Review of Methods for Evaluating Important Change in Patient-Reported Pain Measures

Kathleen W. Wyrwich, Ph.D.
Saint Louis University

Brief Outline

1. Underlying Issues of Interpreting Group vs. Individual Change

3. Perspective of Different Stake Holders

5. Methods for Interpreting Change
   - Anchor-Based
   - Distribution-Based

6. Challenges Ahead
   - Should We Put the “C” in MICD...and How?

Underlying Issues of Interpreting Group vs. Individual Change

4 Points

Group vs. Individual Change Point 1

1. Group change evaluation methods are often based on mean differences that satisfy a statistical criterion ($p < .05$)

   - paired t-tests
   - repeated measures
   - ANOVA and ANCOVA
   - general estimating equations (GEE)

Group vs. Individual Change Point 2

1. Achieving the statistical significance standard ($p<.05$) is dependent on

   – the variation (σ^2) and
   – sample size (n).

Group vs. Individual Change Point 3

1. Meaningful individual change cannot be extracted from statistically significant group change because:

   we cannot infer that each individual in the “changed” group uniformly experienced the group mean change

Group vs. Individual Change Point 4

1. Meaningful individual change cannot be extracted from statistically significant group change because:

the statistical threshold for a significant group change may have no relation to a meaningful or clinically relevant difference for individual patients

An Example

A Comparison of Osteopathic Spinal Manipulation with Standard Care for Patients with Low Back Pain

G. Andersson, T. Lucente, A. Davis, R. Kappler, J. Lipton, S Leurgans

NEJM 1999

Andersson et al.

Intervention Group
(n = 83)
- osteopathic manual therapy

Control Group
(n = 72)
- standard medical therapies

Outcomes--Change on:
- Roland–Morris Questionnaire
- Oswestry
- VAS pain scale

Andersson et al.

Outcomes

Baseline to 12 week follow-up on

- Roland–Morris Questionnaire (0– 24)
- Oswestry Questionnaire (0 – 50)
- VAS pain scale (0 – 100)

All better(=0) to worse scales

Andersson et al.

	Baseline		Completion	
Scales	Inv.	Con.	Inv.	Con.
RMQ	7	7	2	1
Oswestry	25	23	12	10
VAS	49	45	16	19

No differences significant at the $p < .05$ level

What Do We Know About Change in Pain From These Results?

Did everyone in both groups change about 23-26 mm on the VAS Pain Scale?

Was the statistically non-significant change in these scales meaningful or important to the enrollees?

If only a few more patients had been enrolled would change on any of these scales reached statistical significance?

If 1000 patients were enrolled in this trial, how small could the pre-post change be and still achieve statistical significance?

What Do These Results Tell Us About Meaningful Change Among the Patients Enrolled?

Not Very Much!

Foundation of Clinical Significance vs. Statistical Significance

Statistically significant (or non-Significant!) group change does not necessarily imply a meaningful difference for patients

But how big is a "meaningful differences"?

Why are Individual Change Standards Needed?

- To meaningfully interpret how interventions and treatments effect HRQoL, and to improve the quality of patient management
- To classify a patient's change, based on the standard, as:
  - improved
  - stable
  - declined

- To improve estimation of the likelihood of HRQoL change through event modeling
  - polytomous regression
  - logistic regression
  - proportional hazards regression

Who are the Stake Holders in Know the Magnitude of an Important Change?

Stake Holders

General Population

Clinicians

Pharmaceutical and Medical Device Developers

Insurance Payers

Patients

Governments

Regulators

Clinically Significant Change: Patient Perspective

A change which ***patients perceive as beneficial*** (or detrimental) and important, and which may ***prompt them to seek healthcare or request changes in their treatment*** *in the absence of troublesome side effects and excessive costs*

Clinically Significant Change: Clinician Perspective

The smallest difference or change that leads ***the clinician to recommend a treatment or therapy*** to their patient

Clinically Significant Change: Population Perspective

Allocation of resources to maximize measurable benefits to society as a whole

Other Stakeholders?

Pharmaceutical and Medical Device Developers

Hope to demonstrate the value of their products and market these interventions in a way that improves the lives of patients

Insurance Payers

Have a financial responsibility to all of their members to understand the value of covered treatments

Regulators

Seek to understand the consequences of new therapies

Governments

Seeks to monitor changes in health status of populations and identify the impact of treatments on populations

So…When Determining a Clinically Important Change Standard…

Perspective can influence the assessment approach and the way in which a clinically important difference is determined

Brief Outline

1. ~~Underlying Issues of Interpreting Group vs. Individual Change~~

3. ~~Perspective of Different Stake Holders~~

5. Methods for Interpreting Change
   - Anchor-Based
   - Distribution-Based

6. Challenges Ahead
   - Should We Put the “C” in MICD...and How?

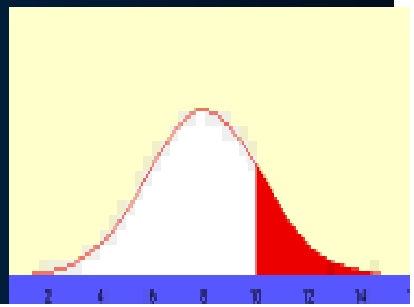
Interpretation of quality of life changes

Lydick, E. and Epstein, R.

Quality of Life Research 1993

Lydick and Epstein, 1993

Anchor-Based **Distribution-Based**



Anchor-Based Methods

- Within-Person Change
- Between-Persons Differences
- Relevant Anchors

Within-Person Studies

Measurement of health status: ascertaining the minimal clinically important difference

Jaeschke R, Singer J, Guyatt G.

Controlled Clinical Trials 1989

Jaeschke et al.

CHQ

- Dyspnea (5)
- Fatigue (4)
- Emotional Function (7)

CRQ

- Dyspnea (5)
- Fatigue (4)
- Emotional Function (7)
- ~~Mastery (4)~~

1. Define a Minimal Clinically Important Difference (MCID)

"the smallest difference in a score of a domain ...that patients perceive to be beneficial and that would mandate...a change in the patient's management."

2. Convene a Clinical Consensus Panel

3 point $\triangle$ in Dyspnea (.6 per item)

2 point $\triangle$ in Fatigue (.5 per item)

4 point $\triangle$ in Emotional (.57 per item)

3. Measure Within-Patient Global Change Ratings

Patients are asked a global change question for each dimension.

