

Expanding Our Understanding of Meaningful Change from a Patient Perspective

SEVENTH ANNUAL PATIENT-REPORTED OUTCOME (PRO) CONSORTIUM WORKSHOP

April 27 - 28, 2016 ■ Silver Spring, MD



Disclaimer



The views and opinions expressed in the following slides are those of the individual presenters and should not be attributed to their respective organizations/companies, the U.S. Food and Drug Administration, the Critical Path Institute, the PRO Consortium, or the ePRO Consortium.

These slides are the intellectual property of the individual presenters and are protected under the copyright laws of the United States of America and other countries. Used by permission. All rights reserved. All trademarks are the property of their respective owners.

Session Participants



Moderator

- *Cheryl D. Coon, PhD* – Principal, Outcometrix

Presenters

- *Mona Martin, RN, MPA* – Executive Director, Health Research Associates
- *Allison Martin Nguyen, MS* – Sr. Principal Scientist, Patient Reported Outcomes & Study Endpoints Group, Merck & Co., Inc.
- *Katarina Halling, MSc* – Global Head, Patient Reported Outcomes, AstraZeneca and Co-Director, PRO Consortium
- *Karon Frances Cook, PhD* – Research Professor in Medical Social Sciences, Northwestern University Feinberg School of Medicine

Panelists

- *Wen-Hung Chen, PhD* – Reviewer, COA Staff, OND, CDER, FDA
- *Tara Symonds, PhD* – Strategic Lead, Clinical Outcomes Assessments and Partner, Clinical Outcome Solutions
- *Kathleen (Kathy) Wywrich, PhD* – Executive Director, Center of Excellence for Outcomes Research, Evidera

1. Determining a meaningful outcome through patient interviews (15 min.)
 - Example from psoriasis
2. Supporting a clinical trial endpoint by considering meaningfulness data (15 min.)
 - Example from dysmenorrhea
3. Influencing study endpoints based on feedback from patient interviews (15 min.)
 - Example from gastroesophageal reflux disease
4. Quantifying meaningful score change with modified bookmarking (20 min.)
 - Example from multiple sclerosis
5. Panelist discussion/Q&A (20 min.)

Recap of the 2015 Session

Demonstration of Anchor- based Methods

- Logistic regression to interpret change in pain in fibromyalgia
- Linear regression to interpret change in itch severity in plaque psoriasis

Introduction to Novel Methods

- Bookmarking
- Exit interviews
- Conjoint analysis

Productive Panel Discussion

- FDA perspective on methods
- Responder analysis as an endpoint vs. supportive analysis
- Selection of an appropriate anchor
- Thresholds for evaluating improvement versus worsening

Activities Since 2015 Session



- Numerous publications advancing this topic
 - Summary of 2015 session¹
 - Bookmarking extended to interpreting change²
 - Exit interviews used to support responder definition³
 - Novel scale-judgment method applied to PROMIS⁴
- C-Path webinar summarizing 2015 COMPASS* discussion
- COMPASS has this topic on their next meeting agenda to move towards consensus on methods

* Consensus Panel for Outcomes Measurement and Psychometrics: Advancing the Scientific Standards

1. Coon CD, Cappelleri JC. Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Ther Innov Regul Sci*. 2016;50(1):22-29.

2. Cook KF, Kallen MA, Victorson D, Miller D. How much change really matters? Development and comparison of two novel approaches to defining clinically important differences in fatigue scores. *Qual Life Res*. 2015; 24 (Suppl 1):157-158.

3. Gelhorn HL, Kulke MH, O'Dorisio T, et al. Patient-reported symptom experiences in patients with carcinoid syndrome after participation in a study of telotristat etiprate: a qualitative interview approach. *Clin Ther*. In press.

4. Thissen D, Liu Y, Magnus B, et al. Estimating minimally important difference (MID) in PROMIS pediatric measures using the scale-judgment method. *Qual Life Res*. 2016;25(1):13-23.

Objectives of the 2016 Session



- This year's session will build on the momentum over the past year to move beyond numbers to actual meaning from the patient perspective
- Learning objectives:
 1. Recognize how an understanding of *outcomes that are meaningful to patients* can aid in understanding what score changes are meaningful
 2. Understand what insight *emerging approaches* might provide to complement or supplement traditional methods
 3. Explain *how the approaches presented may be implemented* in future instrument development projects to strengthen interpretation of scores on PRO measures

Meaningful Terminology

Meaningful Concept

- Does the instrument measure the concepts that are important to patients with this condition?

Meaningful Outcome

- What outcome do patients want to see to know that a treatment is beneficial?

Meaningful Change

- How much change should be observed in the PRO scores to know that a patient has experienced a meaningful outcome?

Defining an outcome meaningful to patients

Mona Martin, RN, MPA

Executive Director, Health Research Associates

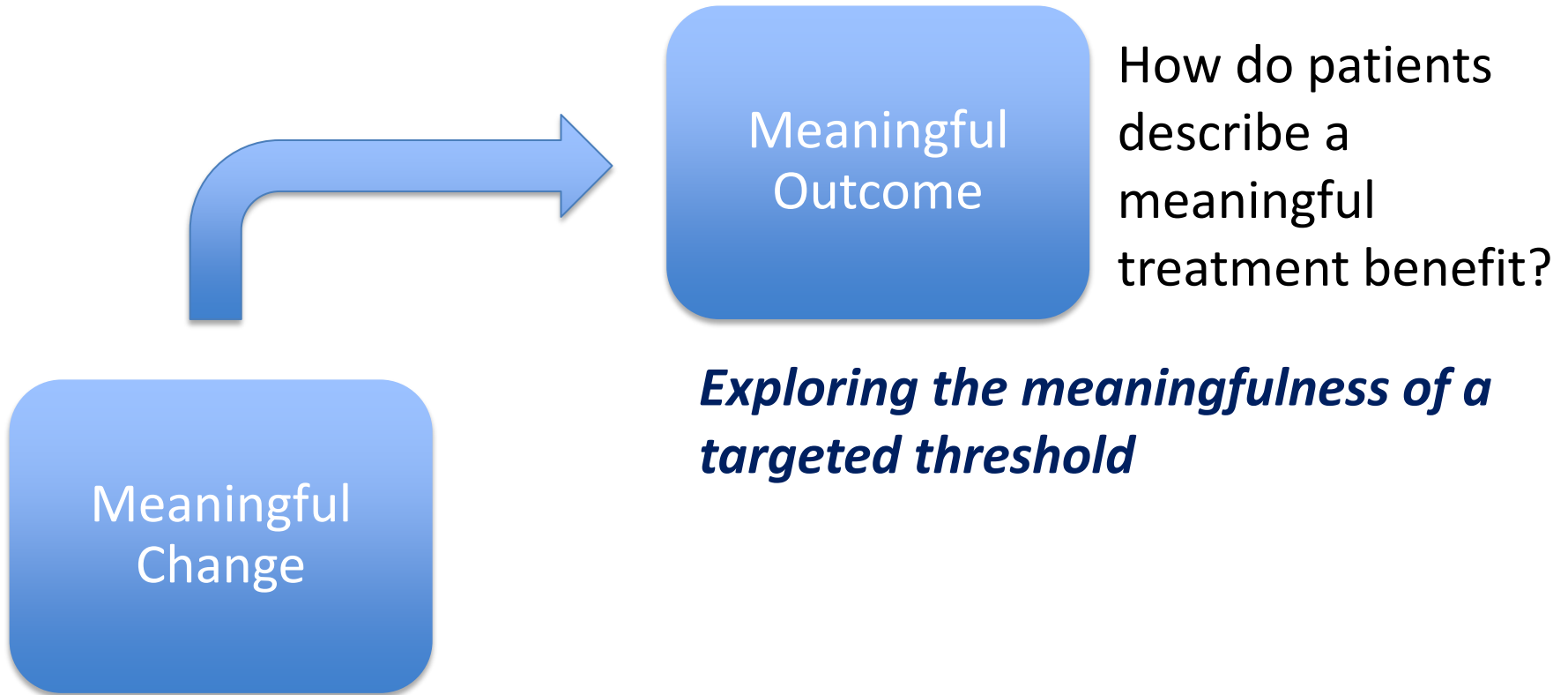
Brian G. Ortmeier, PharmD, PhD

Executive Director and Therapeutic Area Lead, Inflammation Global Health Economics

Neurology Global Health Economics, *AMGEN* Inc.

- To provide an example from the therapeutic area of psoriasis to show
- how the responder definition for a PRO measure can reflect both a clinically established threshold,
 - and an outcome (or state) that is meaningful to patients undergoing treatment for their psoriasis.

Presentation Objectives



In this case, the meaningful change can be considered the change required to achieve the target threshold or state

Background: Psoriasis

- Chronic plaque psoriasis is a common inflammatory skin disease that significantly impacts health-related quality of life (HRQoL) and leads to impairments in physical functioning and well-being
- Psoriasis signs: redness, cracking, scaling, flaking
- Non-observable symptoms: sensations of burning, stinging, pain, and itch.
- The Psoriasis Symptom Inventory (PSI) is a newly developed patient-reported outcome (PRO) measure designed to reflect the **patient's perception of symptom severity** of their chronic plaque psoriasis.

Psoriasis Symptom Inventory (PSI)

Score Range: 0-32



For the following group of questions, the “last 24 hours” means from right now - back to yesterday at this time.	Not at all	Mild	Moderate	Severe	Very Severe
1) Overall, during the last 24 hours, how severe was the <u>itch</u> from your psoriasis?					
2) Overall, during the last 24 hours, how severe was the <u>redness</u> of your skin lesions?					
3) Overall, during the last 24 hours, how severe was the <u>scaling</u> of your skin lesions?					
4) Overall, during the last 24 hours, how severe was the <u>burning</u> of your skin lesions?					
5) Overall, during the last 24 hours, how severe was the <u>stinging</u> of your skin lesions?					
6) Overall, during the last 24 hours, how severe was the <u>cracking</u> of your skin lesions?					
7) Overall, during the last 24 hours, how severe was the <u>flaking</u> of your skin lesions?					
8) Overall, during the last 24 hours, how severe was the <u>pain</u> of your skin lesions?					

Meaningful Target Thresholds

- Assessment of psoriasis severity has traditionally been conducted by a small group of clinician reported outcomes.
 - Psoriasis Area and Severity Index (PASI)
 - An estimate of the body surface area (BSA) affected by psoriasis
 - Physician's Global Assessment (PGA)
- Until relatively recently, a 75% improvement in a PASI score was the standard target for improvement and supported by clinicians as a “meaningful” amount of change.
- Improved efficacy of new biologics push potential improvement to PASI 90%, and even PASI 100%.

These new **target thresholds** for clinician determined improvement are commonly referred to as “**clear**” or “**almost clear**”

Physician’s global assessment scores:

- | | | | |
|----------|---------------|--|------------------|
| 0 | Clear |  | PASI 100% |
| 1 | Almost Clear |  | PASI 90% |
| 2 | Mild Moderate | | |
| 3 | Moderate | | |
| 4 | Severe | | |
| 5 | Very Severe | | |

(each with clinician determination of plaque elevation and amount of scaling and redness)

The Question of Meaningfulness

Based on score distributions from Phase 2 studies
clinical experts suggested that:

“Clear” **PSI=0**

“Almost Clear” **PSI= <8 (with no single item >1)**

Targeted Thresholds on the PSI

CLEAR

ALMOST
CLEAR

For the following group of questions, the "last 24 hours" means from right now - back to yesterday at this time.	Not at all	Mild	Severe	Very Severe
1) Overall, during the last 24 hours, how severe was the <u>itch</u> from your psoriasis?	✓ = 0	✓ = 1		
2) Overall, during the last 24 hours, how severe was the <u>redness</u> of your skin lesions?	✓ = 0	✓ = 1		
3) Overall, during the last 24 hours, how severe was the <u>scaling</u> of your skin lesions?	✓ = 0	✓ = 1		
4) Overall, during the last 24 hours, how severe was the <u>burning</u> of your skin lesions?	✓ = 0	✓ = 1		
5) Overall, during the last 24 hours, how severe was the <u>stinging</u> of your skin lesions?	✓ = 0	✓ = 1		
6) Overall, during the last 24 hours, how severe was the <u>cracking</u> of your skin lesions?	✓ = 0	✓ = 1		
7) Overall, during the last 24 hours, how severe was the <u>flaking</u> of your skin lesions?	✓ = 0	✓ = 1		
8) Overall, during the last 24 hours, how severe was the <u>pain</u> of your skin lesions?	✓ = 0	✓ = 1		

But WHAT PSI SCORE represents a threshold or a state that is meaningful to patients?

- How do patients describe their symptom state when they reach a target threshold?
- Are threshold differences between 90% (almost clear) and 100% (all clear) meaningful to patients?

To help understand and to clarify *patient perceived meaningfulness* of “clear” and “almost clear” score thresholds on the PSI,

A subgroup of 30 subjects were identified from the 220 who had been enrolled in a larger observational study in 8 different treatment centers across the US.

These 30 adult subjects had been treated with one of the new biologics (STELARA[®] HUMIRA[®] or ENBREL[®])

- 15 had clinician global assessment scores of 0 (clear)
- 15 had clinician global assessment scores of 1 (almost clear)

Qualitative interviews were conducted to obtain patient language to explore the meaningfulness of these incremental treatment benefits.

Sample quotes from subjects scoring PSI=0

- *I'm just like **normal skin***
- *After the second injection I was... **clear of all my psoriasis.***
- *The medicine's **completely cleared me up.***
- *My **skin is clear right now...** no symptoms*
- *Only on my right side, my elbow. And it's **only like three little spots, It doesn't bother me at all.***
- *Right now,... the **tiny red mark over there. It doesn't bother me at all.***

Main Characteristic of descriptions for PSI=0

(symptoms either totally gone, or very minor and causing “no bother”)

Sample quotes from subjects scoring PSI=0.17 to 1.0

- *I have not one skin lesion... But I would say there's **still some pain**.*
- *Very **slight** [discoloring on the knee] and that's almost completely gone.*
- *I have a toenail that still has a, it's kind of...**pulled back from my skin** from the psoriasis. But really that's the only problem I've got from it.*
- *where now you know it's very light [on the forehead]...but now it's **just real light** [on the stomach], there's no irritation and it's **very light**, it's like a birthmark.*
- *A small plaque maybe once every two months and **barely noticeable**.*
- *I got **just a smudge** (redness) on my elbow... **lighter now**....*

Main Characteristic of descriptions for PSI=PSI=0.17 to 1.0
(variety of residual symptoms, “less noticeable”)

Sample quotes from subjects scoring PSI=3.0 to 6.86

- *I have **a few** [lesions] on my heels.*
- *some of my **nails are separating again** and getting the white.*
- *You could see a couple of you know, not big patches but you know, **fifty cent sized patches**.*
- *when I use my elbows it wants to come back; it's just really crazy...**It's in my scalp and it's just it's minor**.*
- *the one on my right knee was really bad and now you can look there and now there is **just a slight redness** in that area that is **hardly even noticeable**...*
- *It will get a **little bit itchy** between my eyes and forehead and like on the sides of my nostrils, if I don't use [cream], then I'll get it through there and **it gets red**.*
- *I'm looking at my legs now and I've got a little bit. They're not bright red like they had been.*

Main Characteristic of descriptions for PSI=3.0 to 6.86
(more symptoms expressed, “hardly noticeable and still mild”)

Are Score Differences Meaningful to Patients?



