Applying Methodologies for Estimating Meaningful Within-person Change Thresholds: Considerations and Alternative Approaches

14th Annual
Patient-Reported Outcome Consortium Workshop

April 19-20, 2023 • 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 or the Critical Path Institute.
- 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 Objectives

- Highlight industry perspective of operationalizing anchor selection and inclusion in clinical trials
- Provide reflections from the Rheumatoid Arthritis Working Group's experience estimating meaningful within-person change for the *PROMIS® Fatigue 10a*
- Introduce an FDA-funded Shareware project that leverages Idioscale Judgment methodology to derive meaningful change thresholds
- Describe the DIA Meaningful Change Working Group and explore challenges associated with meaningful change threshold estimation when considering endpoints derived from sensor data

Session Participants

Moderator

- *Rebecca (Becks) M. Speck, PhD, MPH* – Clinical Outcome Assessment Scientist, Eli Lilly and Company

Presenters

- *Elizabeth (Nicki) Bush, MHS* – Senior Director, Endpoints and Measurement Strategy, Janssen Pharmaceutical Companies of Johnson & Johnson
- *Devin Peipert, PhD* – Assistant Professor of Medical Social Sciences, Northwestern University
- *Karon Cook, PhD* – Research Professor (Retired), Feinberg School of Medicine, Northwestern University
- *Bill Byrom, PhD* – Vice President, Product Intelligence and Positioning, and Principal, eCOA Science, Signant Health, UK

Additional Panelists

- *Selena Daniels, PharmD, PhD* – Clinical Outcome Assessment Team Leader, Division of Clinical Outcome Assessment, Office of Drug Evaluation Sciences, Office of New Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration
- *Monica Morell, PhD* – Patient-Focused Statistical Support Reviewer, Division of Biometrics III, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Clinical Outcome Assessment Score Interpretation: Anchors and Anchor-based Methods

Elizabeth (Nicki) Bush – Senior Director, Endpoints and Measurement Strategy

Janssen Pharmaceutical Companies of Johnson & Johnson

Topics Covered

- Brief definition of anchor-based methods
- Selecting or developing an anchor
- Industry perspective

Anchor-based Methods

Purpose

- Anchor-based methods are frequently used to aid clinical outcome assessment (COA) score interpretation in terms of clinical meaningfulness

Methods

- Identify one or more easily-interpretable external criteria (anchors) related to the concept of interest (COI) of the COA itself to assess change in that concept
- Explore associations between COA score changes and anchor-based classifications and score changes indicative of meaningful change
- Combine findings with other information (e.g., cumulative distribution function curves) to establish interpretation guidelines (e.g., range of meaningful change thresholds)

Attributes of "Good" Anchors

- Plainly understood, easier to interpret than the COA itself
- "Sufficiently associated" with the COA or COA endpoint
  - "Sufficiently associated" has no universal, quantitative definition, though some guidelines have been suggested (some references on last slide)
- Based on the same recall period as the COA-supported endpoint
- Administered at the same timepoints as the COA of interest
- Administered (when feasible) just after the COA of interest (with the exception of performance outcome assessments where it should be administered right before)
- One among many (or, at least, a few) (i.e., use multiple anchors!)
- NOT an afterthought (consider including in cognitive interviews)

Anchor Archetypes

Please choose the response below that best describes the overall severity of your <symptoms of [Condition]> < [over the past 24 hours/7 days/ at this time]>. (Select one response)

- ❑ No symptoms
- ❑ Mild
- ❑ Moderate
- ❑ Severe
- ❑ Very severe

PGIS

Please choose the response below that best describes the overall change in your <[Condition] symptoms> < [compared to X days ago]/ [compared to when you started the study]>. (Select one response)

- ❑ Much better
- ❑ A little better
- ❑ No change
- ❑ A little worse
- ❑ Much worse

PGIC

Meaningful Change on an Anchor

- Evidence of what constitutes a “meaningful change” on an anchor is needed—so that we can use it to indicate what constitutes a meaningful change on a COA score.
  - Qualitative (e.g., in-trial interviews)
  - Empirical data (if items asking about meaningfulness of change were included)
  - Literature (ensure fidelity to concept of interest and context of use)

Anchor-based Methods Are Not Perfect

Conclusions may be influenced by

- Magnitude of correlation between anchor(s) and COA
- Time (recall bias, within-patient changes in perspective)
- Sample size
- Variability in sample distribution and change scores
- Variability in rate of response to treatment
- Floor and ceiling effects
- Triangulation methods

A lot of ongoing work exploring how to identify and address these issues!

Additional Methods (Also Not Perfect)

- New methods for interpreting COA scores emerging, some with “anchor-based” characteristics, some with “distribution-based” characteristics, and some with both.
- No need to select one approach; methods can be complementary
- A few examples
  - Qualitative interviews
  - Idioscale Judgment
  - Reliable Change Index
  - Likely Change Index
  - Mediation Analyses
  - Item Response Theory-based

Operational Considerations

Any data collection tool added to a trial has implications, and anchor items or scales are no exception

- Translations (time, cost)
- Programming, testing (eCOA)
- Data standards and management
- Length of study protocols (limited real estate)
- Administrative (version tracking and control, duplication)

Operational Considerations

- Hypothetical Phase 2 Trial evaluating a new treatment for heart failure in adults
- Randomized Clinical Trial
- Duration 52 weeks
- Time between clinical visits varies (2 weeks to 12 weeks)
- Schedule of Assessments (COAs only)
  - New York Heart Association (NYHA) Functional Classification (ClinRO measure)
  - 6-Minute Walk Test (6MWT) (PerfO assessment)
  - Kansas City Cardiomyopathy Questionnaire (KCCQ) (PRO measure)
  - SF-36 (PRO measure)
  - EQ-5D-5L (PRO measure)

Hypothetical Trial: Objectives (subset)

Primary

- Compare reduction of rate of composite endpoint: cardiovascular death and total number of heart failure-related hospitalizations between experimental treatment (ET) and standard of care (SOC) over course of 52 weeks

Secondary

- Compare change in heart failure severity from baseline to 6 months between ET and SOC as assessed by NYHA Functional Classification
- Compare change in heart failure symptoms from baseline to 6 months between ET and SOC as assessed by KCCQ Total Symptom Score (TSS)
- Compare change in heart failure-related daily functioning from baseline to 6 months between ET and SOC as assessed by KCCQ Physical Limitations Domain
- Compare change in functional capacity and ability from baseline to 6 months between ET and SOC as assessed by 6-minute walk test (6MWT)
- Compare change in physical function from baseline to 6 months between ET and SOC as assessed by SF-36

Exploratory

- Compare change in quality of life (QOL) from baseline to 6 months between ET and SOC as assessed by EQ-5D-5L
- Compare changes in severity of shortness of breath (SOB) between ET and SOC as assessed by KCCQ SOB item
- Explore impact of SOB on functional capacity from baseline to 6 months between ET and SOC as assessed by association of change in KCCQ SOB and 6MWT

Hypothetical Visit 3 (Baseline+14 weeks)

Arrive at clinical site, begin with PRO measures in the waiting room

Starting with the KCCQ, including

3. Over the <u>past 2 weeks</u>, on average, how many times has **fatigue** limited your ability to do what you wanted?

All of the time	Several times a day	At least once a day	3 or more times a week but not every day	1-2 times a week	Less than once a week	Never over the past 2 weeks
❑	❑	❑	❑	❑	❑	❑

4. Over the <u>past 2 weeks</u>, on average, how many times has **shortness of breath** limited your ability to do what you wanted?

All of the time	Several times a day	At least once a day	3 or more times a week but not every day	1-2 times a week	Less than once a week	Never over the past 2 weeks
❑	❑	❑	❑	❑	❑	❑

Hypothetical Visit 3 (Baseline+14 weeks)

Then onto the SF-36, which includes

During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of your physical health**?

	Yes	No
13. Cut down the **amount of time** you spent on work or other activities	○ 1	○ 2
14. **Accomplished less** than you would like	○ 1	○ 2
15. Were limited in the **kind** of work or other activities	○ 1	○ 2
16. Had **difficulty** performing the work or other activities (for example, it took extra effort)	○ 1	○ 2

Hypothetical Visit 3 (Baseline+14 weeks)

Then the EQ-5D-5L, including

USUAL ACTIVITIES *(e.g. work, study, housework, family or leisure activities)*

I have no problems doing my usual activities	❑
I have slight problems doing my usual activities	❑
I have moderate problems doing my usual activities	❑
I have severe problems doing my usual activities	❑
I am unable to do my usual activities	❑

Hypothetical Visit 3 (Baseline+14 weeks)

Changes in impact of heart failure on daily functioning (as assessed by the KCCQ Physical Limitations Domain) is important and we would like to explore interpretation options prior to Phase 3, so we add an anchor (PGIS).

Please choose the response below that best describes the overall impact of your heart condition on your ability to complete your usual activities over the past two weeks. (Select one response)

- ❑ None
- ❑ Mild
- ❑ Moderate
- ❑ Severe
- ❑ Very severe

Hypothetical Visit 3 (Baseline+14 weeks)

We are also planning on assessing impact of heart failure on physical function (as assessed by the SF-36). It is a function-related concept, so we could likely use the same global assessment (ability to complete usual activities)—but the KCCQ has a recall period of 2 weeks and SF-36 has a recall period of 4 weeks. Do we need to add this anchor?

Please choose the response below that best describes the overall impact of your heart condition on your ability to engage in physical activity over the past four weeks. (Select one response)

- ❑ None
- ❑ Mild
- ❑ Moderate
- ❑ Severe
- ❑ Very severe

Hypothetical Visit 3 (Baseline+14 weeks)

It’s advisable to use both “static” anchors, and anchors that assess perception of change, so we also include a PGIC item

Please choose the response below that best describes the overall change in your heart condition’s impact on your ability to complete your usual activities, compared to when you started the study. (Select one response)

- ❑ Much better
- ❑ A little better
- ❑ No change
- ❑ A little worse
- ❑ Much worse

Hypothetical Visit 3 (Baseline+14 weeks)

The mechanism of action suggests that shortness of breath (SOB), a key symptom, will improve beyond what would be expected—so we plan to explore this symptom and its impacts, specifically, (as measured by the KCCQ). We consider including a concept (SOB)-specific anchor.

Please choose the response below that best describes the overall impact of your shortness of breath on your ability to complete your usual activities over the past two weeks. (Select one response)

- ❑ None
- ❑ Mild
- ❑ Moderate
- ❑ Severe
- ❑ Very severe

Hypothetical Visit 3 (Baseline+14 weeks)

And, of course, change.

