Patient-Reported Outcomes (PRO’s) in Clinical Trials: Overview of Issues in Measurement

Neil K. Aaronson

Educational Program on PRO’s in Clinical Trials

College voor Zorgverzekeringen

Diemen, December 1, 2006

A note on terms

- Health outcomes
- Health status
- Quality of life
- Health-related quality of life
- Patient-reported outcomes (PROs)

Patient-reported ouctomes (PROS’s)

- Physical and psychosocial functioning
- Symptoms
- Health-related quality of life
- Satisfaction with care

fact

n. a thing known to be true || a statement about something which has occurred, *he got the facts distorted* || (*law*, in certain phrases only) a crime as a matter of fact, in point of fact, the fact of the matter is… (introductory phrases used to emphasize an explanation or confession) to tell you the truth in fact (usually in contradistinction to some supposed state of affairs) in truth, actually (fr. L. *factum*, a thing done)

The New International Webster Dictionary of the English Language, 1995

fiction

n. A literature consisting of invented narrative, esp. the novel and short story || a falsehoood (e.g., that there exists a ‘man in the street’) conventionally accepted as true because it is useful to make the assumption

The New International Webster Dictionary of the English Language, 1995

fact or fiction?

The term “(health-related) quality of life,” is well defined and widely understood.

Fact – if you keep things simple

Fiction – if you dig deeper

"Quality of life is a vague and ethereal entity, something that many people talk about, but which nobody clearly knows what to do about.“ Campbell et al., 1976

“The idea has become a kind of umbrella under which are placed many different indexes dealing with whatever the user wants to focus on.” Feinstein, 1987

“Quality of life is an ill-defined term…it means different things to different people, and takes on different meanings according to the area of application.” Fayers & Machin, 2000

4 criteria for evaluating clinical effectiveness of chemotherapeutic agents in lung cancer

D.A. Karnofsky et al., *Cancer* 1:634 , 1948

- subjective improvement
- objective improvement
- performance status
- length of survival

Subjective improvement

“The patient’s subjective improvement is measured or described in terms of:

- improvement in his mood and attitude
- his general feeling of well-being,
- his activity, appetite, and the alleviation of distressing symptoms such as pain, weakness, and dyspnea.”

WHO definition of health, 1949

“A state of complete physical, mental and social well-being, and not merely the absence of disease and infirmity.”

Key dimensions of quality of life as defined by ASCO, 1995

Physical	Symptoms commonly caused by cancer and the toxicities of treatment
Psychologic	Effects of cancer and its treatment on cognitive function and emotional state
Social	Effects of cancer and its treatment on interpersonal relationships, school, work and recreation

Attributes of QL definitions

- non-specific vs. health-related
- health states (or status) versus personal evaluation of those states (e.g., expectations, discrepancies, satisfaction)
- scope of concerns (e.g., spirituality or existential issues)
- polarity of concerns (well-being vs. dysfunction and its resolution)

Does it matter?

- Yes, because the content of QL questionnaires reflects the underlying definition.
- It may be less important in clinical trials, where group comparisons will be internally valid, regardless of the definition used.
- It is more important in comparing results across trials and in descriptive (e.g., prevalence) studies.

Examples of QL definitions

“The difference between the hopes and expectations of the individual and the individual’s present experience.”

Calman, 1987

“The functional effect of an illness and its consequent therapy upon a patient, as perceived by the patient.”

Schipper et al. 1996

Covinsky et al. Am J Med 1999; 106:435-440

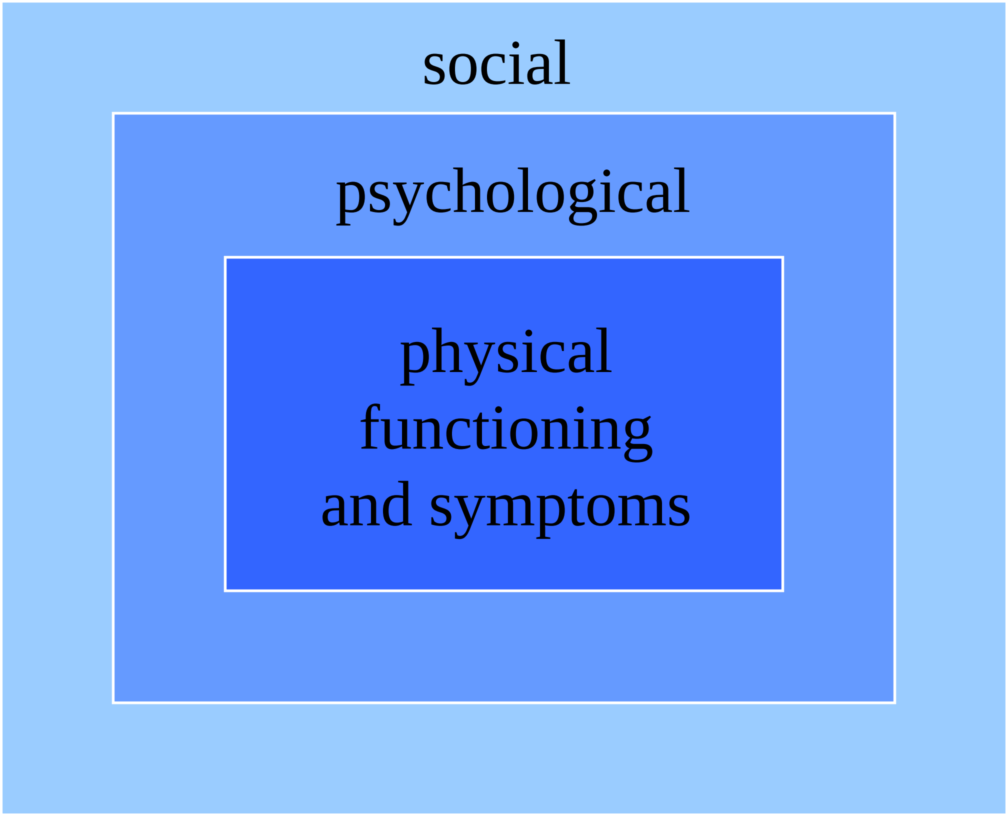
- study of 493 older patients
- QL rated as good/excellent by 43% of those with worst physical functioning and 47% with highest levels of psychological distress
- QL was rated as poor by 15% of those with the best physical functioning and 21% with the lowest levels of psychological distress

fact or fiction?