Statements from Patients who were designated “clear” (PGA=0)
identified the following states as meaningful

- Being **cleared up**, having no lesions.
- Being **completely clear**
- **Not having psoriasis**, able to wear shorts
- **Not feeling constant itch**
- Being **99% psoriasis free**
- **Not having to worry about redness** or soreness
- **Its cleared up**, no itching
- *Better* appearance and comfort
- Feeling *better*, itching *less*
- *Less* pain, *lessen* appearance
- Feeling *better* about self

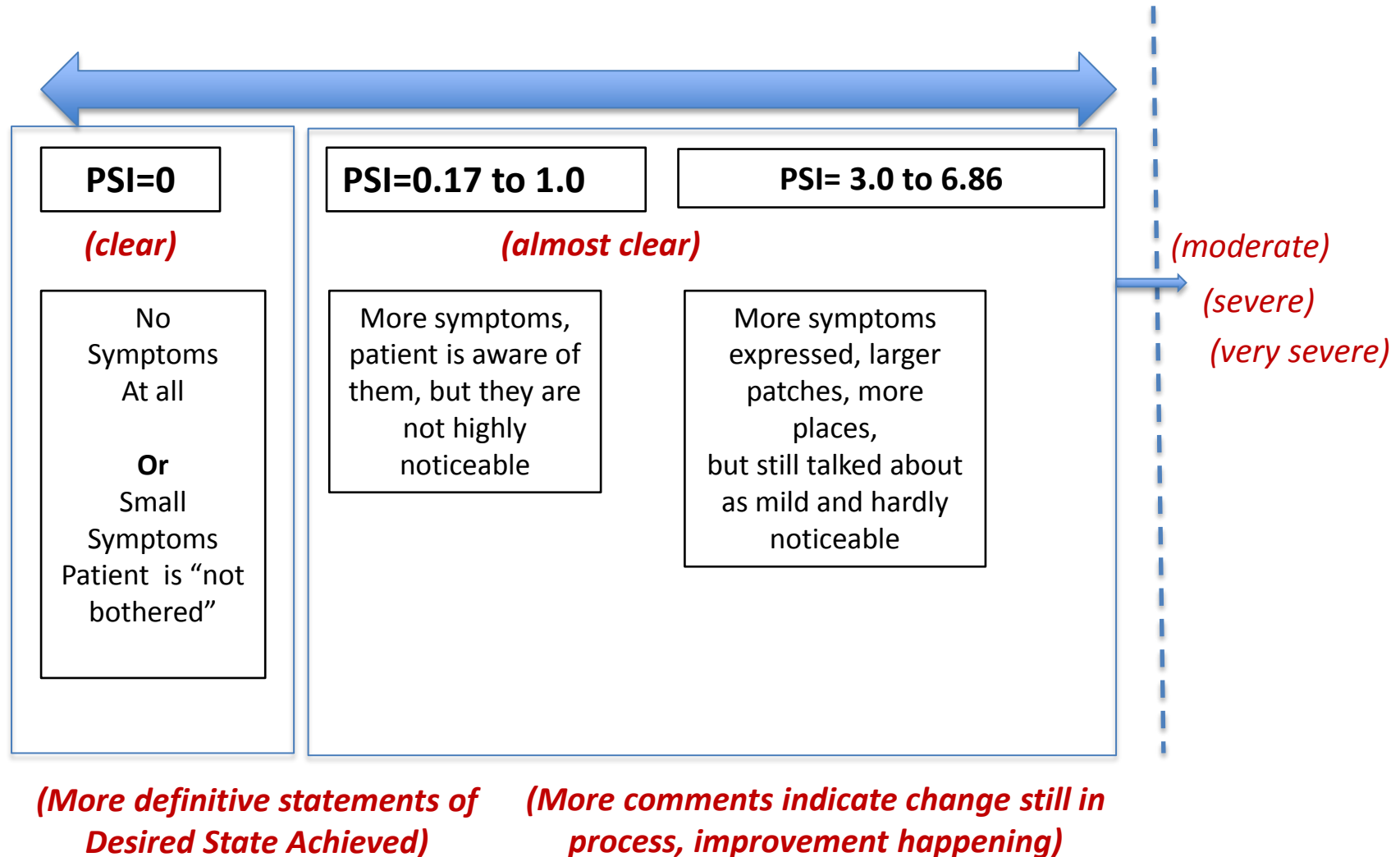
(More definitive statements of Desired State Achieved)

Statements from Patients who were designated “almost clear” (PGA=1)
Identified the following states as meaningful

Don't have cracking
Peace of mind, **irritant gone**
Makes me *feel better*
Have *more energy*
Increased mobility
Seeing *progressive improvement*
Improved appearance
Have *small lesions*
You *feel normal again*
Feel better about yourself
See it disappearing
Not scratching constantly

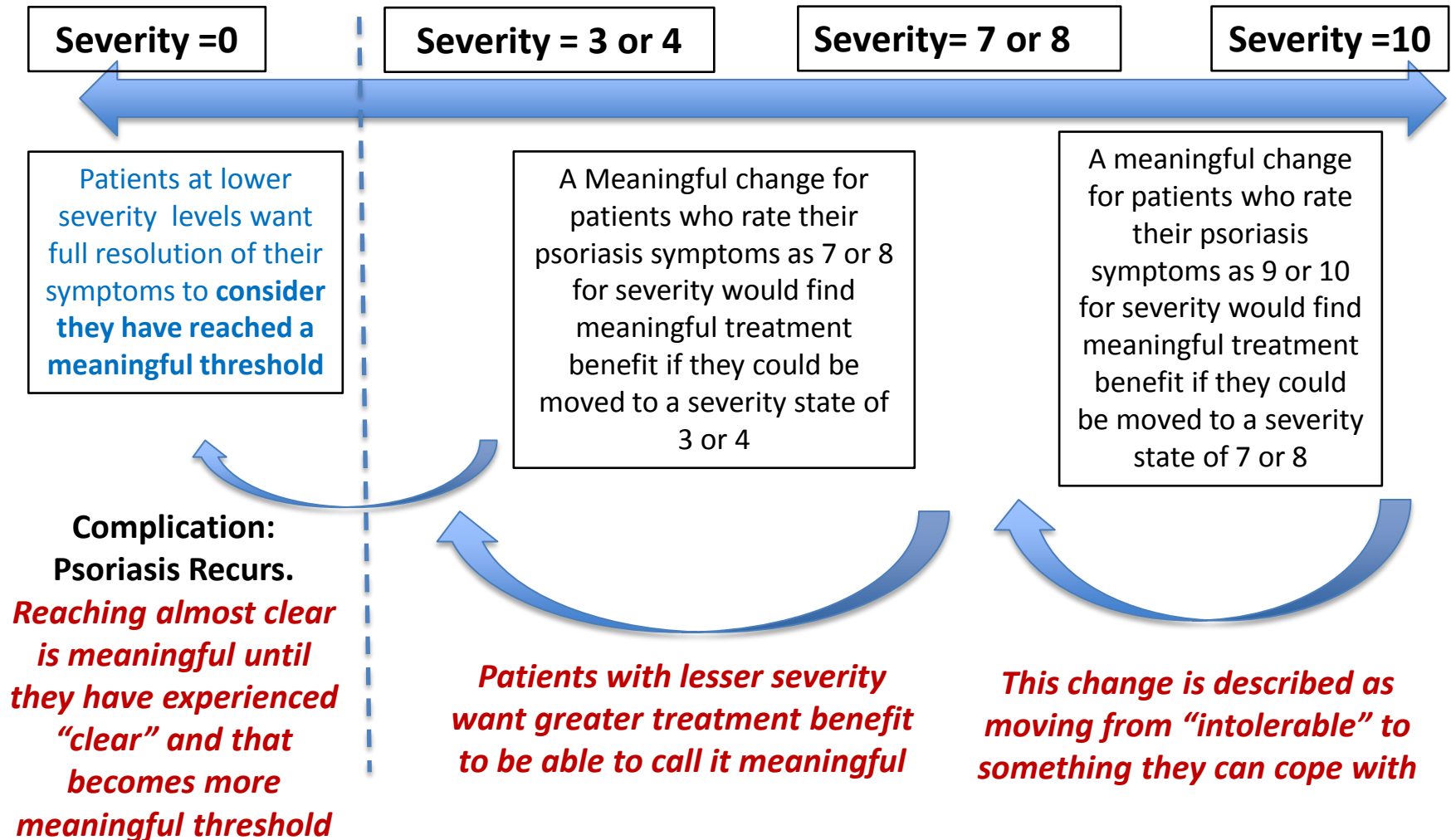
(More comments indicate change still in process, improvement happening)

Patient Perceptions of Symptom Status across Target Score Thresholds



What do we know about Thresholds that Psoriasis Patients Consider to be Meaningful?

We know what patients with more severe psoriasis consider meaningful from previous qualitative work



How does this work enhance our understanding of a meaningful score?



Understanding how patients assign meaning to PRO score thresholds can:

- Provide more information about how the PRO scoring system is working
 - *(provided confirmation that patients are understanding the conceptual structure of the instrument)*
- Increasing clarity about what threshold or target scores are actually representing
 - *(define symptom severity)*
- Providing greater context for interpretation of PRO data
 - *(expanded definitions into new territories of achievable treatment benefit)*

What is a meaningful change?

Ask the patient!

Allison M Nguyen¹, Tjeerd Korver¹, Fang Chen¹, Rob Arbuckle², Alice Turnbull², Josephine M Norquist¹

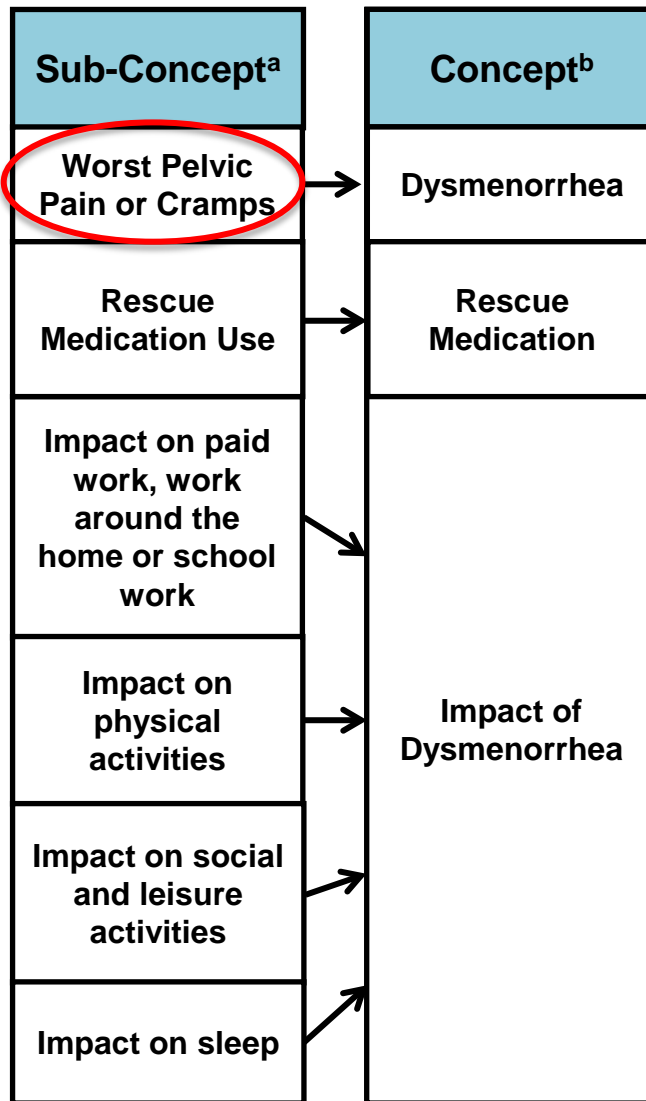
¹Merck Research Laboratories, Whitehouse Station, NJ (USA); ² Adelphi Values, Bollington, Cheshire (UK)

- Primary dysmenorrhea refers to the syndrome of painful menstruation with no organic cause
- The Dysmenorrhea Daily Diary (DysDD), a newly developed measure, was found to have good content validity and to be a valid, reliable and responsive measure for assessing primary dysmenorrhea
 - Nguyen et al. *Qual Life Res* 2015; 24:181-191; Norquist et al. *ISOQOL* 2015
- Although methods for interpretation of clinical trial endpoints have been developed and debated for more than a decade, challenges remain in determining the degree of change considered important and meaningful to patients, physicians, payers, and regulators

- **Objective:** To summarize methodologies, including those based on direct patient input, to determine the degree of change in dysmenorrhea (pelvic pain or cramps) that can be considered meaningful and clinically important
- **Data Source:** Data from a phase IIb, multinational, randomized, blinded, placebo-controlled clinical trial to evaluate the effect of a vaginal ring on primary dysmenorrhea were used for the analyses

Dysmenorrhea Daily Diary (DysDD)

Conceptual Framework



- The DysDD is a 10-item daily disease-specific ePRO measure with a 24-hour recall period
- The DysDD was completed every day over two treatment cycles
- Pelvic pain (i.e., DysDD Item #3) assessed using a 0-10 numeric rating scale with 0='no pain or cramps', 10='extreme pain or cramps'

^a Sub-concepts reflect item-level measurements

^b Concepts reflect the broader categories items map to

Directly Asking Subjects to Rate the Significance of Their Change

- Global Assessment of Change (GAC) -
A single item capturing subjects' ratings of change in their 'pain or cramps' since the start of the study ("much worse", "worse", "a little worse", "the same", "a little better", "better", or "much better")
- 'Meaningful' question
 - Those who reported *worsening* on GAC were then asked:
'Was this increase in pelvic pain or cramps an important change for you?' (Yes/No)
 - Those who reported *improvement* on GAC were then asked:
'Was this decrease in pelvic pain or cramps an important change for you?' (Yes/No)
- Both completed once at the end of Treatment Cycle 2

Directly Asking Subjects to Rate the Significance of Their Change

- The GAC provided the opportunity for subjects to rate both their degree of change

AND

report whether they felt the change was meaningful

The addition of this ‘meaningful’ question is a new approach which was useful for gaining a better understanding of what degree of change the subjects perceive as important

- Responsiveness traditionally assessed by analyzing changes between active and control treatment groups
- Analyses here are based on pooled treatment groups, therefore responsiveness was assessed by analyzing changes between those who improved vs. didn't improve defined as:
 - ≥ 1 point mean improvement on the Menstrual Distress Questionnaire (MDQ) 'cramps' score
 - Score of ≥ 4 ("a little better", "better" or "much better") on the GAC
 - Score of ≥ 5 ("better" or "much better") on the GAC

Clinically Important Responder (CIR) Analyses

- Distribution-based methods to identify the smallest change that would exceed measurement error
 - Based on 0.5 SD, standard error of measurement (SEM) and the minimal detectable change (MDC) with 90 and 95%CI
 - $SEM = [SD * \sqrt{1-r}]$
 - $MDC = SEM * 1.65$ or $SEM * 1.96$
- ROC curve analysis to assess the ability of change in pelvic pain or cramps to discriminate between subjects scoring
 - ≥ 4 (“a little better”, “better” or “much better”) vs. < 4 on GAC
 - ≥ 5 (“better” or “much better”) vs. < 5 on GAC
 - The point on the curve that maximizes sensitivity and specificity is considered the optimal level of change in DysDD pain score that differentiates between responders and non-responders (i.e., CIR)

Interpretation of Results

- All results were qualitatively compared to identify a peak pelvic pain or cramps change score that can be considered meaningful and important to patients
- Input from patients, based on the ‘meaningful’ question used to support the identified threshold for CIR

Results

Responsiveness: Peak change⁺ in DysDD item 3 (pain or cramps rating) in responder groups from baseline to treatment cycle 2

Responder Definition	Responder Status	N	Change in Peak item 3 score (SD) [95%CI]
≥1 point mean improvement on the MDQ cramps score	Responder	231	-4.9 (2.77) [-5.25, -4.54]*
	Non Responder	74	-0.6 (1.66) [-1.00, -0.21]
≥4 on the GAC**	Responder	227	-4.5 (2.94) [-4.87, -4.10]*
	Non Responder	50	-1.1 (2.33) [-1.78, -0.46]
≥5 on the GAC**	Responder	174	-5.0 (2.75) [-5.37, -4.54]*
	Non Responder	103	-2.1 (2.86) [-2.62, -1.50]