Please choose the response below that best describes the overall change in the impact of your shortness of breath on your ability to complete your usual activities compared to when you started the study. (Select one response)

- ❑ Much better
- ❑ A little better
- ❑ No change
- ❑ A little worse
- ❑ Much worse

Hypothetical Visit 3 (Baseline+14 weeks)

Then, the coordinator administers the 6MWT. It's a PerfO assessment, so the recall period for the previous PGIS anchor(s) assessing function-related ability (2 weeks and/or 4 weeks) is not ideal for a momentary assessment (such as a PerfO assessment). Before the 6MWT, the participant is asked

> Please choose the response below that best describes the overall impact of your heart condition on your ability to complete your usual activities today (Select one response)
>
> ❑ None
>
> ❑ Mild
>
> ❑ Moderate
>
> ❑ Severe
>
> ❑ Very severe

Hypothetical Visit 3 (Baseline+14 weeks)

We are still interested in exploring the specific impact of shortness of breath prior to Phase 3, so participant is also asked

Please choose the response below that best describes the overall impact of your shortness of breath on your ability to complete your usual activities today (Select one response)

- ❑ None
- ❑ Mild
- ❑ Moderate
- ❑ Severe
- ❑ Very severe

Didn’t I already answer that question?

New York Heart Association (NYHA) Functional Classification (ClinRO measure) is based on similar concepts of function and generally involves the clinician asking the participants very similar questions

1 ClinRO measure

1 PerfO assessment

53 PRO measure items

6 (or 7) patient-reported anchor items turns 53 items into 59 or 60 items

Anchors Can Add Up

Endpoint Hierarchy

- Primary
  - Walking Endurance – performance outcome (PerfO) assessment; 1 task
- Secondary
  - Function (mobility) – patient-reported outcome (PRO) measure; 7 items; 2-week recall
  - Impact on daily activities – PRO measure; 9 items; 7-day recall
  - Fatigue – PRO measure; 4 items; daily diary with "7-day" endpoint
  - Pain – PRO measure; 2 items; daily diary with "7-day" endpoint
- Exploratory
  - Anchor – "Function" PGIS; recall period of "today" (PerfO task anchor)
  - Anchor – "Function" PGIS; recall period of 2 weeks (Function PRO items anchor)
  - Anchor – "Fatigue" PGIS; recall period of 7 days (Fatigue PRO items anchor)
  - Anchor – "Pain" PGIS; recall period of 7 days (Pain PRO items anchor)
  - Anchor – "Function," "Fatigue," "Pain," PGICs; recall period of "since beginning treatment"

1 PerfO task

22 PRO measure items

8 anchor items

Summary

- Must balance the need for the anchoring information with the participant and administration burden
- No anchor is perfect, so there may be imperfect options already included in the study plan
- Do not add a “better” anchor if the improvement in ability to interpret the COA score(s) is minimal
- Remember that numerous methods exist and consider what is appropriate for the analytic objectives

Resources

Abugov, R., Clark, J., Higginbotham, L., Li, F., Nie, L., Reasner, D., . . . Sharretts, J. (2023). Should responder analyses be conducted on continuous outcomes? *Pharm Stat, 22*(2), 312-327. doi:10.1002/pst.2273

Bjorner, J. B., Terluin, B., Trigg, A., Hu, J., Brady, K. J. S., & Griffiths, P. (2022). Establishing thresholds for meaningful within-individual change using longitudinal item response theory. *Qual Life Res*. doi:10.1007/s11136-022-03172-5

Cappelleri, J. C., & Bushmakin, A. G. (2014). Interpretation of patient-reported outcomes. *Stat Methods Med Res, 23*(5), 460-483. doi:10.1177/0962280213476377

Carrasco-Labra, A., Devji, T., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N., . . . Guyatt, G. H. (2022). Serious reporting deficiencies exist in minimal important difference studies: current state and suggestions for improvement. *J Clin Epidemiol, 150*, 25-32. doi:10.1016/j.jclinepi.2022.06.010

Carrasco-Labra, A., Devji, T., Qasim, A., Phillips, M. R., Wang, Y., Johnston, B. C., . . . Guyatt, G. H. (2021). Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol, 133*, 61-71. doi:10.1016/j.jclinepi.2020.11.024

Cocks, K., & Buchanan, J. (2022). How scoring limits the usability of minimal important differences (MIDs) as responder definition (RD): an exemplary demonstration using EORTC QLQ-C30 subscales. *Qual Life Res*. doi:10.1007/s11136-022-03181-4

Coon, C. D., & Cappelleri, J. C. (2016). Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Ther Innov Regul Sci, 50*(1), 22-29. doi:10.1177/2168479015622667

Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Qual Life Res, 27*(1), 33-40. doi:10.1007/s11136-017-1616-3

Resources

Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N., . . . Guyatt, G. H. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *Bmj, 369*, m1714. doi:10.1136/bmj.m1714

Griffiths, P., Sims, J., Williams, A., Williamson, N., Cella, D., Brohan, E., & Cocks, K. (2022). How strong should my anchor be for estimating group and individual level meaningful change? A simulation study assessing anchor correlation strength and the impact of sample size, distribution of change scores and methodology on establishing a true meaningful change threshold. *Qual Life Res*. doi:10.1007/s11136-022-03286-w

Hays, R. D., & Peipert, J. D. (2021). Between-group minimally important change versus individual treatment responders. *Qual Life Res, 30*(10), 2765-2772. doi:10.1007/s11136-021-02897-z

Hays, R. D., & Woolley, J. M. (2000). The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics, 18*(5), 419-423. doi:10.2165/00019053-200018050-00001

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials, 10*(4), 407-415. doi:10.1016/0197-2456(89)90005-6

King, M. T., Dueck, A. C., & Revicki, D. A. (2019). Can Methods Developed for Interpreting Group-level Patient-reported Outcome Data be Applied to Individual Patient Management? *Med Care, 57 Suppl 5 Suppl 1*(Suppl 5 1), S38-s45. doi:10.1097/mlr.0000000000001111

Mc Carthy, M., Burrows, K., Griffiths, P., Black, P. M., Demanuele, C., Karlsson, N., . . . Cappelleri, J. C. (2023). From Meaningful Outcomes to Meaningful Change Thresholds: A Path to Progress for Establishing Digital Endpoints. *Ther Innov Regul Sci*. doi:10.1007/s43441-023-00502-8

Resources

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol, 61*(2), 102-109. doi:10.1016/j.jclinepi.2007.03.012

Staunton, H., Willgoss, T., Nelsen, L., Burbridge, C., Sully, K., Rofail, D., & Arbuckle, R. (2019). An overview of using qualitative techniques to explore and define estimates of clinically important change on clinical outcome assessments. *J Patient Rep Outcomes, 3*(1), 16. doi:10.1186/s41687-019-0100-y

Terwee, C. B., Peipert, J. D., Chapman, R., Lai, J. S., Terluin, B., Cella, D., . . . Mokkink, L. B. (2021). Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Qual Life Res, 30*(10), 2729-2754. doi:10.1007/s11136-021-02925-y

Trigg, A., Lenderking, W. R., & Boehnke, J. R. (2023). Introduction to the special section: "Methodologies and considerations for meaningful change". *Qual Life Res*. doi:10.1007/s11136-023-03413-1

Weinfurt, K. P. (2019). Clarifying the Meaning of Clinically Meaningful Benefit in Clinical Research: Noticeable Change vs Valuable Change. *Jama, 322*(24), 2381-2382. doi:10.1001/jama.2019.18496

Wyrwich, K. W., & Norman, G. R. (2022). The challenges inherent with anchor-based approaches to the interpretation of important change in clinical outcome assessments. *Qual Life Res*. doi:10.1007/s11136-022-03297-7

Wyrwich, K. W., Norquist, J. M., Lenderking, W. R., & Acaster, S. (2013). Methods for interpreting change over time in patient-reported outcome measures. *Qual Life Res, 22*(3), 475-483. doi:10.1007/s11136-012-0175-x

Resources: Special Section

Quality of Life Research
https://doi.org/10.1007/s11136-023-03413-1

EDITORIAL

Introduction to the special section: "Methodologies and considerations for meaningful change"

Andrew Trigg[1] · William R. Lenderking[2] · Jan R. Boehnke[3]

References: Hypothetical Trial Example

THE CLASSIFICATION OF CARDIAC DIAGNOSIS. (1921). *Journal of the American Medical Association, 77*(18), 1414-1415. doi:10.1001/jama.1921.02630440034013

Green, C. P., Porter, C. B., Bresnahan, D. R., & Spertus, J. A. (2000). Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure. *J Am Coll Cardiol, 35*(5), 1245-1255. doi:10.1016/s0735-1097(00)00531-3

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., . . . Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res, 20*(10), 1727-1736. doi:10.1007/s11136-011-9903-x

Lipkin, D. P., Scriven, A. J., Crake, T., & Poole-Wilson, P. A. (1986). Six minute walking test for assessing exercise capacity in chronic heart failure. *Br Med J (Clin Res Ed), 292*(6521), 653-655. doi:10.1136/bmj.292.6521.653

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care, 30*(6), 473-483.

Thank you!