We have adequate conceptual models for studying the underlying associations between HRQL domains, and the factors influencing those associations.

reasonably factual

multidimensional HRQL assessment

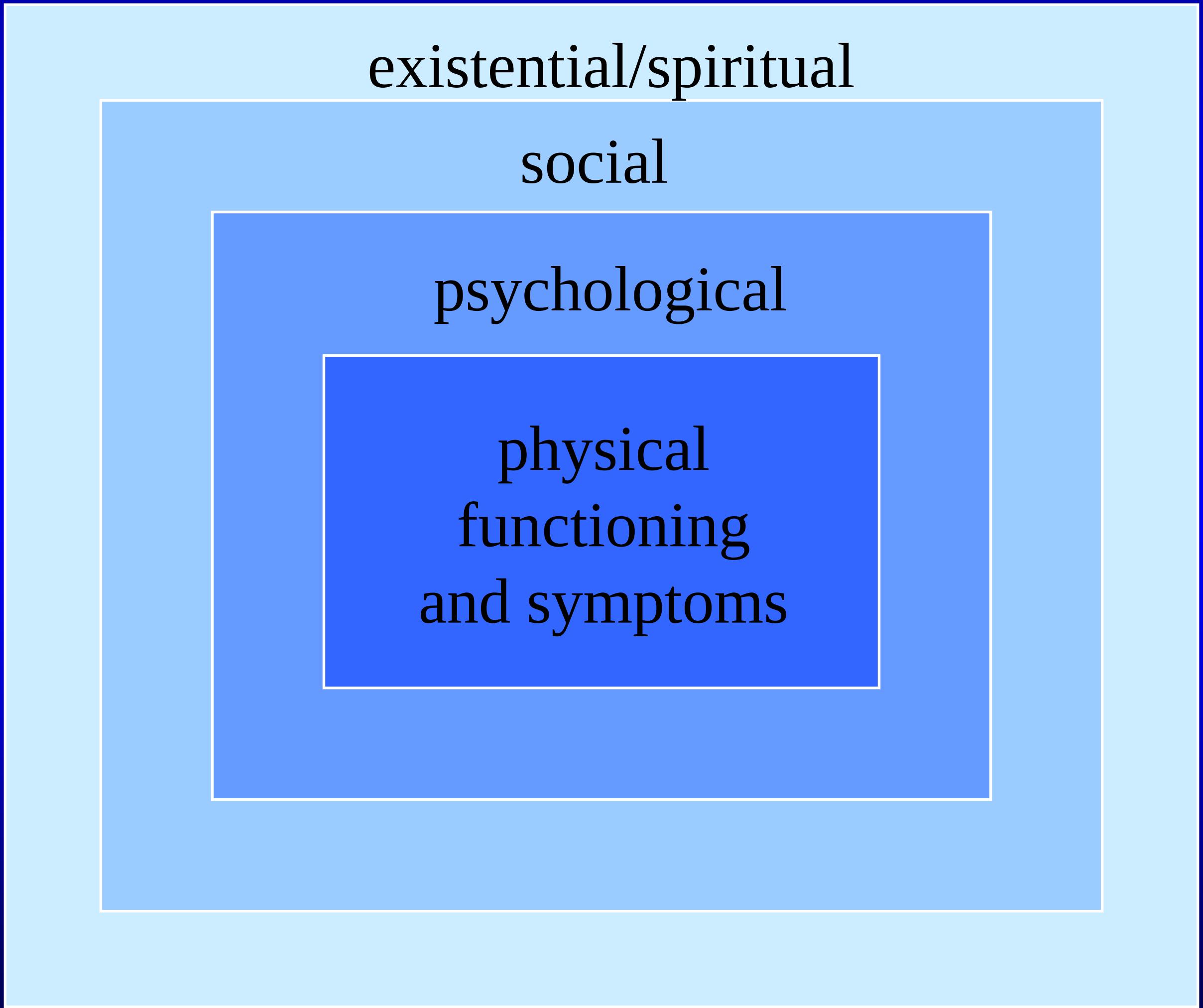


social

psychological

physical
functioning
and symptoms

multidimensional HRQL assessment



existential/spiritual

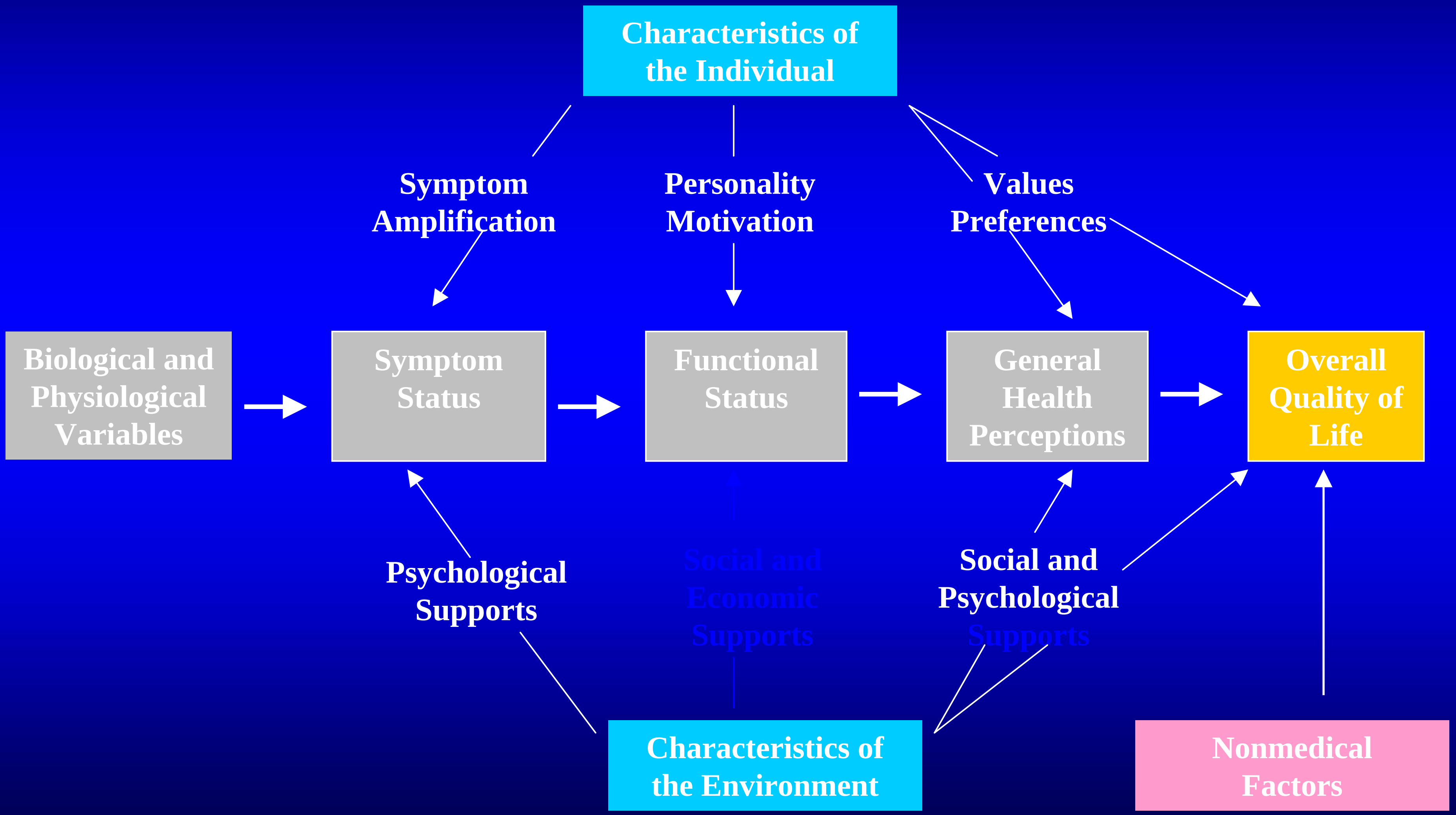
social

psychological

physical
functioning
and symptoms

Relationships among measures of patient outcome in an HRQL conceptual model

(Wilson and Cleary, JAMA 1995; 273(1): 59-65)



