ANEMIA AND HEALTH-RELATED QUALITY OF LIFE MEASURES: PSYCHOMETRIC CHARACTERISTICS OF INSTRUMENTS

Dennis A. Revicki, PhD
Miriam Kimel, PhD

Center for Health Outcomes Research, United BioSource Corporation, Bethesda, Maryland, USA

Prepared for the KDIGO Controversies Conference: Coordination of Clinical Practice Guidelines for Anemia in CKD, New York, NY, October 15, 2007

OVERVIEW

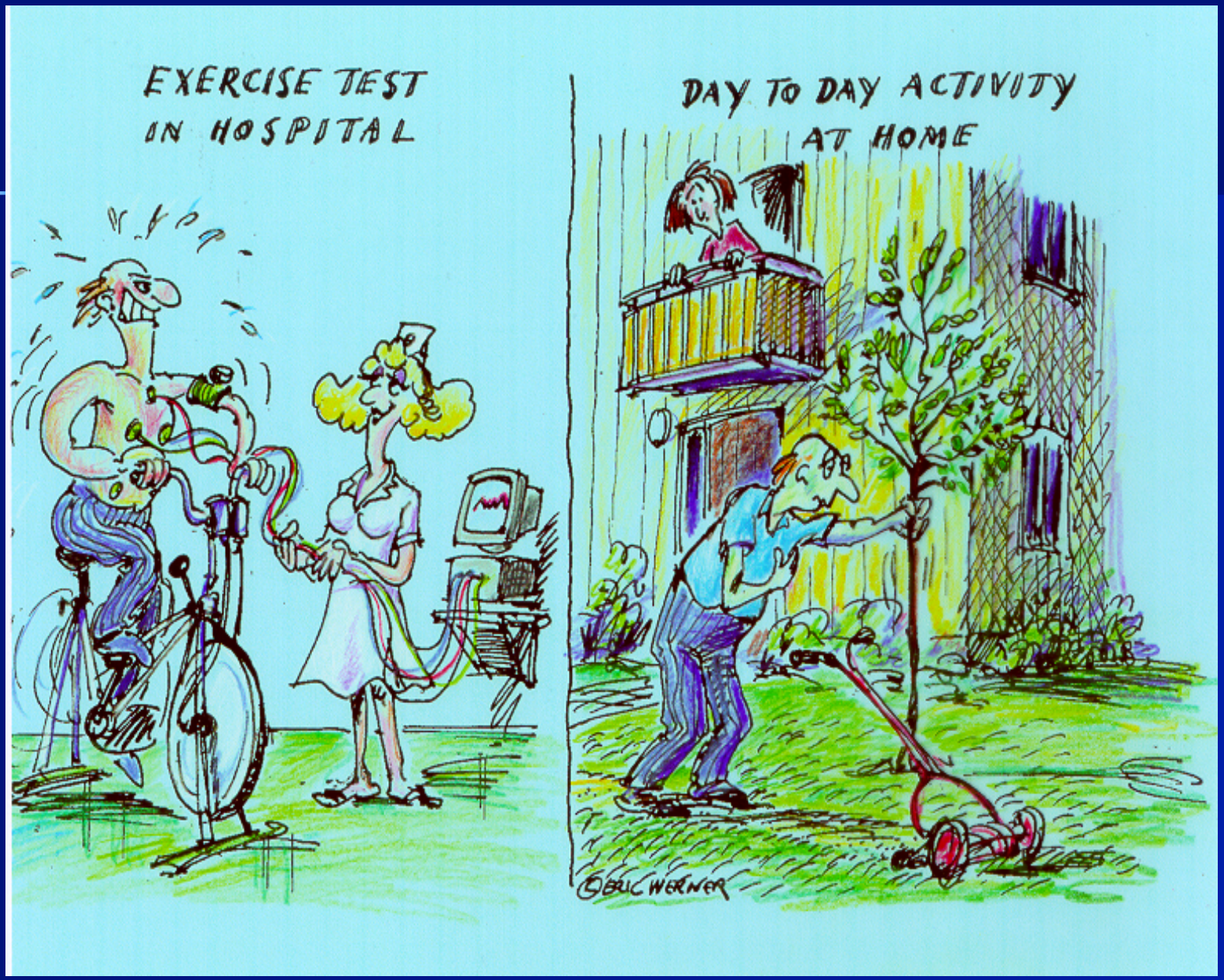
- Why ask patients about their health status?
- Development and psychometric evaluation of health status measures
- Summary of psychometric qualities of frequently used HRQL measures
  - Content coverage
  - Measurement qualities
- Future of HRQL measurement
  - NIH PROMIS initiative

WHY ASK PATIENTS ABOUT THEIR HEALTH STATUS?

- HRQL data describe the impact of treatment and disease on symptoms, functioning and well-being.
- Patients provide a unique perspective on the impact of disease and treatment on their functioning and well-being
- Physiologic, laboratory and clinician evaluations are associated with but not identical to HRQL measures
- HRQL measures extend and translate clinical endpoints

KEY CONCEPTS AND ASSUMPTIONS

- Patient's experience provides a unique and valuable contribution to understanding treatment effectiveness
- Information provided by patient is inherently subjective
- Scientific methods for measuring subjective outcomes are well-developed and are foundation of HRQL assessment
- Need scientifically adequate clinical trial designs and statistical analyses



“Objective” **“Subjective”**

Exercise test versus physical functioning, r = 0.40

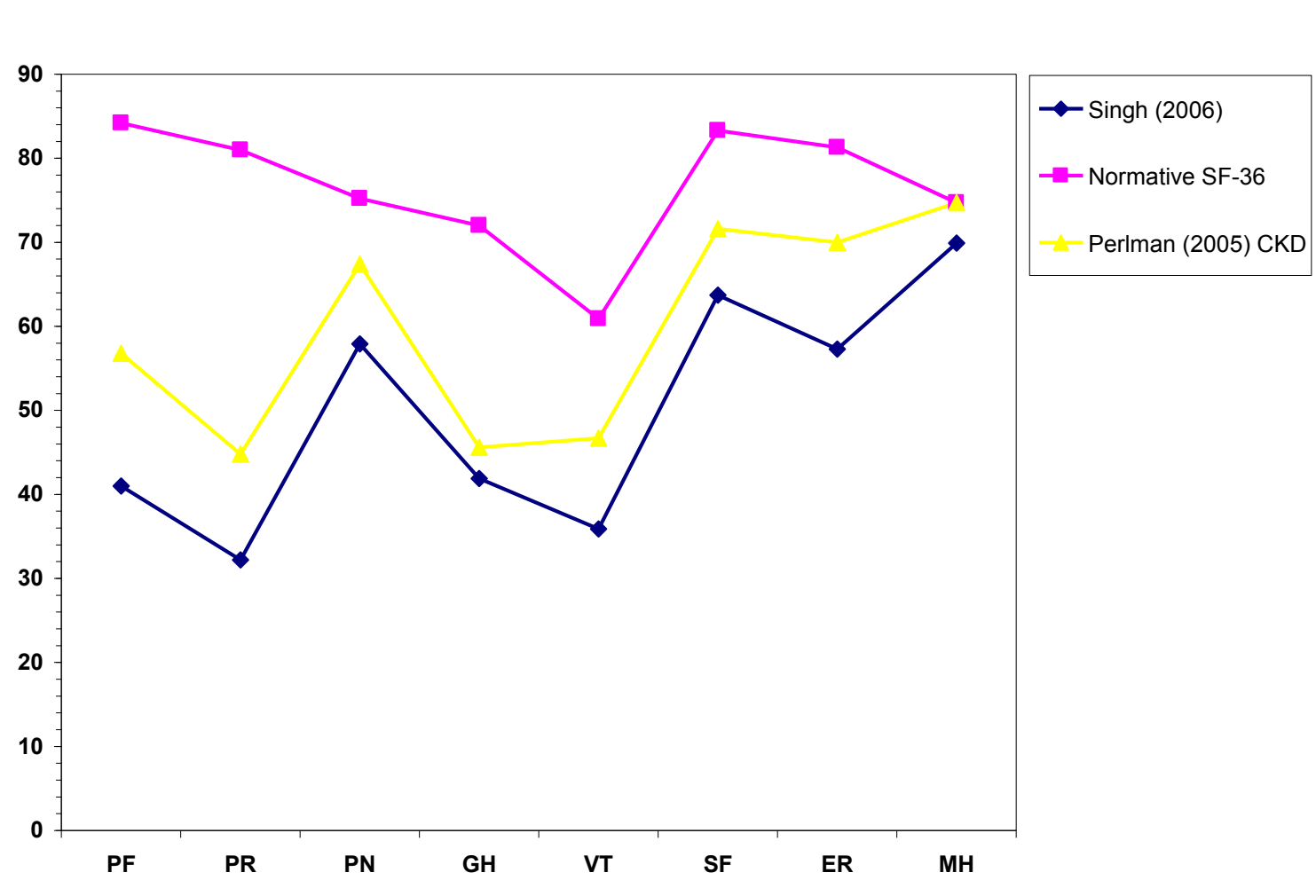
HRQL VERSUS EFFICACY AND SAFETY

- HRQL is the ultimate outcome of health care interventions (implies survival)
- No single outcome adequately represents results of treatment
- HRQL assesses integrated effects of treatment

HRQL AND CHRONIC KIDNEY DISEASE

- CKD is associated with broad and meaningful impairment to HRQL outcomes
- HRQL measures predict mortality in CKD patients, even after adjustment for demographic and clinical variables
- Treatments for anemia have demonstrated impact on symptoms and functioning

HEALTH STATUS IMPAIRED IN CKD PATIENTS



COMPARISON OF MEAN QOL SCORES FOR PATIENTS WITH CKD, END-STAGE RENAL DISEASE, AND THE GENERAL POPULATION

	RRI-CKD Study	Dialysis Patients*	General Population†
PCS	37.3 (11.6)	33.1 (10.7)‡	50.0 (10.0)‡
MCS	50.0 (10.3)	46.6 (11.9)‡	50.0 (10.0)
Physical Function	56.8 (29.4)	40.8 (29.4)‡	84.2 (23.3)‡
Physical Role	44.9 (42.6)	31.7 (39.3)‡	81.0 (34.0)‡
Physical Pain	67.4 (27.1)	59.0 (29.2)‡	75.2 (23.7)‡
General Health	45.6 (20.0)	40.2 (22.1)‡	72.0 (20.3)‡
Mental Health	74.7 (17.4)	67.3 (21.7)‡	74.7 (18.1)
Emotional Role	70.0 (40.4)	51.8 (44.8)‡	81.3 (33.0)‡
Social Function	71.6 (28.2)	62.1 (29.1)‡	83.3 (22.7)‡
Vitality	46.7 (22.7)	42.9 (23.2)‡	60.9 (21.0)‡

NOTE: All scales are from 0 to 100, with higher numbers indicating better QOL.
*Data from the DOPPS (n = 2,855).
†Data from the SF-36 manual (n = 2,474).
‡$P < 0.0001$ compared with patients with CKD.

Source: Perlman et al. 2005

SURVIVAL PROPORTIONAL HAZARDS MODEL*

Covariate	Sign of Coefficient	Unit of Analysis	Percent Survival Change Per Unit Change†	95% Confidence Interval for Percent Survival Change Per Unit	*P* Value‡
Albumin	-	0.1 g/dL	+10.0	6.2 to 14	<0.0001
Age	+	1 yr	-2.8	1.4 to 4.1	0.0002
nPCR	-	0.1 g/kg/d	+17.2	5.4 to 27	0.0053
PCS	-	5 points	+10.4	1.1 to 18	0.0226
Kt/V	-	0.1 Kt/V	+10.8	0.6 to 19	0.0373
Is diabetic	+				0.1739
Is not white	-				0.1773
Is male	+				0.4492
MCS	-	5 points	+1.4	-6.5 to 8.9	0.7280

* For the model, $P < 0.0001$ (Wald)
† The percent change in the probability of survival per unit change of the covariate.
‡ Chi-squared.