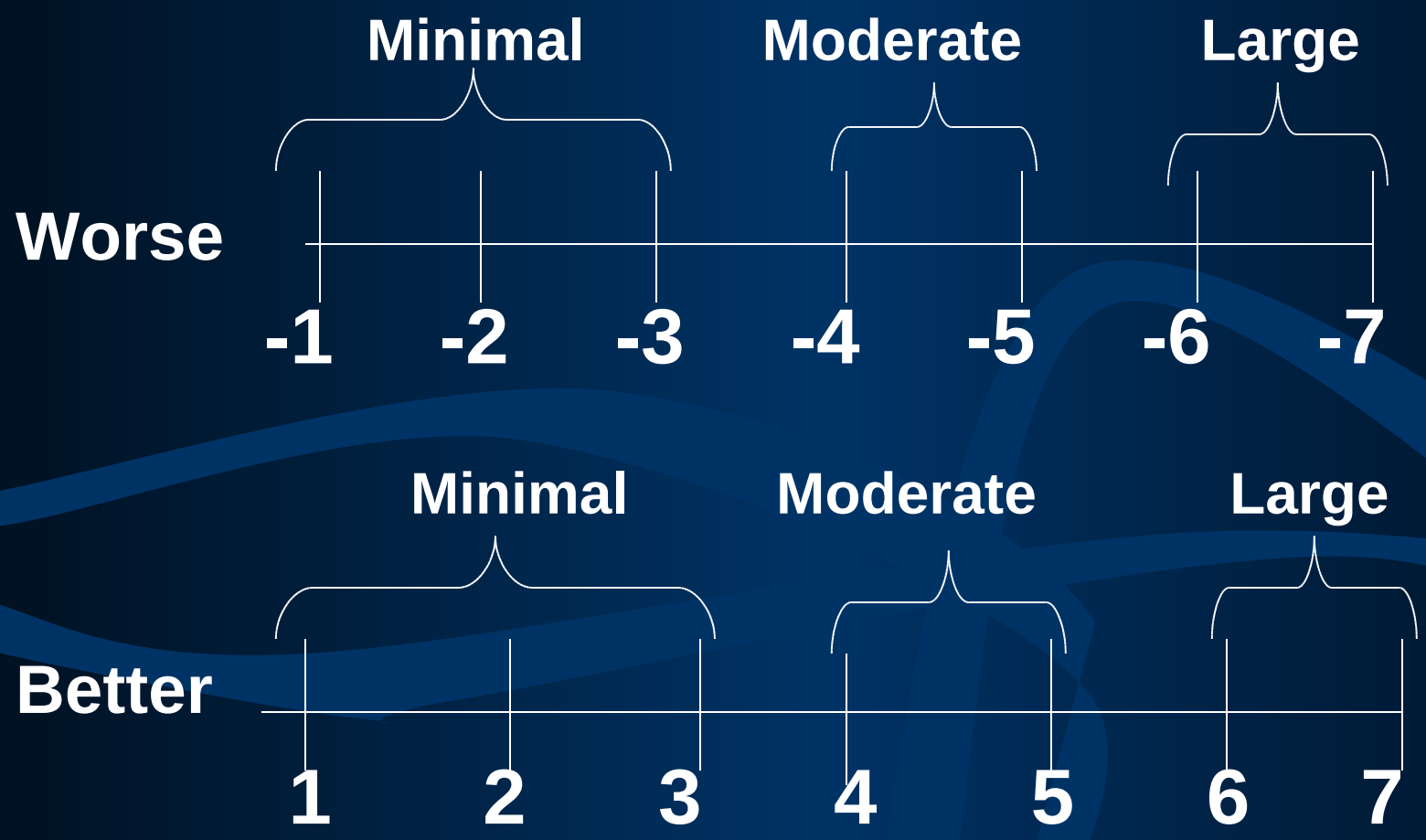
"Has there been a change in your level of fatigue since your last visit?"

Worse **About the same** **Better**

4. If “worse” or “better” rate change on the following response scale

-7	A very great deal worse	7	A very great deal better
-6	A great deal worse	6	A great deal better
-5	A good deal worse	5	A good deal better
-4	Moderately worse	4	Moderately better
-3	Somewhat worse	3	Somewhat better
-2	A little worse	2	A little better
-1	Almost the same, hardly any worse at all	1	Almost the same, hardly any better at all

5. Determine Global Change Classifications



6. Determine the Mean of the Change Scores for Patients with a Minimal Change

Average the dimension changes scores among those subjects with:

a minimally better change or
a minimally worse change

MCID Results

.43 per item in Dyspnea

.64 per item in Fatigue

.49 per item in Emotional Function

MCID Conclusion

Consensus Panel	Global Change Ratings
.6	.43
.5	.64
.57	.49

.5 per item in each dimension

Advantages of Within-Person Change Methods

- Light-Weight and Portable
- Easy to Calculate Results

Problematic Aspects of Within-Person Change

Methodological problems in the retrospective computation of responsiveness to change: the lessons of Cronbach

Norman G, Stratford P, Regehr G.

Journal of Clinical Epidemiology 1997

Norman et al.

Reconstructive memory is poor

- systematic underestimation of initial state
- highly correlated with present state