Characteristics of the Individual

Symptom Amplification

Personality Motivation

Values Preferences

Biological and Physiological Variables

Symptom Status

Functional Status

General Health Perceptions

Overall Quality of Life

Psychological Supports

Social and Economic Supports

Social and Psychological Supports

Characteristics of the Environment

Nonmedical Factors

fact or fiction?

The patient is the sole legitimate source of information about his/her HRQL. Other “proxy” raters (e.g., family members, health care providers) are, at best, poor substitutes.

(partial) fiction

Self-report can be limited by:

- age (very young or old)
- cognitive impairment
- communication problems
- symptom distress
- physical disability
- emotional distress

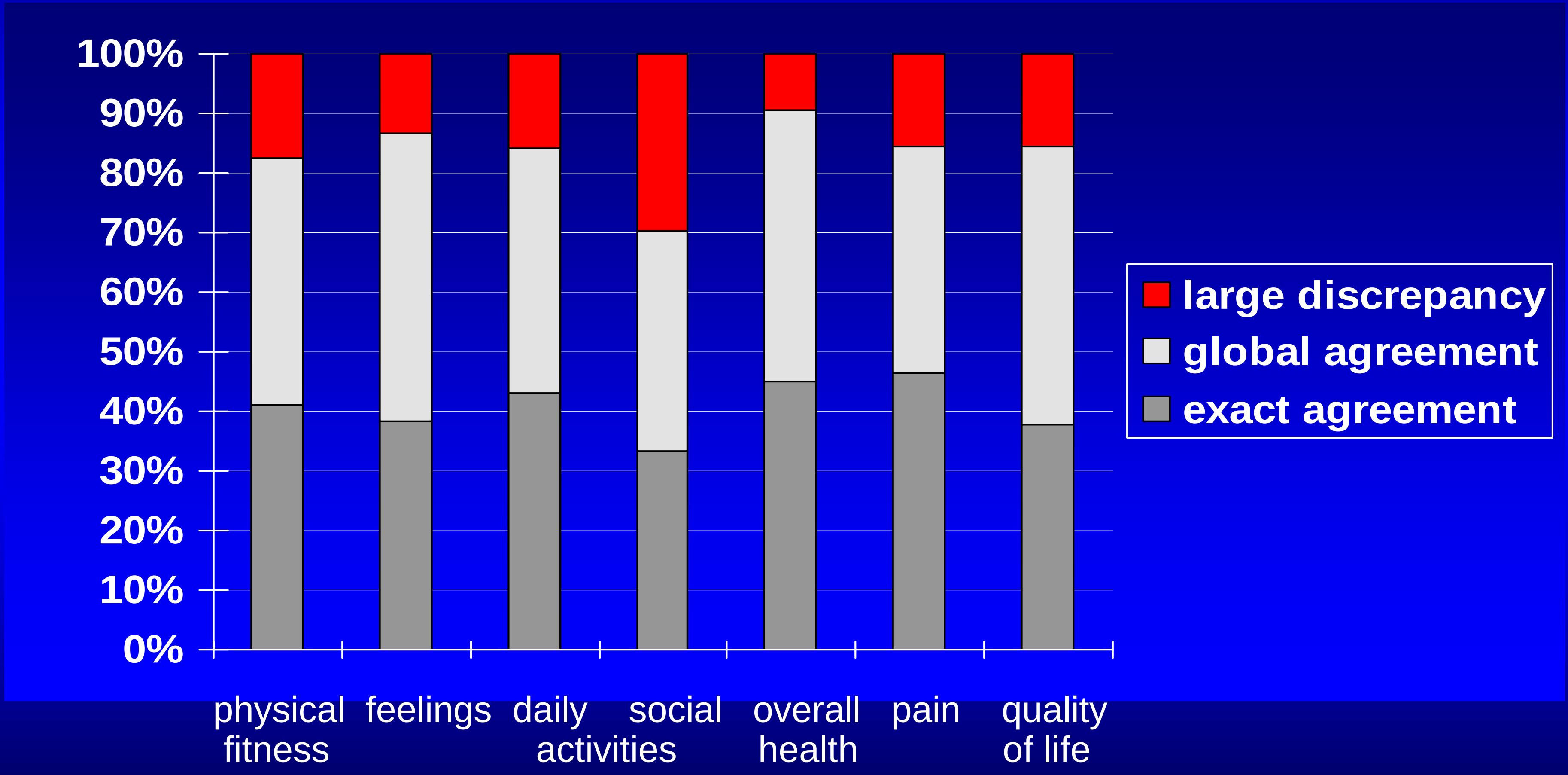
Exclusion of highly relevant subgroup of patients can result in biased study outcomes

The role of health care providers and significant others in evaluating the QL of patients with chronic disease

