = \frac{N \times rel}{1 + (N - 1) \times rel}$$

- This formula can be re-arranged to get N, that is

$$N = \frac{rel^* \times (1 - rel)}{rel \times (1 - rel^*)}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Practical recommendations regarding reliability

- Large random error may be associated with global questions.
- Multi-item scales are usually more reliable than single-item tests
- Reliability may be improved by clarity of expression or lengthening the measure
- The reliability of the scale is increased by including and averaging a number of items, where each item is associated with an independent *random error term*.
- Analogy with measurement of length of the table – using different rulers or multiple measurements.
- Spearman-Brown formula can be used to predict reliability.

Sensitivity

- The usefulness of a measure is dependent upon its ability to detect clinically relevant differences
- *Sensitivity* is the ability to detect differences between groups
- The more sensitive an instrument, the smaller the sample size that is necessary to detect relevant differences
- The ‘floor’ and ‘ceiling’ effects limit sensitivity
- Sensitivity is usually assessed by cross-sectional comparison of groups of patients in which there are expected to be differences in a measured construct.

Relationships between reliability, validity and sensitivity

- Reliability does not imply validity.

- Within CTT, predictive or concurrent validity (correlation between the predictor and the predicted) cannot exceed the square root of the correlation between two versions of the same measure — that is, reliability limits (attenuates) validity.

- Sensitive measurements are usually reliable but reliable instrument may lack sensitivity

Practical 1

Reliability assessment in R

Generalizability theory (G-theory)

- Liberalizes and extends traditional notions of reliability
- Enables investigators to identify sources of inconsistencies (i.e. error)
- Can be viewed as extension of CTT through an application of ANOVA procedures (especially variance components)

Framework of G-theory

- Universes of admissible observations and G-studies
- Universes of generalization and D-studies
- Facets - sets of similar conditions of measurement.

Facets and designs

- X = crossed
- : = nested
- Each letter in the design represents facet

- Examples of design: p x i

  p : t

  p x i x t

  p x (i:c)

Example of Design

- Say our design is **p x i x t**
  - We say it is 2-facet design because objects of measurement (persons) are not usually called „facet“

- Then any observable score can be represented as:

$$X_{pit} = \mu + \upsilon_p + \upsilon_i + \upsilon_t + \upsilon_{pi} + \upsilon_{pt} + \upsilon_{ti} + \upsilon_{pit}$$

and variance of the scores

$$\sigma^2(X_{pit}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(t) + \sigma^2(pi) + \sigma^2(pt) + \sigma^2(ti) + \sigma^2(pit)$$

D-study, universe score and universe score variance

- D-study emphasizes interpretation of variance components from G-study.
- Universe score of respondent is expected score over all instances of measurement (i.e. all instances in the universe of admissible observations
  - it is theoretical score)
- Variance of all persons in the population of universe scores is called universe score variance.
- Universe score variance (denoted as $\sigma^2(\tau)$) is simply $\sigma^2(p)$ for the aforementioned design
- $\sigma^2(\tau)$ is conceptually similar to true score variance
- All other variance components contribute to error variance

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Error variances

- Absolute error – difference between person´s observed and universe score. Its variance (denoted as Δ_p) is sum of all variance components except $\sigma^2(p)$

- Relative error is defined as difference between a person´s observed deviation score and his or her universe deviation score. Variance of relative errors (denoted as $\sigma^2(\delta)$) is similar to error variance in CTT.

Generalizability and dependability coefficients

- Generalizability coefficient (rho): $$\rho = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)}$$

- Dependability coefficient (phi): $$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)}$$

Advantages and disadvantages of G-theory

- Pros
  - Can handle many types of designs and identify source of measurement error
  - Provides useful info on strategy to improve reliability

- Cons
  - Does not take into account validity of items; treats all items to be equally good
  - Based on CTT

References

Validity

- Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist, 50*(9), 741-749.

Reliability

- Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistics in Medicine*, 21(22), 3431-3446.
- McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah: Lawrence Erlbaum Associates, Inc.

Cronbach´s alpha, McDonald´s omega

- Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98-104.
- Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123-133.
- Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating Generalizability to a Latent Variable Common to All of a Scale's Indicators: A Comparison of Estimators for ωh. *Applied Psychological Measurement*, 30(2), 121-144.

Generalizability

- Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre