Duke | Robert J. Margolis, MD
Center *for* Health Policy

Developing Personalized Clinical Outcome Assessments

4/5/17

Developing Personalized Clinical Outcome Assessments

Duke Margolis Expert Workshop
April 5, 2017

Laura Lee Johnson, Ph.D.
Office of Biostatistics
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

"The primary endpoint for determining that a drug is effective should encompass one or more of the important features of a disorder and should be clinically meaningful."

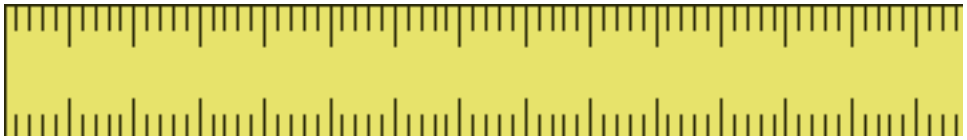
– Lines 510-511, FDA Multiple Endpoints in Clinical Trials Draft Guidance, January 2017

Personalized COA

- What does this mean for today's discussion
- Examples
- Questions to answer/discuss

EVERYTHING OLD IS NEW AGAIN

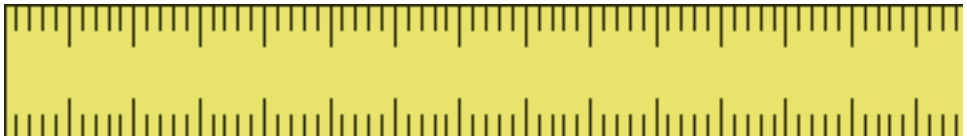
One Endpoint, One Concept



Borrowing graphics from Karon Cook and MS Clip Art

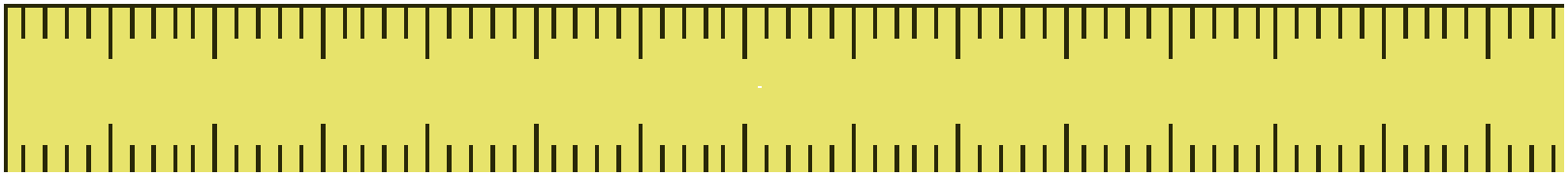
Several ‘Related’

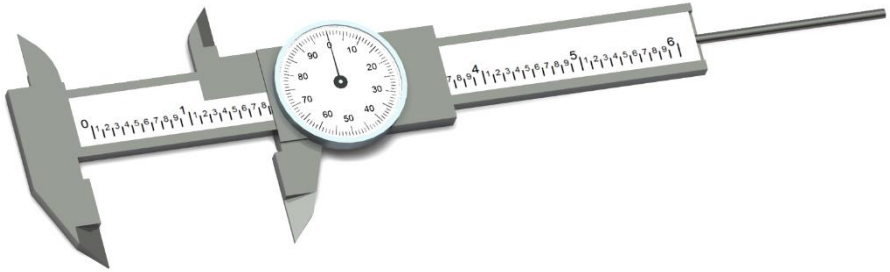
Example: Sleep



Different Rulers, Related Concepts

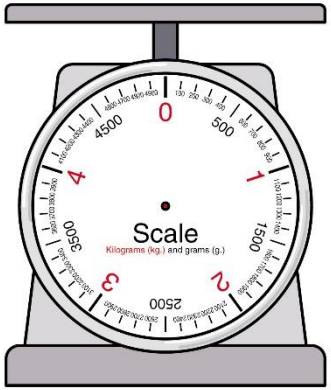
Example: Abdominal pain, Bloating

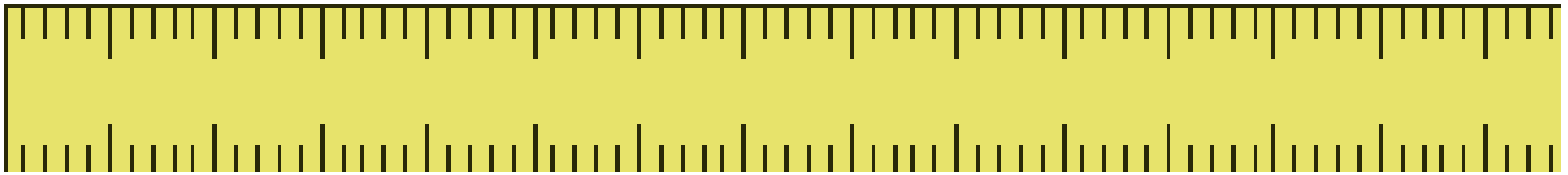




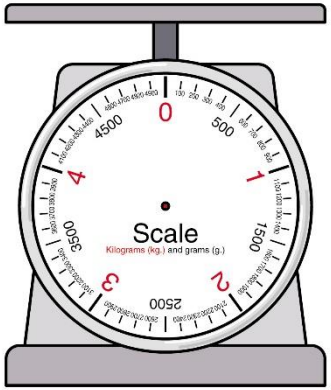
One Common Symptom, but Otherwise Heterogeneity

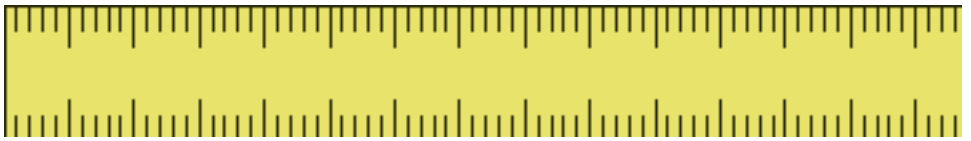
Example: Migraine (Pain, plus one of 3)





Many Measures, Many Concepts, One Endpoint?





10 Symptoms
Only one had 40% Acknowledging It

- Abdominal pain
- Cramping
- Diarrhea
- Rectal bleeding
- Fatigue
- Physical function
- …..

SEIQoL

- The Schedule for the Evaluation of Individual Quality of Life (SEIQoL)
- Judgement analysis
- Not valid, reliable (in some ways) Moons et al 2004
- Very time consuming and logistically difficult
  - As are some other methods

Do we ask about all 10?

- What if only 1 is relevant for Subject A and 3 for Subject B?
  - Need all to resolve?
  - Variance for each?
- What is “numeric worsening”?
  - Is 0.1 unit change on average on a 0-10 score really worse?

Personalized Responder Definitions
'Why not do a responder analysis?'
Just ask a Global?

- How many responder analyses?
- How do I define the responder?
- Power (Type II error)
- What is on the labeling
- Representative sample?

What is Personalized?

- Different items asked of each patient but about the same concept?
- Different outcomes for each patient?

What ails ya?

MOST BOTHERSOME SYMPTOM

GOAL ATTAINMENT SCALING

COMPUTERIZED ADAPTIVE TESTING (WITHIN A DOMAIN)

What is on the Labeling?