Sneeuw KCA et al. 2002; J Clin Epidemiol 55:1130-43

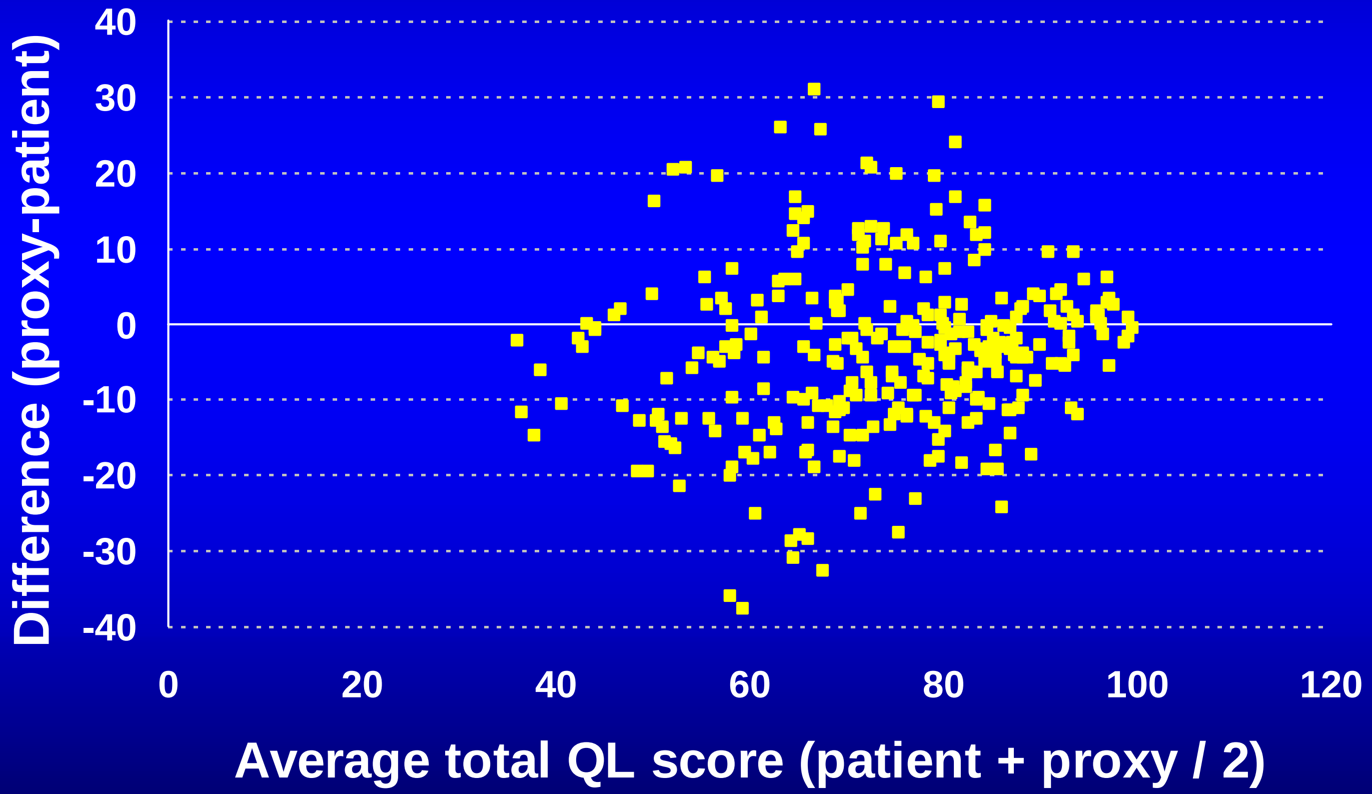
- 23 studies published between 1991 - 2000
- Moderate/high patient – proxy agreement
- Proxies tended to rate patients as having more problems than did patients themselves
- Magnitude of differences was small (median standardized difference 0.20)

Proportion of agreement
by HRQL domain (WONCA charts)



	physical fitness	feelings	daily activities	social	overall health	pain	quality of life
number of comparisons	270	269	270	265	267	268	270

Bland-Altman plot for total HRQL score



fact or fiction?

Although there are many HRQL questionnaires from which to choose, the dust is settling and a “best bet” can be identified based on a comparison of psychometric characteristics and performance.

fiction

Generic HRQL instruments

- Sickness Impact Profile (SIP)
- Nottingham Health Profile (NHP)
- Spitzer QL Index
- COOP/WONCA Charts
- MOS 36-Item Health Survey (SF-36)
- World Health Organization (WHOQoL)

Cancer-specific HRQL questionnaires

- Functional Living Index – Cancer (FLIC)
- Cancer Rehabilitation Evaluation System (CARES)
- Rotterdam Symptom Checklist (RSCL)
- EORTC QLQ-C30
- Functional Assessment of Cancer Therapy (FACT-G)

Key psychometric attributes of HRQL instruments

- measurement model
- reliability
- validity
- responsiveness
- interpretability
- cultural adaptability
- burden

Choice of HRQL instrument should be driven by:

- the research question(s) to be addressed
- the population under study
- the conceptual basis of candidate questionnaires
- the specific content and wording of candidate questionnaires

Negative affect items

SF-36	“Have you felt so down in the dumps that nothing could cheer you up?”
	“Have you felt downhearted and blue?”
FACT-G	“I feel sad”
QLQ-C30	“Did you feel depressed?”

Future perspective items

SF-36	"I expect my health to get worse."
FACT-G	"I worry about dying."
CARES-SF	"I worry about whether the cancer will progress."
QLQ-C30	--

Translations

Involvement of native speakers is mandatory, otherwise spelling or other minor mistakes may go unnoticed

HUNGARIAN

QLQ-C30 (missing 'N')

Agymértékben	Nagymértékben
Brain Sized	*Very Much*

LC-13 (missing 'it')

Menny	Mennyit
Heaven	*How Much*

fact or fiction?

Given the plethora of HRQL questionnaires currently available, there is little or no need for continued efforts at instrument development.

fiction

- Condition-specific questionnaires tend to be more sensitive to group differences and responsive to inter- and intra-individual changes over time

supplemental modules/scales

- combine “core” instrument with condition-specific modules/scales
    - EORTC “modules”
    - FACT subscales
    - NCIC symptom checklists

EORTC QL modules

- body image
- breast
- bladder
- brain
- colorectal
- esophageal
- high dose chemo
- leukemia (adult)
- lung
- opthalmic
- ovarian
- peripheral neuropathy
- prostate
- supportive care