Source: DeOreo et al. 1997

CORRELATIONS BETWEEN CHANGES IN HCT AND HRQL SCORES

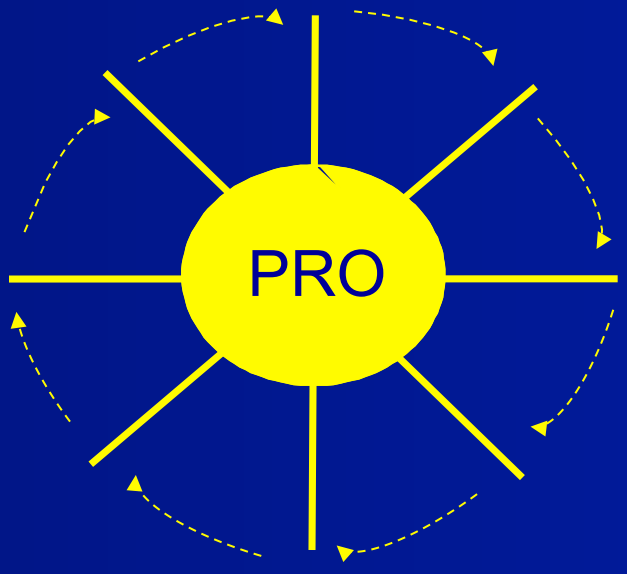
	CHANGES IN HCT	
Score	Week 16	Week 48
Energy	0.35*	0.37*
Physical function	0.37*	0.35*

* $P < 0.05$

Source: Revicki et al. 1995

A. Identify Concepts & Develop Conceptual Framework

Identify concepts and domains.
Identify intended application and population
Hypothesize expected relationships among concepts



D. Modify Instrument

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

B. Create Instrument

Generate items
Choose data collection method
Choose recall period
Choose response options
Evaluate patient understanding
Develop instructions
Identify scoring
Format instrument
Assess burden
Confirm conceptual framework
Finalize items & instrument

C. Assess Measurement Properties

Evaluate reliability, validity, and ability to detect change
Propose methods for interpretation

MEASUREMENT ATTRIBUTES AND REVIEW CRITERIA FOR HRQL INSTRUMENTS

Attribute	Criteria
1. Conceptual and measurement model	Content validity and framework for concept to be measured
	Conceptual and empirical basis for item content and subscales
2. Reliability	Internal consistency (homogeneity)
	Reproducibility (test-retest reliability)
	Inter-rater reliability
3. Validity	Degree to which the instrument measures what it intends to measure.
	Construct-related
	Criterion-relayed

MEASUREMENT ATTRIBUTES AND REVIEW CRITERIA FOR HRQL INSTRUMENTS (CONTINUED)

4. Responsiveness	An instrument's ability to detect change over time
5. Interpretability	Degree to which one can assign easily understood meaning to an instrument's quantitative scores.

RESPONSIVENESS AND MID

Recommended approach, and evolving consensus:

- Estimate the MID based on several anchor-based methods, with relevant clinical or patient-based indicators.
- Examine various distribution-based estimates (i.e., effect size, standardized response mean, etc.) as supportive information.
- Triangulate on a single value or small range of values for the MID.
- Confidence in a specific MID value evolves over time and is confirmed by additional research evidence, including clinical trial experience.

Source: Revicki et al. (in press)

HRQL MEASURES USED IN CKD

- Kidney Disease Questionnaire
  - Physical symptoms, fatigue, relationships, depression, frustration
- SF-36 Health Survey
  - Physical function, pain, vitality, role-physical, role-emotional, social function, general health, mental health
- Kidney Disease Quality of Life Questionnaire
  - Includes SF-36
  - Kidney disease-specific domains

Properties of HRQL Measures in Anemia in CKD

Conceptual and Measurement Model	KDQ	SF-36	KDQOL-SF
Concept to be measured described	++	++	++
Content validity based on literature review	0	++	++
Content validity based on focus groups or cognitive debriefing interviews with patients with chronic renal disease and anemia	++	++	++
Content validity based on clinician or expert review	++	++	++
Specific conceptual framework which identifies concept and unique items (e.g., exploratory factor analysis or via literature)	++	++	++
Evidence of scale variability (i.e., item and scale distributions, frequencies)	0	++	++
Intended level of measurement (e.g., ordinal, interval, ratio)	+	+	+
Record of item development (i.e., rational for item retention and deletion)	++	++	++
Rationale for recall period	0	0	0
Reliability			
Internal consistency reliability	+	++	++
Reproducibility	++	+	+

Properties of HRQL Measures in Anemia in CKD (continued)

Conceptual and Measurement Model	KDQ	SF-36	KDQOL-SF
Validity			
Content-related (see above)	++	++	++
Construct-related	++	++	++
Criterion-related	0	+	0
Responsiveness			
Anchor-based	+	+	+
Distribution-based methods (i.e., effect size, SEM)	+	+	+
Interpretability			
MID estimates	0	0	0
Responder analysis	0	0	0
Respondent Burden			
Time needed to complete	0	++	++
Reading and comprehension levels	0	0	0
Special requirements	0	0	0
Degree of missing data	0	0	0