- What is the endpoint?
- What is the statistical analysis?
- How is it interpreted?
- How is it discussed in promotional materials?

Multiple Endpoints Draft Guidance*

- January 2017
  - https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM536750.pdf
  - Multiplicity
- When Demonstration of Treatment Effects on **All** of Two or More Distinct Endpoints Is **Necessary** to Establish Clinical Benefit (Co-Primary Endpoints) **Not typically personalized**
- When Demonstration of a Treatment Effect **on at Least One of** Several Primary Endpoints Is **Sufficient** **Not typically personalized**
- Composite Endpoints **Maybe but still need to know how to combine**
- Other Multi-Component Endpoints
- Clinically Critical Endpoints Too Infrequent for Use as a Primary Endpoint

*highlights are my own

Care Models and Regulatory Decision Making

Issues: How to

- Sufficiently evaluate for use in trials to support approval and labeling
- Determine when they are most appropriate
- Adequately reconcile personalization of COAs with need for standardization in clinical trial setting

Statistical Challenges: How to

- Analyze the data derived from individualized measures
  - Major issue: can the data can be pooled and if so, how
- Goal Attainment Scaling and most bothersome symptoms approaches
  - Challenging to address intra-patient variability in symptom progression or severity over time
  - Most “relevant” or “bothersome” symptoms for a patient may change over time, making it difficult to accurately measure and evaluate symptoms that matter most to patients
- Computer Adaptive Testing
  - Not always evident how to ensure that endpoints derived from CATs are equivalent across and within patients in a trial so that scores and interpretation are compatible across patients

Great Potential, but

- Potential: personalized COAs measure outcomes most important and meaningful to patients
- Challenges underscore need for more clarification and consensus in the field(s)

No Clear Answer

- Development, implementation, evaluation of personalized COA approaches
- Best practices?
- Small and heterogeneous study populations
- What about non-PROs? Observer-reported outcomes, clinician-reported outcomes, and performance outcomes.

End of the Day
"for use in drug development"

- *When* a personalized COA approach is appropriate and feasible for use in drug development (e.g., special populations such as pediatric, rare disease, etc.)
- *Which* personalized COA approaches are more appropriate in a given context
- Advantages and disadvantages not discussed
- Key analytical and methodological implications of personalized COA approaches
- How apply personalized COA approaches to other types of COA for use in drug development
- What other (emerging) personalized COA approaches warrant further exploration for use in drug development

Questions to Address (1)

- When is an approach useful in clinical trial settings (e.g., special populations such as pediatric, rare disease, etc.)?
  - How do we operationalize this approach?
  - How feasible is this approach?
  - How do we establish baseline scores?
- How do we determine meaningful change under the approach?
- How do we ensure scores and interpretation are compatible across patients? Between study arms?

Questions to Address (2)

- How do we 'best' analyze data?
- What are some best practices when analyzing such data?
- How do we handle heterogeneity within a patient over time?
  - When symptoms naturally relapse and remit so the "goal" or "most bothersome symptom" at baseline has changed?
  - How do we ensure other symptoms do not worsen?
- What are the advantages and disadvantages of an approach for use in drug development?


FDA
U.S. FOOD & DRUG
ADMINISTRATION

Duke | Robert J. Margolis, MD
Center *for* Health Policy

Developing Personalized Clinical Outcome Assessments

4/5/17

Evidera

Evidence. Value. Insight.

Personalized Clinical Outcome Assessments: Most Bothersome Symptom Approach

Prepared by:
Dennis Revicki, PhD
Outcomes Research, Evidera
Bethesda, MD, USA

Prepared for presentation at the Developing Personalized Clinical Outcome Assessments Meeting
Duke Margolis Center for Health Policy, Washington, DC
April 5, 2017

Introduction

Personalized medicine question:

- **What treatment, by whom, is most effective for this individual with that specific problem, and under what circumstances? *Paul Gordon (1967)***

Personalized outcome assessment question:

- **How do we best measure clinical outcome assessments to answer the personalized medicine question?**

Introduction:
Previous Approaches to Personalized Outcome Assessments

- **Asthma Quality of Life Questionnaire (Juniper et al. 1993)**
  - Physical function scale allowed patients to select their own physical activities
- **Schedule for Evaluation of Individual Quality of Life (O'Boyle et al. 1993)**
  - Aspects of QOL important to the individual are elicited (structured interview)
  - Current functioning/satisfaction with each aspect is rated by the individual.
  - Relative importance of each aspect of QOL is measured by deriving the weight the individual assigns to each in judging overall QOL.
- **Migraine Outcome Assessments**
  - Headache pain severity and most bothersome migraine-related symptom (FDA 2014)
- **Various Gastrointestinal Diseases**
  - Most predominant symptom

Most Bothersome Symptom Approach

Evidera

Most Bothersome (Troublesome) Symptom Approach

General Method

- **Identify relevant symptoms associated with target medical disorder**
  - Literature
  - Clinicians
  - Patients
- **Develop standardized rating questionnaire**
  - Severity/frequency scale (various rating scales)
  - Bothersome/importance/troublesome scale (various rating scales)
  - Develop mechanism for identifying and confirming individual's most bothersome symptom (or symptoms)
  - Develop scoring system (based on severity ratings) based on set of most bothersome symptoms
  - Measure all relevant symptoms as secondary endpoints

Most Bothersome Symptom Approach: Depression Example

Objective

- **Evaluate efficacy of adjuvant treatment for treatment-resistant depression**
- **Develop measure focusing on most troublesome symptoms of depression**
  - Focus on residual depressive symptoms

Patient-Rated Most Troubling Symptom Scale for Depression (PaRTS-D): Rationale

- **Between 30% and 40% of patients with MDD never achieve symptom resolution with standard antidepressant therapy**
- **Patient-Rated Troubling Symptoms for Depression (PaRTS-D) instrument was developed to provide a more individualized assessment of the relevant symptoms of depression from the patient's perspective**
- **8 symptoms related to MDD: sadness, feeling tense or uptight, reduced sleep, reduced appetite, trouble concentrating, reduced involvement in things that usually interest the subject, inability to feel emotion, and negative thoughts**
- **Patient rates the severity of each individual component using a 0 (resolved) to 10 (extreme) NRS scale**
- **PaRTS-D scores are determined for each patient:**
  - Sum of 8 symptom severity scores (total global score)
  - Sum of 4 highest ranked baseline symptoms (total score)