Reflections on the Rheumatoid Arthritis Working Group's Experience Estimating Meaningful Within-person Change for the *PROMIS® Fatigue 10a*

John Devin Peipert, PhD – Assistant Professor of Medical Social Sciences

Northwestern University

Rheumatoid Arthritis (RA) Working Group

Eli Lilly
April Naegeli (Chair)
Laure Delbecque

AbbVie
Pankaj Patel
Elektra Papadopoulos
Patrick Zueger

Boehringer Ingelheim
Tristan Gloede

GSK
Wen-Hung Chen

EMD Serono
Paul Kamudoni
Christian Henke

UCB Pharm
Ann-Christin Mörk

C-Path
Stephen Joel Coons (Senior Advisor)
Sonya Eremenco
Maria Mattera
Cheryl Coon
Theresa Griffey
Tarryn Ho

American Institutes for Research
San Keller

Johns Hopkins University
Clifton O. Bingham

McGill University
Susan J. Bartlett

OMERACT
Lee Simon
Vibeke Strand

Patient Representative
Amye Leong

Northwestern University
David Cella
George J. Greene
John Devin Peipert
Xiaodan Tang

Importance of Fatigue in RA

- Fatigue affects ~90% of individuals with RA[1]
- Changes in fatigue track closely with RA disease control[2]
- Fatigue is prioritized among RA patients' most important symptoms[3] and affects other elements of health-related quality of life[4,5]
- For these reasons, fatigue has become a priority target to demonstrate therapeutic benefit in RA

PROMIS® Fatigue 10a

FACIT Fatigue Scale		Retained for *PROMIS Fatigue 10a*
HI7	I feel fatigued	Yes
HI12	I feel weak all over	No
An1	I feel listless (“washed out”)	No
An2	I feel tired	Yes
An3	I have trouble starting things because I am tired	Yes
An4	I have trouble finishing things because I am tired	Yes
An5	I have energy	Yes
An7	I am able to do my usual activities	Yes
An8	I need to sleep during the day	Yes
An12	I am too tired to eat	No
An14	I need help doing my usual activities	Yes
An15	I am frustrated by being too tired to do the things I want to do	Yes
An16	I have to limit my social activity because I am tired	Yes

FDA Drug Development Tool (DDT) COA Qualification Submission

DDT COA #000015: PROMIS® Short Form Fatigue 10a in Rheumatoid Arthritis

- Context of use
  - Adult patients (>18 years) with RA based on a score of ≥ 6 on the American College of Rheumatology/European League Against Rheumatism 2010 Rheumatoid Arthritis Classification Criteria
- Concept of interest
  - Fatigue severity among adults with RA
  - Fatigue: "an overwhelming, debilitating, and sustained sense of exhaustion that decreases one's ability to carry out daily activities, including the ability to work effectively and to function at one's usual level in family or social roles"[6]
- Requestors: PRO Consortium's RA Working Group
- Status: Full Qualification Package submission completed on January 10, 2023

Psychometric Evaluation for DDT COA Submission

- DDT COA submission involves extensive psychometric evaluation, including:
  - Content validity
  - Reliability
  - Construct validity
  - Ability to detect change
  - Estimation of meaningful within-person change (MWPC)

Approach to Estimating MWPC

To the greatest extent possible, we took the FDA-recommended approach to estimating MWPC[7]:

- Anchor-based approach
- Goal is to estimate a change that an individual patient would find meaningful
- Distribution-based approaches do not define meaningfulness but can provide supportive evidence
- Visualize with empirical cumulative distribution function (eCDF) and probability density function (PDF)

Anchor Selection[7]

- Anchors should meet certain criteria to be used for MWPC estimation:
  - Concept-related
  - Sufficiently correlated
  - Represent meaningful changes and be easier to interpret than the PRO measure
  - Recall period matches PRO-based endpoint
- Can show change using:
  - Retrospective reports [e.g., patient global impression of change (PGIC)]
  - Static reports [e.g., change in patient global impression of severity (PGIS) categories]
- Multiple anchors should be used to provide an accumulation of evidence

Dataset for Estimating MWPC

- *A Randomized, Double-Blind, Placebo- and Active-Controlled, Phase 3 Study Evaluating the Efficacy and Safety of Baricitinib in Patients with Moderately to Severely Active Rheumatoid Arthritis Who Have Had an Inadequate Response to Methotrexate Therapy* (RA-BEAM; NCT01710358)
  - Included ambulatory adults with 1) moderately to severely active RA; 2) insufficient response to methotrexate; 3) never been treated with a biologic disease-modifying antirheumatic
  - Randomized to receive placebo, baricitinib, or adalimumab
  - PRO measures (including *PROMIS Fatigue Short Form 10a*) assessed at Baseline, Week 12, and Week 24
- N = 1305 for PRO measure analysis
- Did not include PGIS or PGIC

Anchor Approach

- Searched for candidate anchors that met previously stated criteria
  - SF-36 Vitality Acute (7-day) score and single items
  - Severity of Worst Tiredness item score (numeric rating scale; 11-point)
- Interpretation aided by analysis of SF-36 Vitality Acute score and Severity of Worst Tiredness item data from a previous study of 282 stable RA participants[8]
  - Administered target anchors and a fatigue-specific PGIC with responses of “A lot better,” “A little better,” “Same,” “A little worse,” and “A lot worse”
  - Calculated mean change scores of SF-36 Vitality and Severity of Worst Tiredness item within each PGIC category and used these as thresholds to apply in RA-BEAM

Deriving Anchors in RA-BEAM

Anchor	Operationalized for Analysis	Correlation with Δ in *PROMIS Fatigue 10a*	MWPC Criteria
SF-36 Vitality Score (acute) - Multi-item scale score ranging from 0 to 100; higher scores indicating higher vitality; reflects past 7 days	Change scores categorized as "A lot better" (≥12 pts), "A little better" (≥5, <12 pts), "Same" (< 5 pts, > -4 pts), "A little worse" (≤ -4 pts, >-12 pts), "A lot worse" (≤ -12 pts) from baseline to 24 weeks.	**r = -0.62**	A little vs. no change, A lot vs. a little
Severity of Worst Tiredness item - Patient-administered, 11-point horizontal scale anchored at 0 and 10, from "no tiredness" to "as bad as you can imagine" in last 24 hours	Scores averaged over 7 days prior to baseline and 12-week time points. Then, categorized as: "A lot worse" (>1.5 pts), "A little worse" (>0.5, ≤1.5 pts), "Same" (≤ 0.5 pts, > -1 pt), "A little better" (≤ -1 pt, >-2.5 pts), "A lot better" (≤ -2.5 pts) from baseline to 12 weeks	**r = 0.33**	A little vs. no change, A lot vs. a little
SF-36 VT9e (acute) - "Past 7 days: Did you have a lot of energy," "All of the time" (1) - "None of the time" (5)	Change in response categories baseline to 24 weeks. >0 = improved (1 category, 1-4 categories), <=0 = not improved; <0 = worsened (1 category, 1-4 categories), >=0 = not worsened.	**r = -0.50**	Change of 1 category vs. no change, Change of 1-4 categories vs. no change
SF-36 VT9i (acute) - "Past 7 days: Did you feel tired, "All of the time" (1) to "None of the time" (5)	Change in response categories baseline to 24 weeks. >0 = improved (1 category, 1-4 categories), <=0 = not improved; <0 = worsened (1 category, 1-4 categories), >=0 = not worsened.	**r = -0.57**	Change of 1 category vs. no change, Change of 1-4 categories vs. no change

Statistical Approach to Anchor-Based MWPC Analyses

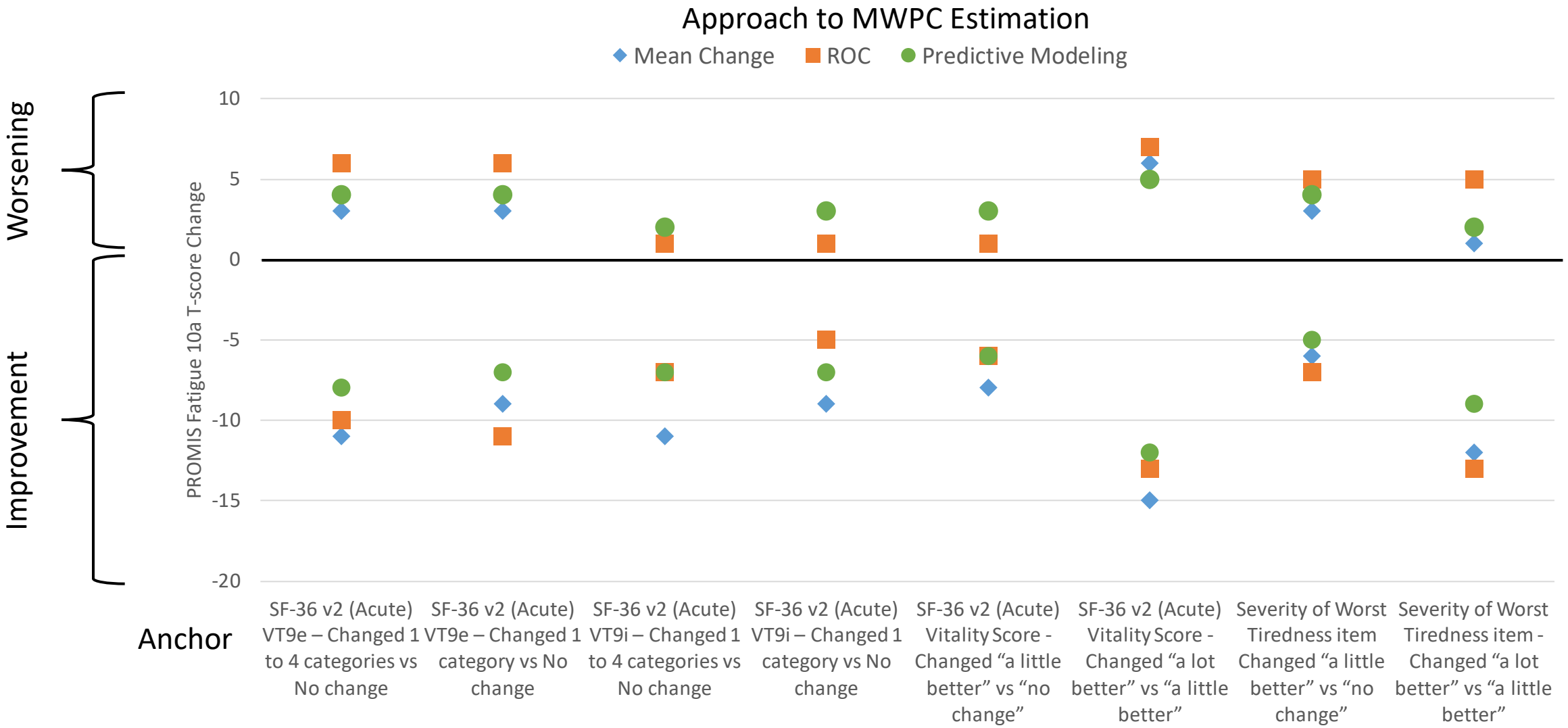
- Mean change approach was primary
- Potential limitations of mean change approach[9,10,11]
  - Biased if the anchor is not accurate in its classification of participants as changed vs. not changed
- In addition to mean change approach:
  - eCDF and PDF curves
  - Receiver operator characteristic (ROC) curve
  - Predictive modeling[9,10]
  - These can improve precision and be adjusted for unequal sample sizes of changed vs. not changed[10]
- Stratification by treatment arm to account for differing magnitude of fatigue change in placebo arm vs. experimental arm (adalimumab + baricitinib)
  - This approach would likely not be appropriate for other settings, including new drug submissions
  - Placebo vs. experimental group (combining adalimumab and baricitinib arms)