Clinical change levels arbitrarily defined

No test-retest reliability data

Other Problems with Within-Person Studies

Small samples

Clinical Consensus Panel

- Abstract reference to patients
- Pooling of the CHQ and CRQ

No ratings made by the patients' own physicians

Anchor-Based Methods

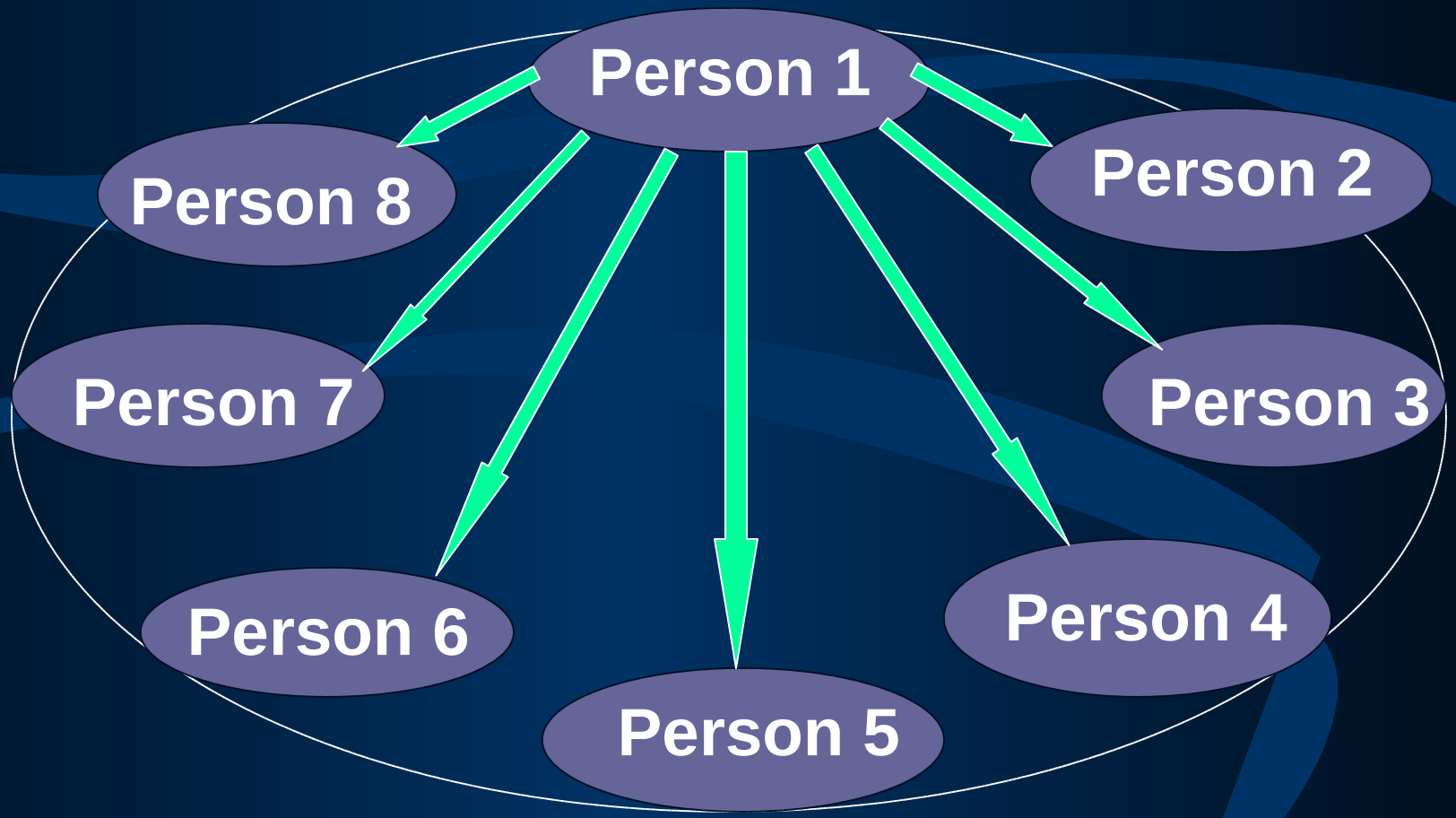
- Within-Person Change
- Between-Persons Differences
- Relevant Anchors

Assessing the minimal important difference in symptoms: a comparison of two techniques

Redelmeier D, Guyatt G, Goldstein R

Journal of Clinical Epidemiology, 1996

Between-Persons Differences



Between-Persons Differences

Compared to this person,
your energy is _____.

much better
somewhat better
a little better
about the same
a little bit worse
somewhat worse
much worse

Between-Persons Results

CRQ Dimensions	MID per item
Dyspnea	.09
Fatigue	.50
Emotional Function	.83
Mastery	.23

Between-Persons Results

- Excluding the dyspnea results
- Pooling the remaining 3 dimensions

CRQ MID Estimate

.53

95% CI (.39 to .67)

Advantages of Between-Persons Methods

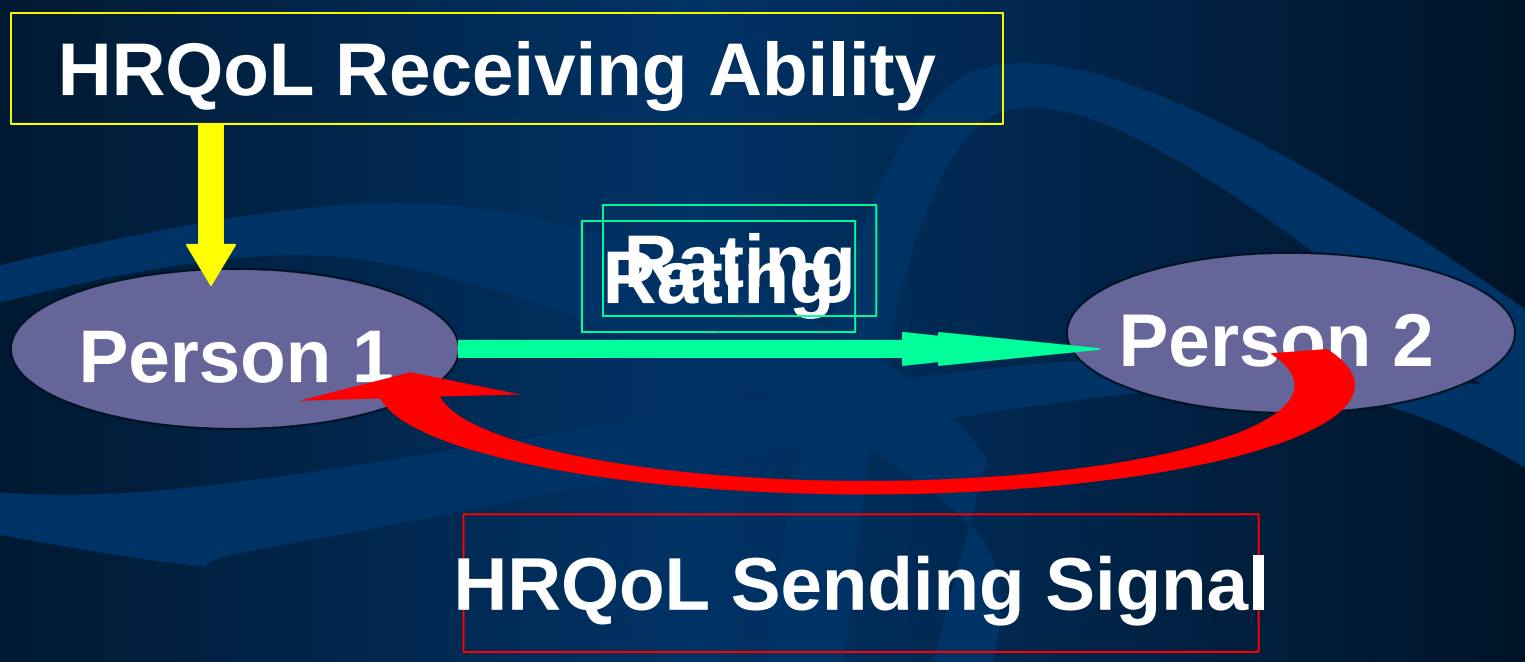
"original and innovative study… in an area that is methodologically challenging and complex"

Wright 1996

Problematic Aspects

- A cross-sectional difference is not individual longitudinal change
- “Double counts” for each pair of pts.
- Possibility of additional sources of measurement error

New Sources of Measurement Error



Anchor-Based Methods

- Within-Person Change
- Between-Persons Differences
- Relevant Anchors

Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis

Kosinski M, Zhao S, Dedhiya S, Osterhaus J, Ware J

Arthritis and Rheumatism 2000

Kosinski et al.