Patient-Rated Most Troubling Symptom Scale for Depression (PaRTS-D): Development

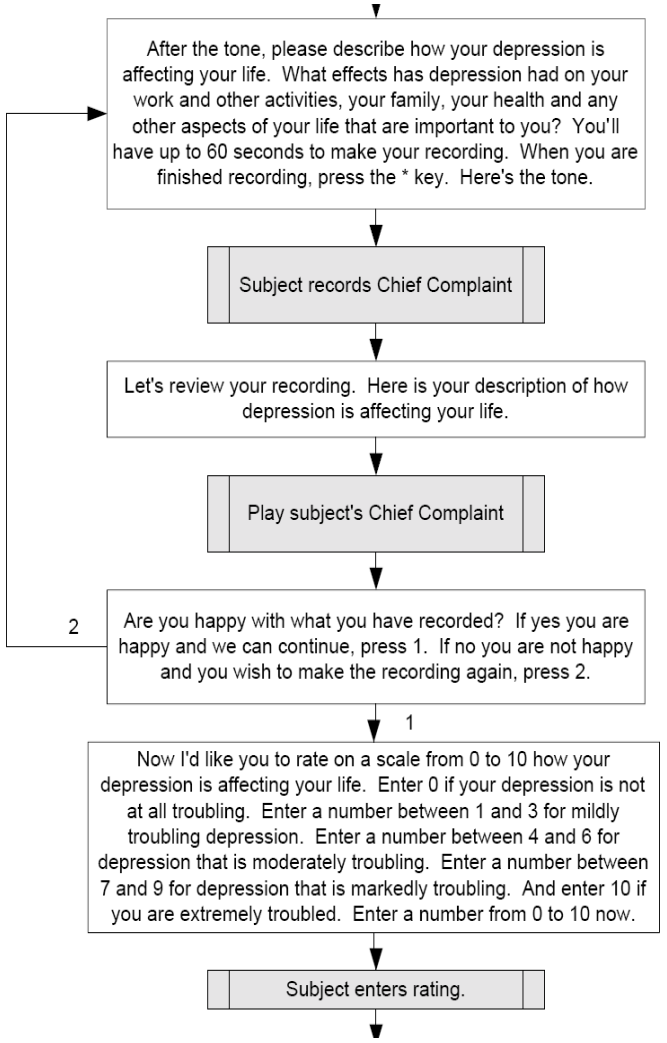
PaRTS-D content determined based on:

- Review of depression literature (residual symptoms)
- DSM-IV depression diagnosis symptom criteria
- Clinician review and recommendations
- Patient clinical trial data review (no direct patient involvement)

Patient-Rated Most Troubling Symptom Scale for Depression (PaRTS-D): Assessment

- **IVRS Call System: Abbreviated Script and Sample Question**

  Welcome to the Telephone Assessment Call. During this study, you will be asked to rate your symptoms, how those symptoms affect your work and social life, and any change you may have felt. For today's call, it may be helpful to review your 0-10 visual scale.



<<System continues to review each symptom and prompt patient's for their response.>>

PaRTS-D Psychometric Characteristics

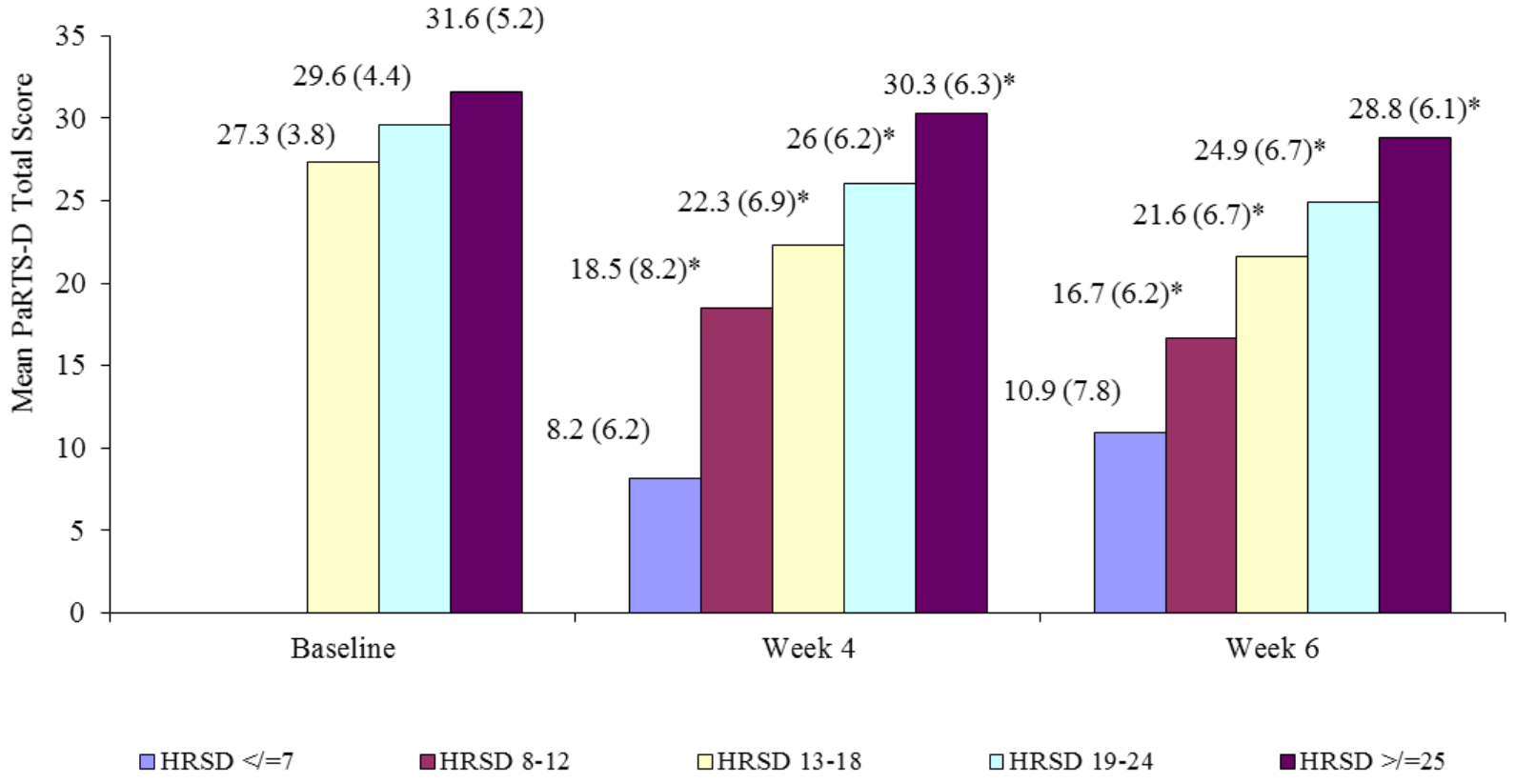
- **Secondary analysis of clinical trial data**
- **Unidimensionality**
  - Exploratory factor analysis supports two factors (mood, somatic)
  - IRT analyses support unidimensional scales for mood and somatic symptoms
- **Reliability**
  - Internal consistency good (>0.80) at weeks 4 and 6
  - Test-retest reliability (ICC=0.55)
- **Construct validity**
  - Moderate to strong correlations with clinician-rated and PRO measure
  - Good evidence supporting known groups validity

Relationship between PaRTS-D Scores and Other Measures

	PaRTS-D Total	PaRTS-D Global
Q-LES-Q		
Baseline	-0.47	-0.48
6 weeks	-0.68	-0.65
SDS		
Baseline	0.65	0.64
6 weeks	0.82	0.81
HRDS		
Baseline	0.33	0.36
6 weeks	0.67	0.65

Abbreviations: PaRTS-D = Patient-Rated Most Troubling Symptom Scale for Depression; Q-LES-Q = Quality of Life Enjoyment and Satisfaction Questionnaire; SDS = Sheehan Disability Scale; HRDS = Hamilton Rating Scale for Depression