Triangulating between Many MWPC Estimates

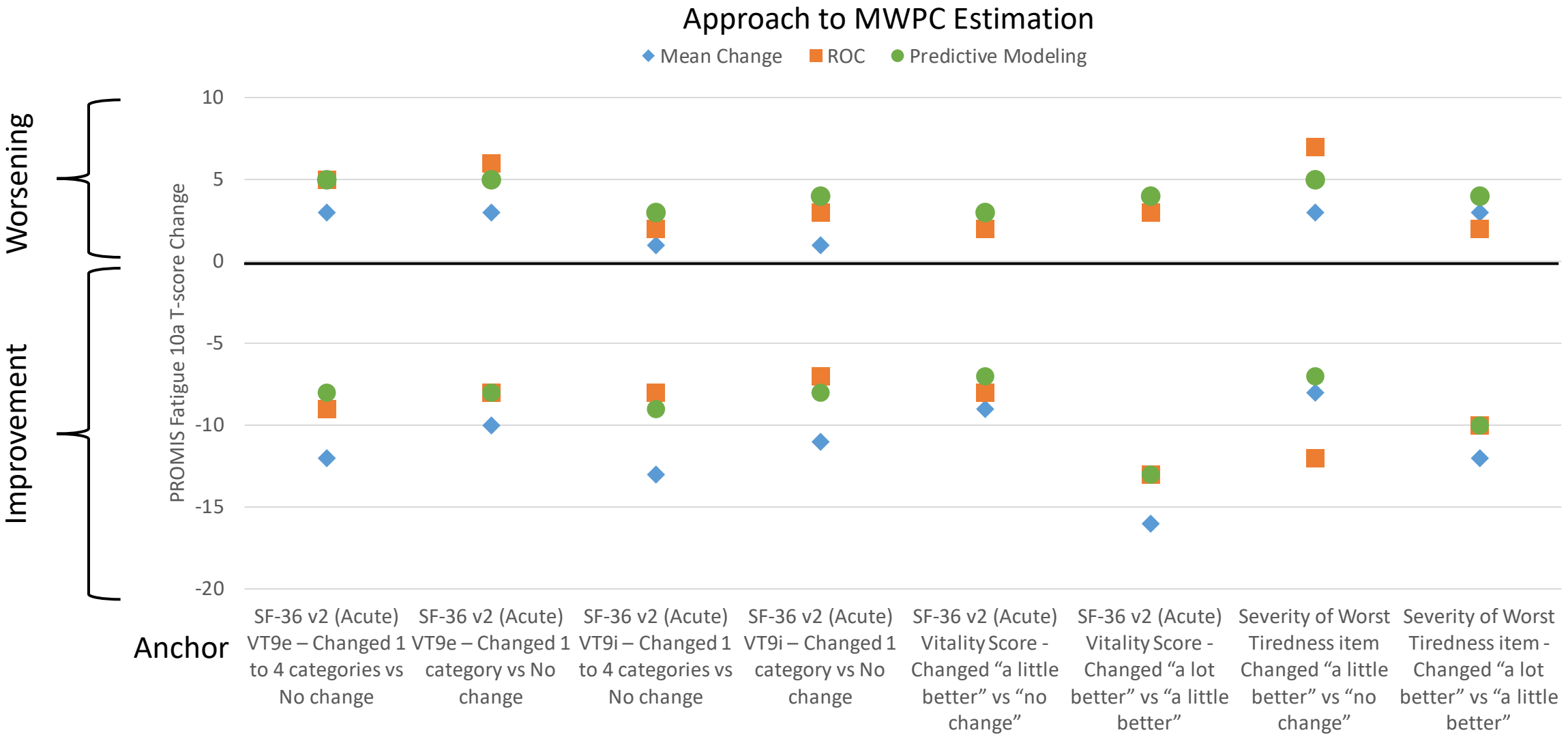
- Our approach generated **96** MWPC estimates
  - **4** anchors x **2** MWPC thresholds per anchor x **3** ways of estimating x **2** directions of change (improvement, worsening) x **2** arms (experimental, placebo)
- Having a lot of information is great!
  - Allows for comparison of the quality of different methods
  - Can build confidence in a final threshold range

- Having a lot of information is a challenge!
  - How do you condense this information into a usable summary?
  - Several approaches triangulating across estimates are available: plot reviews, averaging, correlation-weighted averaging[12]
- We used a process of expert consensus to define a MWPC threshold range (not stratified by arm)

MWPC Estimates: Placebo Group



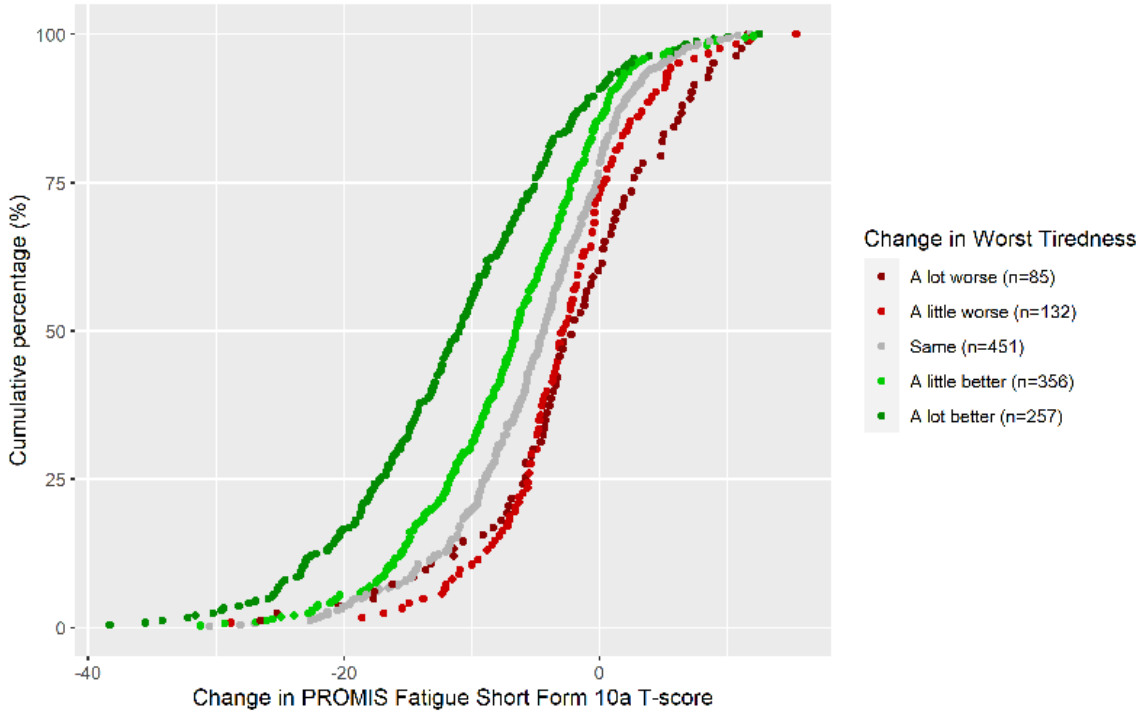
MWPC Estimates: Experimental Group



eCDF Curves Stratified by Anchor Categories (1)