Outcome Measures: Scales of the SF-36

Relevant Anchor: Number of tender joints in arthritis patients

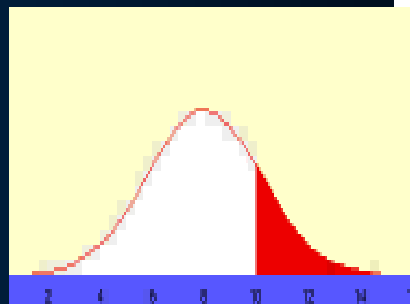
No Improvement: < 1% decrease in number of tender joints

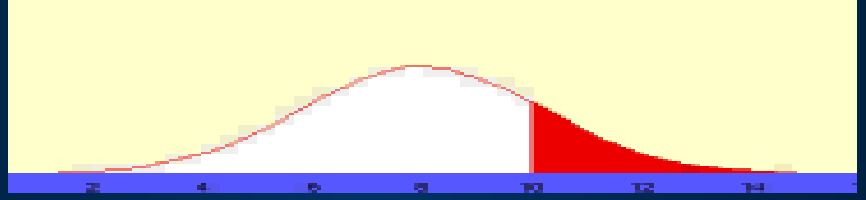
Improvement: 1-20% decrease in the number of tender joints

Calculation: MCID = mean change score in each SF-36 scale among patients meeting the improvement criterion

Lydick and Epstein, 1993

Anchor-Based **Distribution-Based**





Distribution-Based Methods

- Effect Size
- Standard Error of Measurement

Effect Size$_{\text{Group}}$ (δ_g)

$$\delta_g = \frac{m_2 - m_1}{s_1}$$

where

m_1 = mean at baseline
m_2 = mean at follow-up
s_1 = standard deviation at baseline

Effect Size$_{\text{Individual}}$ (δ_i)

$$\delta_i = \frac{x_2 - x_1}{s_1}$$

where

x_1 = score at baseline
x_2 = score at follow-up
s_1 = standard deviation at baseline

Effect Size Standards

	Group (Cohen, 1977)	Individual (Testa, 1986)
Small Change	.2	.2
Moderate Change	.5	.6
Large Change	.8	1.0

Advantages of Individual Effect Size Standards

- Easy to calculate
- Easy to communicate

Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation

Norman GR, Sloan JA, Wyrwich KW

Medical Care, 41(5): 582-592, 2003.

Literature Search

Intersection of “quality of life” with:

- meaningful change, meaningful difference
- important change, important difference
- relevant change, relevant difference
- effect size
- minimally important change
- clinical significance

Criteria

- Baseline Standard Deviation
- Anchor-Based approach to determining MID or MCID
- 38 studies filled the criteria, resulting in 62 computed effect sizes

Results

- The MID estimates were remarkably close to one half a standard deviation (Mean = 0.495; SD = 0.155)
- There was no clear relationship between the magnitude of the estimate (~.50) and factors such as disease-specific or generic instrument or the nature of the response scale
- Negative changes were not associated with larger effect sizes

WHY?

A possible explanation for the consistency in these results derives from a classic paper 1956 in *Psychology Review*

"The Magic Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information"

by George Miller

- Across a wide range of unidimensional discrimination tasks (saltiness of tastes, points on a line, pitch and loudness of sounds, etc.), the limit of people's abilities to make absolute discriminations turned out to be very consistent
- People were capable of identifying the category of a particular stimulus (loudness of sounds, saltiness of tastes) accurately until the number of categories reached about 7 (with a range from about 5 to 9)

Miller argued that this uniformity derives from a fundamental characteristic of human information processing that he called ‘channel capacity’, related indirectly to limits on short-term memory

- First convert “1 part in 7" to standard deviation units
- In the original (Miller) tasks, the stimuli were sampled from a rectangular distribution with a finite range
- It can be shown that for a uniform rectangular distribution 7 units wide, the standard deviation equals 2.16, so 1 part in 7, expressed in SD units, is 1 / 2.16 or **0.46**

- Similarly, accounting for Miller's "+/- 2", for a rectangular distribution of 5 levels the SD is 1.58 and 1 in 5 is an effect size of 0.63 ; for a distribution 9 units wide, the SD is 2.73 and effect size is 0.36

- Thus, based on Miller's review, the limit of human discrimination is equivalent to an effect size between 0.36 and 0.63

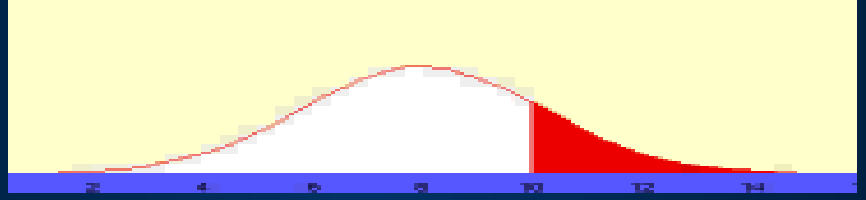
- The effect sizes observed in in 38 studies have a range (+/- 1 sd) from 0.34 to 0.64

- The range of estimates for the minimally important difference from the MID studies, expressed in SD units, corresponds almost exactly to the limit of human discrimination identified by Miller over 40 years ago

Since nearly all of the MID measures we examined are based, one way or another, on the notion of a threshold between essentially undetectable and minimally detectable patient change, it may not be a coincidence that these disparate methods, conducted on diverse clinical populations with a wide range of instruments and different criteria, almost always arrive at a similar value