PARTS-D Scores by HRDS Defined Groups

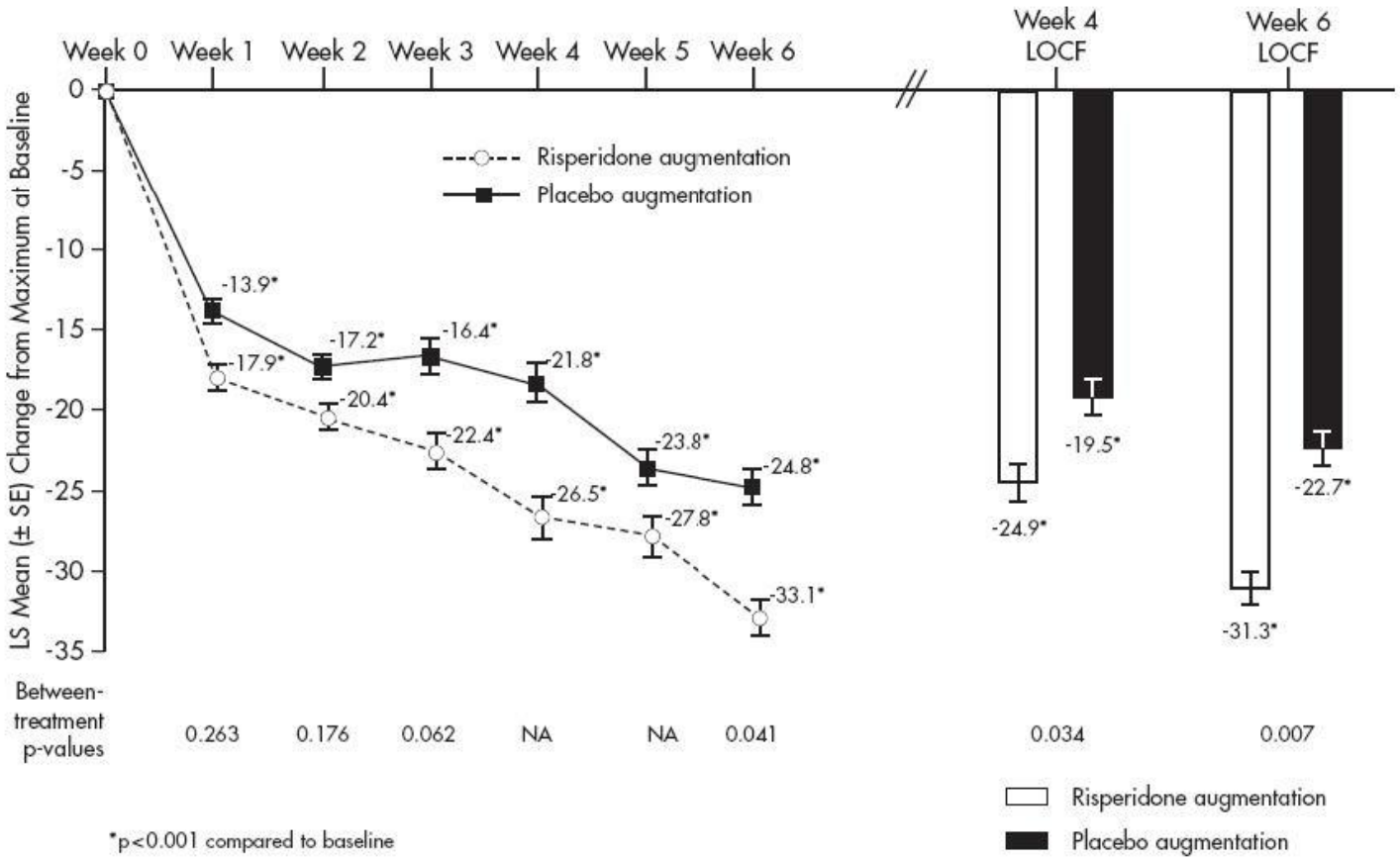


*p<0.005 from the LS mean multiple comparisons with Scheffe adjustment and the first level of HRSD as the reference group.:

Source: Padina et al. 2010

Baseline to Endpoint Changes in PARTS-D Scores by Treatment Group

Figure 2. Change from Maximum PaRTS-D Total Score (LS Mean ± Standard Error [SE])



Source: Padina et al. 2009

PaRTS-D: Conclusions

- **PaRTS-D score focused on the individual's most bothersome symptoms at baseline**
  - Limited involvement of patients in identifying relevant symptoms
- **PaRTS-D scores demonstrated good evidence supporting reliability, and concurrent and known groups validity**
- **PaRTS-D scores detected statistically significant differences between active treatment and placebo**
- **Clinical efficacy findings were comparable to those based on clinician ratings of depression symptoms**
- **PaRTS-D may be used as an adjunct to clinician-rated instruments to assess response to antidepressant treatment in future clinical trials**

Assessing Most Bothersome Symptoms for Clinical Trials

- **Depends on good understanding of symptom experience of patients with the target condition**
  - Evidence based on literature, clinicians and patients
  - Patient information is critical (.e., patient engagement)
  - No different than developing any other PRO measure
- **Challenges associated with changing severity ratings and constellation of bothersome symptoms**
  - Symptom experience may change over course of study
    - Effectiveness of treatment
    - Some symptoms may be more resistant to change than other symptoms

Assessing Most Bothersome Symptoms: Challenges for Regulatory Agencies

- By definition, different patients identify and outcomes are based on different sets of symptoms
- How to describe the COA endpoint for product labels when using personalized assessments?
- If evidence supporting unidimensionality of concept for relevant symptoms, endpoint not so difficult a challenge
- How to handle reviews where individual determined most bothersome symptoms improve, but other relevant symptoms remain stable or worsen?
- How to summarize treatment group experience when COA endpoint consists of different clusters of the pool of relevant symptoms?

Assessing Most Bothersome Symptoms: Advantages and Disadvantages

- **Advantages of most bothersome symptom approach**
  - Patient-centric outcome, reflects what is bothering the patient
  - Flexible method for evaluating diseases with variable presentation
    - Not all patients experience all symptoms
  - May fit necessary outcomes for selected diseases
    - Rare disorders (e.g., FOP)
    - Disorders with heterogeneous presentations (e.g., MS)
- **Disadvantages of most bothersome symptom approach**
  - Changing patterns of symptoms over time within and between study subjects
  - Challenges for summarizing endpoints across subjects
  - Challenges for statistical analyses of treatment differences
  - Challenges for psychometric analyses (i.e., reliability)
  - Determining clinically meaningful interpretation guidelines

Evidera

BOSTON | BUDAPEST | HARRISBURG | LONDON | MONTREAL | SAN FRANCISCO | SEATTLE | WASHINGTON, DC

www.evidera.com

Duke | Robert J. Margolis, MD
Center *for* Health Policy

Developing Personalized Clinical Outcome Assessments

4/5/17

Goal Attainment Scaling, an individualized instrument with potential for outcome measurement in rare diseases

Hanneke van der Lee

Charlotte Gaasterland

Susanne Urach

Academisch Medisch Centrum
Universiteit van Amsterdam

asterix

MEDICAL UNIVERSITY OF VIENNA

Patient centered outcomes in rare diseases

- Generic outcome measures usually not responsive
- Development and validation of disease-specific outcome measures in rare diseases problematic
- Heterogeneity among rare disease trial participants

- Looking for an individual outcome measure: Goal Attainment Scaling (GAS)

  Kiresuk and Sherman, Community Mental Health Journal 1968; 4 (6): 443.

asterix

GAS in practice (1)

Heterogeneous patients, different goals

‘I want to walk’

‘I want to eat independently’

‘I want to breathe independently’

How do we measure effect of intervention?