FACT subscales

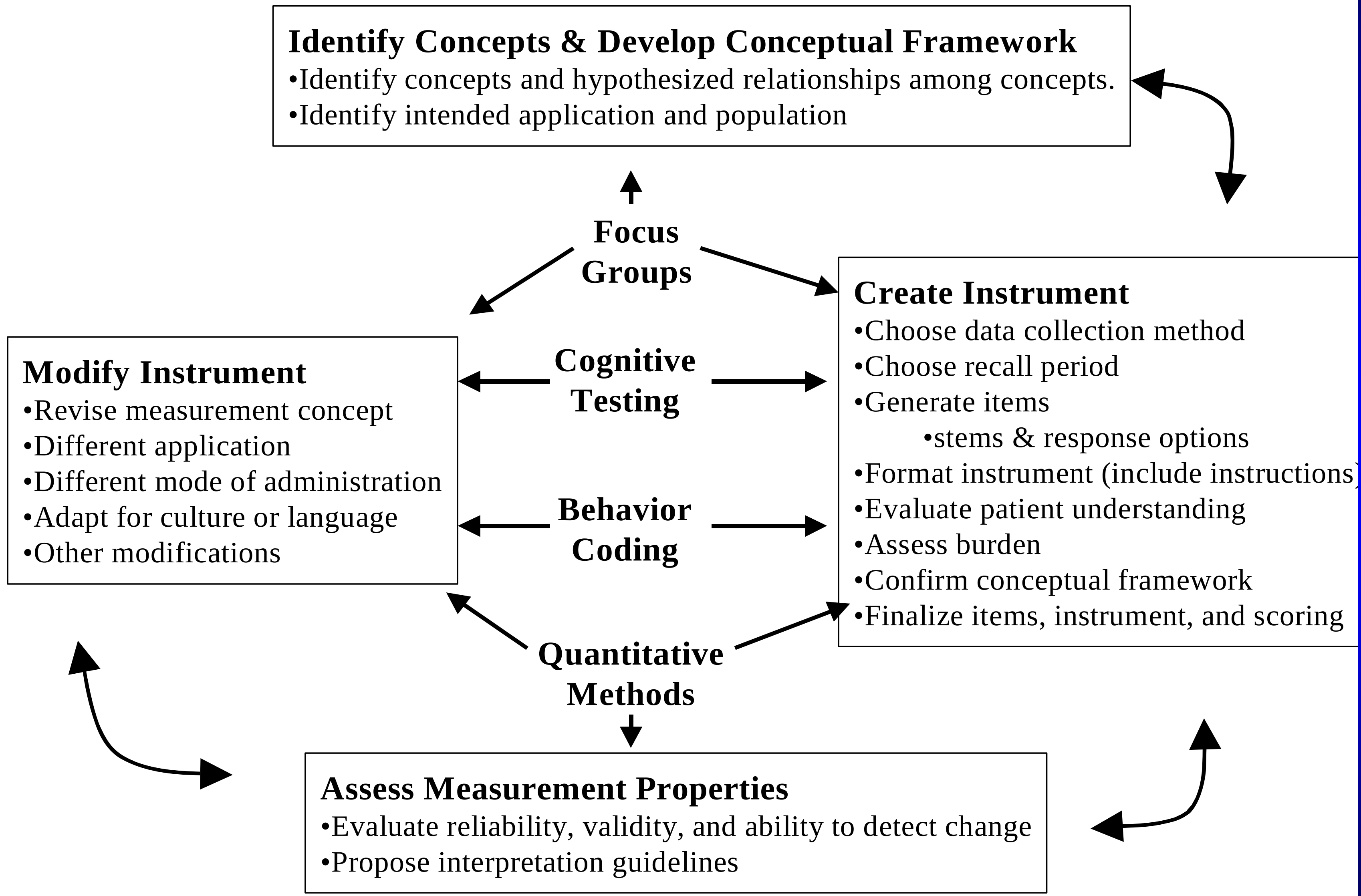
- breast
- bladder
- brain
- cervical
- colorectal
- esophageal
- head & neck
- HIV
- lung
- ovarian
- pancreatic
- prostate
- anorexia
- BMT
- fatigue
- fecal incont
- neurotoxicity
- taxane toxicity
- urinary incont
- spirituality

Advantages of core + module approach to HRQL assessment

- facilitates comparison of results across studies
- provides sufficient flexibility to address questions specific to a given patient population or treatment

Develop a new instrument at your own risk and only as a last resort

- Labor intensive
- Time consuming
- Long “probationary” period before psychometrics are assessed and verified
- Translation and cultural adaptation process for use in international settings adds layer of complexity



Identify Concepts & Develop Conceptual Framework

- Identify concepts and hypothesized relationships among concepts.
- Identify intended application and population

Focus Groups

Cognitive Testing

Behavior Coding

Quantitative Methods

Create Instrument

- Choose data collection method
- Choose recall period
- Generate items
  - stems & response options
- Format instrument (include instructions)
- Evaluate patient understanding
- Assess burden
- Confirm conceptual framework
- Finalize items, instrument, and scoring

Modify Instrument

- Revise measurement concept
- Different application
- Different mode of administration
- Adapt for culture or language
- Other modifications

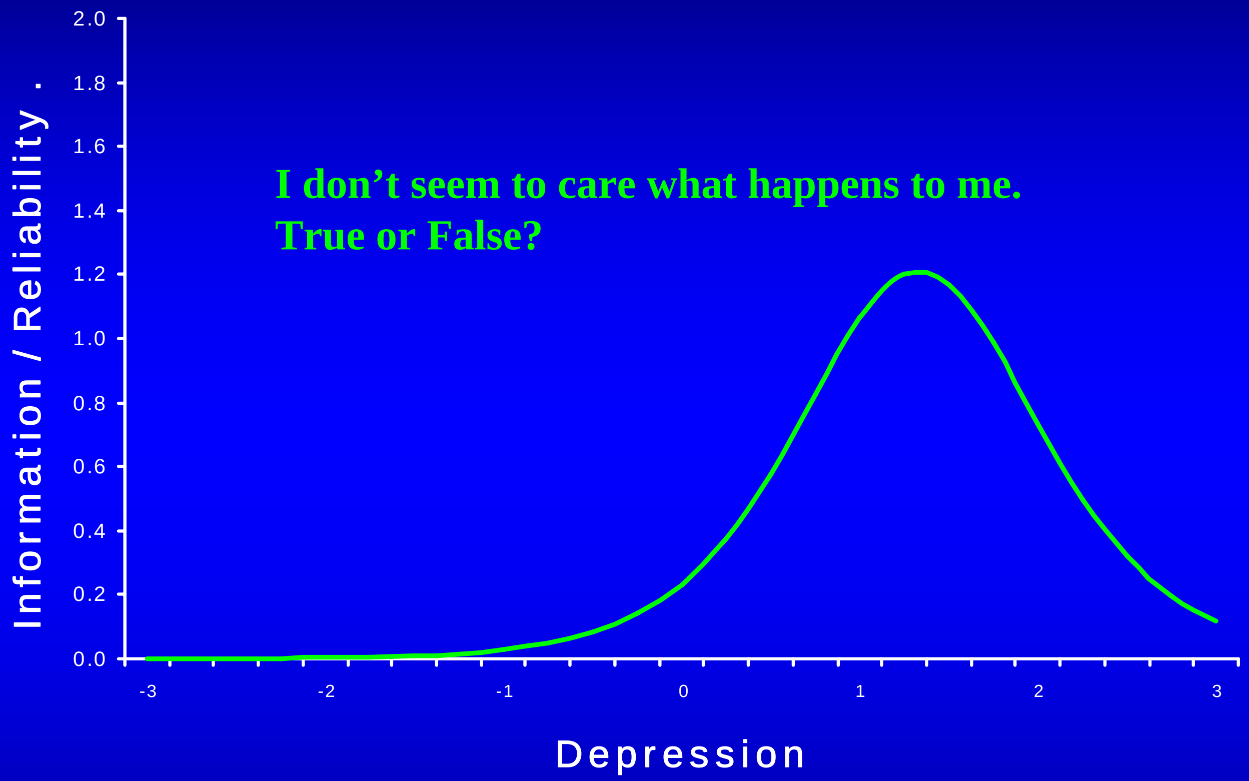
Assess Measurement Properties

- Evaluate reliability, validity, and ability to detect change
- Propose interpretation guidelines

“Modern” Psychometrics

- Item response theory (IRT)
- Item banking
- Dynamic or computer-adaptive testing