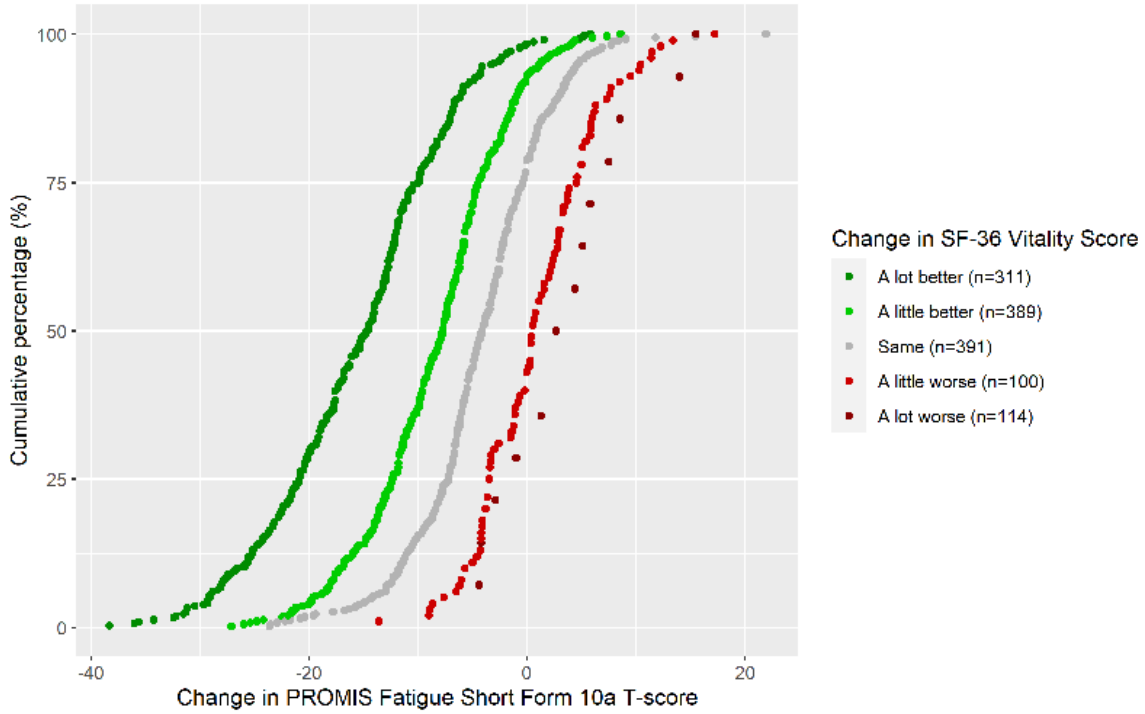
PRO CONSORTIUM CRITICAL PATH INSTITUTE

Anchor: Severity of Worst Tiredness Item



Negative change scores indicate improvement

Anchor: SF-36 Vitality Score

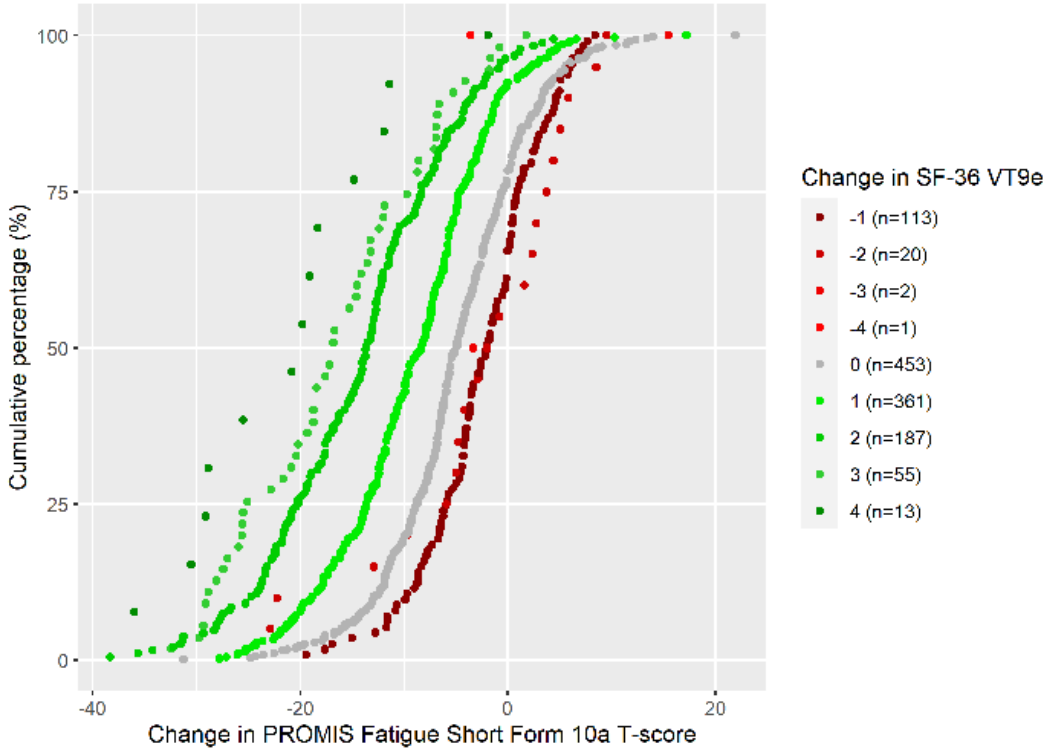


Negative change scores indicate improvement

eCDF Curves Stratified by Anchor Categories (2)

Anchor: SF-36 VT9e

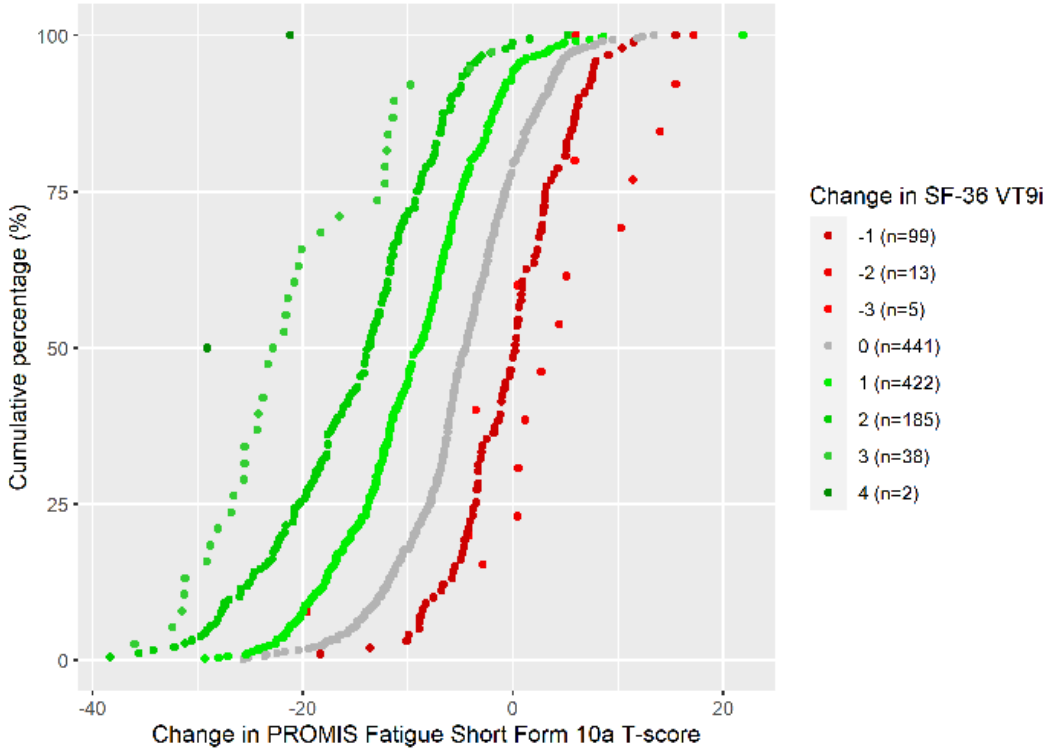
Did you have a lot of energy?



Negative change scores indicate improvement

Anchor: SF-36 VT9i

Did you feel tired?



Negative change scores indicate improvement

Conclusions and Discussion

- Final MWPC Ranges
  - 5 to 8 T-score points for improvement
  - 2 to 5 T-score points for worsening
- Psychometric evaluations often conducted on existing trials
  - Pros: More likely that change has occurred in an interventional setting
  - Cons: May not have anchors we want
  - Can use anchor guidance and enhanced statistical approaches for MWPC estimation to help address anchor limitations
- It was helpful to put these results into context with MWPC estimates for RA from other PROMIS Fatigue measures
  - Vignette-based study of PROMIS Fatigue with RA patients and clinicians found consensus of 10 T-score points for improvement but disagreement for worsening with clinicians suggesting 5 points and patients suggesting 10-15 points[13]

References

1. Piper BF. *Fatigue: current bases for practice*. Key Aspects of Comfort: Management of Pain, Fatigue and Nausea 1989.
2. Bingham CO, Bartlett SJ. Qualification of a measure of fatigue for use in the assessment of treatment benefit in rheumatoid arthritis clinical trials. A report prepared for the Critical Path Institute, PRO Consortium, RA Working Group. 2016. June 15, 2016.
3. Minnock P, Bresnihan B. Pain outcomes and fatigue levels reported by women with established rheumatoid arthritis. *Arthritis & Rheumatology*. 2004;50:S471.
4. Hewlett S, Chalder T, Choy E, et al. Fatigue in rheumatoid arthritis: time for a conceptual model. Rheumatology (Oxford). Jun 2011;50(6):1004-6.
5. Carr A, Hewlett S, Hughes R, et al. Rheumatology outcomes: the patient's perspective. *Journal of Rheumatology*. Apr 2003;30(4):880-3.
6. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . PROMIS Cooperative Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Medical Care, 45(5 Suppl 1), S3-S11.
7. US Food and Drug Administration. Discussion Document for Patient-Focused Drug Development Public Workshop on Guidance 4: Incorporating Clinical Outcome Assessments into Endpoints for Regulatory Decision-Making. Silver Spring, MD: United States Department of Health and Human Services; December 6, 2019.
8. Bartlett SJ, Haque U, Bykerk V, Curtis JR, Jones M, Bingham C. Identifying Meaningful and Detectable Change from the Patient Perspective across Common Fatigue Measures in Rheumatoid Arthritis. presented at: EULAR Congress 2021; June 5th 2021; Virtual. Session Poster Tour.
9. Terwee CB, Peipert JD, Chapman R, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res. 2021.
10. Terluin B, Eekhout I, Terwee CB, de Vet HC. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. J Clin Epidemiol. 2015;68(12):1388-1396.
11. Griffiths P, Sims J, Williams A, et al. How strong should my anchor be for estimating group and individual level meaningful change? A simulation study assessing anchor correlation strength and the impact of sample size, distribution of change scores and methodology on establishing a true meaningful change threshold. Qual Life Res. 2022 Nov 19. doi: 10.1007/s11136-022-03286-w. Epub ahead of print.
12. Trigg A, Griffiths P. Triangulation of multiple meaningful change thresholds for patient-reported outcome scores. *Qual Life Res.* 2021;30(10):2755-2764.
13. Bingham CO, III, Butanis AL, Orbai AM, et al. Patients and clinicians define symptom levels and meaningful change for PROMIS pain interference and fatigue in RA using bookmarking. Rheumatology. 2021;60(9):4306-4314.

Shareware for Deriving Thresholds for Meaningful Change

Karon F. Cook, PhD (karon.cook@feralscholars.net)

Learning Objectives

- Introduce an FDA-funded (CERSI*) Shareware project
- Explain Idioscale Judgment Studies
- Describe the Shareware functions and potential impact

*Centers of Excellence in Regulatory Science and Innovation

Stanford/USCF CERSI* PROJECT

Award Number: 2U01FD005978-06

The Threshold Project

Development, Implementation, and Evaluation of an Open Source Software Program to Support Patient-based Estimation of Clinically Meaningful Levels and Change Scores for Patient-Reported Outcome Measures.

*CERSI - Centers of Excellence in Regulatory Science and Innovation

Collaborators

Sean Mackey (PI)

Sophia You (Co-I)

Maisa Ziadni (Co-I)

Corinne Jung (PM, Pain Division Research Manager)

Juliette Hong (SME, Biostatistician)

Wendy M. Schadle (CERSI PM)

Garrick Olson (Infrastructure & Architecture Lead)

Karon Frances Cook (SME; Feral Scholars)

Nan Rothrock (SME, Northwestern University)

Arthur Stone (SME, University of Southern California)

Chris Veasley (Patient Representative, Chronic Pain Research Alliance)

Office of Regulatory Science and Innovation (ORSI)

Kinnera Chada (UCSF-Stanford CERSI Program Official)
Rebekah Zinn (FDA-CERSI Program Team Lead)

Center for Devices and Radiological Health (CDRH)

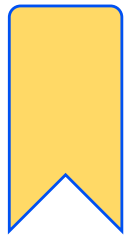
Fraser Bocell (FDA Lead)
Michelle Tarver (SME)
Caiyan Zhang (SME)

Abbreviation Key

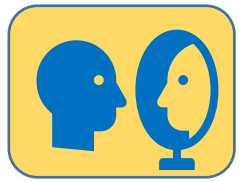
PI – Principal Investigator
Co-I – Co-Investigator
PM – Project Manager
SME – Subject Matter Expert