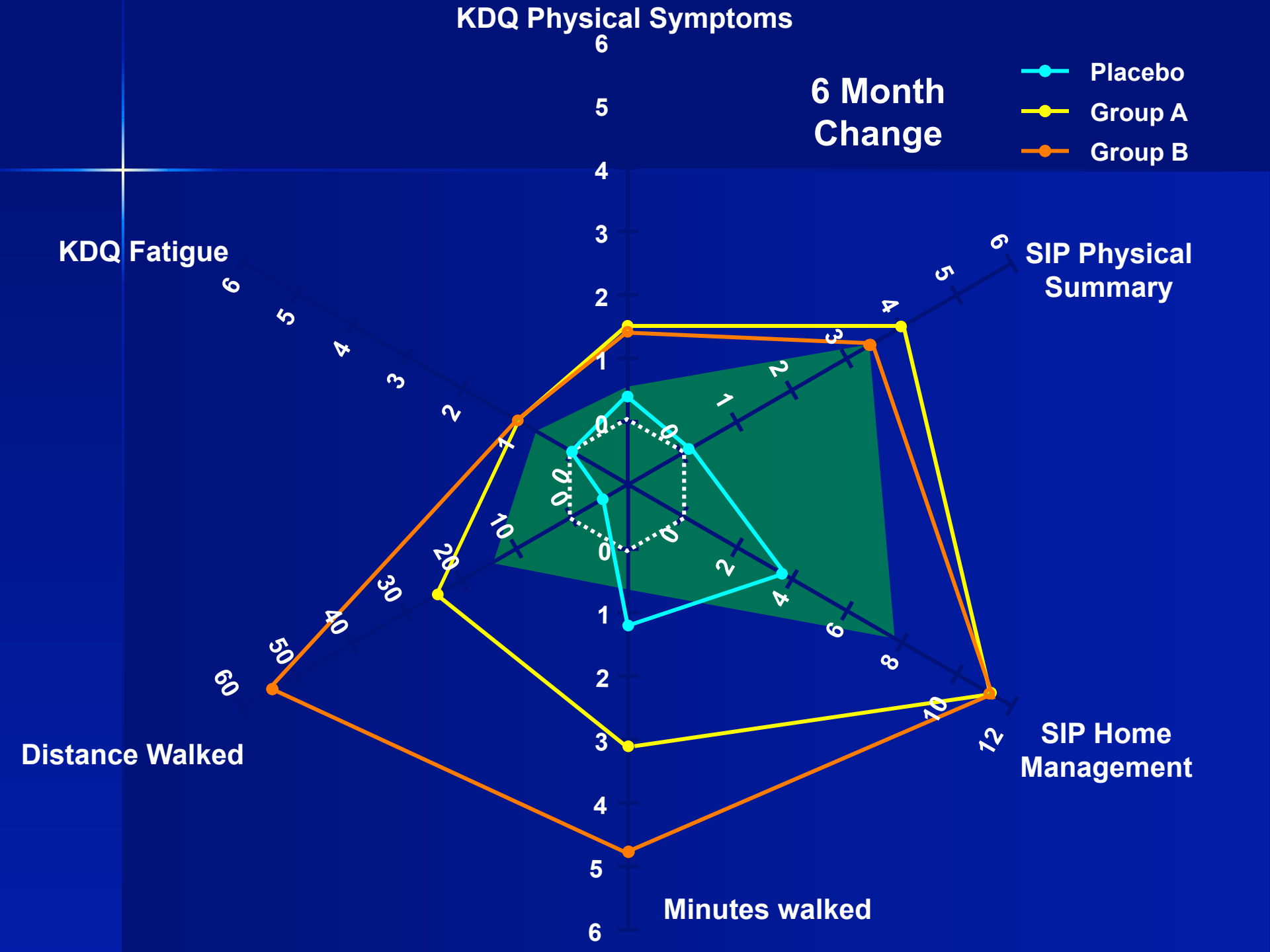
Properties of HRQL Measures in Anemia in CKD (continued)

Conceptual and Measurement Model	KDQ	SF-36	KDQOL-SF
Alternate modes of administration			
Self-report	++	++	++
Interviewer-administered	0	++	0
Cultural and language adaptations or translations			
# of available countries with cultural and linguistic translations	?	22	22
# of available translations with evaluations of measurement properties	?	6	6