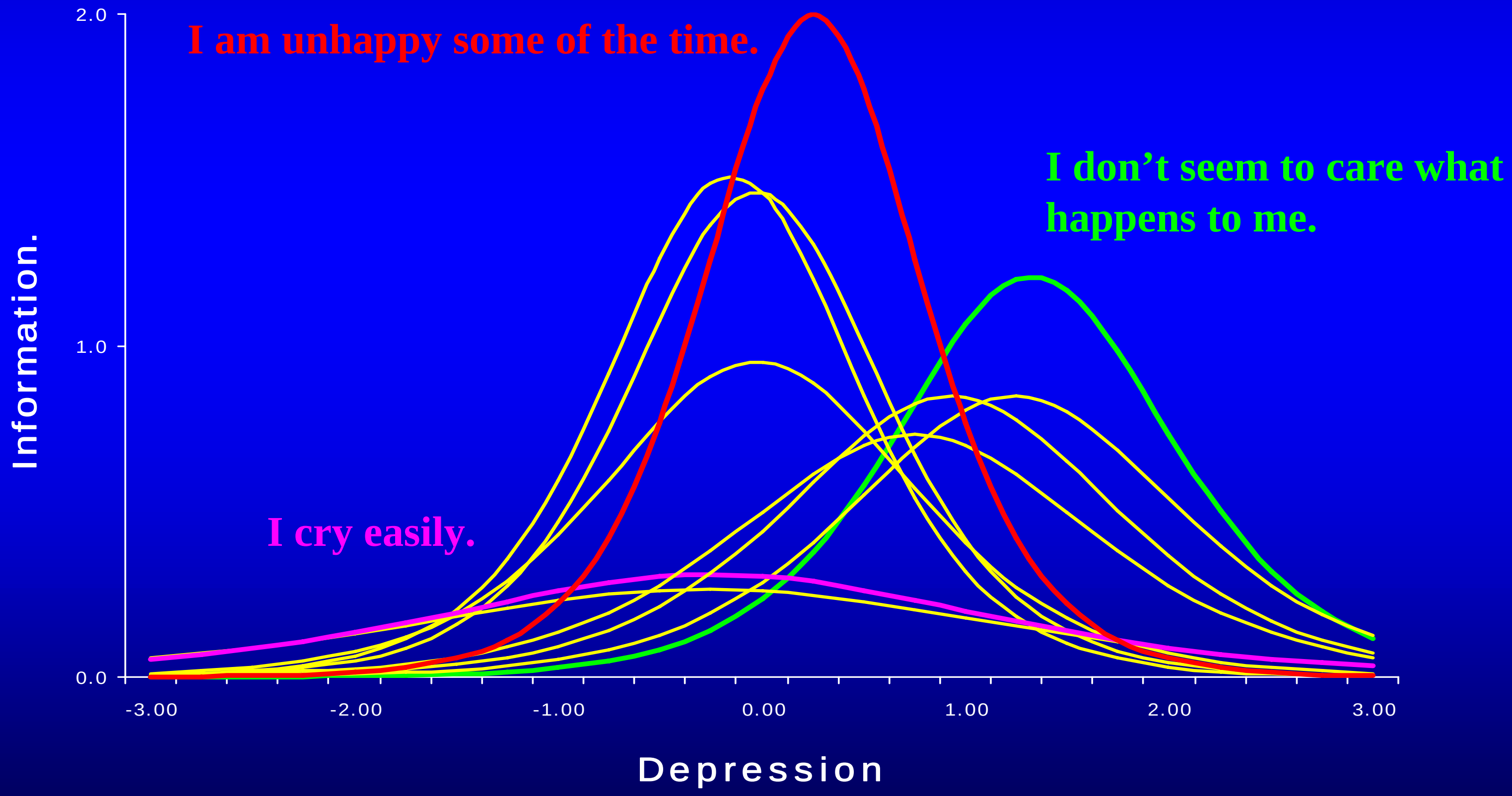
Item Response Theory
Information Curve



**I don’t seem to care what happens to me.
True or False?**

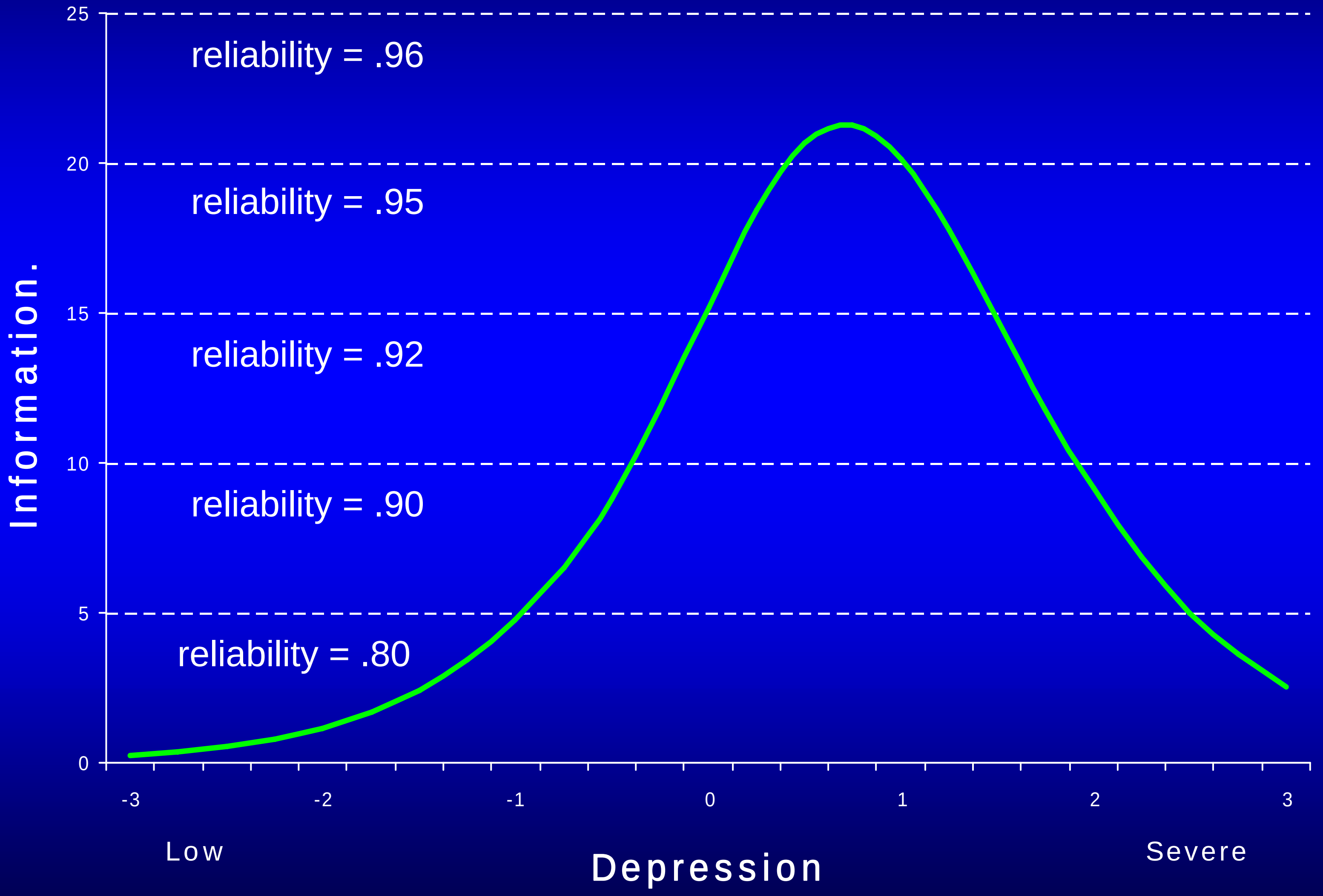
Indicates the range over the measured construct where an item is best at discriminating among individuals. Higher information denotes more precision (or reliability) for measuring a person’s trait level.