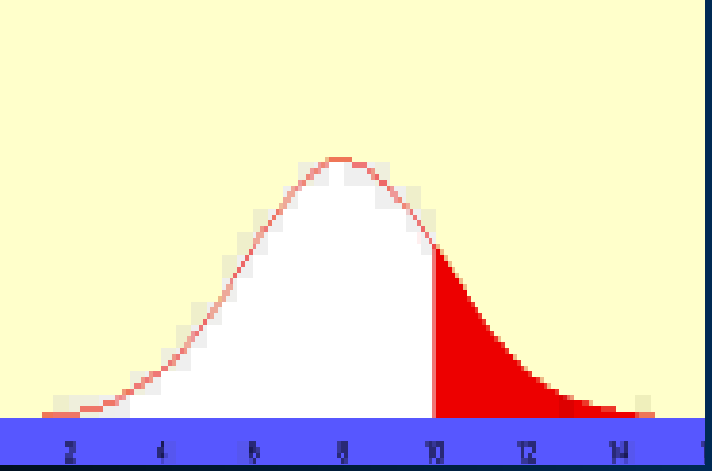
Some Important Exceptions

- Stratford Studies
- Schwartz-2 days after chemotherapy



Distribution-Based Methods

- Effect Size
- Standard Error of Measurement

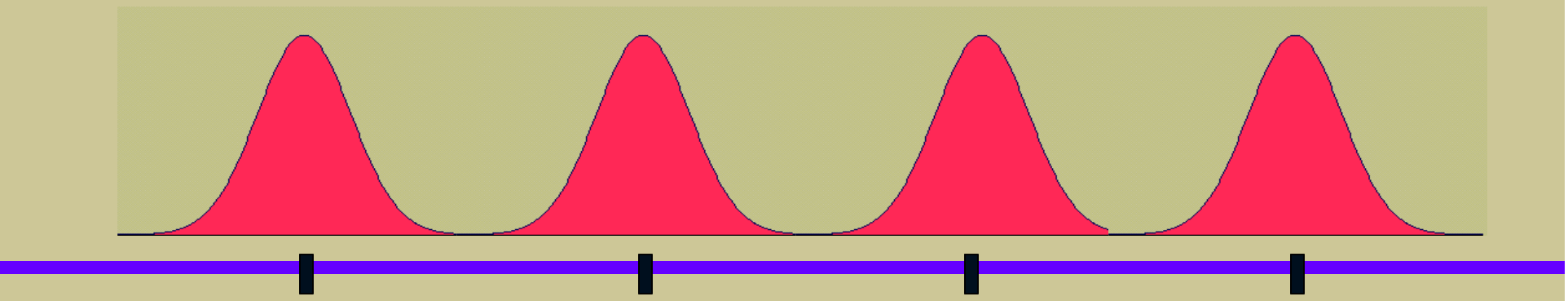


The Standard Error of Measurement (SEM)

$$\text{SEM} = s_x \sqrt{1 - r_{xx}}$$

- Fixed characteristic of a measure that is not sample-dependent
- Expressed in the **original metric** of the measure

What is a SEM?

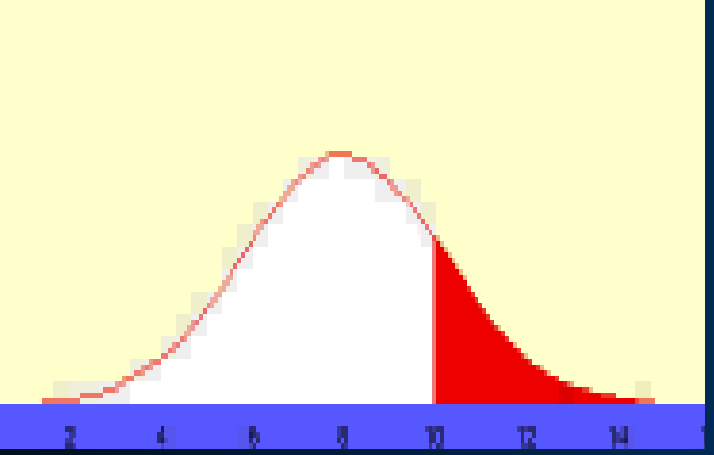


Mary's True Score

Jim's True Score

Gary's True Score

Kim's True Score



How Many SEMs = Important Individual Change?

1 SEM

1.96 SEM

2.77 SEM

Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life

Wyrwich K, Nienaber N, Tierney W, and Wolinsky F

Medical Care, 1999

Further evidence supporting a SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life

Wyrwich K, Tierney W,
and Wolinsky F.

Journal of Clinical Epidemiology, 1999

Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire.

Wyrwich K, Tierney W,
and Wolinsky F.

Quality of Life Research 2002

What is a clinically meaningful change on the Functional Assessment of Cancer Therapy - Lung (FACT-L) questionnaire? Results from the Eastern Cooperative Oncology Group Study

Cella D, Eton DT, Fairclough DL, Bonomi P, Heyes AE, Silbermans C, Wolf MK, Johnson DH

Journal of Clinical Epidemiology, 2002

Is This Really All Connected?

Relationship Between One-SEM Criterion & Cohen's effect size standards

- reflects a minimal change (.2-.5)
- rewards highly reliable scales

Effect Size For A One-SEM Change

If r_{xx} = .95 → $\square_{individual}$ = .22

If r_{xx} = .90 → $\square_{individual}$ = .32

If r_{xx} = .85 → $\square_{individual}$ = .39

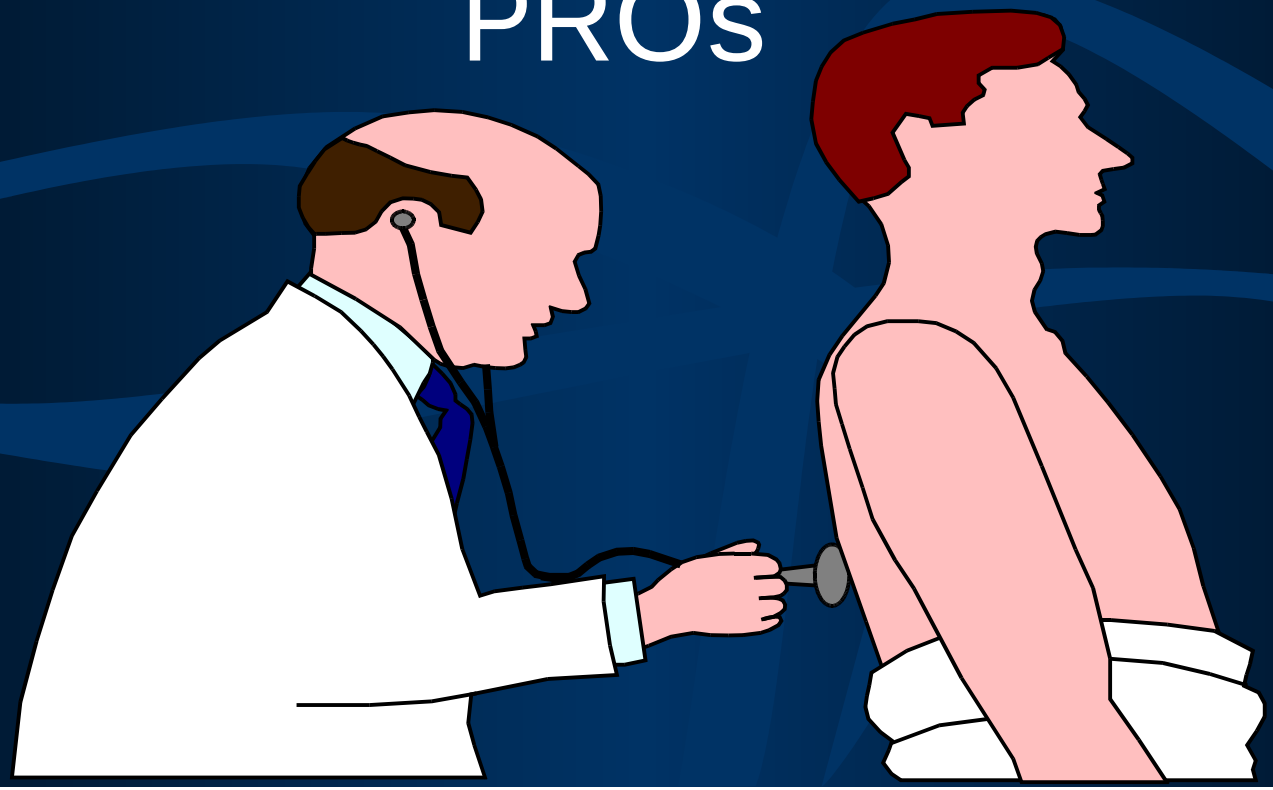
If r_{xx} = .80 → $\square_{individual}$ = .45

