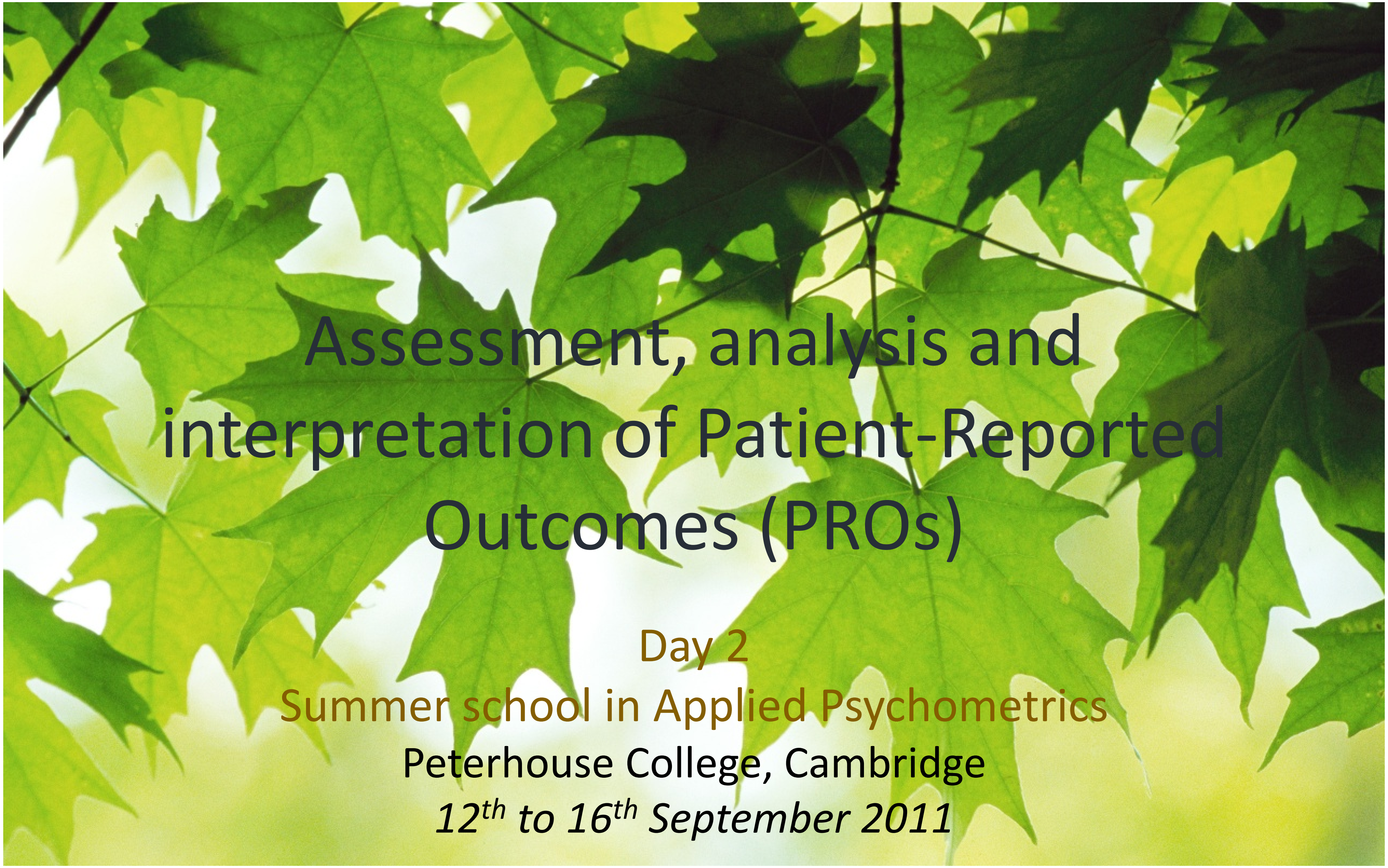


Assessment, analysis and interpretation of Patient-Reported Outcomes (PROs)

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

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

RDI Researcher Development Initiative

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Anna Brown

4. DEVELOPING A QUESTIONNAIRE

UNIVERSITY OF
CAMBRIDGE
The Psychometrics Centre

General requirements

- Good theory (sound constructs, items, and response process)
- Knowledge and experience in psychometric principles of questionnaire design
- Validation
- Documentation
- Requires time, resource and patience
  - So if there is tool that does the job, it might be wise to stick with it

QoL is multi-dimensional?