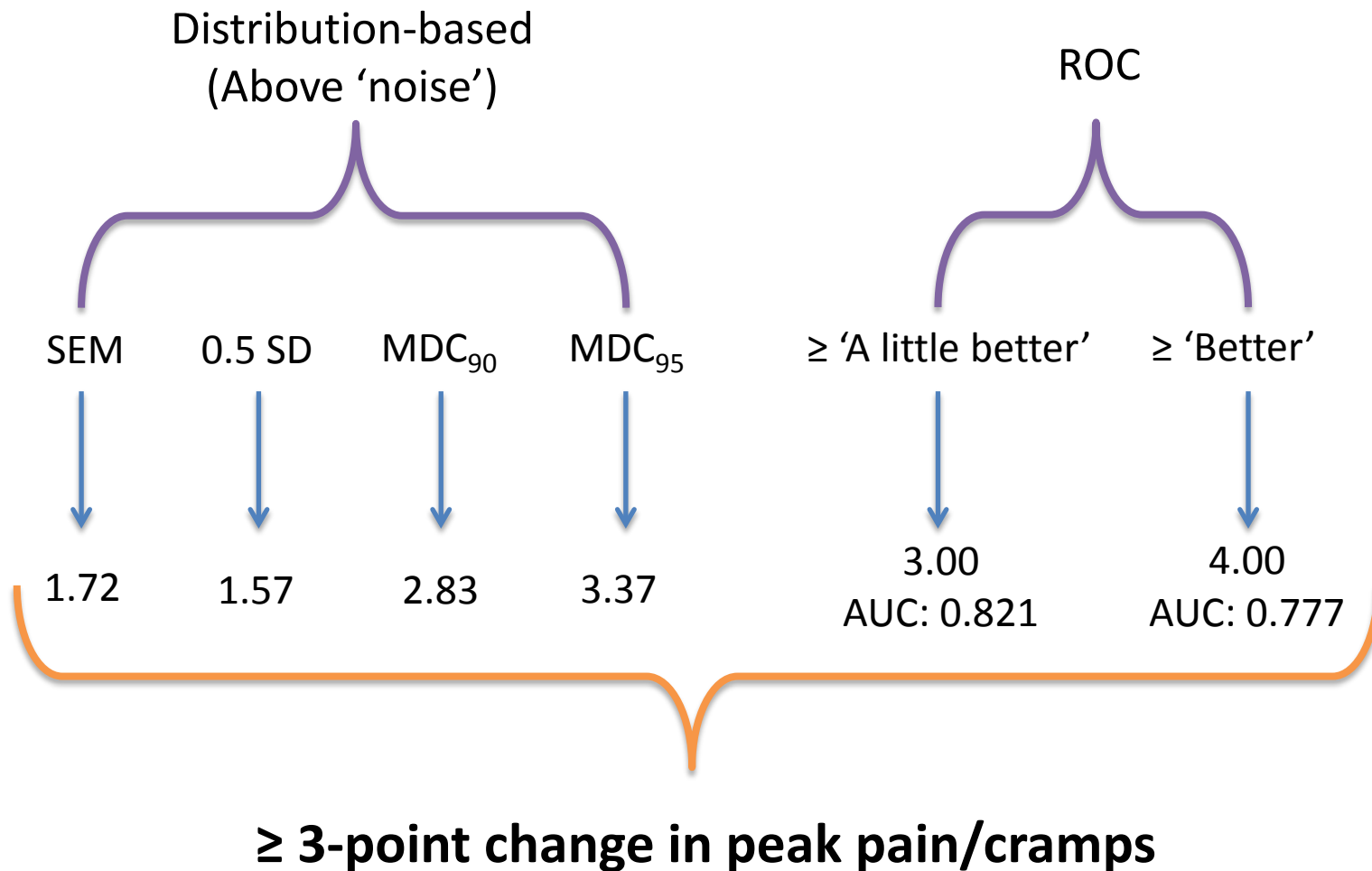
+Change from baseline to treatment cycle 2 for the mean pain scores calculated on the day in the menstrual bleeding period with the highest pain rating

*p<0.001

**≥4: “a little better”, “better”, “much better”; ≥5: “better” or “much better”

The DysDD pain or cramps was shown to be responsive to change over time

Triangulation of Results: Change in Item #3 Pain/Cramps Score



‘Meaningful’ Data Used to Support CIR

Among subjects who experienced at least a 3-point reduction in peak pelvic pain (n=171)

Global Assessment of Change

A little better
(n=25)

Better
(n=55)

Much Better
(n=86)

Was this decrease important for you?

No

Yes

4 (16%)

21(84%)

No

Yes

1 (1.8%)

54 (98%)

No

Yes

2 (2%)

84 (98%)

93% reported the change to be meaningful

- ROC curve analyses typically anchor on GAC ratings of “better” or “much better” with responses of “a little better” included in the “no change” or worsening groups
 - ‘Meaningful’ data indicate that subjects who achieve the 3-point change and rate themselves as “a little better” consider that change important
 - Categorizing those subjects with subjects who reported “no change” or worsening is therefore not appropriate
- Evidence to support a cut-point of greater than or equal to 3 should not be driven by subjects achieving changes greater than 3
 - Among subjects who experienced a change equal to 3 points (n=20), 70% reported the decrease to be important

- A 3-point or greater reduction in the peak pelvic pain or cramps NRS score was found to be a consistent degree of change that is both meaningful and above any inherent “noise” of the measure
- Among subjects who experienced ≥ 3 -point reduction in their peak pelvic pain or cramps, the vast majority (93%) reported that the decrease in their pain was important

How does asking the patient enhance our ability to determine meaningful score changes?



- Traditional analyses of responsiveness, anchor-/distribution-based methods, and ROC curves, combined with the 'meaningful' data was a novel approach to substantiate a CIR for use in future clinical trials
- The 'meaningful' question provided insight into the degree of change patients feel is meaningful and argues against categorizing “a little better” with “no change” or “worsening”
- Consistent with the emphasis on greater 'patient-centricity', this approach directly utilizes patient input to define an important treatment effect

- Moos RH. The development of a menstrual distress questionnaire. *Psychosom Med* 1968;30(6):853-67
- Nguyen AM, Humphrey L, Kitchen H, Rehman T, Norquist JM. A qualitative study to develop a patient-reported outcome for dysmenorrhea. *Qual Life Res* 2015; 24:181-191.
- Norquist JM, Korver T, Chen F, Arbuckle R, Turnbull A, Nguyen AM. Validation of a New Disease-Specific ePRO Measure to Support Dysmenorrhea Clinical Trials: The Dysmenorrhea Daily Diary (DysDD). Poster presented at ISOQOL Conference 2014
- U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009 Dec.
- Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36); I. Conceptual Framework and item selection. *Med Care* 1992;30(6):473-83

Frontloading clinical programs with patients' perspective of meaningful change.

Example of GERD

Katarina Halling, MSc

Global Head, Patient Reported Outcomes,
AstraZeneca and Industry Co-Director, PRO
Consortium

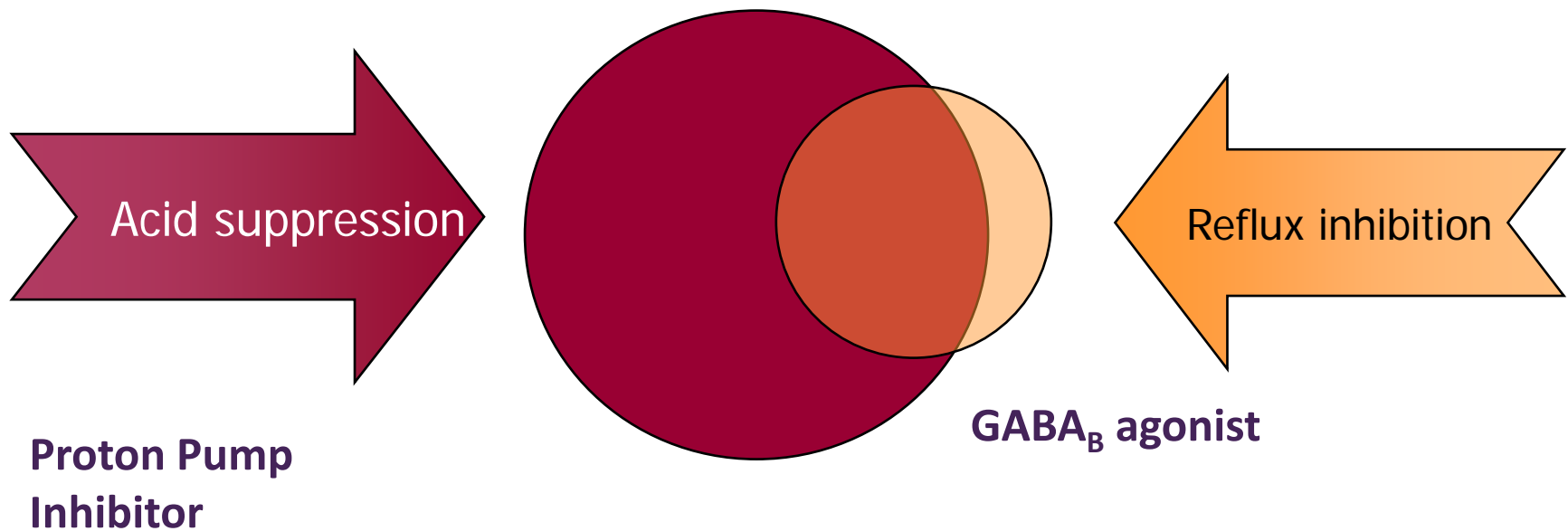
This research was performed with significant contributions from:



- Anna Rydén, AstraZeneca
- Maria Karlsson
- Jean Paty, Quintiles
- Mona Martin, HRA
- Nimish Vakil, Aurora Wilkinson Medical Clinic

- Proton pump inhibitors (PPIs) are the most commonly prescribed class of medication for the treatment of heartburn and acid-related disorders.
- Partial response to a PPI is a problem.
- AZ has a compound in development that targets transient lower esophageal sphincter relaxations.
- The primary endpoint was GERD symptoms.

GERD Treatment Strategy



The primary treatment goal for lesogaberan is to provide 24-hour symptom relief in patients with persistent GERD symptoms, who experience a partial response to PPI treatment.

Change in regulatory landscape: old vs newer expectations in what to measure for a similar label

What concept to
measure...

Then

Heartburn



... to obtain the label

GERD Symptom
relief

Now

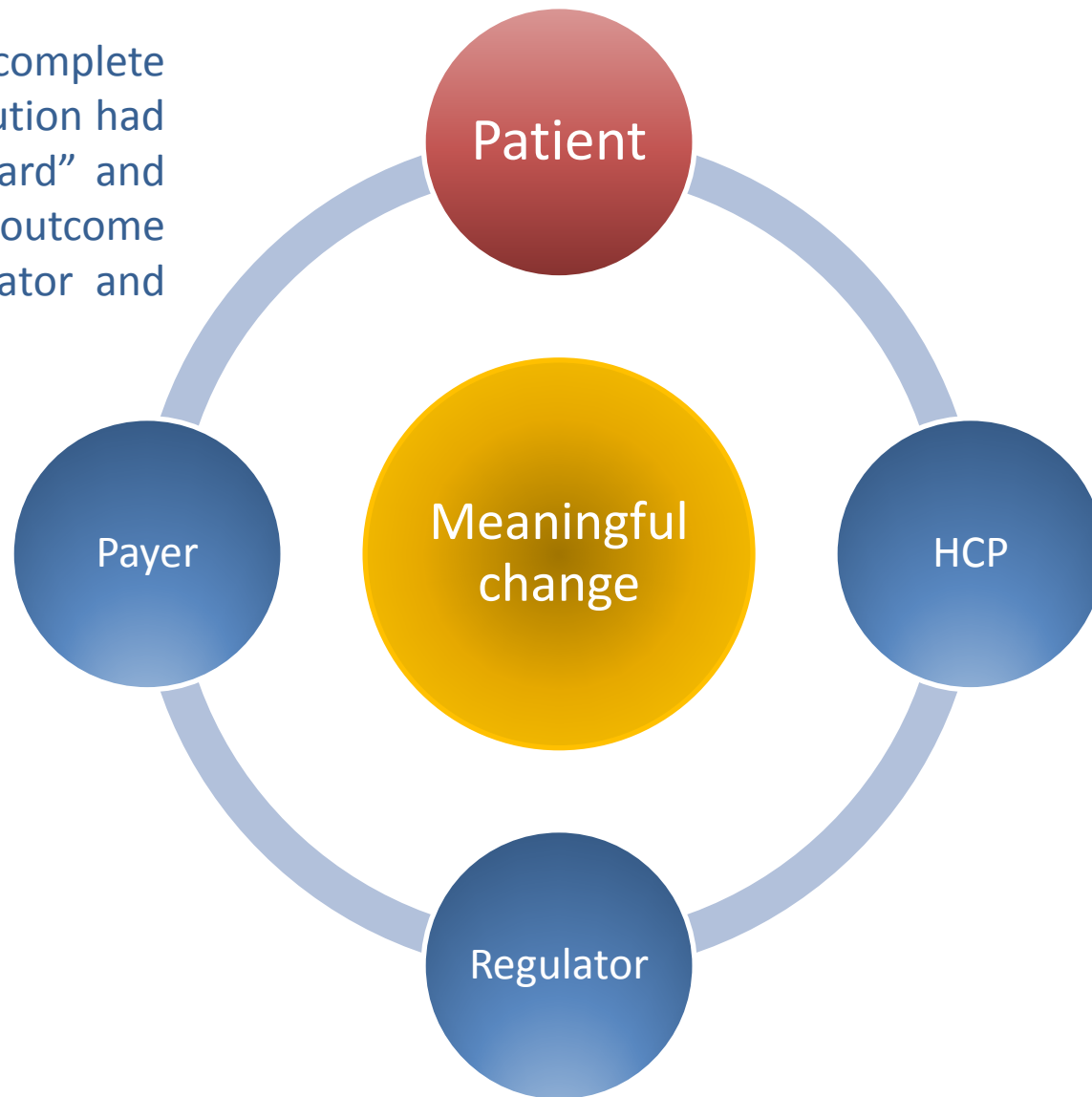
"All relevant and
important GERD
symptoms"



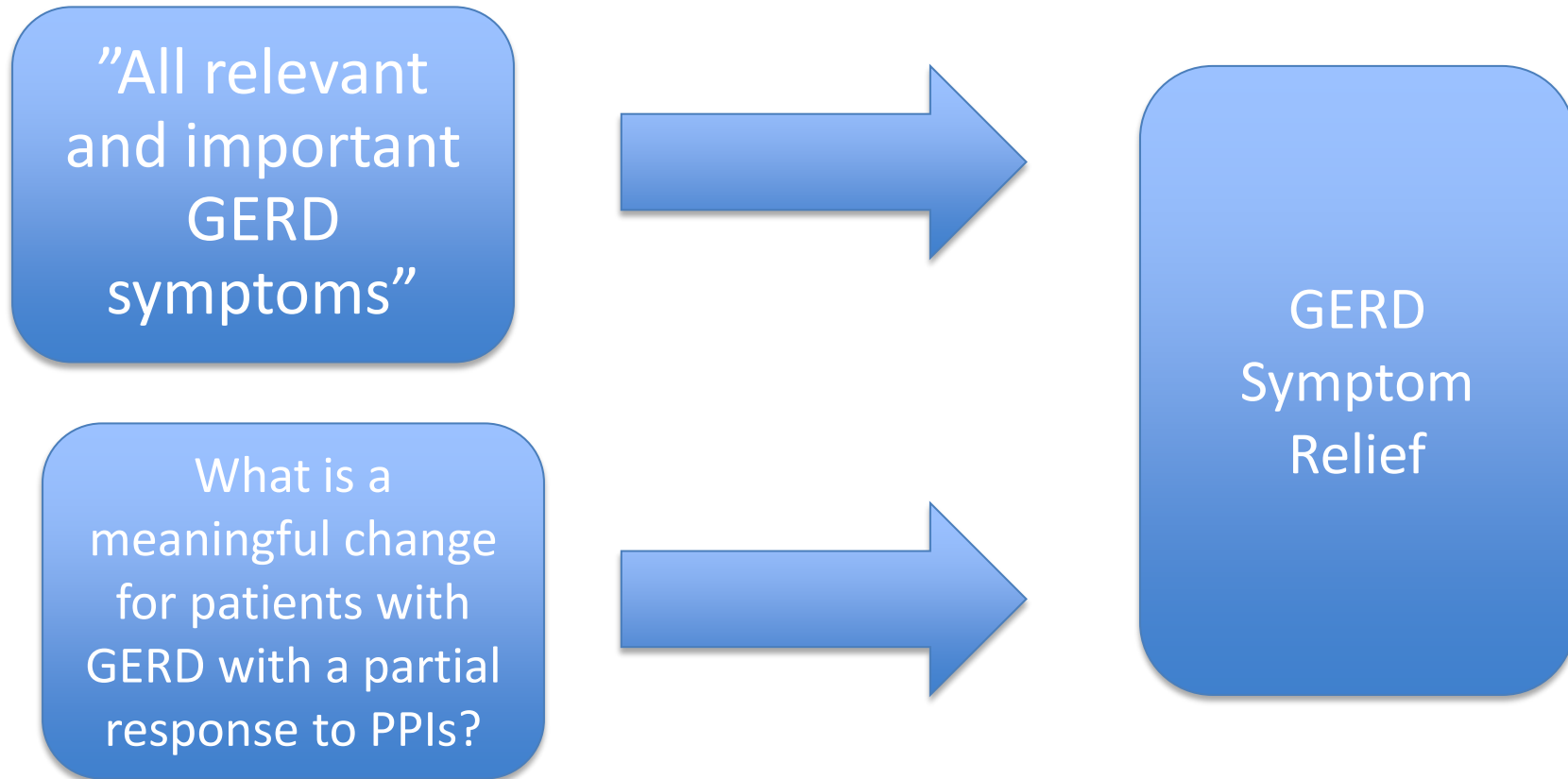
GERD Symptom
relief

How much relief?