GAS in practice (2)

1. What are your goals?
2. Definition of 5 levels of attainment per goal
3. Which goals are most important to you (weights)?
4. *Intervention*
5. Independent assessment:
   At what level is each goal attained?

GAS in practice (3)

At baseline:

1. Selection of goals (1 or more)
2. Definition of attainment levels for each goal,

e.g.

-2	unable to walk
-1	can take 3 steps
0	can walk for 5 minutes
+1	can walk for 15 minutes
+2	can walk for a longer period

3. Goals may be weighted

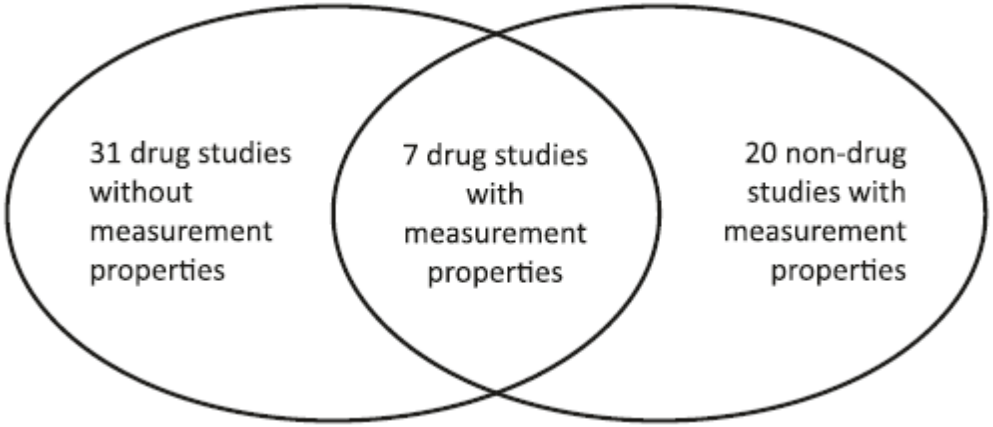
Post intervention:

1. Assessment of goal attainment levels
2. Kiresuk T-score: weighted sum across all goals

$$T = 50 + \frac{10 \Sigma w_i x_i}{\sqrt{(1-\rho)\Sigma w_i^2 + \rho(\Sigma w_i)^2}}$$

Systematic review

- Has GAS been used in drug trials?
- For what (drug) interventions has GAS been used?
- What is known about the measurement properties?



Mostly investigated:

Botox and Baclofen in patients with Cerebral Palsy

Donepezil and Galantamine in Alzheimer Disease patients

Conclusions SR

- Validation is mainly done in geriatrics/rehabilitation
- Usually in non-drug trials
- Insufficient information about validity

Gaasterland et al. BMC Medical Research Methodology (2016) 16:99.

When is GAS useful?

Useful:

- Chronic disease
- Effect of intervention expected on behavioral ability, that can be assessed independently
- Concurrent blinded controls

Not useful:

- Acute, episodic or unpredictable diseases
- Cross-over trials

A statistical approach for the efficient design of GAS studies

asterix

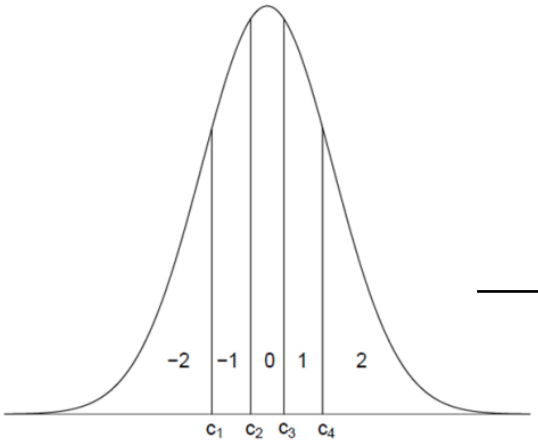
- How is a statistical analysis of GAS studies affected by
  - Maximum number of goals
  - Correlation between the goals
  - Proportion of goals affected by the treatment
  - Number of attainment levels

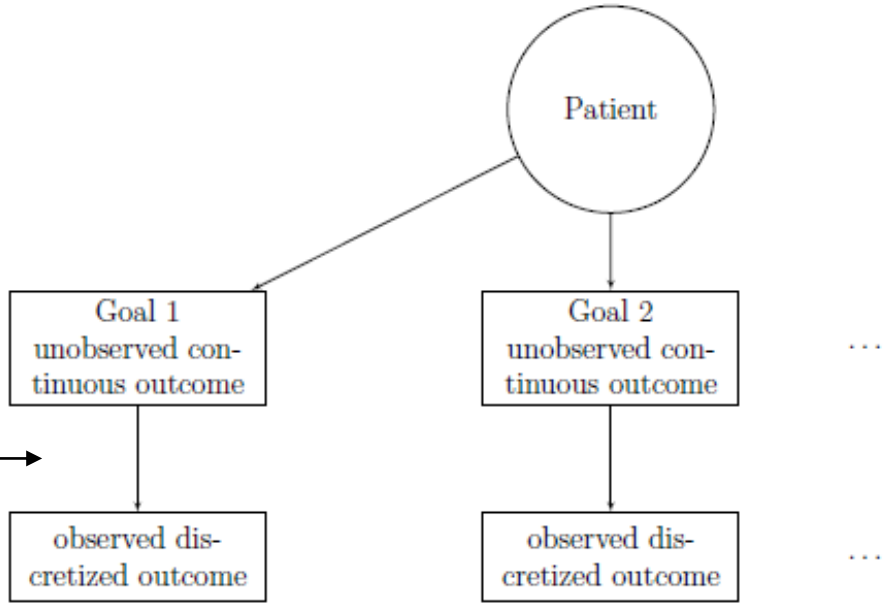
- How should the aggregated scores be analysed best?

- What kind of weights should be applied to the individual goals?

A model to simulate GAS data

- The treatment potentially affects several correlated goals of a patient.
- The observed ordinal attainment level for each goal is the result of a discretization of a continuous normal variable.
- The means of the continuous normal variables shift due to the treatment effect.





The use of GAS to demonstrate treatment effects

- Some aggregation is needed because the number of goals per patient varies and the goals are not directly related to one another.

- To demonstrate treatment effects, mean Kiresuk T scores between treatment groups can be compared. The interpretation of mean T scores in single arm trials is challenging.

- Treatment effects can only be estimated on the scale of the Kiresuk T scores. For the clinical interpretation, the goals and weights chosen by the patients have to be taken into account.

- The use of parametric test procedures to compare mean Kiresuk T scores is justified because of the robustness to non-normality of tests of central tendency such as the t-test.

Designing trials with GAS outcome

- The power increases with the number of goals affected by the treatment per patient, but levels off.

- For weak correlation between goals, there can be substantial power increase up to about 5 goals.

- Including goals that are not affected by the treatment can lead to a substantial loss in power.

- A scale with 5 levels appears to be sufficient.

Analysis of GAS data

- Improvement in power is possible if a GEE approach is used instead of the suggested Kiresuk formula.