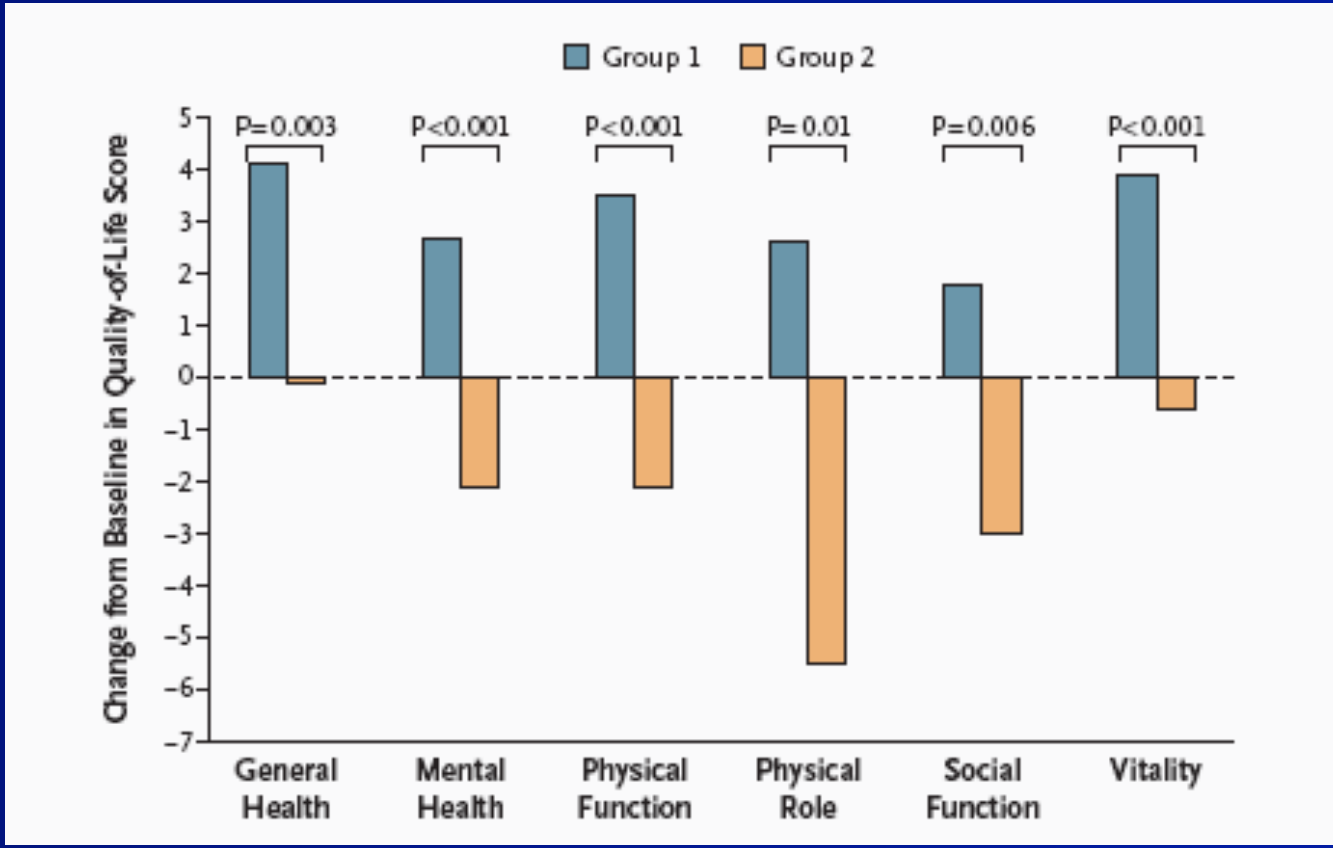
RESULTS OF CESG ITT ANALYSES: TREATMENT VERSUS PLACEBO OVER TIME

MEASURE	Mixed Model p-value	LOCF p-value
Exercise Capacity		
Treadmill Stress Test	0.0001*	0.0001*
6-Minute Walk	0.0498	0.0508
Physical Function		
SIP Physical Summary	0.0015*	0.0004*
Ambulation	0.0077	0.0127
Body Care & Movement	0.0068	0.0016*
SIP Home Management	0.0291	0.0387
Symptoms		
KDQ Fatigue	0.0001*	0.0001*
KDQ Energy Symptom	0.0118	0.0314
KDQ Weakness Symptom	0.0110	0.0187
KDQ Physical Symptoms	0.0001*	0.0001*
KDQ Shortness of Breath Symptom	0.7969	0.7961

* Statistically significant after application of Bonferroni adjustment


KDQ Physical Symptoms
6 Month Change
Placebo
Group A
Group B
KDQ Fatigue
SIP Physical Summary
Distance Walked
SIP Home Management
Minutes walked

CHANGES IN HRQL SCORES IN HIGH AND LOW HGB GROUP



Source: Drueke et al 2006

PHYSICAL FUNCTION SUPPORTING EVIDENCE

MEASURE	STUDY	DESIGN	THRESHOLD*	CHANGE	P-value
Physician-assessed Karnofsky	Evans (19900	Single-arm	10	5.0	<0.001
	Delano (1989)	Single-arm	10	10.6	Not evaluated
	Harris (1991)	Single-arm	10	12.0	<0.0001
Patient-reported Karnofsky	Moreno (1996)	Controlled	10	12.6	<0.0001
	Moreno (2000)	Single-arm	10	2.8	<0.01
SIP Physical Function	McMahon (1992)	Cross-over	5.1	7.4	<0.01
	Moreno (1996)	Controlled	6.8	5.8	<0.0001
	McMahon (2000)	Cross-over	3.0	2.7	<0.01
KDQ Physical Symptoms	Muirhead (1992)	RCT	0.5	0.9	<0.005
	Foley (2000)	RCT	0.5	1.1	Not evaluated
	Furuland (2003)	RCT	0.5	0.7	<0.05
SF-36 Physical Functioning	Beusterien (1996)	Controlled	8	3.7	<0.05
	Besarab (1998)	RCT	8	Not evaluable	<0.05
Other: “Physical Activity”	Barany (1990)	Single-arm	1	1	<0.05
Other: “Physical Activity”	Barany (1993)	Controlled	0.04	0.06	<0.01

Clinically Meaningful or Statistically significant

Not Clinically Meaningful or Statistically significant

*Threshold indicates established clinically meaningful difference as defined in literature, or minimally important effect size of ½ SD baseline value

ENERGY SUPPORTING EVIDENCE

MEASURE	STUDY	DESIGN	THRESHOLD*	CHANGE	P-value
KDQ Fatigue	Muirhead (1992)	RCT	0.7	0.8	<0.05
	Foley (2000)	RCT	0.71	0.04	<0.01
Fatigue Symptoms	Evans (1990)	Single-arm	0.19	0.26	<0.001
	Harris (1991)	Single-arm	0.87	1.66	<0.0001
NHP: Energy	Evans (1990)	Single-arm	Not evaluable	27	<0.001
NHP: Energy (%)	Auer (1990)	Single-arm	0.24	0.5	<0.0005
	Auer (1992)	Single-arm	0.22	0.52	<0.0001

Clinically Meaningful or statistically significant

Not clinically meaningful or statistically significant

*Threshold indicates established clinically meaningful difference as defined in literature, or minimally important effect size of ½ SD baseline value

EXERCISE CAPACITY SUPPORTING EVIDENCE

STUDY	PROTOCOL	BASELINE	POST	CHANGE	P-value
VO_2 (ml/kg/min)					
Mayer (1988)	Cycle Ergometer Test†	16.0	23.2	7.2	<0.02
Baraldi (1990)	Cycle Ergometer Test†	24.1	32.6	8.5	<0.05
Grunze (1990)*	Cycle Ergometer Test†	1.19	1.37	0.18	<0.05
Robertson (1990)	Cycle Ergometer Test†	15.3	17.8	2.5	<0.0005
Lundin (1991)	Cycle Ergometer Test†	15.1	22.7	7.6	<0.003
Metra (1991)	Cycle Ergometer Test†	21.4	26.6	5.2	<0.001
Lewis (1993)	Weber Treadmill Protocol	18.7	25.1	6.4	<0.05
Marrades (1996)	Cycle Ergometer Test†	25.4	33.1	7.7	0.003
Treadmill Test (minutes walked)					
Robertson (1990)	Cycle Ergometer Test†	6.45	7.60	1.15	<0.0005
Lundin (1991)	Maximal Treadmill Test	6.0	9.1	3.1	<0.001
Hase (1993)	Bruce Treadmill Protocol	4.63	6.40	1.77	<0.01
Lewis (1993)	Weber Treadmill Protocol	15.2	21.4	6.2	<0.05
Metra (1991)	Cycle Ergometer Test†	9.62	11.9	2.32	<0.05
6-minute walk					
Harris (1991)	6 Minute Walk Test‡	400	600	200	<0.001