- Some QoL instruments focus upon a single concept, such as emotional functioning
- Other instruments regard these individual concepts as aspects or dimensions (of QoL) and therefore include items relating to several aspects
- There is some disagreement about what aspects
- Most agree that
  - a number of [the above] dimensions should be included in QoL questionnaires
  - and that QoL is multi-dimensional

Global measures : Single-item scales

- Single Item Global Assessments
  - "How good is your overall quality of life?"
  - "How do you rate your overall health?"
- Considered a useful adjunct, but questions are often regarded as too vague and non-specific [to be used on their own]
- Most instruments include one or more global items alongside a number of other items covering specific issues
  - EQ5D: asks parsimonious 5 questions before using a single global question that enquires about 'your health'

Multi-item scales

- Multiple items can cover a construct more fully, by asking questions about different aspects of the construct
- More specific questions are less prone to subjective biasing effects
- Multi-dimensional, multi-item assessments
  - Physical v Mental Health
  - Social vs Role Functioning
  - Pain (by location)
  - Vitality/Energy vs Tiredness and Fatigue

Single or multi-item?

- Many individual concepts (e.g. emotional functioning) lack a formal agreed definition that is universally understood
- In many cases the problem is compounded by the language differences and some concepts do not readily translate
- There are also cultural differences regarding the importance of issues

- Single item questions on these aspects of quality of life, as for global questions about overall quality of life, are likely to be ambiguous and unreliable, therefore
- it is usual to develop questionnaires that consist of multi-item measurement scales for each concept

Decisions, decisions…

- Target population
  - Age range
  - Range of symptoms (from healthy to very ill)
- Purpose
  - Clinical trials, i.e. discriminative
  - Individual patient evaluation
  - Screening
- Dimensions
- General or disease/treatment specific
  - Dimensions and their bandwidth
- Precision required

Item generation stages

1. Literature search
   - Any existing instruments that address the same or related areas of QoL assessment
2. Inclusion/exclusion of issues

   (a) overlap with other issues that are included

   (b) relevance to the target group of patients

   (c) importance to QoL evaluation

   (d) their prevalence and proportion of patients they affect
3. Semi-structured interviews or focus groups
   - Subject matter experts (physicians and nurses, also psychiatrists or social workers - 3 to 5 initially)
   - Patients (5 to 10 patients from each treatment group, disease stage or symptom severity)

Item stem and response options

- Item stem is the actual stimulus to respond to.
- Respondent has to judge how well the item stem describes their symptoms, or matches their state etc.
- This judgement is typically collected using response options.

Responses to test items

- Binary responses (yes – no, agree - disagree)
- Ordered categorical (ordinal) responses
  - Most often labelled with response categories
  - Most often use 3, 4 or 5 categories
  - Might have many rating categories (for instance, 9) – then the data approaches continuous
  - Rating scales can be symmetrical (agree-disagree) and not (never-always)
- Might use a sliding scale (continuous)

Examples

I take tablets to help me sleep

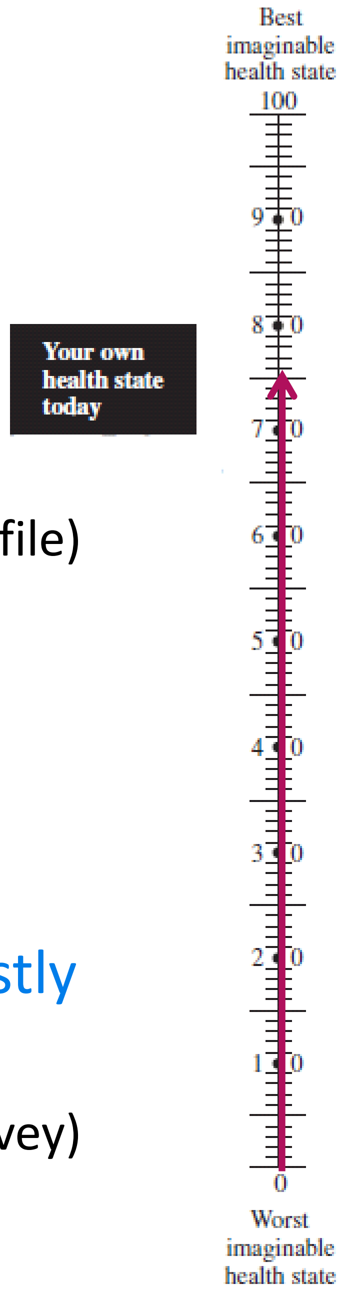
[✓ YES] – NO

(from Nottingham Health Profile)

I seem to get sick a little easier than other people

definitely true – [✓ mostly true] – don’t know – mostly false – definitely false

(from SF-36v2 Health Survey)


Best imaginable health state
100
90
80
Your own health state today
70
60
50
40
30
20
10
0
Worst imaginable health state

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Ordinal response format

- Typically designed to have roughly even spread of labels
- Specific phrasing (for instance, for middle categories) is important
  - “Neutral” or “in-between” is better than “unsure”
- More categories generally mean more information
  - compare binary agree/disagree and 5-point from strongly agree to strongly disagree
  - It has been shown that with 5 categories items approximate continuous scale well
  - It has been shown that respondents cannot reliably discriminate between more than 6 or 7 categories