- Weighting of goals
  - If the weights are not correlated with the treatment effect on the goals, weighting may lead to a substantial loss in power.
  - We are investigating to which extent power can be gained by choosing weights that are correlated with the treatment effect on each goal.

Discussion

1. Validation of GAS faces specific challenges:
   is generic validation across diseases/interventions possible?

2. Randomization and blinding is of paramount importance to address potential sources of bias, as, e.g., the patient‘s and investigator‘s choice of goals.

3. For an efficient application of GAS endpoints in clinical trials, the statistical implications of design choices (as, e.g., the maximum number of goals) should be considered.

4. GAS is a promising instrument for heterogenous patient groups. We propose to develop it as an endpoint for applications in the regulatory context.

References

Kiresuk, T. J. (1973). Goal attainment scaling at a county mental health service. *Evaluation: The International Journal of Theory, Research and Practice*.

Agresti, A., & Kateri, M. (2011). *Categorical data analysis* (pp. 206-208). Springer Berlin Heidelberg.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 933-944.

www.asterix-fp7.eu

Duke | Robert J. Margolis, MD
Center *for* Health Policy

Developing Personalized Clinical Outcome Assessments

4/5/17

Computer Adaptive Testing and Personalized COAs: A Domain Based Approach

David Cella, PhD
Ralph Seal Paffenbarger Professor
Chair, Medical Social Sciences
Feinberg School of Medicine
Northwestern University

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED
PROMIS
NeuroQoL
ASCQ-Me℠
Adult Sickle Cell Quality of Life Measurement Information System
NIH TOOLBOX® Motor Emotion Sensation Cognition

Questions to address

- What are the analyses and evidence needed to demonstrate that the scores and interpretation of those scores are comparable across patients under CAT (i.e., the same concept is being measured at all time points and in all patients)?
- When is this approach useful in clinical trial settings?
- How do we determine meaningful change under this approach?

Essential Components of PROMIS

DOMAIN

The feeling, function, or perception you wish to measure

Cuts across different diseases and settings. E.g., physical function, depressive symptoms

ITEM BANK

Collection of items that each measure the same domain

Used to create different measure types, all producing a score on the same metric

PROMIS®

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

PROMIS Measure Types

SHORT FORMS

- Subsets of item banks
- Focused on a single domain
- Off-the-shelf or custom
- Usually 4-10 items

COMPUTER ADAPTIVE TESTS (CATs)

- Individually tailored electronic questionnaires
- Focused on a single domain
- Next item administered from item bank depends on previous answer

PROFILES

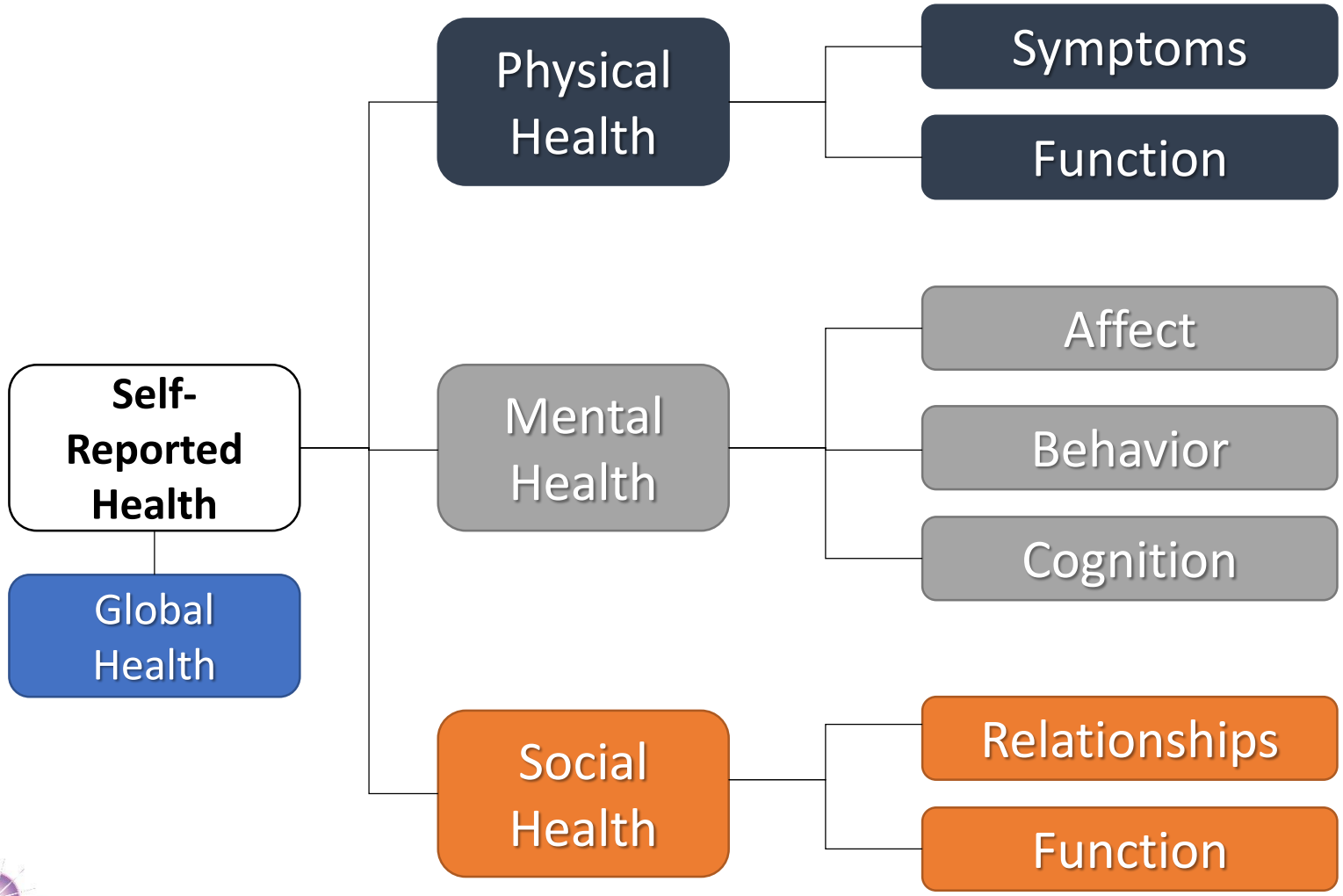
- Collection of 4, 6, and 8-item short forms
- Covers 7 physical, mental, and social health domains
- Also includes a single Pain Intensity item

PROMIS®

HealthMeasures

TRANSFORMING HOW HEALTH IS MEASURED

PROMIS Domain Framework



PROMIS

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Computerized Adaptive Testing (CAT)

PROMIS
NeuroQoL
ASCQ-Me
Adult Sickle Cell Quality of Life Measurement Information System
NIH TOOLBOX Motor Emotion Sensation Cognition
HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