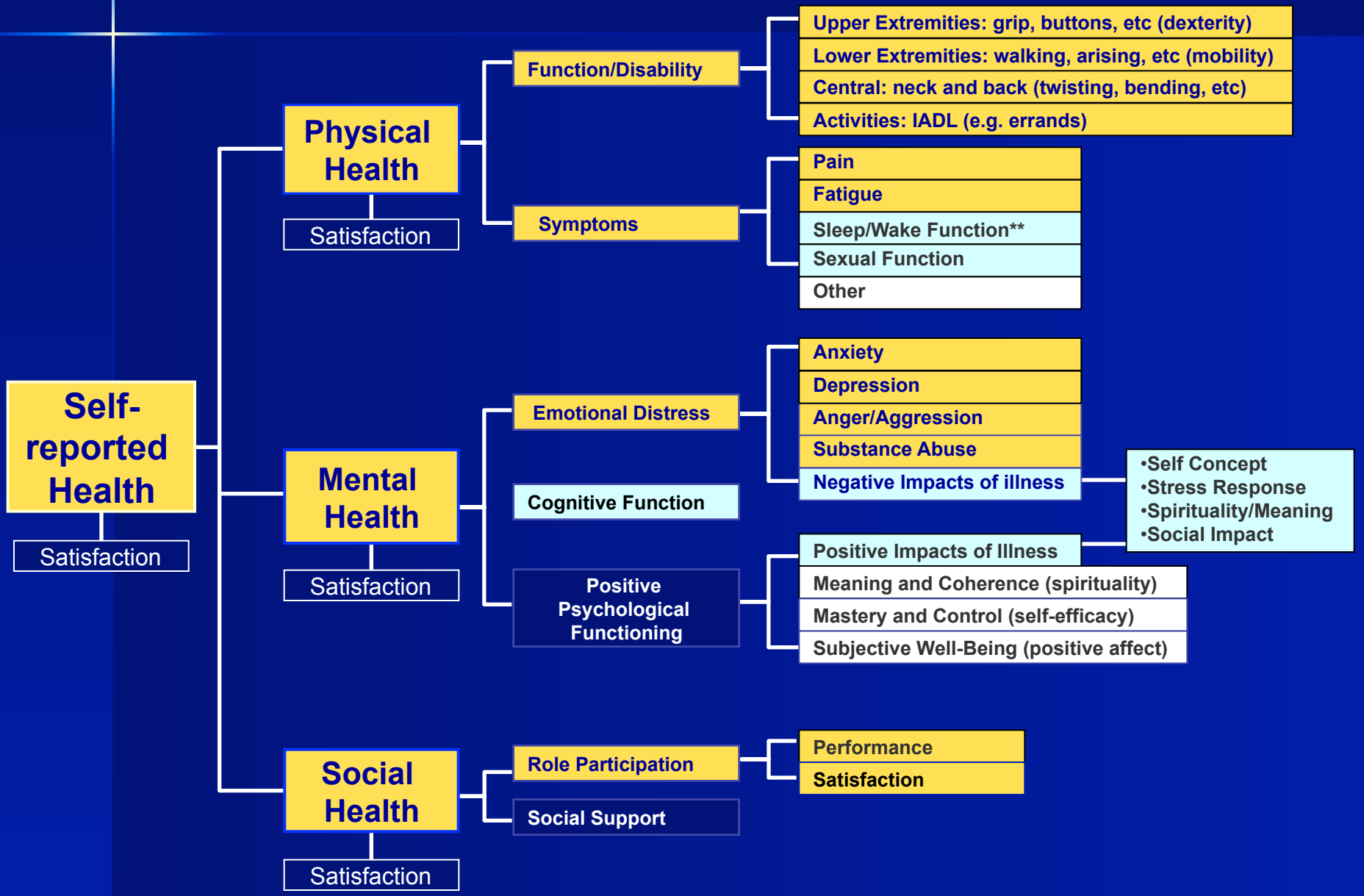
Statistically significant

† = Cycle ergometer tests vary in cycle speed, inclination, and termination ; ‡ = meters walked, * = L/min

FUTURE OF PRO MEASUREMENT: NIH PROMIS

- Improve assessment of self- reported symptoms and domains of HRQL for application across a wide range of chronic diseases
- Develop and test a large bank of items for measuring PROs
- Develop computer-adaptive testing (CAT) for efficient assessment of PROs
- Create a publicly available, flexible, and sustainable system allowing researchers to access to item banks and CAT tools

PROMIS DOMAIN HIERARCHY



Items from Instrument A

Items from Instrument B

Items from Instrument C

New Items

Item Pool

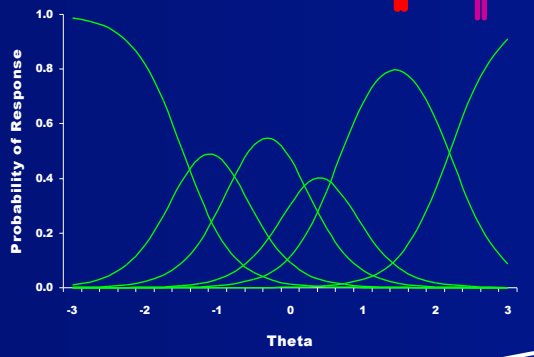
Content Expert Review

Focus Groups

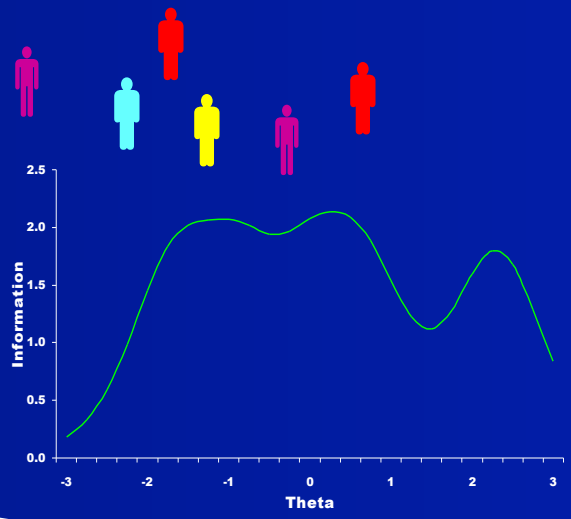
Cognitive Testing

Secondary Data Analysis

Questionnaire
administered to large representative sample



Item Response Theory (IRT)