In GERD, complete symptom resolution had become “standard” and the expected outcome by HCPs, regulator and payers

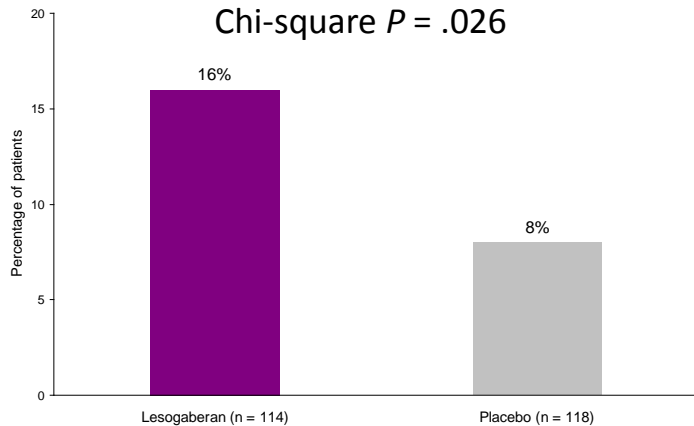


Two critical aspects that had implications for the design

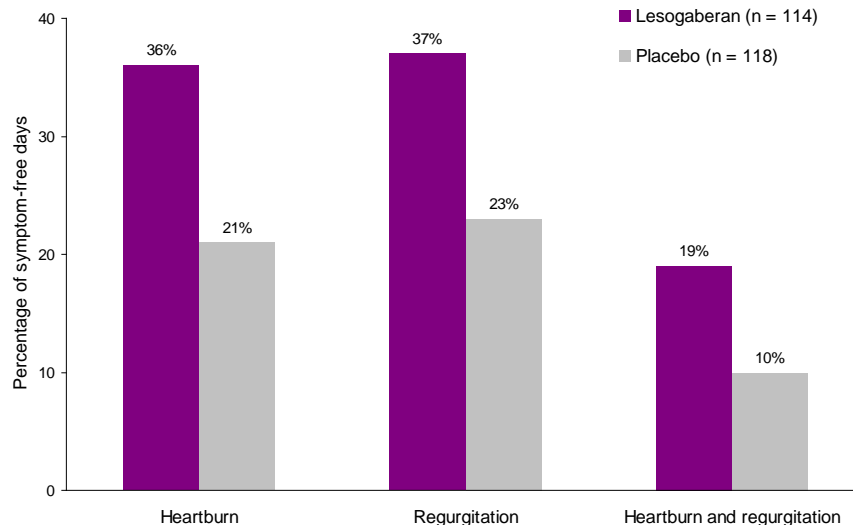


And Implications for the Target Product Profile!

Response rates during 4 weeks treatment with lesogaberan 65mg twice daily



Percentage of patients with treatment response = at most one 24-hour period with heartburn or regurgitation of not more than mild intensity during the last 7 days of treatment



Percentage of symptom-free days during the 4-week treatment period

Patient Reported Symptoms RDQ (Reflux Disease Questionnaire)

Recording of the intensity of the following symptoms in the eDiary every morning and evening;

- A burning feeling behind the breastbone
- Pain behind the breastbone
- A burning feeling in the center of the upper stomach
- A pain in the center of the upper stomach
- An acid taste in your mouth
- Unpleasant movement of material upwards from the stomach

Heartburn dimension

Dyspepsia dimension

Regurgitation dimension

AstraZeneca

2a @RDQ Baseline

a. A burning feeling behind your breastbone?

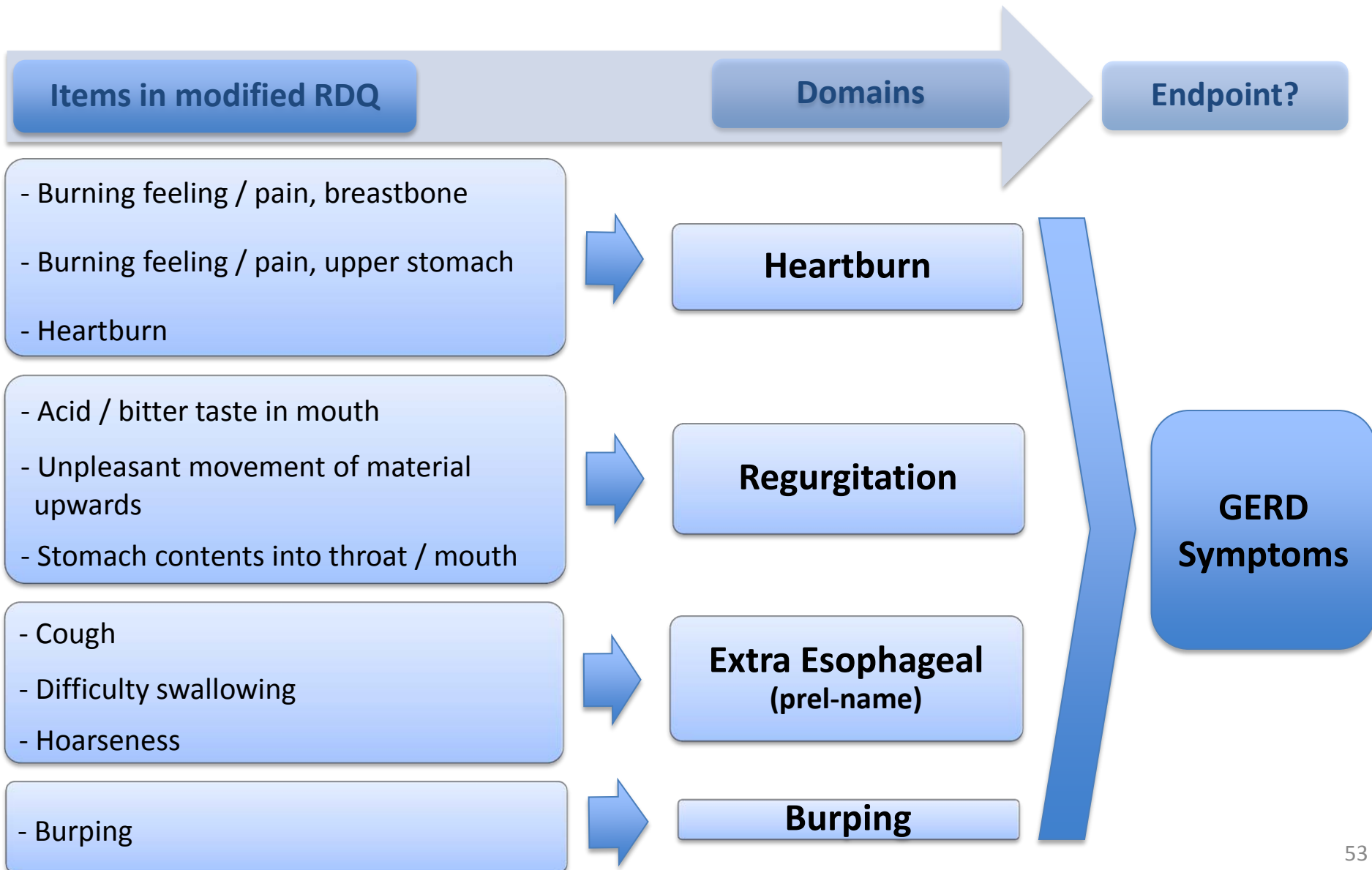
- ☐ Did not have
- ☒ Very mild
- ☐ Mild
- ☐ Moderate
- ☐ Moderately Severe
- ☐ Severe



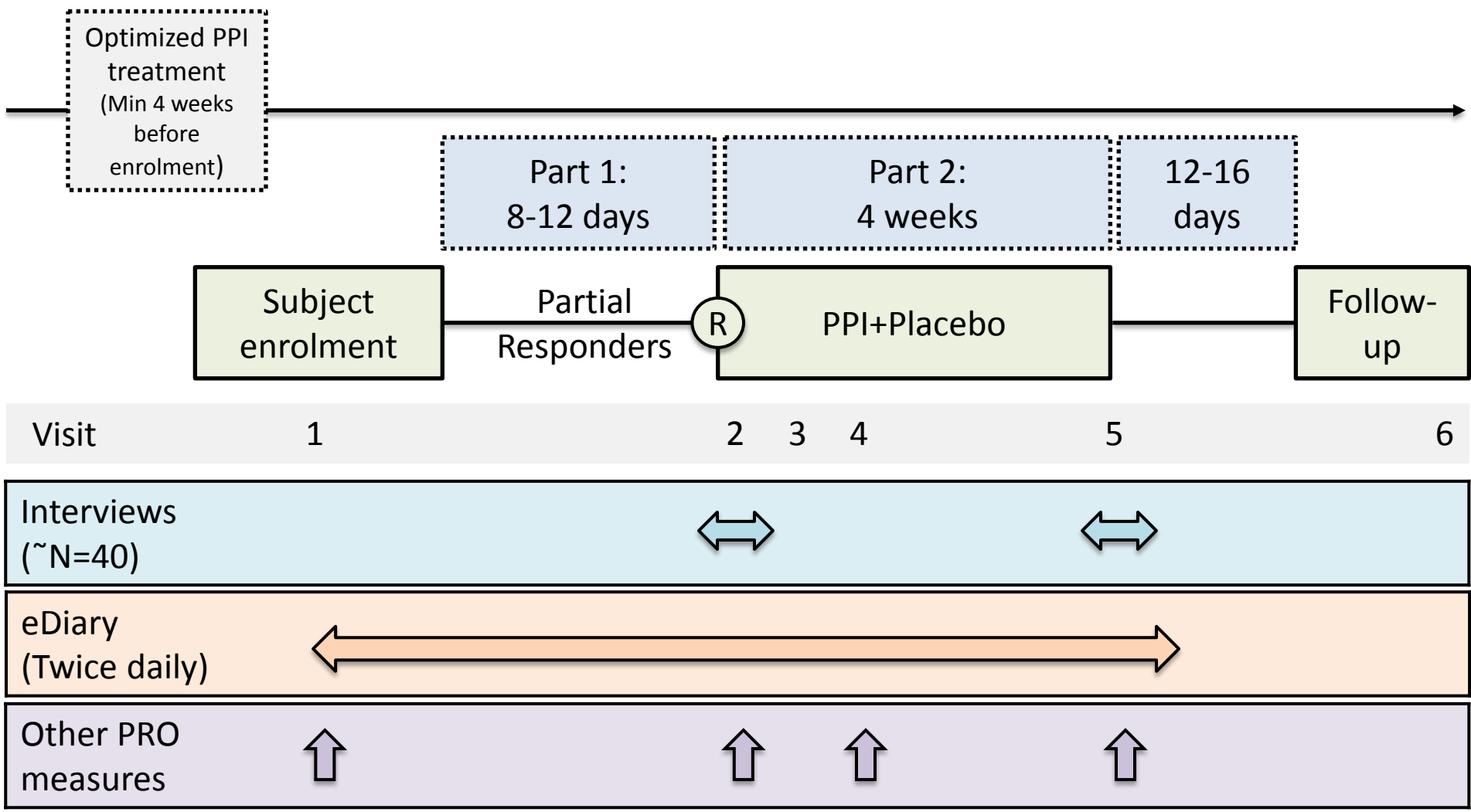
Main objectives of the PRO study

- Assess the psychometric properties of the e-diary in the target population.
- Selection of symptoms for primary end-points in future trials.
- Define clinically meaningful change through quantitative and qualitative methods and establish the responder definition in the target population:
 - In trial interviews before (n=42) and after (n=37) treatment;
 - Triangulation across multiple quantitative methods (e.g. cumulative frequency distribution, anchor and distribution-based);
 - Select PRO anchor that has intuitive meaning and reflects important change for patients to qualifying change according to meaningful/important;

Draft conceptual framework



Study Design



Meaningful Concept

- Does the instrument measure the concepts that are important to patients with this condition?
- Confirmed in Interview 1

Meaningful Outcome

- What outcome do patients want to see to know that a treatment is beneficial?
- Explored in Interview 1 and 2

Meaningful Change

- How much change should be observed in the PRO scores to know that a patient has experienced a meaningful outcome?
- Explored in Interview 2 and quantitatively

Objectives of the interviews

Interview

1

- To understand how patients describe in their own words what type of GERD symptoms they experience and if and how these symptoms are a problem to them
- To confirm that the symptoms and concepts included in the PRO symptom tool are relevant and important in the target patient population
- To understand how patients' lives are impacted by their GERD symptoms
- To explore what patients regard as successful treatment for their GERD or reflux symptoms, and what they would consider to be "symptom control" and "sufficient symptom control"

Interview

2

- To understand what, if any, symptom change the patient has experienced since the add-on treatment with AZD3355 or Placebo started, and how the patient perceives this change
- To explore what patients consider to be a meaningful or important change in their GERD symptoms

Results:

Interview before treatment

Frequency most important aspect of symptoms but severity also mentioned

Question: What is it that troubles / bothers you the most about your symptoms?

Coded Expressions	Number of mentions	% of mentions in total
Frequency	76	62%
Severity	28	23%
Duration	19	15%
Total	123	100%

Patients most often expressed their bother with symptoms in terms of frequency (62% of total responses mentioned).

<u>Definition of symptom control</u>	<u>Concept Code</u>	<u>Total number of responses</u>	<u>% of total responses (47)</u>	<u>Total number of transcripts contributing</u>	<u>% of contributing transcripts (42)</u>
Question: While thinking about your GERD symptoms, what would you consider symptom control?	A band-aid as opposed to symptom elimination	1	2%	1	2%
	Controlling the problem	1	2%	1	2%
	Controlling the symptoms	7	15%	7	17%
	Daily medication	3	6%	3	7%
	Elimination of symptoms / almost all gone	2	4%	2	5%
	I don't experience it / it is very, very minor	1	2%	1	2%
	I don't know	2	4%	2	5%
	It's just hiding	1	2%	1	2%
	Minimisation of symptoms	2	4%	2	5%
	Not as frequent	2	4%	2	5%
	Not eating what I like to eat / being careful what I eat	2	4%	2	5%
	Not fixing it but masking it	1	2%	1	2%
	Noticeable change in the symptoms	1	2%	1	2%
	Preventative type medicine	2	4%	2	5%
	Reducing the frequency / duration of symptoms	1	2%	1	2%
Number of transcripts with no response: 1	Noticeable change in the symptoms	1	2%	1	2%
	Satisfying your discomfort	1	2%	1	2%
	Symptom free	1	2%	1	2%
	Symptoms occur but only within set parameters	1	2%	1	2%
	Symptoms take over the person	1	2%	1	2%
Number of transcripts with 1 response: 36	Symptoms under control 6 out of 7 days	1	2%	1	2%
	Symptoms would be kept at bay	1	2%	1	2%
	Take all that stuff away that's happening	1	2%	1	2%
	Take meds / control my eating	3	6%	3	7%
	The medicine controls it but it does not go away	1	2%	1	2%
Number of transcripts with multiple responses: 5	Under control but not alleviated all together	3	6%	3	7%
	Would not be embarrassed	1	2%	1	2%
	Would not have symptoms	1	2%	1	2%
	You can deal with it	2	4%	2	5%
		47	100%		

The majority of the definitions of “Symptom Control” pertained to partial as opposed to complete, symptom control.

Interview after treatment

- Perception of change in symptoms?
- Was the change important?
- Was the change meaningful?
- Did the medication provide sufficient symptom control?
- Was the treatment successful?

Interview 2: Reports of Sufficient Symptom Control and Treatment Success



- Complete symptom relief (N=4)
- No Relief (N=4)
- Partial Relief throughout the Treatment Period (N=22)
- Relief in the first 2 weeks of the Treatment Period, with decline in the second (N=3)
- Relief only in the second 2 weeks of the Treatment Period (N=3)

NB: Treatment change during the study relates to change irrespective of which treatment (placebo or AZD3355) the patient had received.