Item Response Theory (IRT) Modeling
Item Information Curves



10 Items from the MMPI-2 Depression Scale

Item Response Theory (IRT) Modeling
Scale Information Curve



MMPI-2 Depression Scale

Item banking and “dynamic” or computer-adaptive testing

- Ask a question from an item bank
- Use the response to this question to estimate the ‘level’
- Select next question based on this knowledge
- Use the response to this question and the previous one to re-estimate the ‘level’ - and so on...
- Stop when the desired level of precision is reached

Implications of IRT for HRQL and other PRO assessments

- Increase measurement precision
- Can vary level of precision dependent on task at hand (group versus individual level)
- Can Calibrate scores across existing measures
- Can generate a number of “versions of a questionnaire
- May eliminate (or at least reduce) issues of proprietary interest

fact or fiction?

The major methodological challenges in HRQL analysis – missing data, multiple comparisons, and clinical interpretation of statistical results – have been resolved or are well on their way to being resolved.

Reasonably factual

(2 out of 3 ain’t bad)

Missing data:
Items from questionnaires

- Relatively minor problem (less than 5%)
- For multi-item scales, missing responses can be estimated/replaced
- High level of missing values for a given item may signal problem of appropriateness or acceptability

Missing data:
Questionnaires

- Missing at random (e.g., administrative failure)
    - Largely avoidable
- Systematic loss to follow-up due to illness or death (“informative censoring”)
    - Often unavoidable (e.g., in advanced disease trials)
    - Complex problem with imperfect but workable solutions
        - Mixed effects ANOVA
        - Growth curve analysis

Multiple comparisons

- Inherent problem with multidimensional HRQL measures (health profiles)
- Results in inflated p values
- Three primary solutions
  - Use summary scores, where available
  - Focus on a few “cardinal” (primary) outcomes
  - Apply statistical adjustments

Defining clinical vs. statistical significance in HRQL scores

"A conception not reducible to the small change of daily experience is like a currency not exchangeable for articles of consumption; it is not a symbol, but a fraud."

George Santayana

The Life of Reason (1905-1906)

Clinical significance is NOT Statistical Significance

HSQ before / after scores for 1300 pts. (JCO, 2002)

- - all p-values <0.0001
- stated conclusion: “clinically significant changes in all domains of QOL”
- 80% power to detect a change of 1 unit on 0-100 point scale

The significance of statistical significance

- Medline search identified over 200 letters to journals since 1970 on this topic
- Numerous examples of published statistically significant results that clinicians regard as clinically irrelevant and as having no impact on practice

Definition of clinical significance

“The smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate in the absence of troublesome side-effects and excessive cost, a change in the patient’s management”

(Juniper et al, 1994)

Defining clinical vs. statistical significance in QL scores

- The post-test mean score of the intervention group was 74.3 vs. 67.5 in the control group (p < .001)
- The between-group difference in mean QL scores represented a moderate effect size (Cohen’s d = 0.5).
- 35% of the intervention group experienced a QL benefit (> 10 point change) vs. 22% of the control group

fact or fiction?

HRQL data have contributed significantly to the clinical trial process in oncology and other fields of medicine.

Fact, but we can do better

Baseline HRQL as an independent predictor of survival

Study	N	Patients	QL Measure
Kassa et al. (1989)	10	NSCLC	Study-Specific
Ganz et al. (1991)	40	NSCLC	FLIC
Coates et al. (1992)	226	Breast	LASA/QLI (+ change)
Coates et al. (1993)	152	Melanoma	LASA/QLI
Seidman et al. (1995)	49	Breast	FLIC/MSAS
Earlam et al. (1996)	50	Colorectal	RSCL
Tamburini et al. (1996)	115	Terminal	TIQ
Dancy et al. (1997)	474	Mixed	EORTC

How would you rate your overall health?

excellent good fair poor

In general population studies, self-rated health is one of the most consistent, independent predictors of:

- use of medical and mental health services
- morbidity
- 5 and 10 year mortality

HRQL as prognostic factor: clinical trial applications

- stratification prior to randomization
  - help ensure pretreatment group equivalence
  - increase efficiency of trial
  - facilitate planned subgroup analyses

HRQL data can yield unanticipated results

Sugarbaker et al. Surgery 91:17-23, 1982

- Small RCT (n = 26) in soft-tissue sarcoma:
  - amputation + CT vs. limb-sparing surgery + RT + CT
- HRQL assessed post-surgery
  - SIP, PAIS, Katz ADL, Barthel Index, clinical assessment of mobility, pain, sexuality
- No significant differences between treatment arms, with exception of sexual functioning, which favored amputation group
- Led to improvement in limb-sparing procedure (e.g., better RT shielding)

HRQL data can yield unanticipated results

Hopwood P, Stephens R. Lung Cancer 1994; 11 (Suppl 1) S82

- RCT in SCLC (n = 314)
- 3 cycles of etoposide, cyclofosfamide, methotrexate and vincristine vs. 3 cycles of etopsodie en vincristine only
- RSCL and HADS
- No statistically significant differences in survival or tumor response
- 4 drug combination resulted in less distress, fatigue, dyspnea and cough, and better physical functioning

Clinical trial-based HRQL data in drug approval process

Tannock IF et al. *J Clin Oncol* 1996;14:1756-64.
Osoba D et al *J Clin Oncol* 1999;17:1654-63)

- RCT of daily predinsone ± mitoxantrone (q 3 wks) in metastatic, hormone-resistant prostate cancer (n =161)
- QLQ-C30, PROSQOLI at baseline and q 3 wks
- Pain improved in 29% of men treated with P+M *vs* 12 % treated with P alone, and lasted 43 *vs* 18 wks
- Survival the same in both groups
- P+M ⇒ improvement over time in physical and role function, fatigue, insomnia, drowsiness
- QOL improvements lasted longer in P+M group