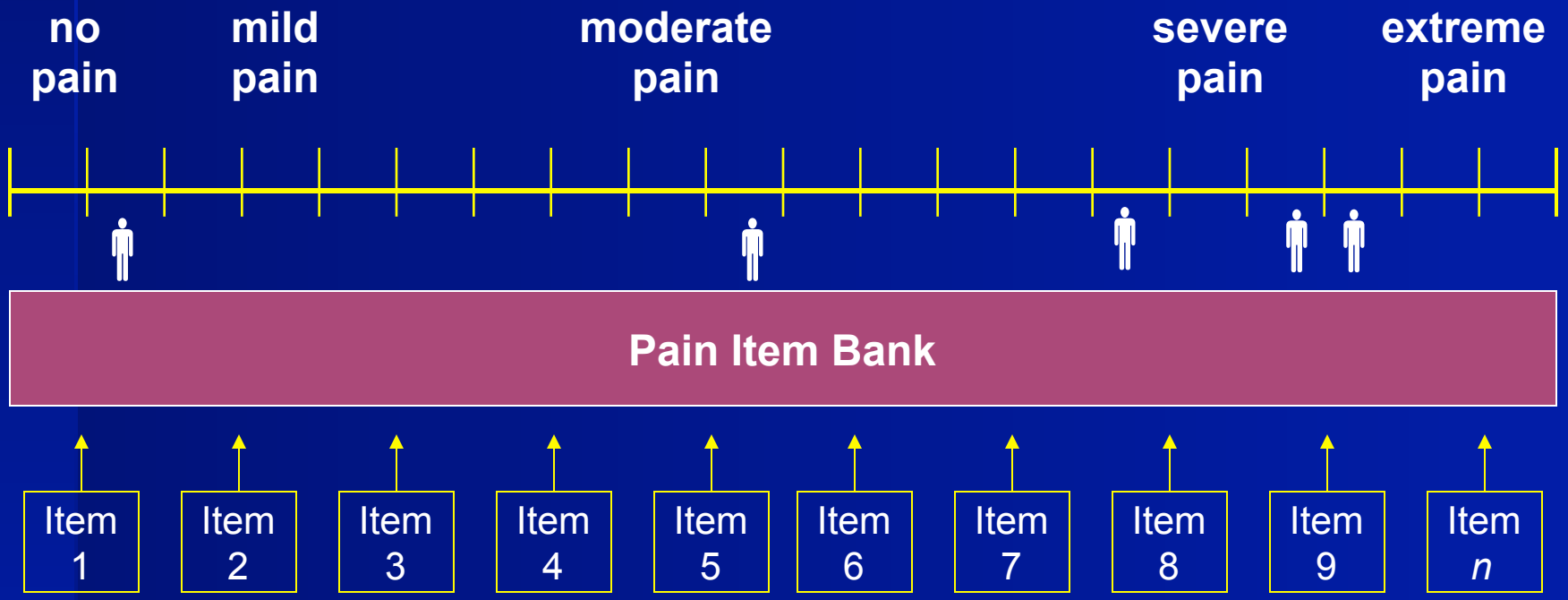
Item Bank
(IRT-calibrated items reviewed for reliability, validity, and sensitivity)

Short Form Instruments

CAT

ITEM BANKS

An item bank is a large collection of items measuring a single domain, e.g., pain…



These items are reviewed by experts, patients, and methodologists to make sure:

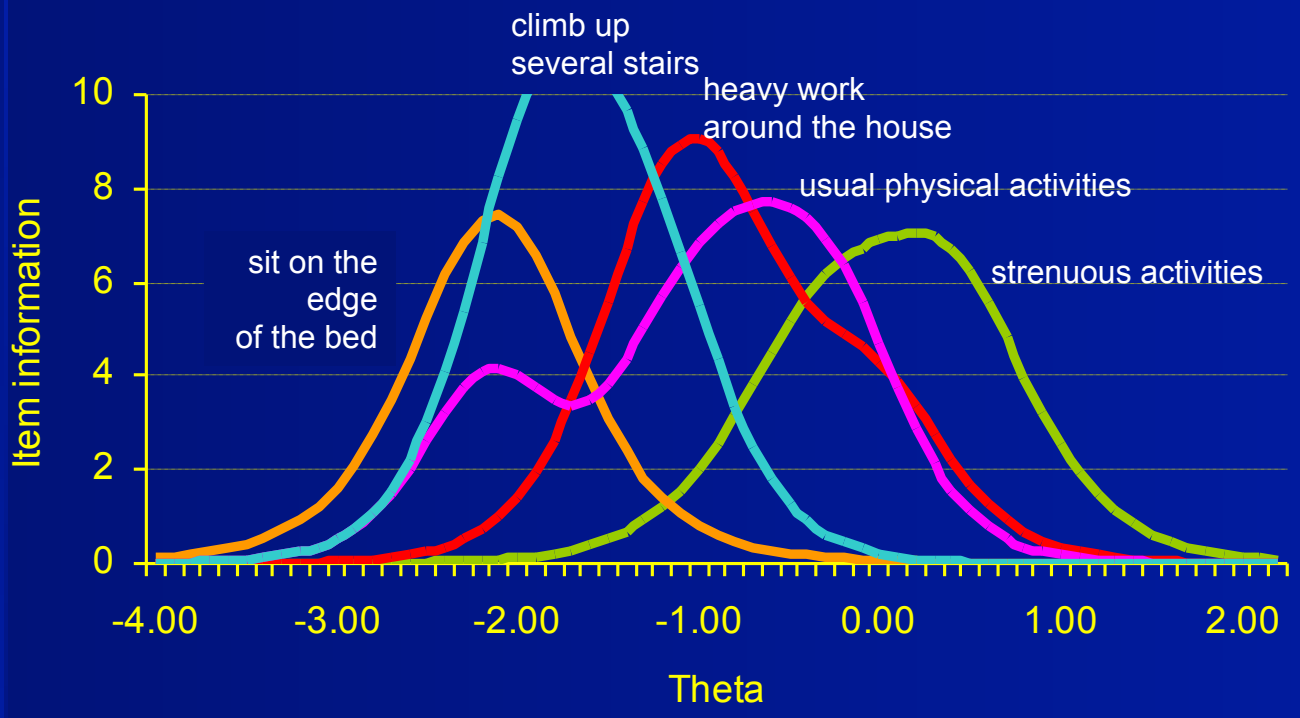
- Item phrasing is clear and understandable for those with low literacy
- Item content is related to pain assessment and appropriate for target population
- Item adds precision for measuring different levels of pain

ITEM RESPONSE THEORY MODELS

- IRT models enable reliable and precise measurement of PROs
  - Fewer items needed for equal precision
  - Makes assessment briefer
- More precision gained by adding items
  - Reducing error and sample size requirements
- Error is understood at the individual level
  - Allowing practical individual assessment

RANGE OF MEASUREMENT

Are you able to …
Does your health now limit you in ...