- Of the 22 patients who reported experiencing partial symptom relief throughout the treatment period, 17 indicated that the medication provided sufficient control of their symptoms and that the treatment was successful:
 - Improvement reported in symptom frequency n=7
 - Improvement reported in symptom severity n=5
 - Improvement reported in both frequency and severity n=5

Patient quotes from interview 2

- [The symptoms are] “extremely mild and very infrequent.”
- “...everything that I’ve felt in the past has lessened intensity and the frequency. It’s not as severe and it’s so much less now than before.”
- “They did seem to lessen the number of times per day, um, that I had probably the most common symptom, which was a burping sensation and acid in my throat and back of my mouth. I did not see as often, either on a daily or during the course of a week.”

Proposed responder definition

Proportion of patients experiencing at least 3 more symptom free* days on average per week compared to baseline (entire treatment period).

Dimension	Intensity defined as symptom free	Change in % symptom free days (at least)	n	PGIC n(%)			
				Unchanged n=174	Small improvement n=68	Moderate improvement n=86	Large improvement n=103
All items	<=Very mild	14.29% (1/7)	161	33 (19%)	23 (34%)	38 (44%)	67 (65%)
		28.57% (2/7)	116	21 (12%)	14 (21%)	27 (31%)	54 (52%)
		42.86% (3/7)	67	11 (6%)	8 (12%)	8 (9%)	40 (39%)
		57.14% (4/7)	39	4 (2%)	4 (6%)	4 (5%)	27 (26%)
		71.43% (5/7)	17	3 (2%)	3 (4%)	1 (1%)	10 (10%)
		85.71% (6/7)	3	0 (0%)	1 (1%)	0 (0%)	2 (2%)
	<=Mild	14.29% (1/7)	234	65 (37%)	38 (56%)	46 (53%)	85 (83%)
		28.57% (2/7)	174	42 (24%)	25 (37%)	35 (41%)	72 (70%)
		42.86% (3/7)	113	25 (14%)	15 (22%)	21 (24%)	52 (50%)
		57.14% (4/7)	71	13 (7%)	8 (12%)	13 (15%)	37 (36%)
		71.43% (5/7)	39	8 (5%)	3 (4%)	3 (3%)	25 (24%)
		85.71% (6/7)	16	3 (2%)	0 (0%)	1 (1%)	12 (12%)

Challenges that complicated definition of meaningful change



- It had to be right and robust as it was the primary endpoint.
- Changing regulatory landscape.
- We went for a GERD symptoms claim.
- Target patient population was defined by lack of response.
- Several key symptoms that had to be taken into consideration.
- Severity and frequency.

How did the qualitative interviews increase the Understanding of Meaningful Change from a Patient Perspective?



- Challenged the previous notion that patients with GERD only expect symptom resolution.
- Increased the understanding of patient definition of “control” and “sufficient control”.
- Enabled triangulation approach to define meaningful change.
- Enabled patients to explain more what the quantitative change in the study meant to them.
- Provided foundation for any similar compound for the future



Using vignettes to establish thresholds for status and change

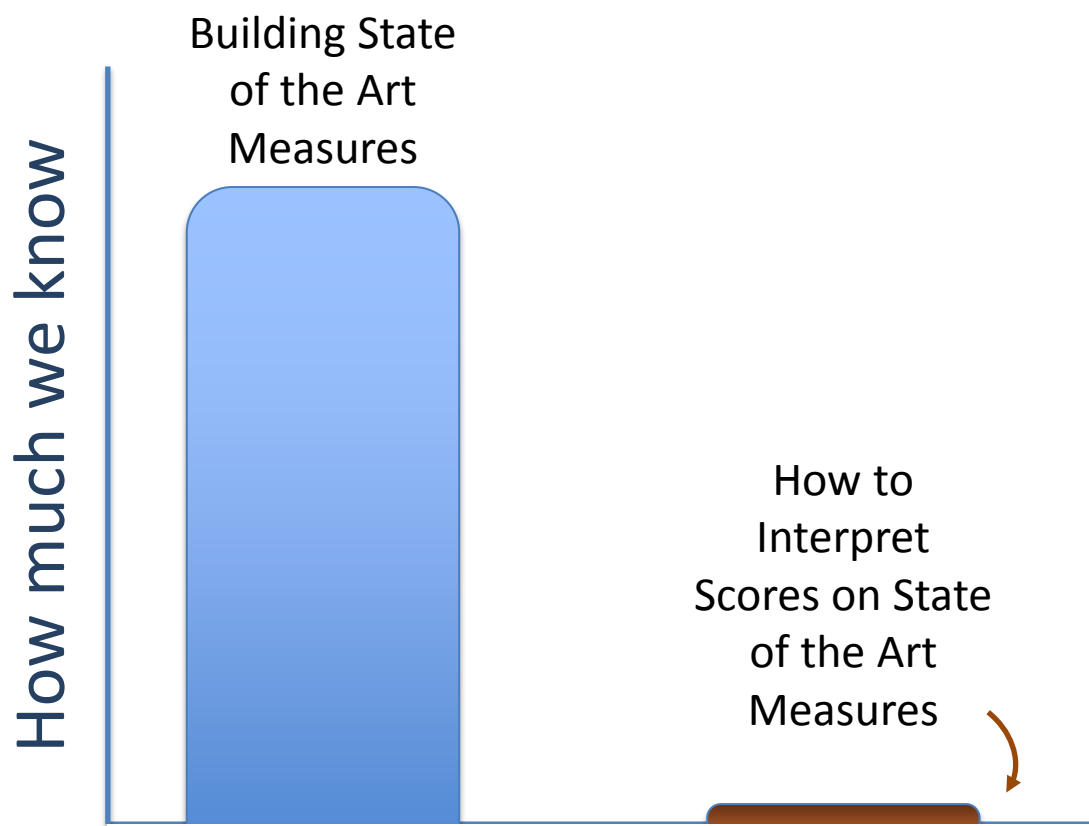
Karon F. Cook, PhD
Northwestern University, Chicago

Cheryl Coon, PhD
Michael Kallen, PhD

THRESHOLDS
THRESHOLDS

- Vignettes to quantify levels of symptoms and outcomes: PROMIS and Neuro-QOL case studies
- Vignettes to quantify change: Case study in MS fatigue
- Keeping it real: extending the method by attending to context.
- ***In conclusion: What insight was gained using this approach versus traditional methods?***

Background



Review Article

Cut Points on 0–10 Numeric Rating Scales for Symptoms Included in the Edmonton Symptom Assessment Scale in Cancer Patients: A Systematic Review

Wendy H. Oldenmenger, RN, PhD, Pleun J. de Raaf, MD, Cora de Klerk, PhD, and Carin C.D. van der Rijt, MD, PhD

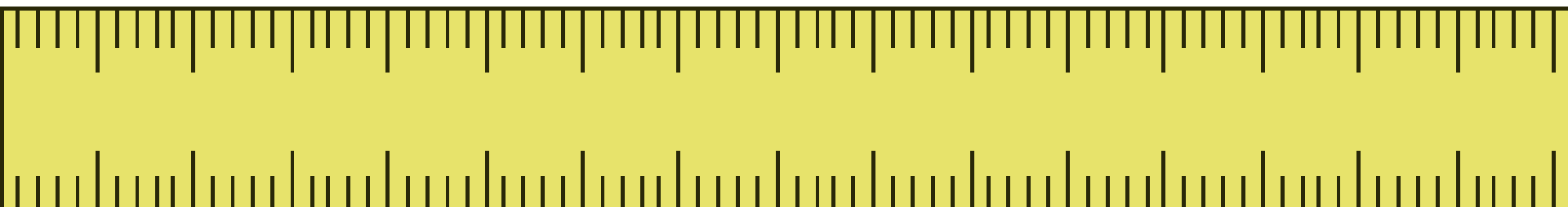
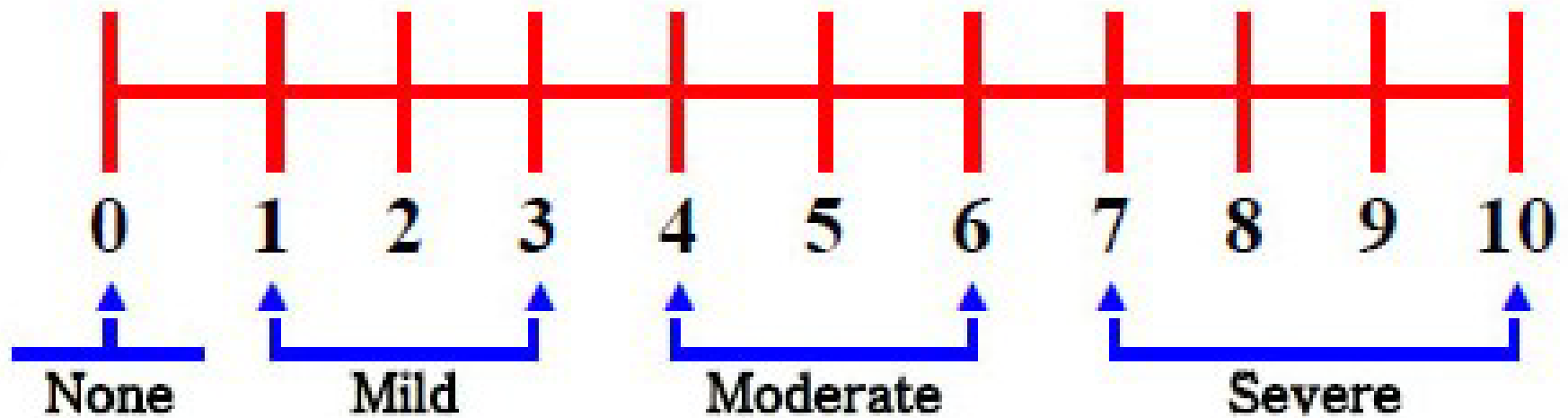
Department of Medical Oncology (W.H.O., P.J.d.R., C.C.D.v.d.R.), Erasmus MC Daniel den Hoed Cancer Center and Department of Medical Psychology and Psychotherapy (C.d.K.), Erasmus MC, Rotterdam, The Netherlands

Abstract

Context. To improve the management of cancer-related symptoms, systematic screening is necessary, often performed by using 0–10 numeric rating scales. Cut points are used to determine if scores represent clinically relevant burden.

Objectives. The aim of this systematic review was to explore the evidence on cut points for the symptoms of the Edmonton Symptom Assessment Scale.

Methods. Relevant literature was searched in PubMed, CINAHL[®], Embase, and PsycINFO[®]. We defined a cut point as the lower bound of the scores representing





Educational and
Psychological
Testing



Health
Outcomes
Assessment

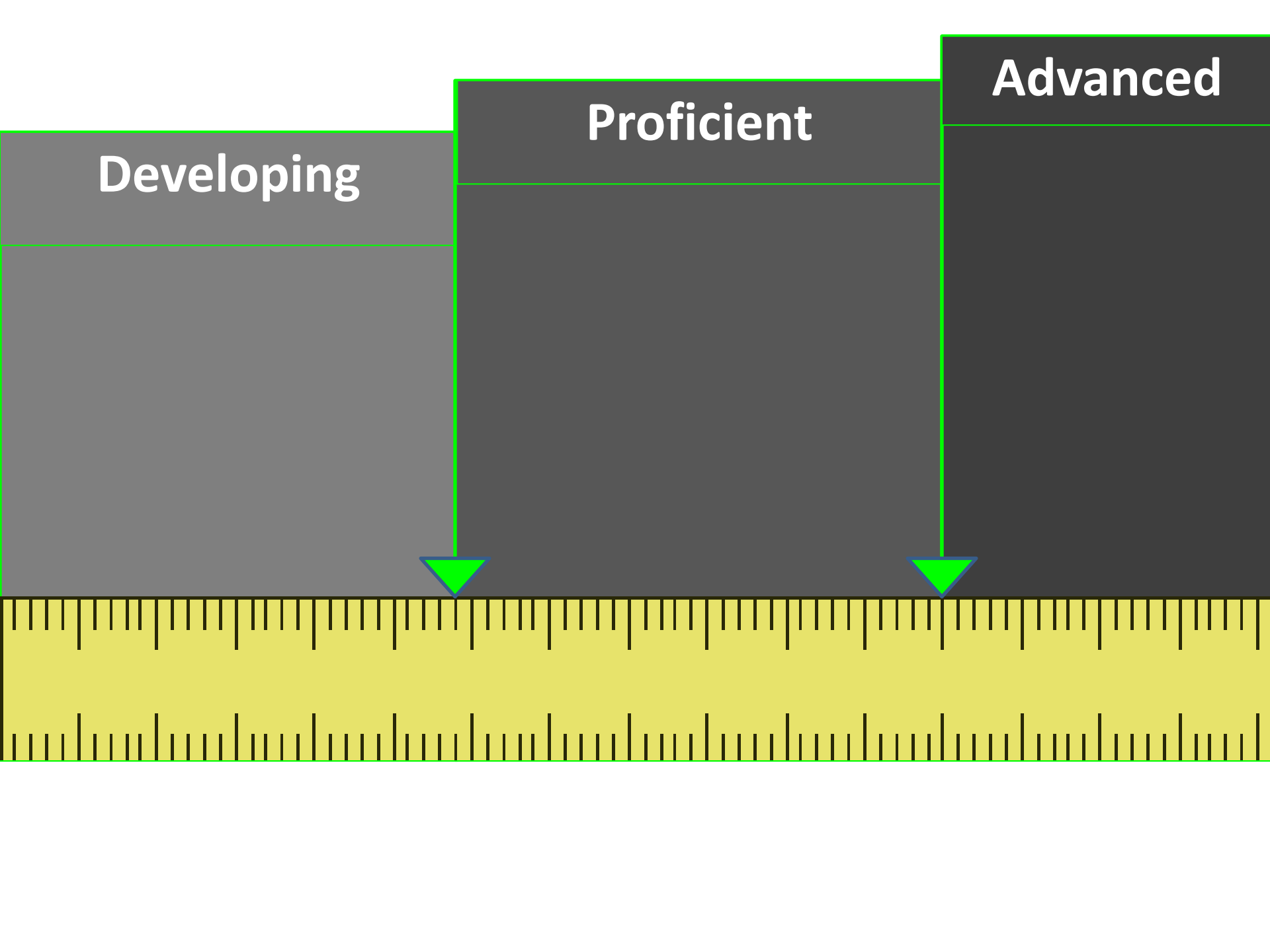
Scaling latent
constructs
Item response theory
Item banking
modeling
self-report



Educational Standard Setting

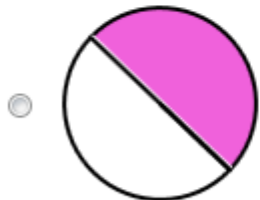
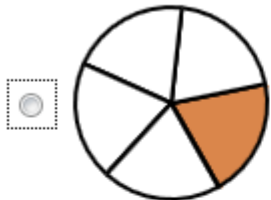
Educational standard setting is, “the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance” (p. 100).