Stanford/USCF CERSI PROJECT

Award Number: 2U01FD005978-06



Bookmarking



Idioscale Judgment*

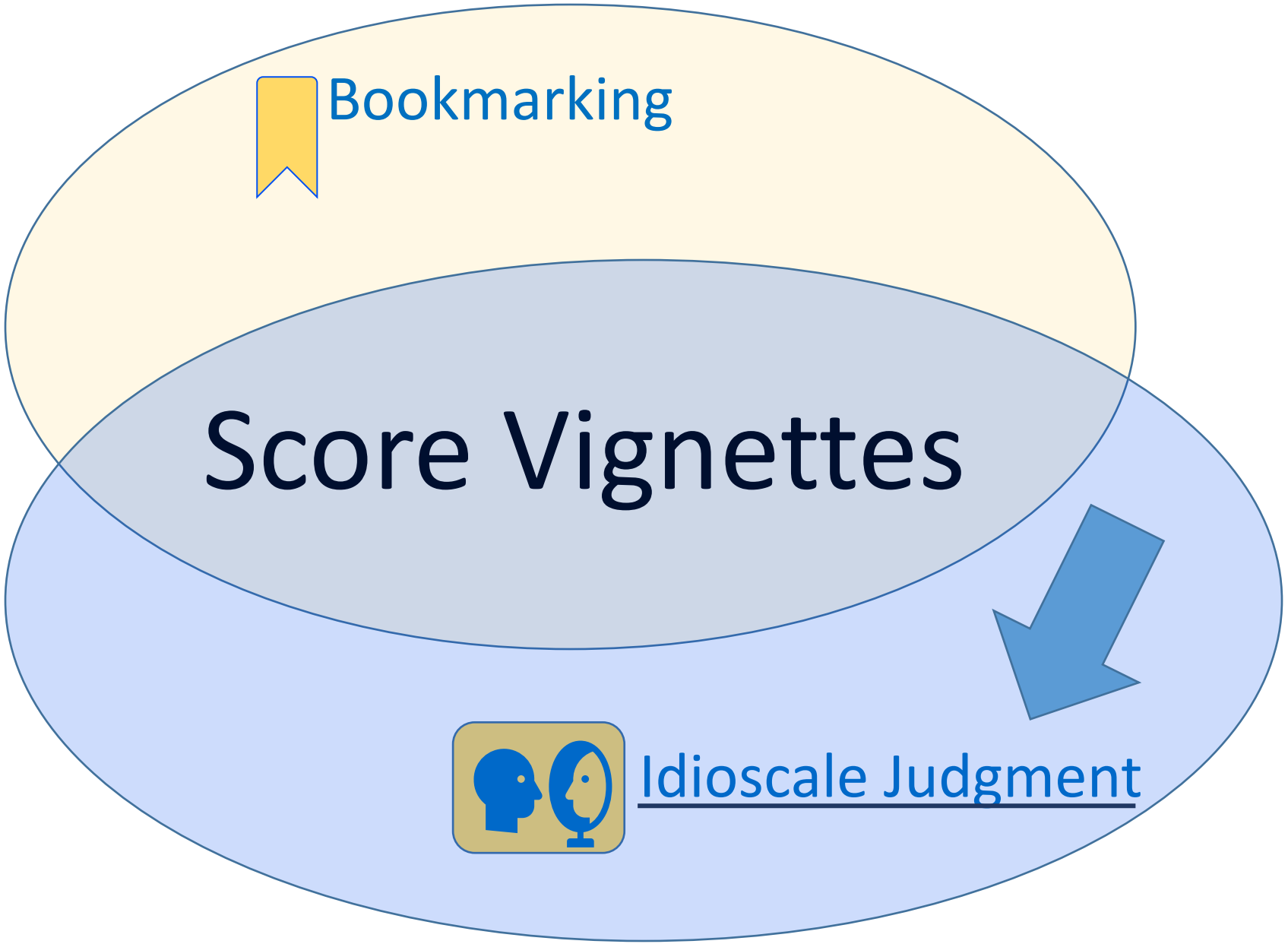
* **Why is it called Idioscale Judgment?**

- L.L. Thurston's developed law of comparative judgment.

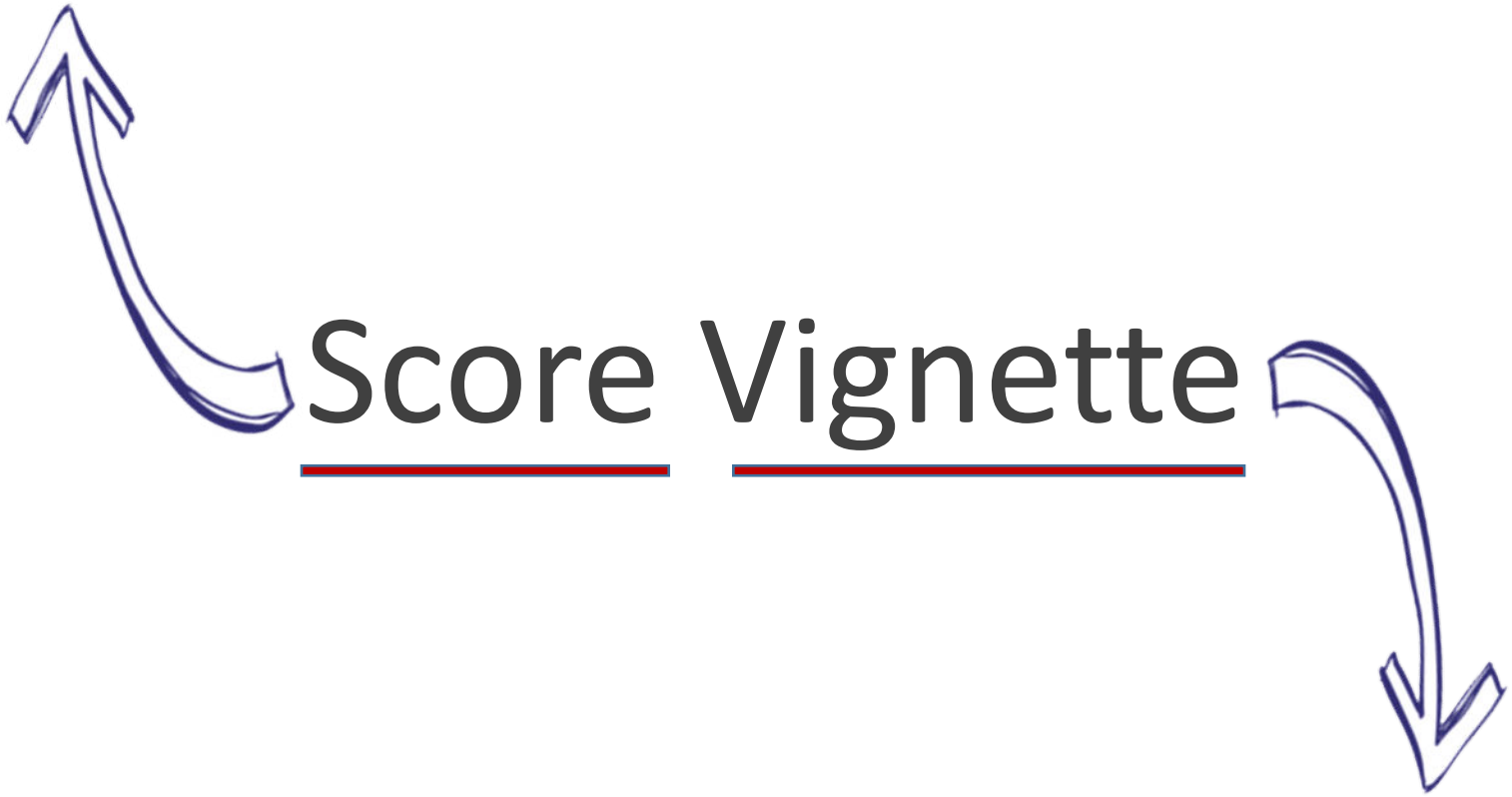
- Idioscale Judgment is a comparative judgment method, in which the comparator is one's own status. The SELF (idio) is the comparator.

Bookmarking and Idioscale Judgment Share this in common.

For today's purpose, we're going to concentrate on Idioscale Judgment.



Patient-Reported Outcome Measure (PROM)



Short story with elements that describe a character, a scene, or a context.

Tell me a story from PROMIS Measurement—The Fatigue Story.

Tell me the one about 55. I love that story.

Ms. Patel

- She was <u>rarely</u> too tired to think clearly.
- Fatigue <u>sometimes</u> interfered with social activities.
- Fatigue interfered <u>a little bit</u> with physical functioning.
- On average, she was <u>somewhat</u> fatigued.
- Fatigue <u>did not</u> make it at all hard to carry on a conversation.

65 is a whole ‘nother story.

Mr. Cruz

- He was <u>sometimes</u> too tired to think clearly.
- His fatigue interfered <u>quite a bit</u> with his physical functioning.
- On average, he was <u>quite a bit</u> wiped out.
- He was <u>sometimes</u> too tired to feel happy.
- His fatigue made him <u>somewhat</u> more forgetful.

Where do these stories come from?

Each of these 6 statements is an item from the PROMIS® Fatigue Item Bank

Mr. Cruz

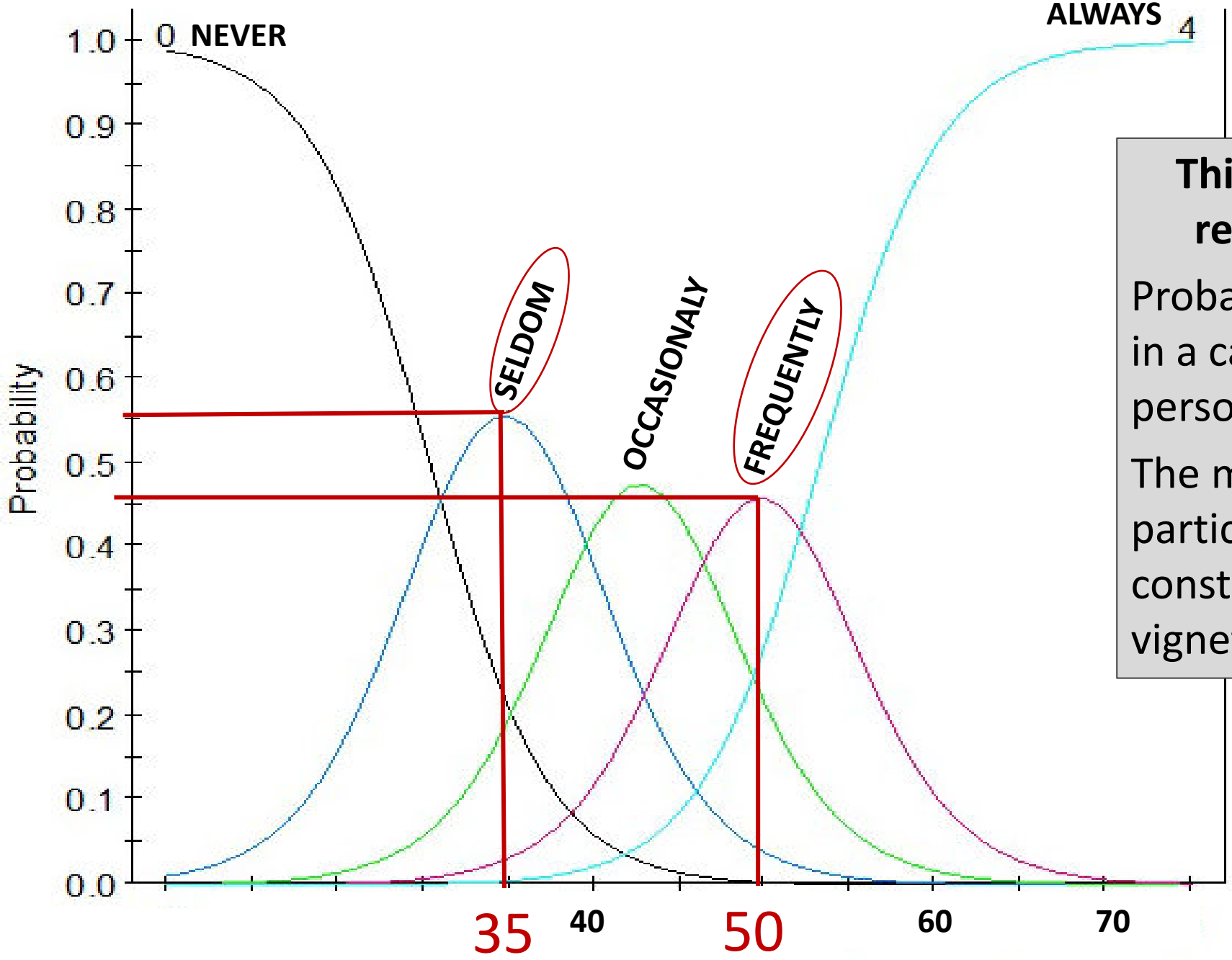
- He was sometimes too tired to think clearly.
- His fatigue interfered quite a bit with his physical functioning.
- On average, he was quite a bit wiped out.
- He was sometimes too tired to feel happy.
- His fatigue made him somewhat more forgetful.