Summated scales

- Ordinal responses are typically coded as consecutive integers and are added to produce the summated test score
- Assumptions need to be made
  1. Category labels are equidistant
  2. All respondents interpret categories in the same way
- Typically, these *ordinal* scales cannot be assumed *interval*
  - However, they have been successfully used over years and do provide useful information for many purposes

Visual analogue scales

- Any position between two extremes
  - Extremes are often labelled
- Assumed to have equal-interval properties
  - Usually not true due to subjective interpretation
- Contradictory claims about their usefulness
  - “easy to complete” versus “difficult to complete”
  - Prone to response biases

Guttman scales

- Consist of several items of increased difficulty
- Rigidly hierarchical scale
  - Assumed that it is not possible to agree with a stronger statement without agreeing to a weaker statement
  - Example:
    1. *Do you have trouble taking a short walk outside your house?*
    2. *Do you have trouble taking a long walk outside your house?*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Writing items – general principles

- Item writing requires:
  - thorough familiarity with the subject;
  - language proficiency (grammar, spelling, and punctuation).
- Good items should be clear, unambiguous, give the respondent fair chance to express their experiences, and more…
- Guidelines on item writing do evolve over time

The importance of item wording

Questions wording influences eyewitness testimonies (Wrightsman et al., 1994)

Subjects watched a car accident video and estimated car speed by different questions:

- How fast were the cars going when they smashed into each other ?
  - 40.8 mph
- How fast were the cars going when they hit ?
  - 34 mph
- How fast were the car going when they contacted ?
  - 31.8 mph

Writing items – some rules

1. Write in the simplest way possible
   - Avoid unnecessary jargon, local and cultural references
   - Do not use complex sentences or conditionals, particularly with less able patients;
   - Carefully balance item length (long items are difficult to follow; short items might be lacking specificity);
2. Be careful with the use of negation
   - Double-negatives should be avoided at all cost
   - Consider antonyms to key verbs rather than using negation
3. Avoid 'loaded' words or phrases (with strong positive or negative emotional appeal)
4. Avoid leading items (making certain assumptions about respondents)
5. Avoid items involving inter-individual comparisons (patients might have different frames of reference)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Some examples

- Ambiguous item

  Independence is important to me. (what exactly is meant by independence?)

- Double-barrelled item

  I keep my emotions under control except when things become too difficult.

- Leading item

  I regularly do exercise to avoid becoming ill.

Practical exercise

- In groups of 4 or 5
- Consider fragments from some established PRO measures
  - Generic
    - A PATIENT GENERATED INDEX OF QUALITY OF LIFE
    - SF-36v2 Health survey standard version
  - Disease specific
    - European Organisation for Research and Treatment of Cancer QLQ-C30 (EORTC QLQ-C30)
    - Functional Assessment of Cancer – General version (FACT-G)
  - Domain specific
    - Hospital Anxiety and Depression Scale (HADS)
    - Multidimensional Fatigue Inventory (MFI-20)
- Discuss their phrasing, rating options, construct coverage
- Where and how can these questionnaires be used best?

Consideration for data quality

- Underline or use boldface for key words
- For items that may not be applicable to some patients, make sure to include “N/A” response option
  - Similarly, give an option to skip embarrassing or very personal questions
- Consider giving additional instructions to negatively keyed items
- Avoid items that will be answered in the same way by most of your target audience (we will talk about this later)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Pre-testing

- Questionnaire should be given to patients and staff (could be the people involved in the interview stage) for comments
- Pilot study
  - Representative sample of new patients (10-30)
  - They complete provisional questionnaire and then are debriefed
  - Feedback on question quality is collected
  - Questions with lots of missing responses are reviewed

Field-testing

- Aim is to confirm the acceptability, validity, sensitivity, responsiveness, reliability and applicability to subgroups
- Should involve a large heterogeneous group of patients representative of all intended responders
- Debriefing questionnaire should accompany the actual PRO questionnaire
  - How long did it take
  - Did anyone help to complete and how
  - Were the questions acceptable
  - Were all questions relevant, or anything important missing

Tabulation of field-test data

- Check if any questions yield unusually high numbers of missing responses
- Examine distribution of item responses
  - – Ceiling or floor effects
    - “indicator” items which do not discriminate between patients are probably better deleted
    - “causal” items (such as symptoms) might display these effects but still be important to retain
      - – Capturing very rare symptoms could be important for sensitivity of the instrument
      - – Very common symptoms might be still important to retain for completeness
  - – Check that the number of response categories is adequate
    - Clustering of responses in one category may indicate that more response options are needed

Further development stages

- Further psychometric analysis will be discussed in our next topics
- Psychometric approaches are used to refine and improve the questionnaire
- Documentation of all analyses is important
- Full development might take years