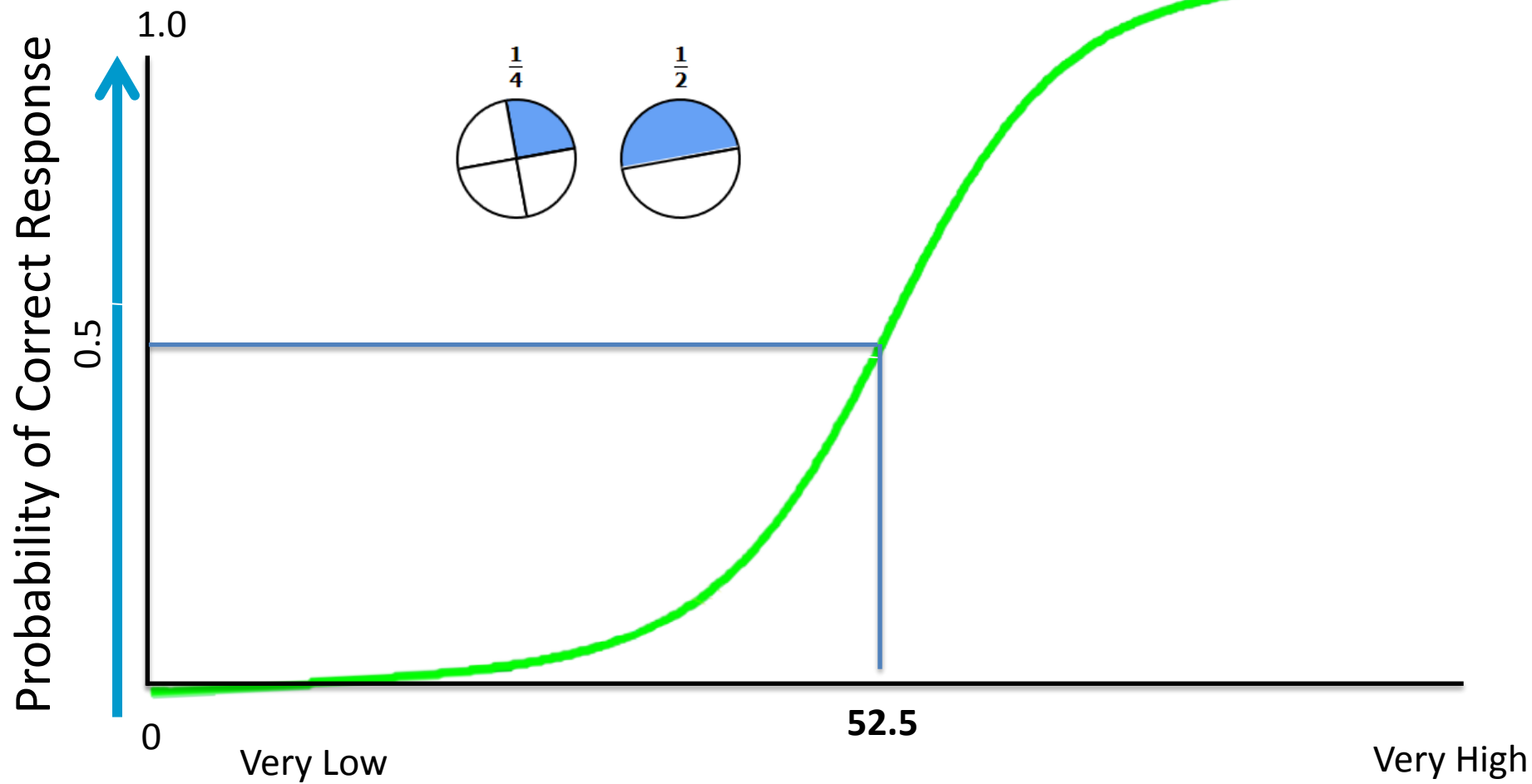
Cizek GJ. Reconsidering standards and criteria. *Journal of Educational Measurement*. 1993;30(2):93-106.

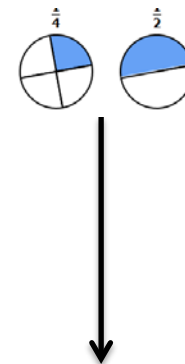
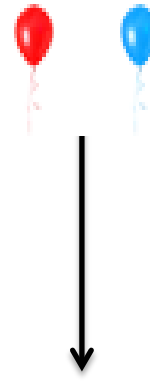
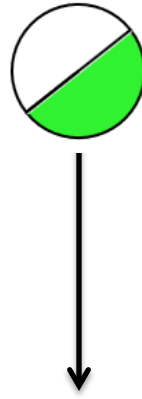
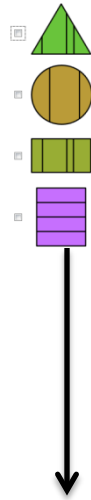




Which shape shows the fraction $\frac{1}{4}$?







42.5

37.5

42.5

47.5

52.5

57.5

Which figure shows halves?

Select all the pictures that show equal parts.

What fraction of the shape is green?

What fraction of the balloons are red?

Which fraction is greater?

Which shape shows the fraction $\frac{1}{4}$?

☒ $\frac{1}{3}$

☐ $\frac{1}{2}$

☐ $\frac{1}{4}$

☐ $\frac{1}{5}$

☐ $\frac{1}{5}$

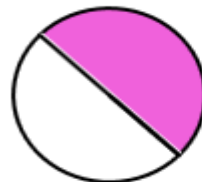
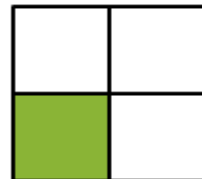
☐ $\frac{1}{3}$

☐ $\frac{1}{2}$

☐ $\frac{1}{4}$

☐ $\frac{1}{4}$

☒ $\frac{1}{2}$



Developing

Proficient

Advanced

50.0

60.0



42.5

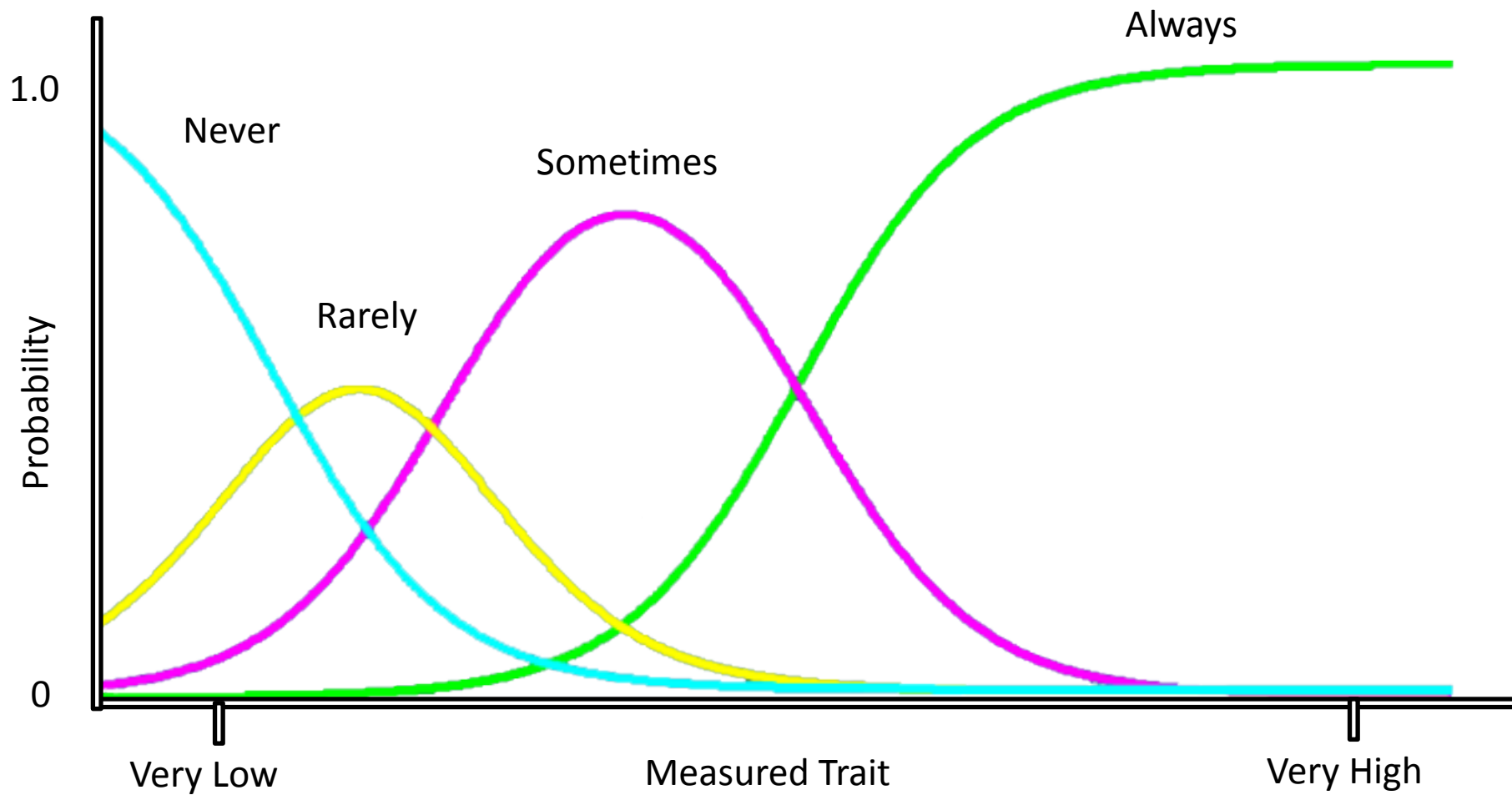
47.5

52.5

57.5

62.5

67.5

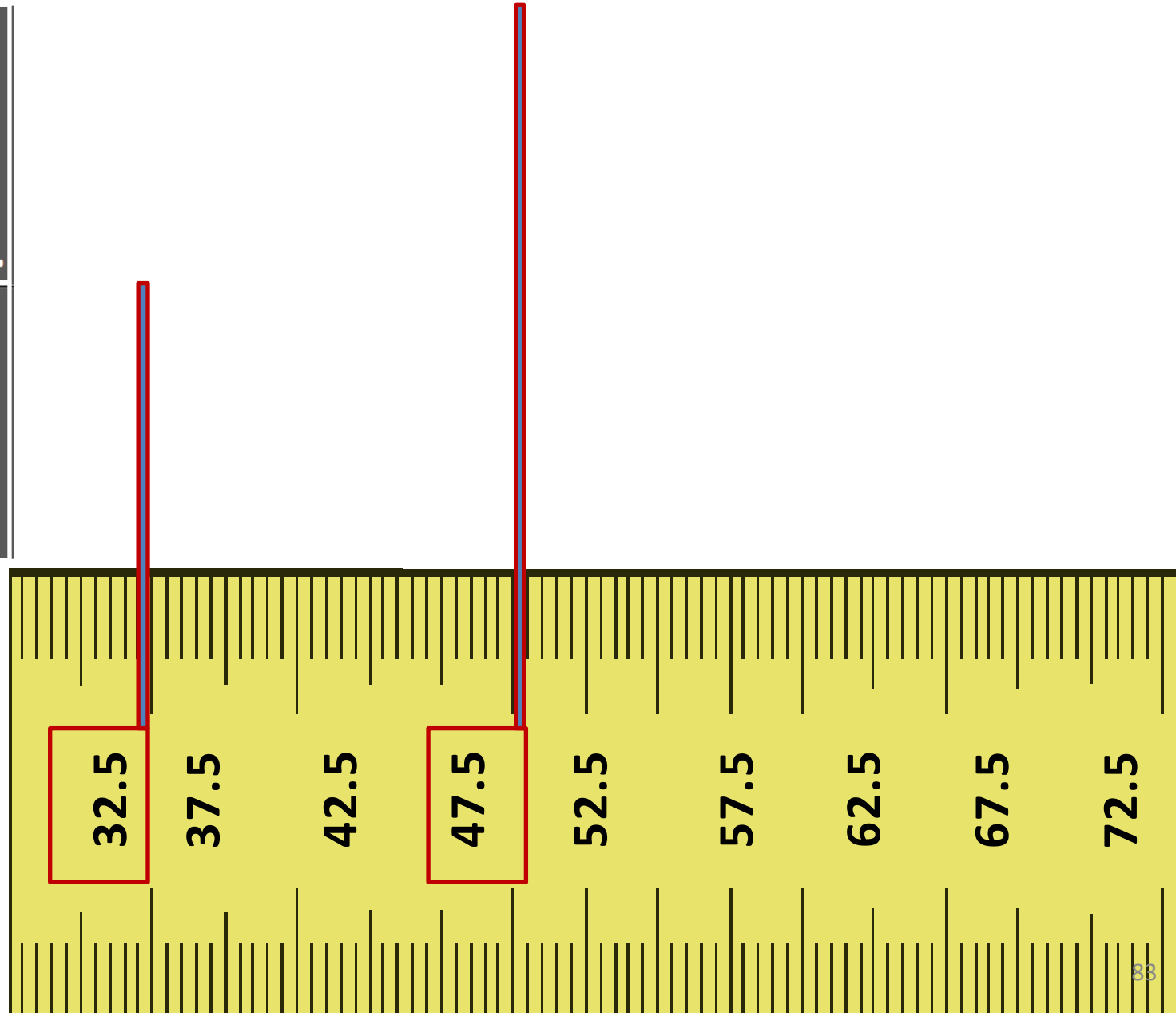


Most Likely Response?

I was too
tired to eat.

I felt tired.

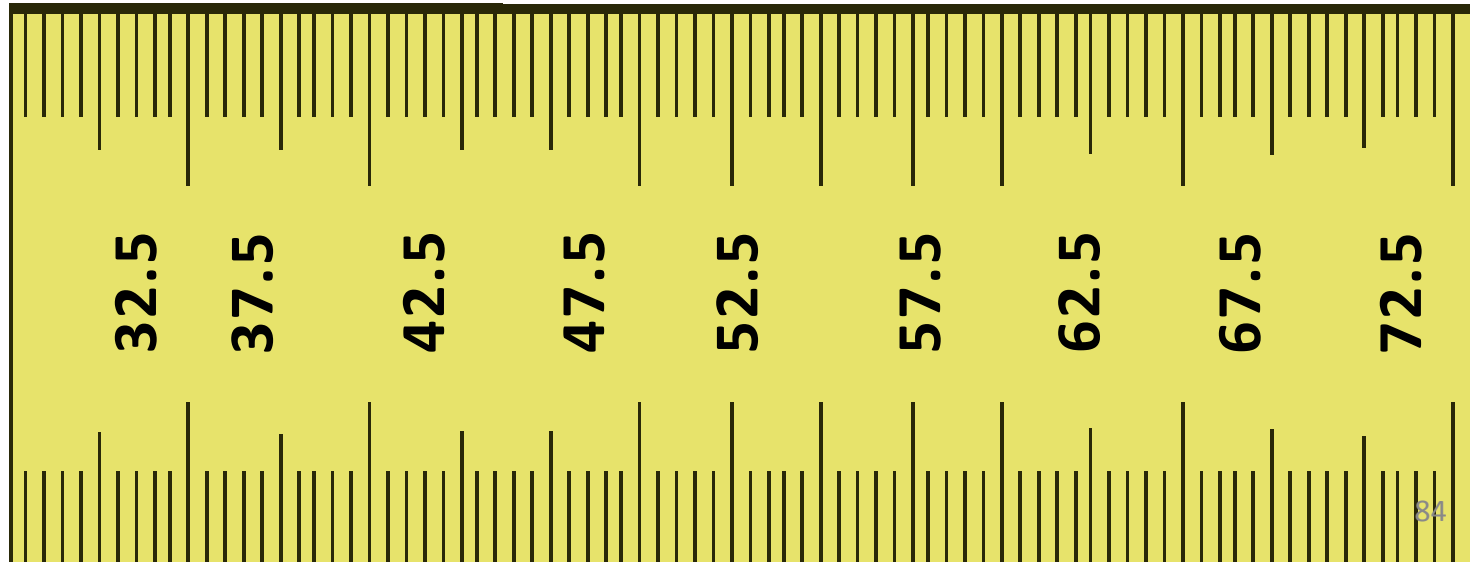
It
depends.



Most Likely Response?

I was too tired to eat.	never	never	never	never
I felt tired.	never			

It
depends.



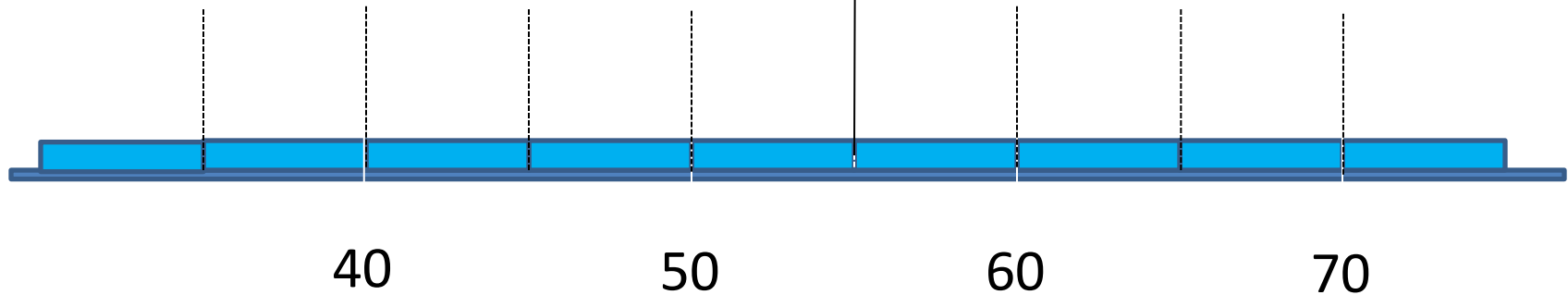
Items chosen to maximize
variation in response.