Do I have to use PROMIS to get score stories?

No, but, using Item Response Theory item banks like PROMIS makes it easier.

Item Response Theory → Around 40 to 100 items available

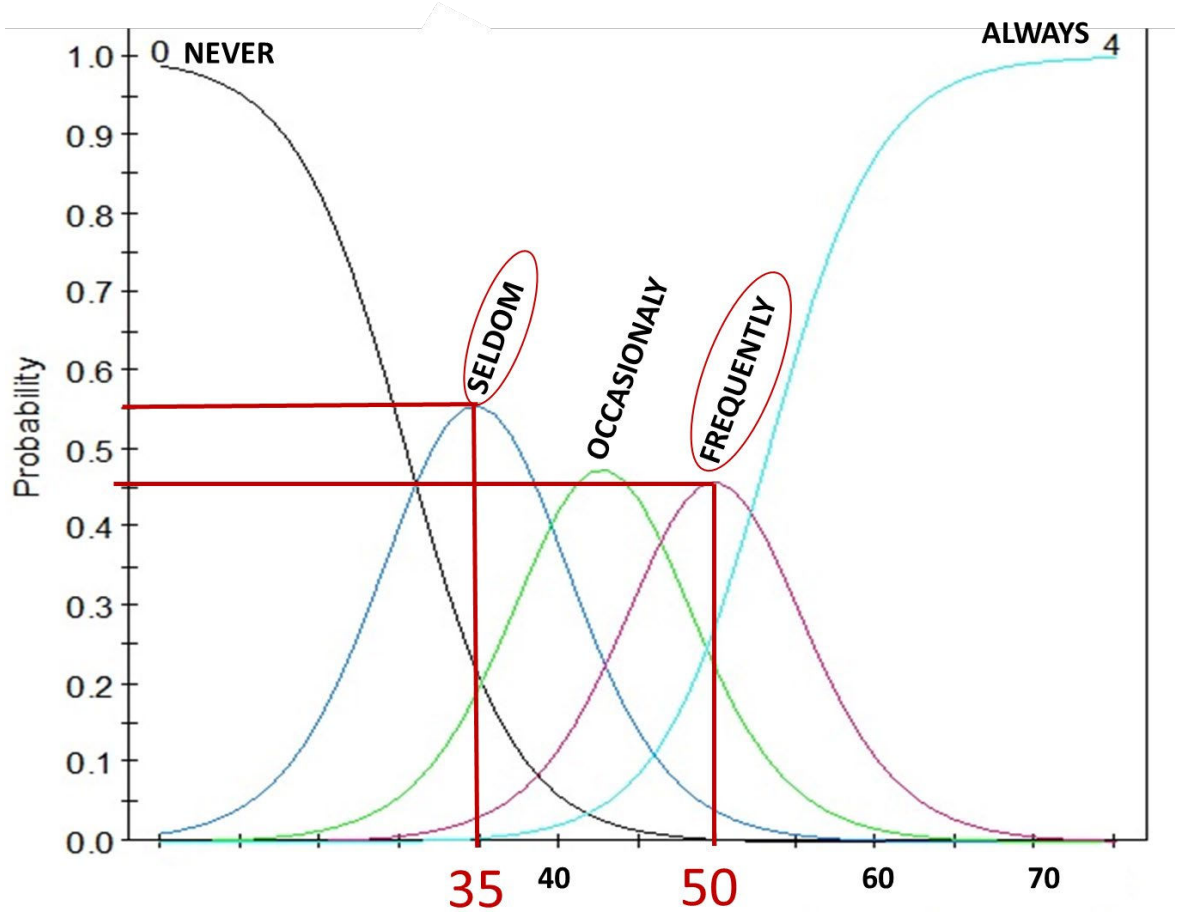
Verbal Rating Scales → Around 4 to 15 items available



This graph is from an item response IRT calibration

Probability (y-axis) responding in a category depends on the person's score.

The most probable item at a particular score is used in construction the score vignette.



Mr. Cruz

- He was sometimes too tired to think clearly.
- His fatigue interfered quite a bit with his physical functioning.
- On average, he was quite a bit wiped out.
- He was sometimes too tired to feel happy.
- His fatigue made him somewhat more forgetful.

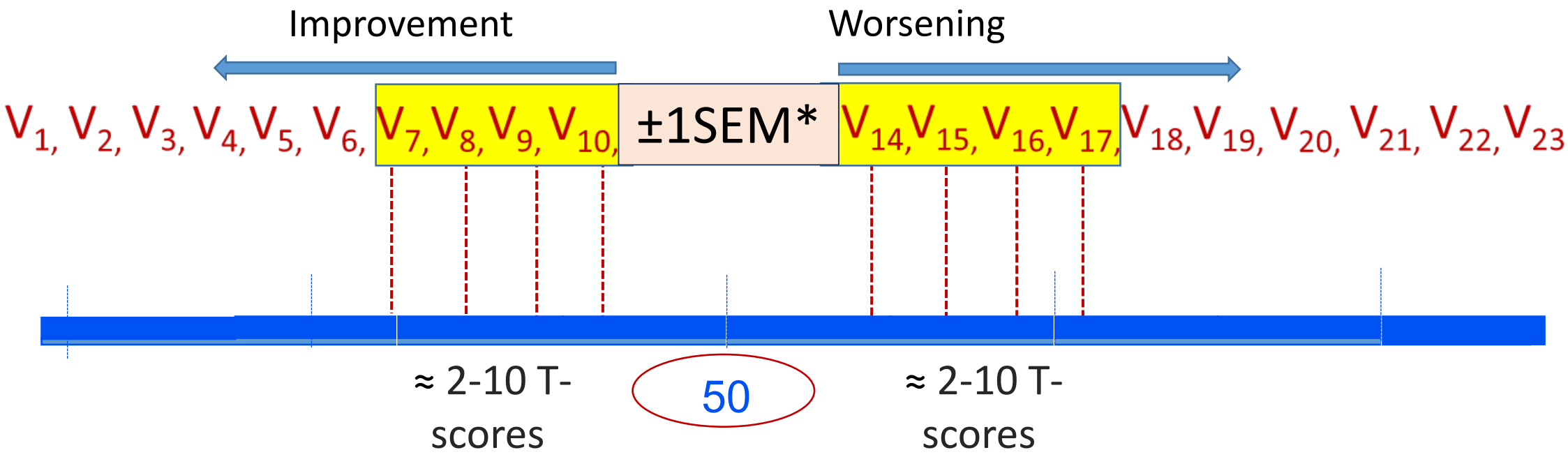
Step 1—Write the Stories

- The first step is to identify, for the full range of scores, what is the most likely response.
- An R-program was developed to provide this based on the item parameters.
- The shown scores are 10 T-score points apart, but you can obtain this result for any T-score value (e.g., 33.5)

	T-Score			
ITEM	40	50	60	70
FATIMP1	1	2	4	5
FATIMP2	1	1	3	4
FATIMP3	1	2	3	5
FATIMP4	1	2	3	4
FATIMP5	1	2	3	4
FATIMP6	1	2	3	4
FATIMP8	1	1	3	3
FATIMP9	1	1	3	4
FATIMP10	1	2	3	4
FATIMP11	1	2	3	4
FATIMP13	1	2	3	4

Etcetera to the end of item bank

Example of an Idioscale Judgment Study Being Conducted by EMD Serono Using PROMIS® Fatigue



- 8 score vignettes are presented to each participant; 4 represent Worsening and 4 represent Improvement.
- No score vignette that is within ± 1 SEM is presented
- Collectively, the presented vignette ranges are ≈ 2-10 T-scores above and ≈ 2-10 T-scores below the participant's score

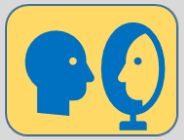
*SEM = Standard Error of Measurement

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

- **rarely was too tired to do her household chores.**
- **rarely needed help doing usual activities because of weakness.**
- **sometimes needed to sleep during the day.**
- **sometimes had to limit social activity because she was tired.**
- **sometimes was too tired to take a short walk.**

Compared to Ms. Howard's , has YOUR FATIGUE been:

- ❑ Greater than Ms. Howard's **[Go to Question that asks about its meaningfulness]**
- ❑ The same as Ms. Howard's **[Go to Next Vignette]**
- ❑ Less than Ms. Howard's **[Go to Question that asks about its meaningfulness]**



Idioscale Judgment

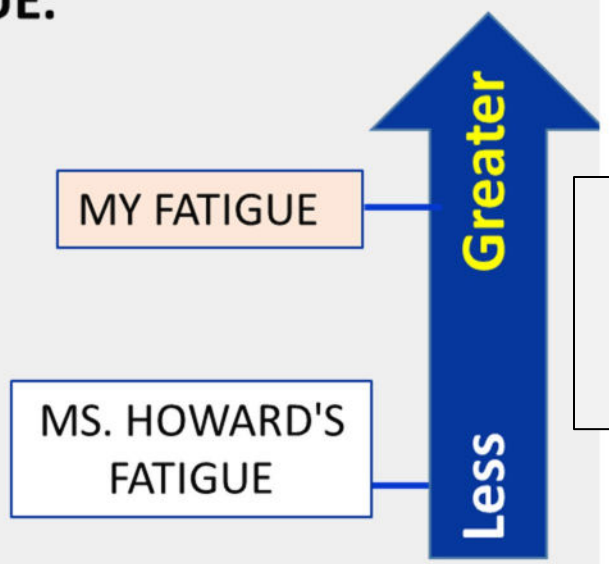
Participants compare their own experiences with the symptom or outcome to that of someone else.