If r_{xx} = .75 → $\square_{individual}$ = .50

Practical Suggestions for The Development Of Clinically Relevant Individual Change Standards


WHERE'S THE
CLINICAL?

Incorporating **Clinically** Into Significant Individual Difference Standards for PROs



Clinically Significant Individual Change Standards

Needed to move PROs outcomes

From

Clinical Trial Research

To

- Routine Clinical Practice
- Clinical Decision-Making

- The value added to the clinician of measuring PROs in clinical practice or research
- How clinicians compare the accuracy and precision of PRO data relative to other clinical measures
- Methods for clinicians to interpret PRO data

Difficult Issues

- Can only *patients* can *report* PROs?
- Advantages of clinician reports
  - Retrospective Overview—beyond the moment
- Dangers of clinician reports
  - Traditional under-reporters of pain and other aspects of patient QOL

Clinically Important Changes in Health-related Quality of Life for Patients with Chronic Obstructive Pulmonary Disease

An Expert Consensus Panel Report

Kathleen W. Wyrwich, PhD, Stephan D. Fihn, MD, William M. Tierney, MD, Kurt Kroenke, MD, Ajit N. Babu, MBBS, MPH, Fredric D. Wolinsky, PhD

OBJECTIVE: **Without clinical input on what constitutes a significant change, health-related quality of life (HRQoL) measures are less likely to be adopted by clinicians for use in daily practice. Although standards can be determined empirically by within-person change studies based on patient self-reports, these anchor-based methods incorporate only the patients' perspectives of important HRQoL change, and do not reflect an informed clinical evaluation. The objective of this study was to establish clinically important difference standards from the physician's perspective for use of 2 HRQoL measures among patients with chronic obstructive pulmonary disease (COPD).**

DESIGN: **We assembled a 9-person expert panel of North American physicians familiar with the use of the Chronic**

KEY WORDS: **quality of life; COPD; important change; consensus panel; RAND method; Delphi process.**
J GEN INTERN MED 2003;18:196–202.

Chronic obstructive pulmonary disease (COPD) is currently the fourth leading cause of death in the world, and a major cause of chronic morbidity.[1] Unfortunately, clinicians cannot currently offer treatments to most COPD patients that will favorably change the course of this highly prevalent condition. Therefore, the goal of clinical management is to improve patients' health-related quality of life (HRQoL) by relieving symptoms and enhancing

Other Practical Developments Suggesting That:

All Points on Pain Scales Are Not Equal

The numeric rating scale and labor epidural analgesia

Beilin Y, Hossain S, Bodian C

Anesthesia & Analgesia, 2003

Labor Pain Study

- A verbal numeric 0-10 rating scale
- In three studies, a verbal NRS score was obtained:
  - before
  - 15 min. after labor epidural analgesia
- At 15 min, the woman was also asked if she wanted more pain medication

Labor Pain Study

Results showed that when:

NRS = 0-1	2% wanted more meds
NRS = 2-3	51% wanted more meds
NRS > 3	93% wanted more meds

Labor Pain Study– Implications for Clinical Differences

Would a change from 6 (before) to 4 (15 min. after) be meaningful among these women?

Would a change from 3 (before) to 1 (15 min. after) be meaningful among these women?

Revisiting IRT and How These Methods Inform Clinical Significance

Not all points on a pain scale are equal!

An Item Response Theory Based Pain Item Bank Can Enhance Measurement Precision

Lai J-S, Dineen K, Cella D, Roenn J.