I (rarely) felt helpless.

I (sometimes) had mood swings.

I (never) felt worthless.

I (sometimes) felt lonely.



I (sometimes) felt worthless.

**I (sometimes) withdrew from
people.**

I (often) felt like crying.

I (often) felt depressed.



Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment

David Cella • Seung Choi • Sofia Garcia • Karon F. Cook •
Sarah Rosenbloom • Jin-Shei Lai • Donna Surges Tatum •
Richard Gershon

Clinical cut score recommendations
mapped onto distribution of actual
patient scores

Qual Life Res (2015) 24:575–589
DOI 10.1007/s11136-014-0790-9

Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers

Karon F. Cook • David E. Victorson •
David Cella • Benjamin D. Schalet •
Deborah Miller

Accepted: 14 August 2014 / Published online: 23 August 2014
© Springer International Publishing Switzerland 2014

Abstract

Purpose To establish clinically relevant classifications of health outcome scores for four Neuro-QOL measures (lower extremity function, upper extremity function, fatigue, and sleep disturbance).

Methods We employed a modified educational standard-setting methodology to identify cut-scores for symptom

Patients and Clinicians

Conclusions The modified bookmarking method is effective for defining thresholds for symptom severity based on self-reported outcome scores and consensus judgments. Derived cut-scores and severity levels provide an interpretative context for Neuro-QOL scores. Future studies should explore whether these findings can be replicated and evaluate the validity of the classifications.



Change

Vignettes for Estimating Important Change

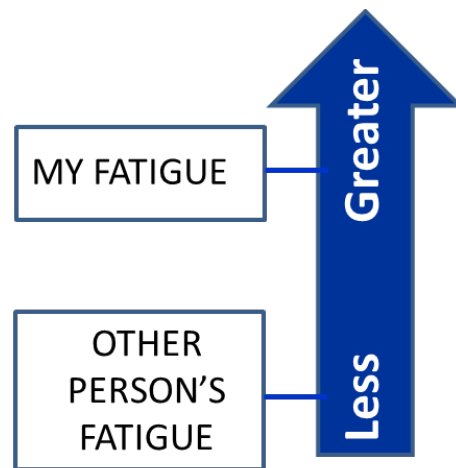
- N= 500 participants with MS
- Internet panel
- Branched exercise

Internet Vignette Exercise

- Collect Demographics and Clinical Information
- Administer the Neuro-QoL Fatigue Short Form
- Branch respondents into 8 Groups based on scores
- Ask patients to rate the severity of each vignette compared to their own level of fatigue (i.e., better, same, worse)
- Ask respondents to indicate whether difference would “make a difference in my daily life”

In PART B, you will

- Look at the fatigue reports of 7 people who have MS
- Compare each person's fatigue to your own fatigue. For example, your fatigue might be greater.



- Or, you might decide your fatigue is the SAME or LESS than that other person's.

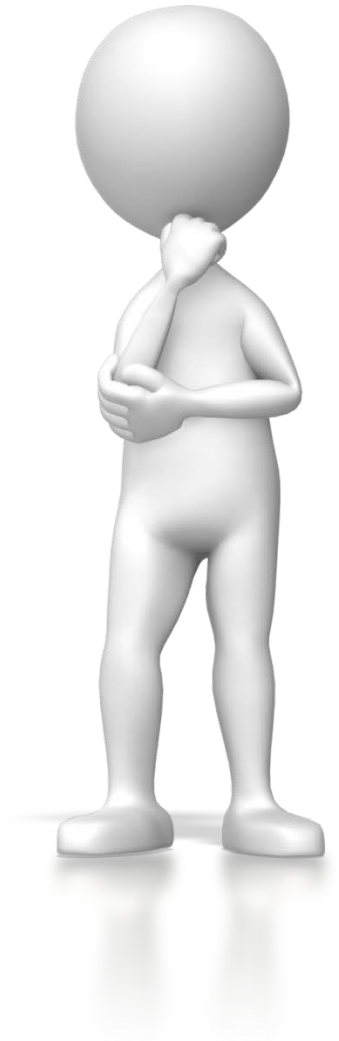
If you decide your fatigue is DIFFERENT from the other person's, you will then

- Consider what it would be like to have this person's fatigue, and
- Decide if the difference would matter to you in your daily life.



Depending on your own fatigue, you may decide that none, some, or all of these people have more, less, or the same amount fatigue.

There are no “right” answers—
just your own thoughtful judgments.



[T Score = 58]

This is what Ms. Anderson said about his fatigue over the last 7 days. She reported that she:

- **sometimes felt weak all over.**
- **often had to limit social activity because she was tired.**
- **sometimes had trouble starting things because she was too tired.**
- **often was too tired to take a short walk.**
- **often had trouble finishing things because she was too tired.**

[Q50]Compared to Ms. Anderson's , has YOUR FATIGUE been:

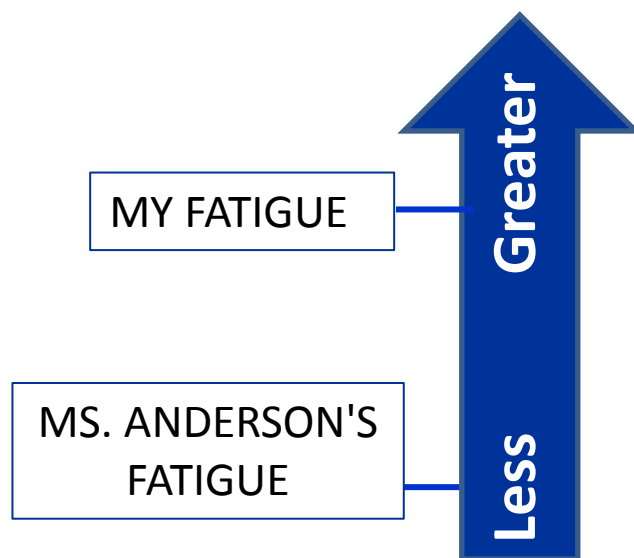
- ☐ Greater than Ms. Anderson's
- ☐ The same as Ms. Anderson's
- ☐ Less than Ms. Anderson's

The logo consists of the word "SCREEN" in a large, blue, sans-serif font, with the word "SHOTS" in a smaller, black, sans-serif font directly beneath it. The entire logo is set against a light blue background that resembles a document or folder icon.

You said YOUR FATIGUE over the past week was Greater than MS. ANDERSON'S FATIGUE.

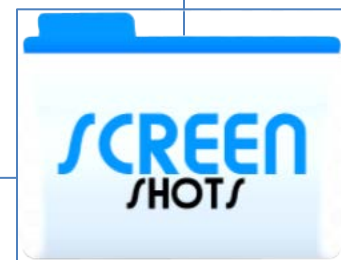
[Q51] If your fatigue IMPROVED to Ms. Anderson's level, would it make a difference in your daily life?

- ☐ It wouldn't really make a difference in my daily life.
- ☐ It would make a difference in my daily life (things I do day-to-day would be easier).



This is what Ms. Anderson said about his fatigue over the last 7 days. She reported that she:

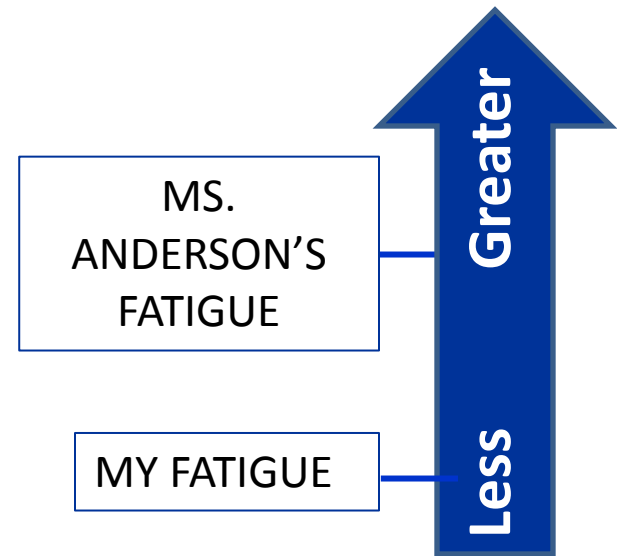
- sometimes felt weak all over.
- often or always had to limit social activity because she was tired.
- sometimes had trouble starting things because she was too tired.
- often was too tired to take a short walk.
- often had trouble finishing things because she was too tired.



You said YOUR FATIGUE over the past week was LESS than MS. ANDERSON'S FATIGUE.

[Q52] If your fatigue WORSENE^D to Ms. Anderson's level, would it make a difference in your daily life?

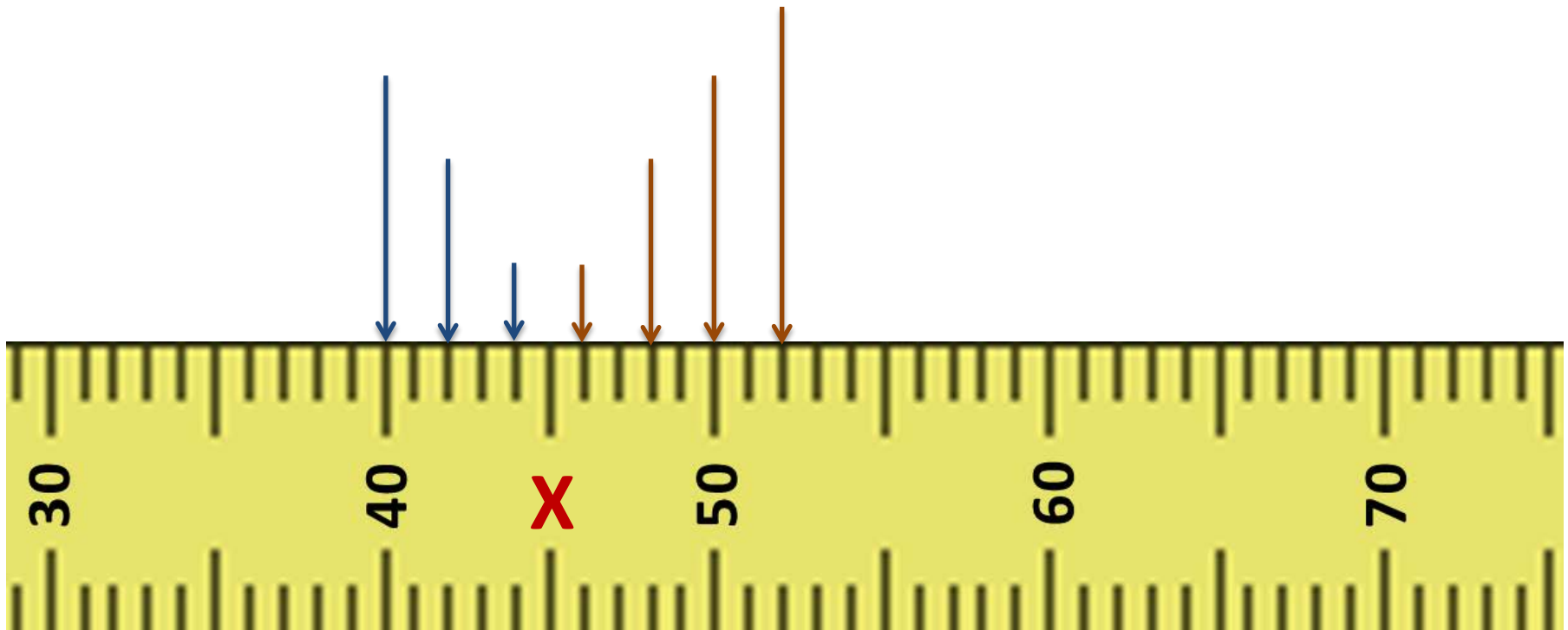
- ☐ It wouldn't really make a difference in my daily life.
- ☐ It would make a difference in my daily life (many of the things I do day-to-day would be harder).



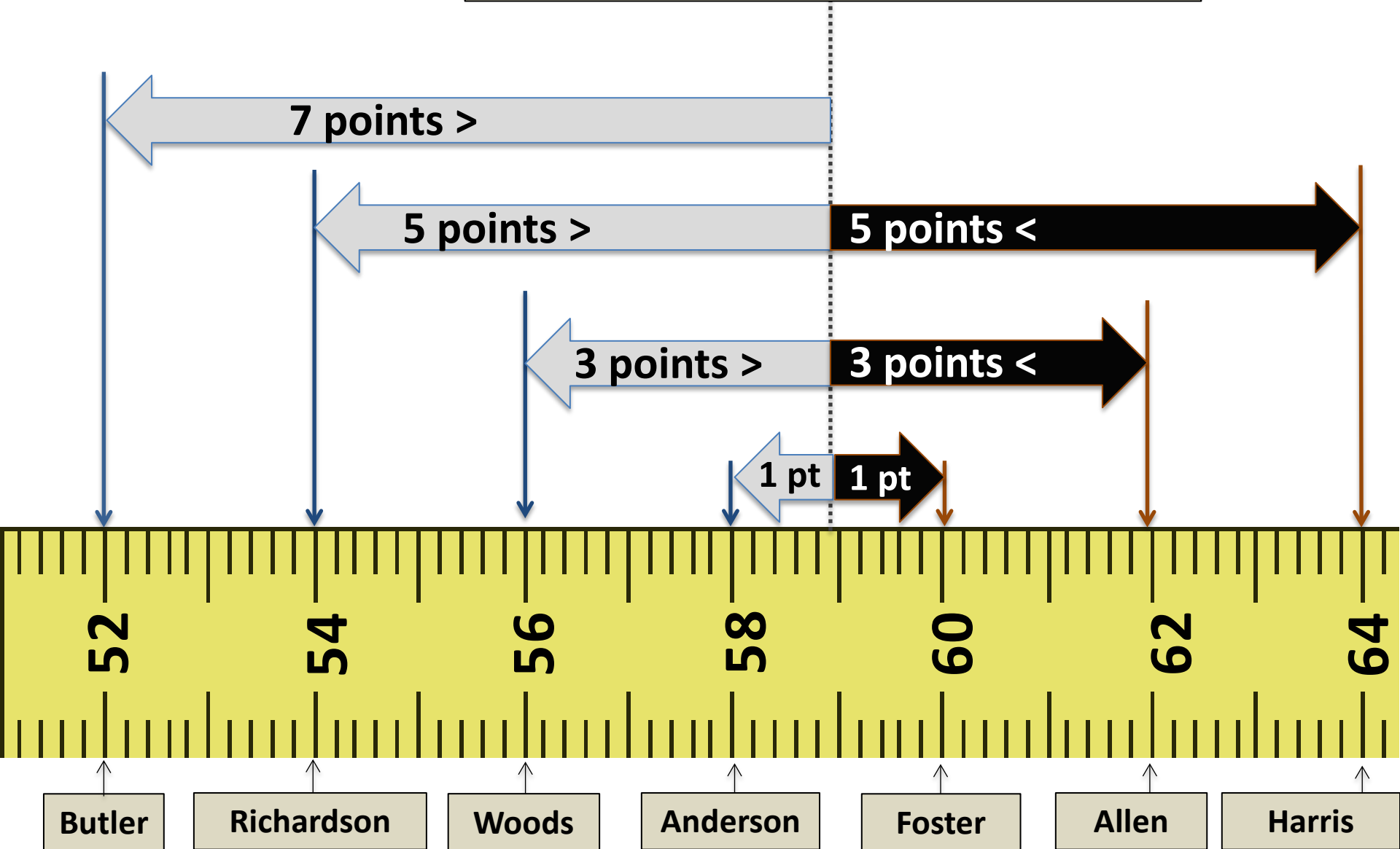
This is what Ms. Anderson said about his fatigue over the last 7 days. She reported that she:

- sometimes felt weak all over.
- often or always had to limit social activity because she was tired.
- sometimes had trouble starting things because she was too tired.
- often was too tired to take a short walk.
- often had trouble finishing things because she was too tired.