PROMIS

Neuro QoL

ASCQ-Me℠
Adult Sickle Cell Quality of Life Measurement Information System

NIH TOOLBOX Motor Emotion Sensation Cognition

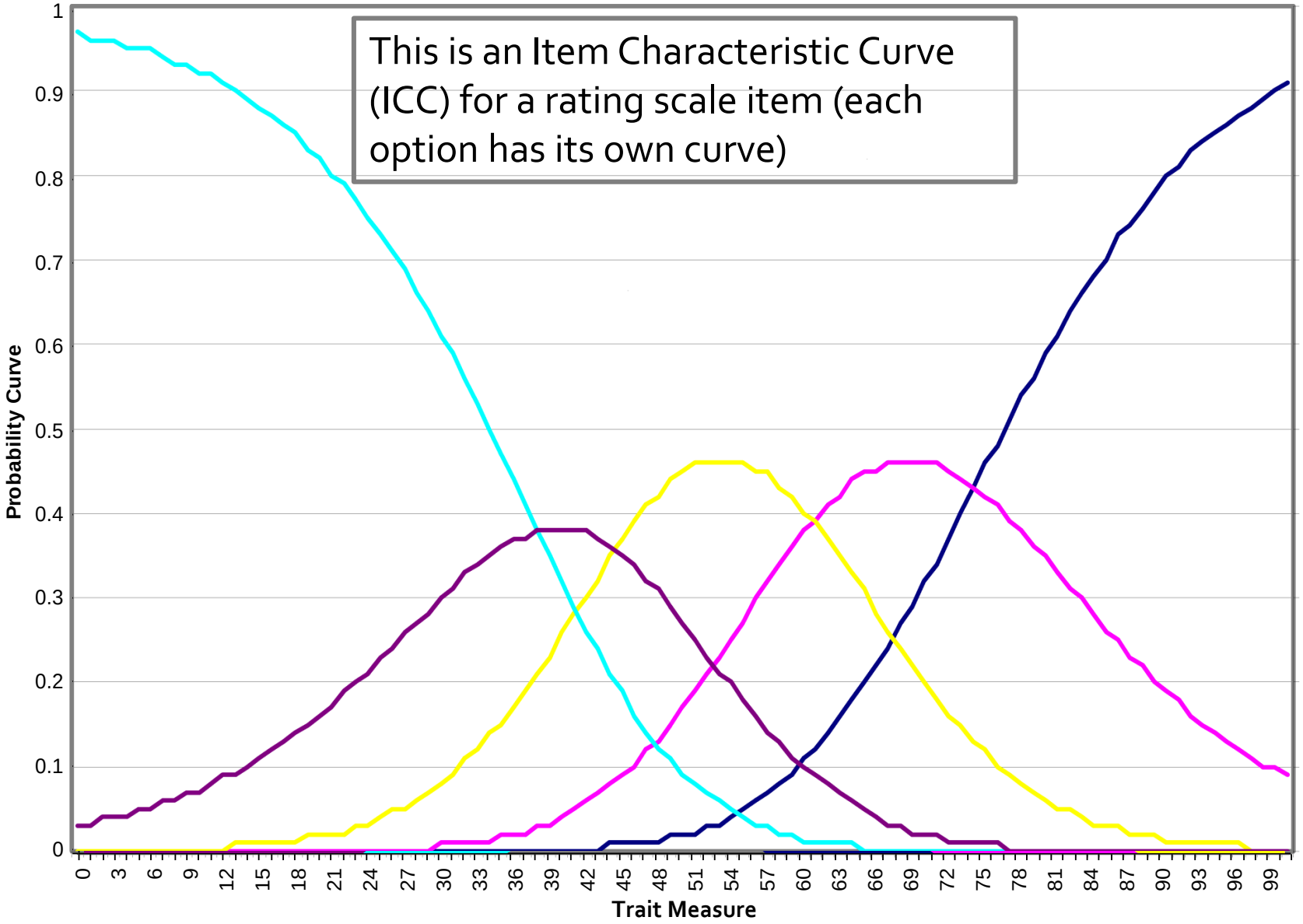
HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

I Have a Lack of Energy

Traditional Test Theory

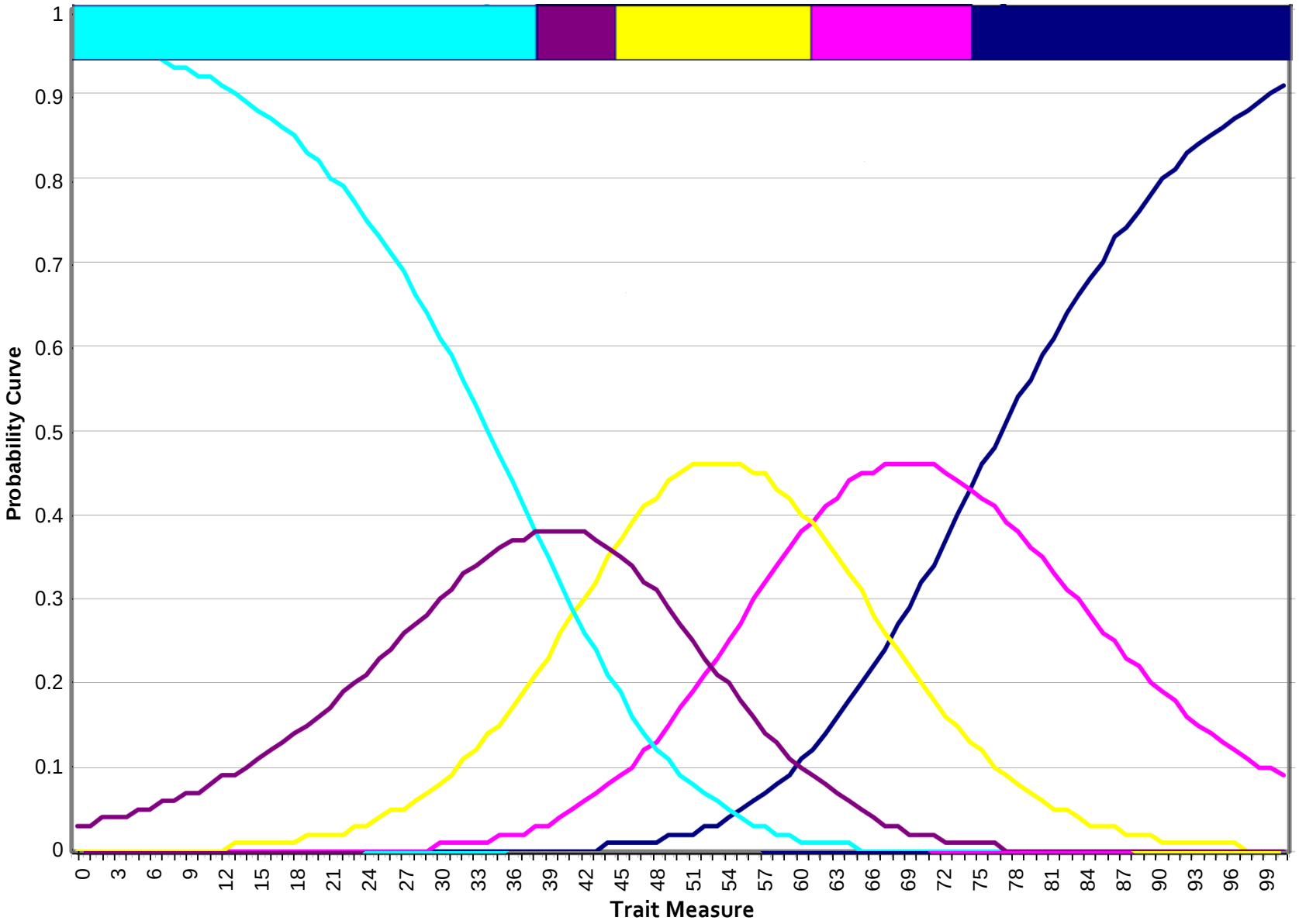
4 = Not at All 3 = A Little Bit 2 = Somewhat 1 = Quite a Bit 0 = Very Much