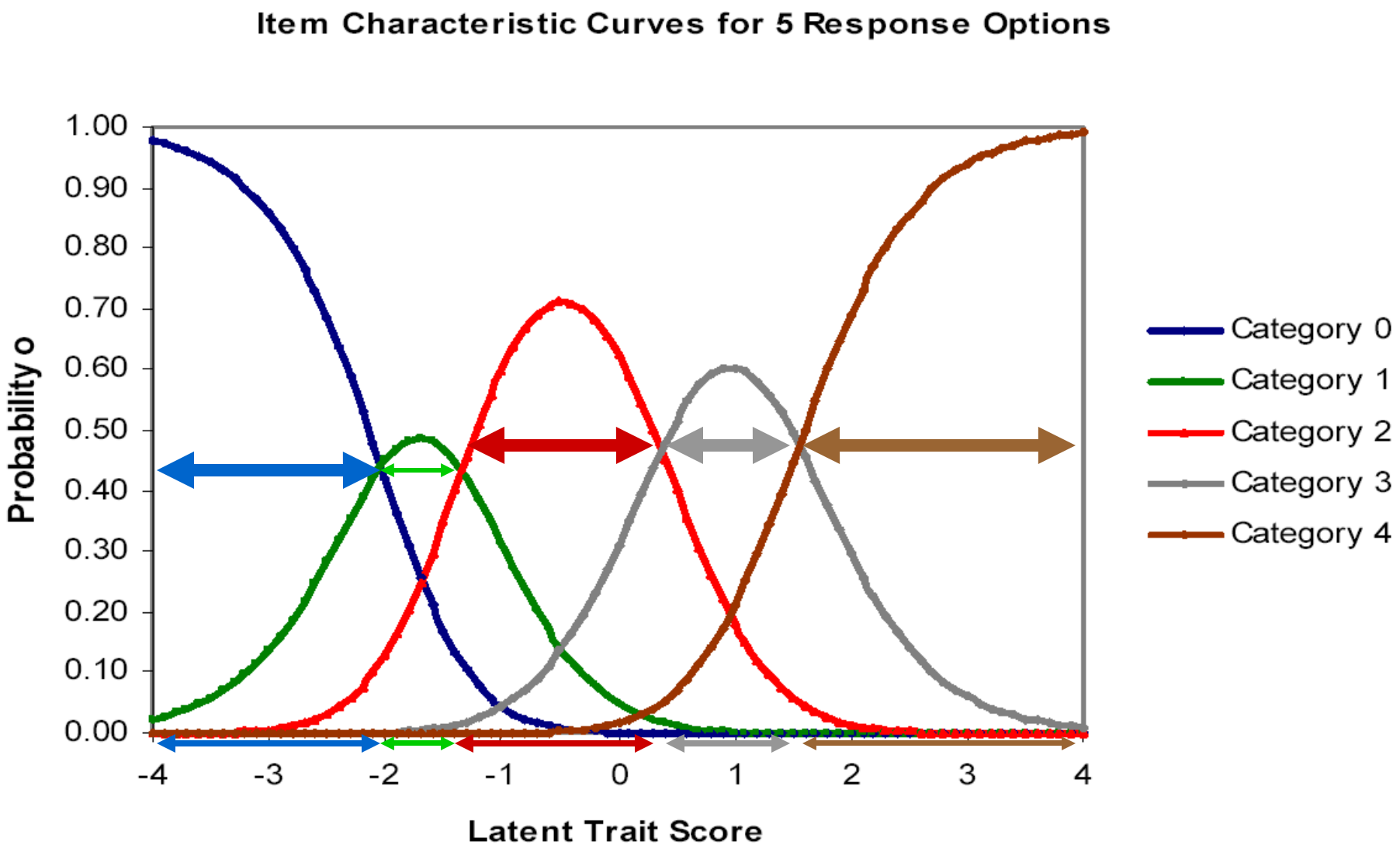
Clinical Therapeutics, 2003

Applying Item Response Theory (IRT) Models to Evaluate the Scaling of VAS Pain Measure

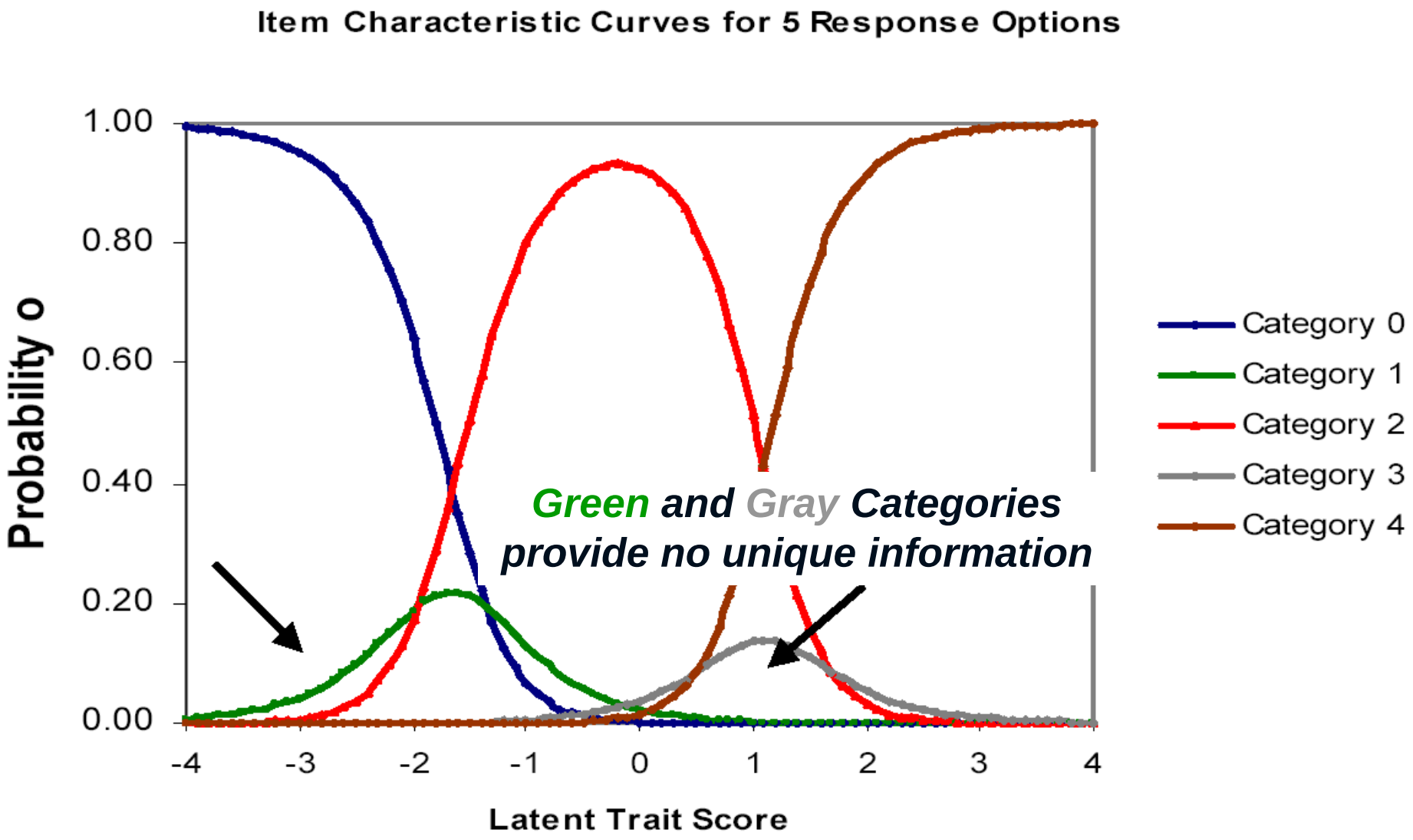
Kosinski, M

Association for Health Services Research Workshop, 2002

Example of Perfectly Functioning Item Characteristic Curves (ICCs)