Clinical trial-based HRQL data: Making tradeoffs explicit

Holmberg L et al. NEJM 2002; 347:781-789

- RCT of radical prostatectomy vs watchful waiting in early stage prostate cancer (N = 695)
  - concealed randomization; blinded adjudication of outcomes; 100% follow-up survival (6 yrs), 87% follow-up HRQL (4 yrs)
- Prostatectomy – trend for reduced all-cause mortality (18% versus 15%; RR 0.83, 0.57 to 1.2, p = 0.31)
- Decrease in prostate-specific death rates (9% versus 5%; RR 0.50, 0.27 to 0.91, p = 0.02)

Clinical trial-based HRQL data: Making tradeoffs explicit

Steineck G et al. NEJM 2002; 347: 7980-796

- Sexual dysfunction
  - 45% waiting; 80% prostatectomy
- Urinary leakage
  - 21% waiting; 49% prostatectomy
- Urinary obstruction (weak stream)
  - 44% waiting; 28% prostatectomy
- Bowel function, anxiety, depression, well-being did not differ

fact or fiction?

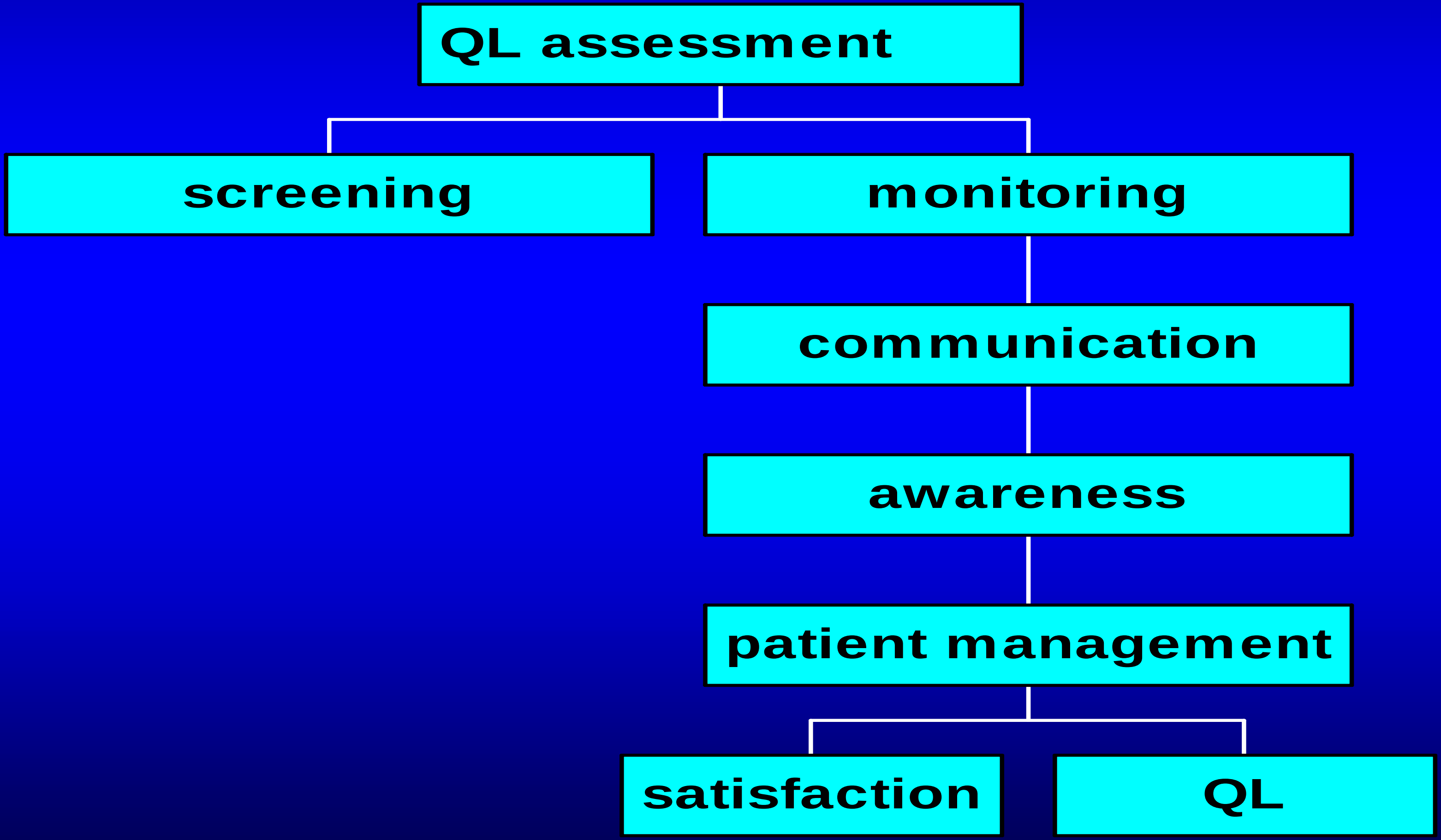
HRQL assessment is ready for prime time as a tool in daily clinical practice.

“faction”

Brodman K. et al. The Cornell Medical Index: An adjunct to medical interview JAMA 1949; 140:531-34

- 195 item self-administered questionnaire on physical and psychological symptoms and medical history
- completed prior to office visit in 10-30 minutes; high compliance rates
- Elicited information not found in medical records

Modeling the use of HRQL assessment in clinical practice



HRQL assessment in daily clinical practice: Feasibility

- Self-administered questionnaires can be completed quickly in office-based practice

- Computer-assisted (e.g., touchscreen) administration is acceptable and efficient

- No evidence that collection of standardized HRQL data interferes with normal clinic routine or lengthens average visit time

HRQL assessment in daily clinical practice:

16 randomized studies published 1987-2004

4 of which were in oncology setting:

(Taenzer et al. 2000; McLachlan et al. 2001; Detmar et al. 2002; Velikova et al. 2003)

• communication	+
• awareness	+
• patient management	+/-
• satisfaction	-
• health outcomes	+/-

Moving things forward (1)

- Make better use of existing conceptual models in:
  - shaping research questions
  - selecting appropriate measures and methodologies
  - guiding analysis strategies

Moving things forward (2)

- Don’t reinvent the HRQL measurement wheel, but rather:
  - refine existing “traditional” HRQL measures
  - invest in the development of additional, condition-specific measures
  - contribute to the collective effort needed to develop “dynamic” (computer-adaptive) measurement systems

Moving things forward (3)

- Continue efforts to translate statistically significant HRQL findings into clinically meaningful terms

- Make substantial investment in a limited number of high profile clinical trials where HRQL is likely to yield added value

Moving things forward (4)

- Move forward with the application of HRQL measures in daily clinical practice, but don’t make promises you can’t keep

- Investigate (and attempt to strengthen) the “weak links” in the putative causal chain between routine HRQL assessment and improved patient HRQL over time