I have a lack of energy



4 = Not at All 3 = A Little Bit 2 = Somewhat 1 = Quite a Bit 0 = Very Much

I have a lack of energy



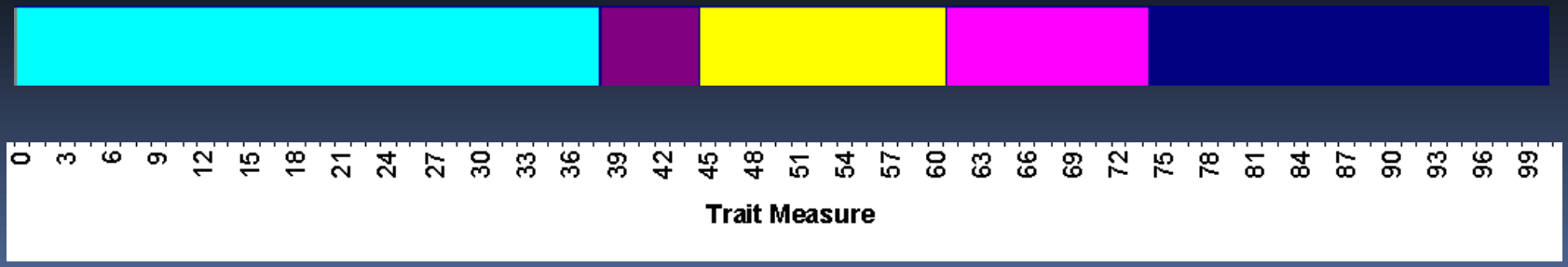
4 = Not at All 3 = A Little Bit 2 = Somewhat 1 = Quite a Bit 0 = Very Much

I Have a Lack of Energy

Traditional Test Theory

4 = Not at All 3 = A Little Bit 2 = Somewhat 1 = Quite a Bit 0 = Very Much

Item Response Theory



Comparing Fixed and Variable (CAT) Assessment

Fixed Short Form

- All patients get same questions
- One metric
- Often requires 6-10 questions per domain

CAT

- Patients get different questions
- One metric
- Usually requires 3-5 questions per domain

PROMIS Neuro-QoL ASCQ-Me™ Adult Sickle Cell Quality of Life Measurement Information System NIH TOOLBOX Motor Emotion Sensation Cognition HealthMeasures TRANSFORMING HOW HEALTH IS MEASURED

Comparison of PROMIS Profile Short Forms & CATs

PROMIS®

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Using PROMIS "Wave 1" Item Calibrations

(Cella et al, J Clin Epi 63 (11): 1179-1194; 2010)

- Compared CAT to 4, 6, and 8-item short forms
- Focus: The 7 PROMIS Profile domains
  - Anxiety, Depression, Physical Function, Fatigue, Sleep Disturbance, Social Function, Pain Interference
- Simulated 10,000 participant responses across the 7 PROMIS Profile domains
  - Normal distribution
  - Mean of theta=1
  - SD=1

Work by Segawa and Schalet

CAT Achieves High Accuracy with Fewer Items than Short Forms

	2	3	4	5	6	7	8	9	10	11	12	Weighted Average
Physical Function	53.4	23.6	12.7	4.7	3.1	1.3	0.5	0.4	0.2	0.1	0.1	**2.9**
Anxiety	0.0	46.8	39.4	8.7	2.9	0.8	0.4	0.4	0.3	0.2	0.1	**3.8**
Depression	17.6	59.0	13.2	3.6	1.8	1.5	1.0	0.4	1.3	0.3	0.1	**3.3**
Fatigue	20.5	64.8	8.5	2.1	0.2	1.4	1.1	1.1	0.2	0.0	0.0	**3.1**
Sleep Disturbance	0.0	4.1	44.1	28.6	10.2	4.6	2.5	2.1	1.4	1.6	0.9	**5.1**
Social Function	41.3	45.9	5.2	2.3	1.6	1.9	0.2	0.2	0.2	1.0	0.3	**3.0**
Pain Interference	76.6	9.9	1.1	6.2	1.5	1.3	0.6	0.6	0.6	0.5	1.0	**2.7**

Percentages of #s of CAT items answered; weighted average of #s of CAT items answered in accurate range

PROMIS®

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Results

- Relative to short forms, CAT delivered a wider range of accurate scores with fewer items per domain
- CAT superiority most evident in individuals with very high or low scores

PROMIS®

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Conclusion

- CAT reduces burden and time
- CAT and Fixed Form measure the same “thing”
- CAT removes option of comparing groups at the level of item content

Question 1

Analyses and evidence needed to demonstrate that the scores and interpretation of those scores are comparable across patients under CAT (i.e., the same concept is being measured at all time points and in all patients)

- Do we know this about static measures?
- Reliability helps but doesn't guarantee against response shift/adaptation.

Psychometric Standards for PROMIS Banks

- PROMIS instruments were developed to permit high comparability across forms
- Analysis standards reflect this:
  - Essential unidimensionality
  - Few local dependencies
  - Minimal DIF with most “trivial”
- Removed or modified items from banks which showed violations of the above
- Methods are detailed in Reeve et al. (2007) and Hansen et al. (2014).

PROMIS®

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Question 2

When is this approach useful in clinical trial settings?

- Multiple domains to assess
- High degree of confidence in stability of the domain(s)
- (Consider custom, fit-for-purpose SFs and branched assessment starting with fixed first item)

PROMIS®

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Question 3

How do we determine meaningful change under this approach?

All the usual approaches, with enhanced ability to develop model-based clinical vignettes

plus something to think about……

PROMIS

HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PHQ-9 to PROMIS

PHQ-9 Score	PROMIS T-score	SE
0	37.4	6.4
1	42.7	5.3
2	45.9	4.8
3	48.3	4.7
4	50.5	4.3
5	52.5	4.0
6	54.2	3.8
7	55.8	3.7
8	57.2	3.6
9	58.6	3.5
10	59.9	3.4
11	61.1	3.3
12	62.3	3.3
13	63.5	3.2
14	64.7	3.2
15	65.8	3.2
16	66.9	3.2
17	68.0	3.1
18	69.2	3.2
19	70.3	3.2
20	71.5	3.2
21	72.7	3.3
22	74.0	3.4
23	75.3	3.5
24	76.7	3.6

17

7

Choi et al, Psych Assessment 26 (2): 513-527, 2014

How Do People Select Domains and Measures?

- Relevance of the questions
- Reliability and Validity ("fitness for purpose")
  - Prior use and performance
- Patient burden
- Cost
- Likelihood of impact on valued audience
  - Academics to academics
  - Industry to regulators and payers
  - Everyone to consumers (patients)
- Various idiosyncratic heuristics
  - Brand affinity/loyalty
  - Recency
  - Familiarity/popularity
  - Superstition

Duke | Robert J. Margolis, MD
Center *for* Health Policy

Developing Personalized Clinical Outcome Assessments

4/5/17