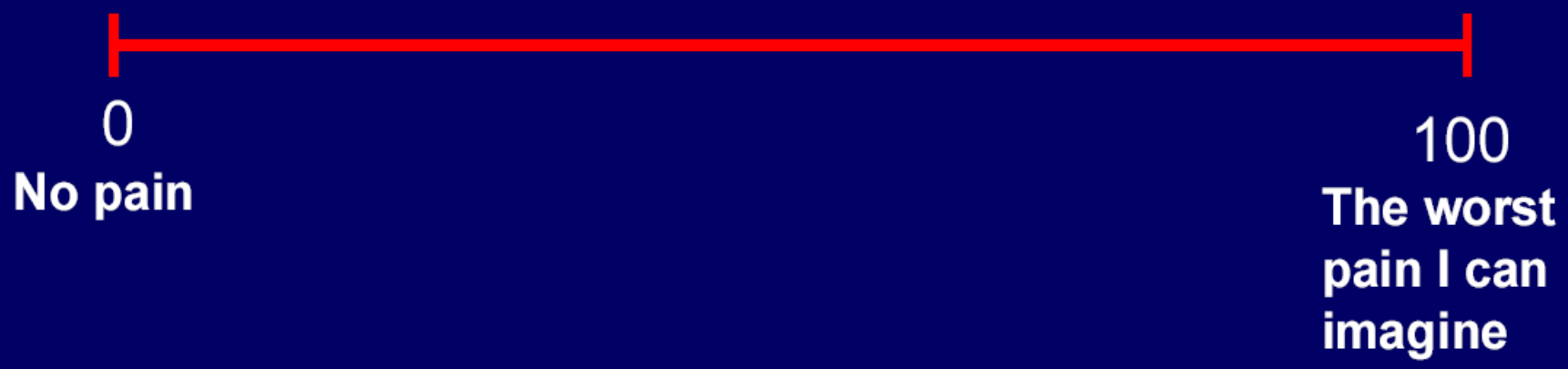
Example of Poorly Functioning Item Characteristic Curves (ICCs)



Continuous Pain Measure

- **Visual Analogue Scale: 100-mm Scale**

How would you rate your pain?



Categorical Pain Measures

What is the worst pain you experienced over the past week? (PAST PAIN)

0	1	2	3	4
none	mild	moderate	severe	extreme

What is your pain now? (CURRENT PAIN)

0	1	2	3	4
none	mild	moderate	severe	extreme

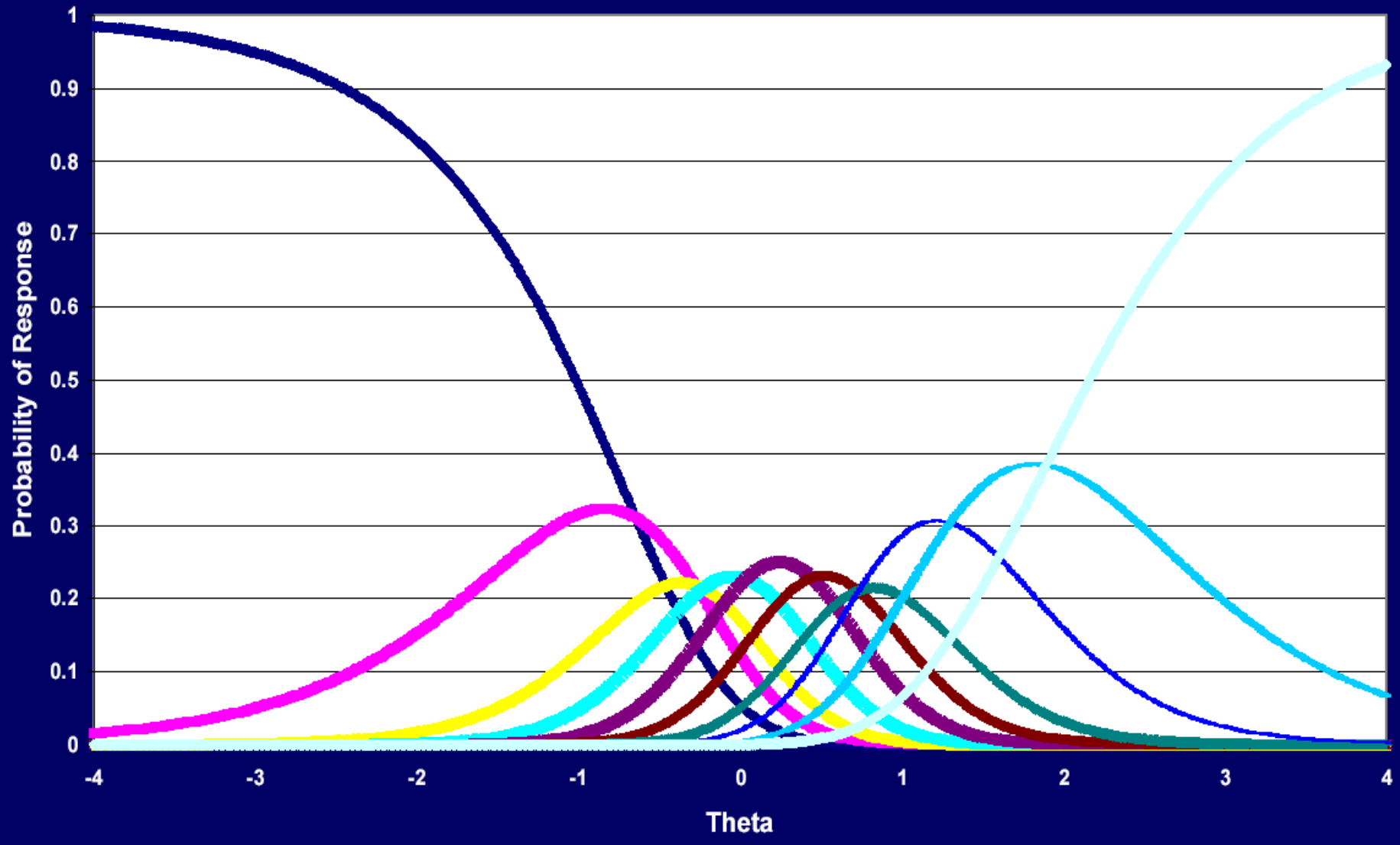
How much bodily pain have you had during the past 4 weeks? (SF-36 BP1)

1	2	3	4	5	6
none	very mild	mild	moderate	severe	very severe

During the past 4 weeks, how much of the time did pain interfere with your normal work (including both work outside the home and housework)? (SF-36 BP2)

1	2	3	4	5
not at all	a little bit	moderately	quite a bit	extremely

ICC's Drawn for 10 VAS Score Categories



Conclusions

- Analysis of ICC showed VAS discriminates well at extremes
- Analysis of ICC showed VAS discriminates poorly in the middle
  - categories did not show unequivocal and unique relation to latent pain score
  - scale did not distinguish between patients differing in the level of the latent pain variable

Outcomes research: measuring the end results of health care

Clancy C. & Eisenberg J.

Science 1998

Clancy & Eisenberg

"additional work to

enhance the interpretability

of outcome measures, particularly

in terms of clinical significance

is needed to increase the

usefulness of these tools."