N=30 Raw Score = 31; T = 58.8



Devil in the Details



Please tell us how much **confidence** you have that your answers to these items reflect how you would judge actual changes in your fatigue:

- 1 = Not at all confident
- 2 = A little bit confident
- 3 = Moderately Confident
- 4 = Very Confident

Judgment Validity

Out of Range Judgments

- Perfect judgments would never be expected
 - identify vignettes \geq current score as important improvement
 - endorse \leq than their current score as important worsening
- Calculated proportion of these “out of range” judgments for improvement and for worsening

Judgment Validity

Out of Range Judgments

- Correlated number of out of range judgments with
 - Education
 - Fatigue
 - Confidence

- Confidence Ratings
 - Mean = 3.3 (SD = 0.71)
 - Between moderately and very confident

RESULTS: Judgment Validity

- Correlation between # of out of range judgments and

Percent	Correlation
Education	-.080
Confidence in Ratings	.074
Fatigue Score	.209

RESULTS

Out of Range Judgments

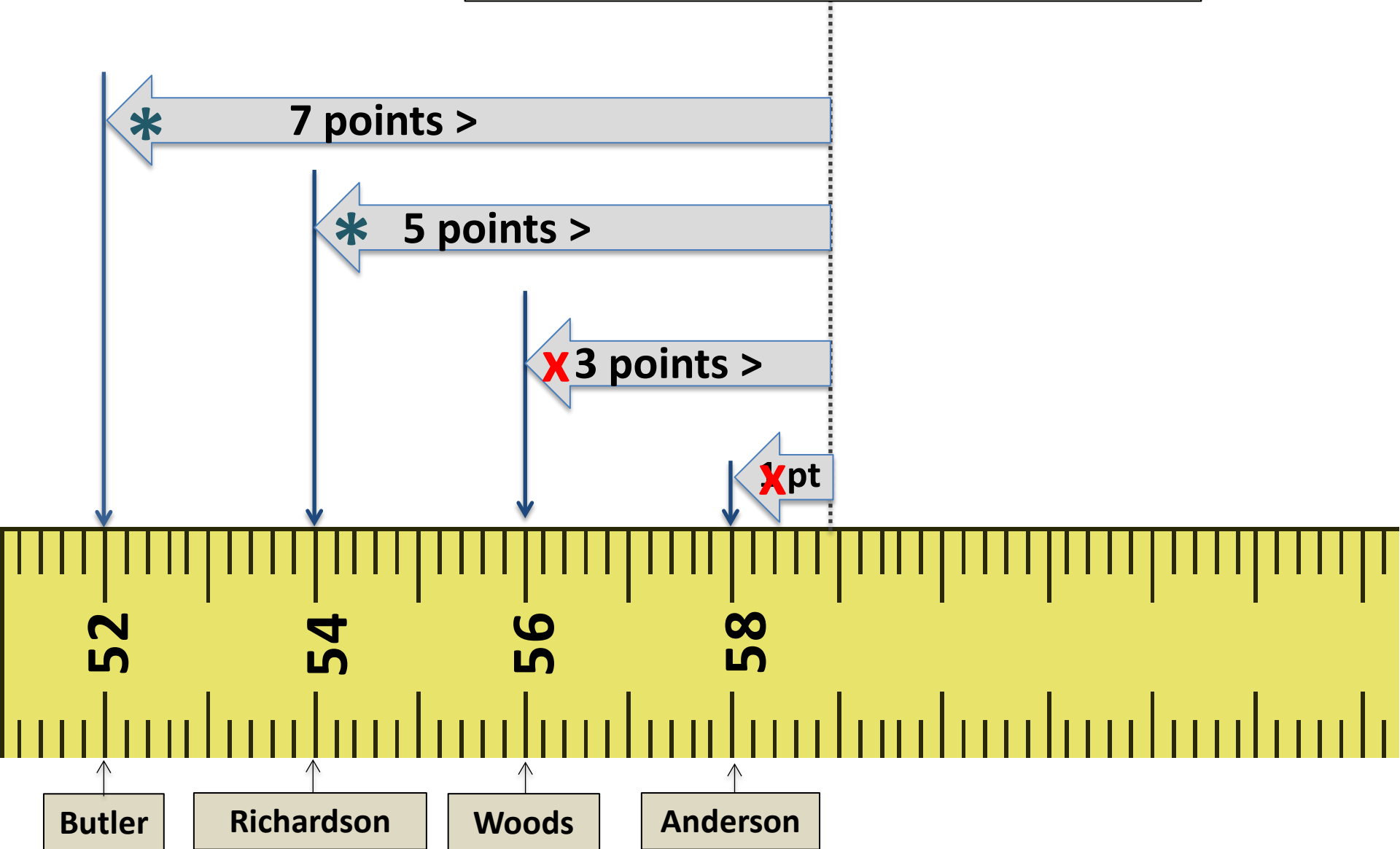
- Person level

# out of range	Frequency	Percent	Cumulative Percent
0.00	212	49.9	49.9
1.00	134	31.5	81.4
2.00	47	11.1	92.5
3.00	22	5.2	97.6
4.00	8	1.9	99.5
5.00	1	.2	99.8
6.00	1	.2	100.0

Analyses to Estimate Thresholds for Interpreting Change

- Mean based on following
 - for each participant identified the shortest distance endorsed as “enough to make a difference in my daily life.”
 - calculation was repeated after dropping “out of range” responses
- Minimum distance endorsed by respondent as meaningful change

N=30 Raw Sore = 31; T = 58.8



Analyses to Estimate Thresholds for Interpreting Change

- Threshold that identifies 50% of those endorsing distance as meaningful change.
- Sensitivity, Specificity, Negative predictive value (NPV), Positive predictive value (PPV)

Thresholds for Interpreting Change



- Identified the minimum distance endorsed as “meaningful” by each respondent
- Calculated the mean, range, and SD of these minimum distances across persons

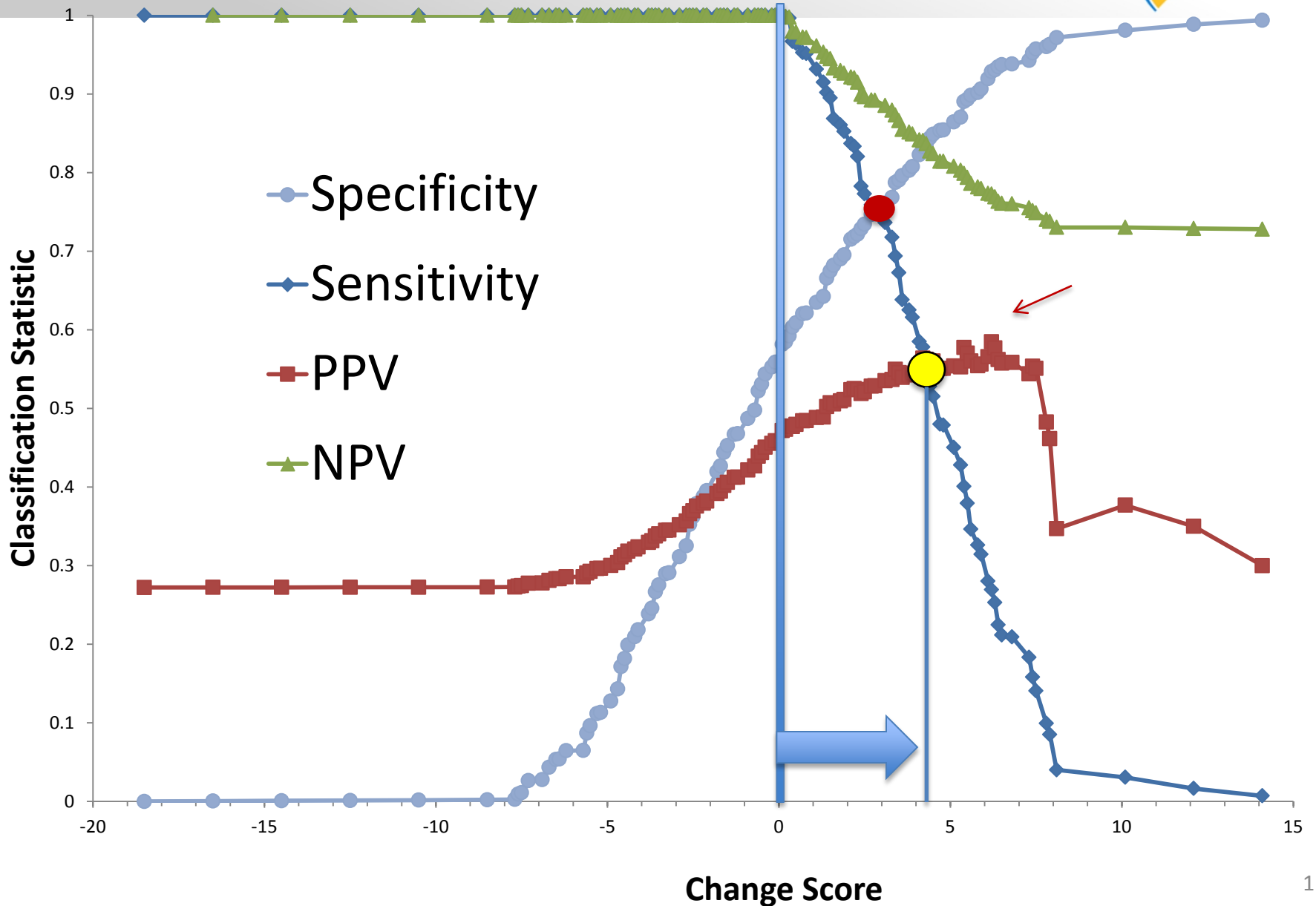
Thresholds for Interpreting Change

At what threshold are 50% of those endorsing important change captured by the threshold

Improvement	Worsening
4.5	-4.2

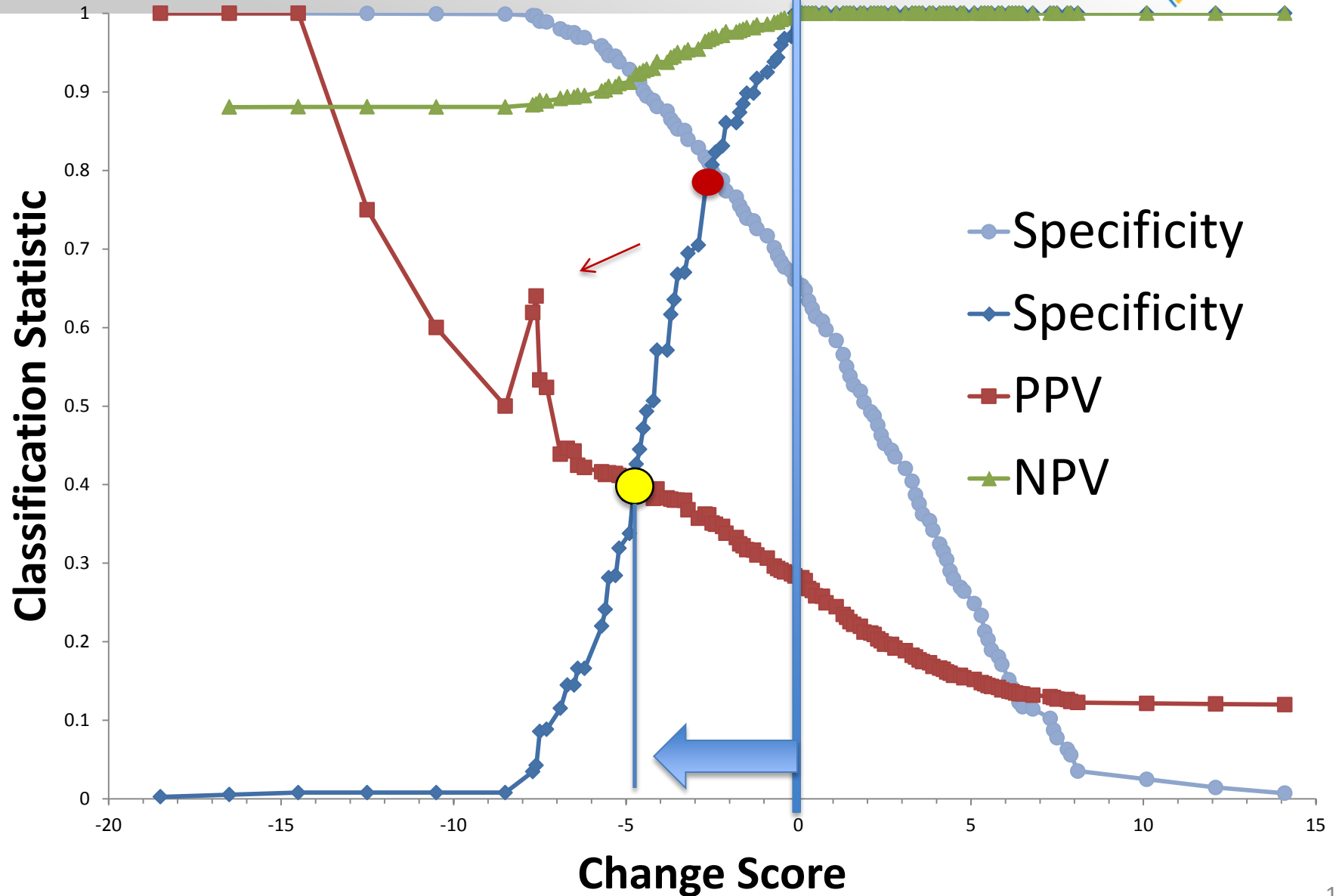
Meaningful Improvement

Classification Statistics Across Range of Change Scores



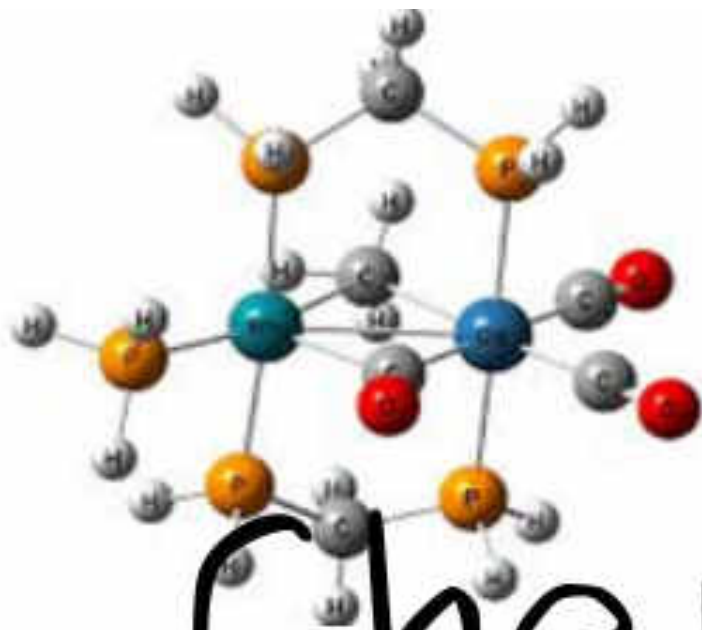
Meaningful Worsening

Classification Statistics Across Range of Change Scores



- Context is everything
- Judging Vignettes Is a Challenging Task
 - Could refine methods using cognitive interviews
 - Conduct exercise as interview
- Some people said no amount of change was different from their level of fatigue.
- More Fundamentally
 - What do patients consider when judging change

- Sources of Vignette error
 - No real gold standard; Self-rating isn't a true gold standard
 - Fatigue scores themselves have error
 - Error in interpreting vignettes
- Sources of error in other methods
 - Retrospective recollections subject to bias
 - Every judgment at every point of time has error



Chemistry



How do vignettes enhance our ability to determine meaningful score change?



- Another strategy for “getting the truth surrounded”
- Allows scores on multiple item measures to be used to estimate thresholds for status and change.
- Another tool

“The process of writing vignettes is a process of discovering what you know

Rebecca Yamin, "Through Many Eyes." *Reconsidering Archaeological Fieldwork*, ed. by Hannah Cobb et al. (Springer, 2012).

Moderator

- *Cheryl D. Coon, PhD* – Principal, Outcometrix

Presenters

- *Mona Martin, RN, MPA* – Executive Director, Health Research Associates
- *Allison Martin Nguyen, MS* – Sr. Principal Scientist, Patient Reported Outcomes & Study Endpoints Group, Merck & Co., Inc.
- *Katarina Halling, MSc* – Global Head, Patient Reported Outcomes, AstraZeneca and Co-Director, PRO Consortium
- *Karon Frances Cook, PhD* – Research Professor in Medical Social Sciences, Northwestern University Feinberg School of Medicine

Panelists

- *Wen-Hung Chen, PhD* – Reviewer, COA Staff, OND, CDER, FDA
- *Tara Symonds, PhD* – Strategic Lead, Clinical Outcomes Assessments and Partner, Clinical Outcome Solutions
- *Kathleen (Kathy) Wywrich, PhD* – Executive Director, Center of Excellence for Outcomes Research, Evidera