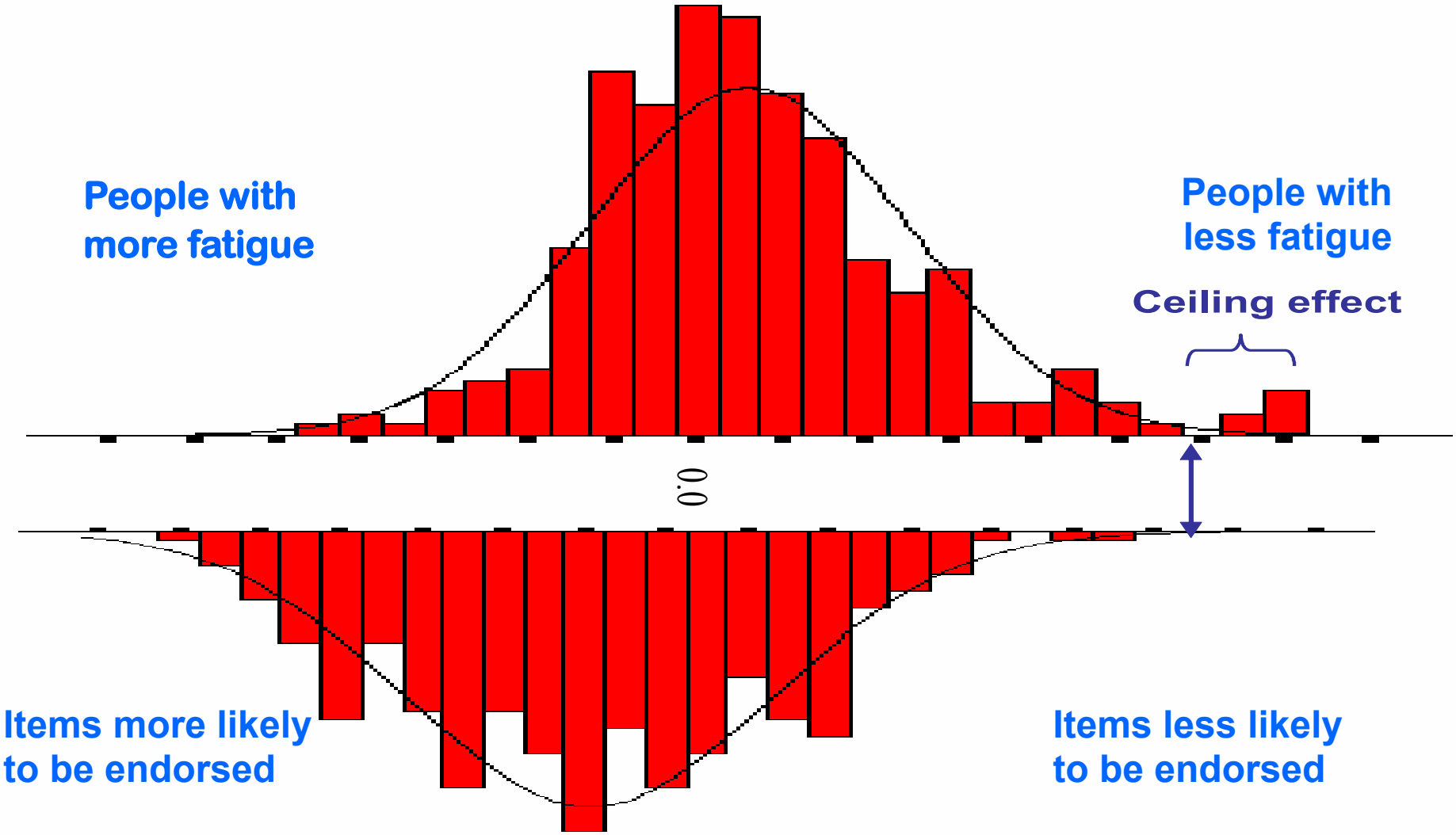
5 = Not at all
4 = Very little
3 = Somewhat
2 = Quite a lot
1 = Cannot do

5 = Without any difficulty
4 = With a little difficulty
3 = With some difficulty
2 = With much difficulty
1 = Unable to do

Disability

Physical Function

PEOPLE AND ITEMS DISTRIBUTED ON THE SAME METRIC: FATIGUE



THE ADVANTAGES OF CAT-BASED ASSESSMENT

- Provide an accurate estimate of a person's score with the minimal number of questions
  - Questions are selected to match the health status of the respondent
- CAT minimizes floor and ceiling effects
  - People near the lower or upper extremes of a scale will receive items that are designed to assess their health status

SUMMARY

- Good availability of HRQL instruments for assessing outcomes in CKD patients with anemia
  - Evaluating treatment effects
  - Monitoring health status
- Good content coverage and psychometrically sound
  - Reliability
  - Validity
  - Responsiveness
- Future research needs to focus more on interpretation and clinical significance
- PROMIS may provide relevant and psychometrically sound measures of pain, fatigue, physical functioning and other domains

CONCLUSION

- Relevancy of HRQL data for regulatory and clinical decision making depends on the strength of the research evidence on added value
- Safety and clinical efficacy data are insufficient for the comprehensive understanding of medical treatments
- HRQL is the ultimate outcome of health care interventions and is the key to assessing effectiveness beyond safety and efficacy
- Patients, clinicians and regulatory agencies need HRQL data to make decisions about the benefit and risk of new therapies

THE GOAL OF MEDICINE (C 1400)

"To cure sometimes, to relieve often, to comfort always"