Thus, the term—Idioiscale Judgment.

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

If your fatigue IMPROVED to Ms. Howard'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).



Graphical Reminder of their judgment

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

- **rarely was too tired to do her household chores.**
- **rarely needed help doing usual activities because of weakness.**
- **sometimes needed to sleep during the day.**
- **sometimes had to limit social activity because she was tired.**
- **sometimes was too tired to take a short walk.**

Reminder of the Comparison

WORSENING

V = 58 — Meaningfully Worse

V = 56 — Meaningfully Worse

V = 54 — Meaningfully Worse

V = 52 — Meaningfully Worse

IMPROVEMENT

V = 48

V = 46 — Meaningfully Better

V = 44 — Meaningfully Better

V = 42 — Meaningfully Better

Individual's Threshold Score

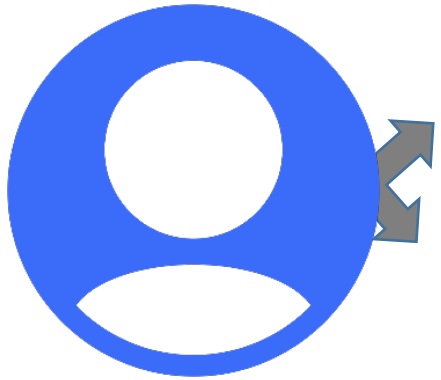
Vignette Score – T-score

Meaningful Worsening = 2

Meaningful Improvement = -4

Where To Go From Here?

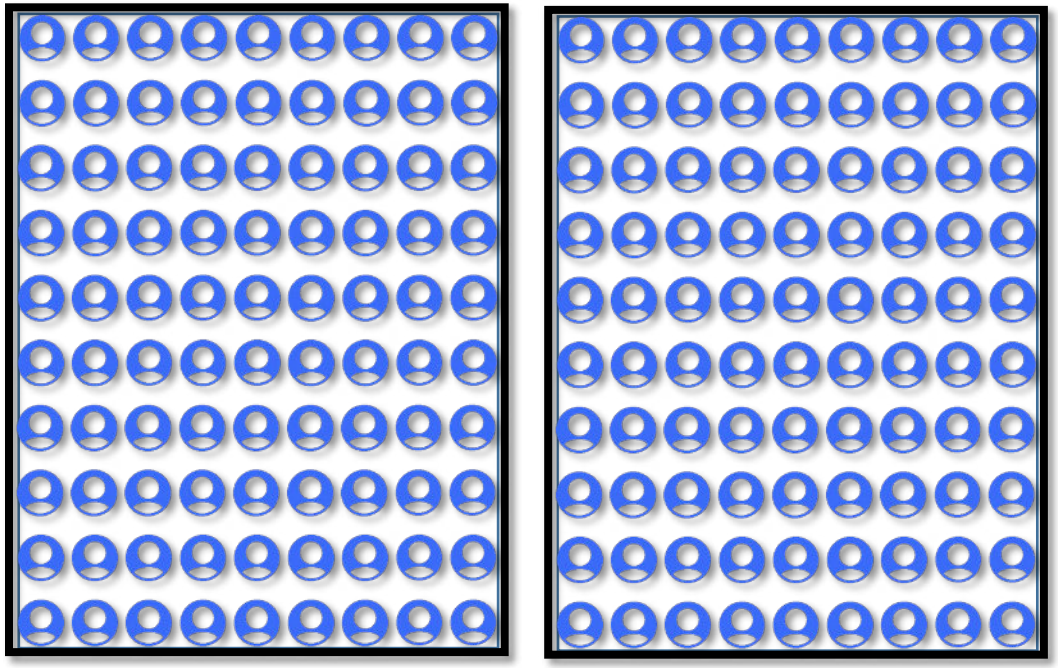
Individual Thresholds



Improvement

Worsening

Aggregated Thresholds



Every Domain Requires
8 judgments per person

We expect errors.

Out-of-Range Judgment (ORJ) Example

- Participant Fatigue score is 60
- Endorses Vignette of 63 as a meaningful improvement
- In calculations, these judgments are not included

Relatively low % of ORJs

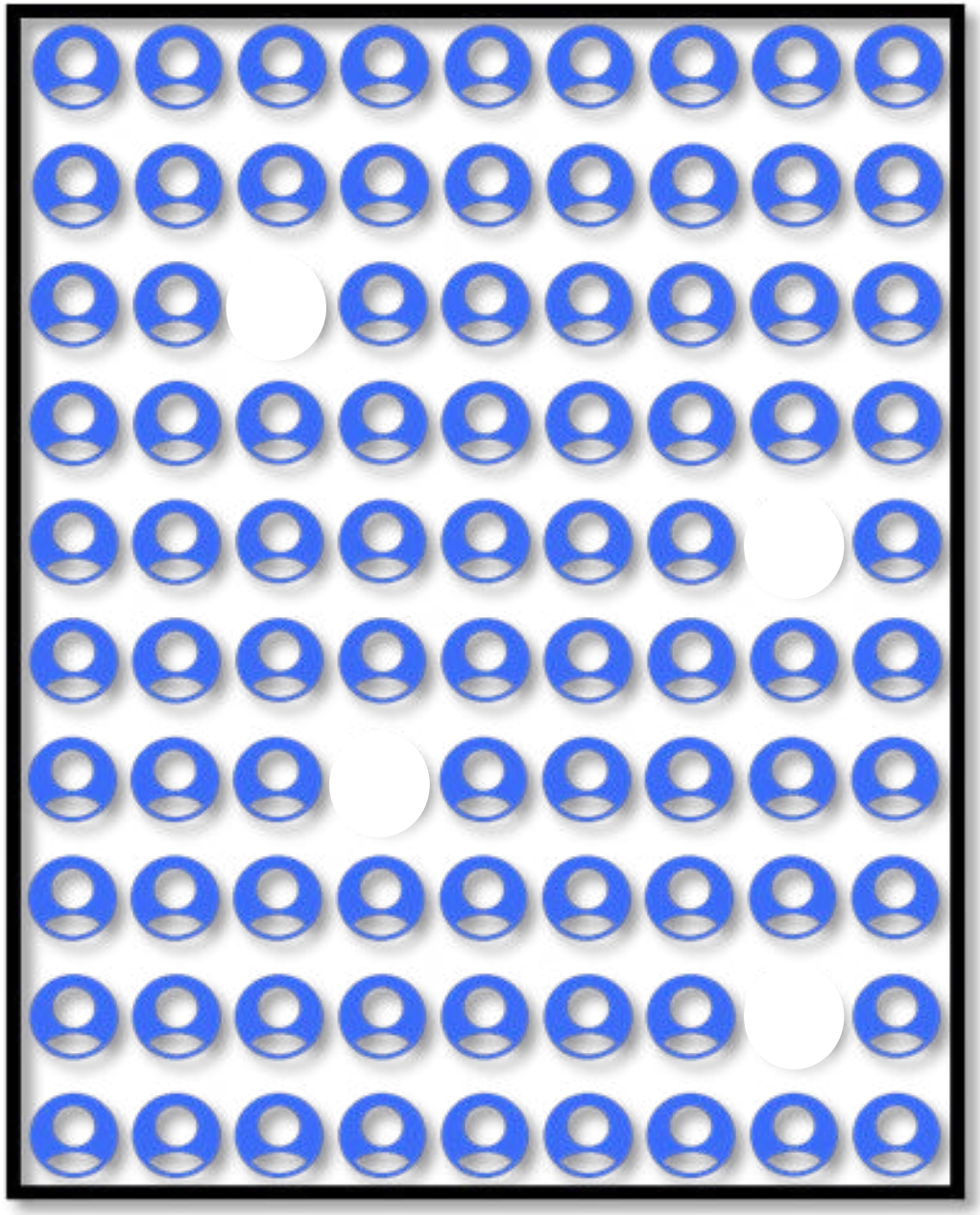
EMD Serono

Physical Function and Fatigue in Multiple Sclerosis

- 4% Made >1 ORJ
- 79%-81% None

How Do We Deal with These Errors?

Individual out-of-range judgments (but not people) are excluded from calculations of thresholds.



How Do You Derive a Meaningful Change Estimate from All This?

In the original study calculated mean and percentile ranks

Future studies should consider applying predictive modeling (See Terwee, 2021 in references)

Idioscale Judgment Method

Known Advantages

No retrospective judgment

Score changes are presented as “stories”

Explicitly separates ***noticeable*** from ***meaningful*** change

Cross-sectional method. Doesn’t require longitudinal follow-up.

Unknowns

Are some items better than others for vignettes?

What’s the stability of thresholds (test/retest and longitudinal analyses)?

What do subgroup differences tell us about judgments?

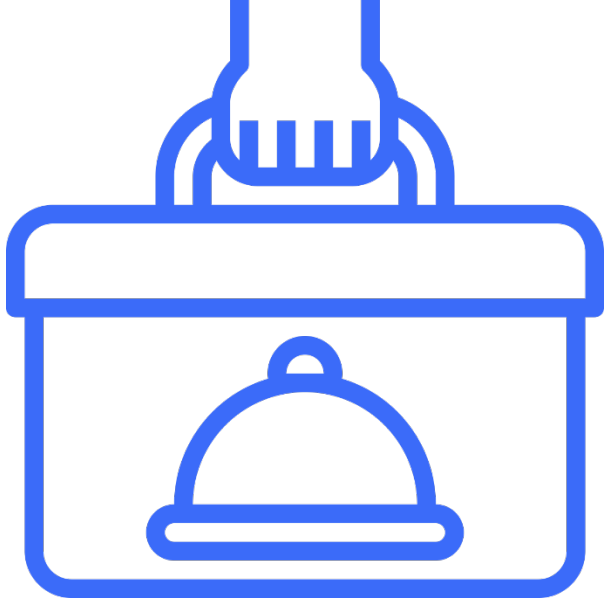
Idioscale Judgment Software

Modifications of studies easier and faster

Efficient platform for methods studies (e.g., are some items better than others for communicating scores?)

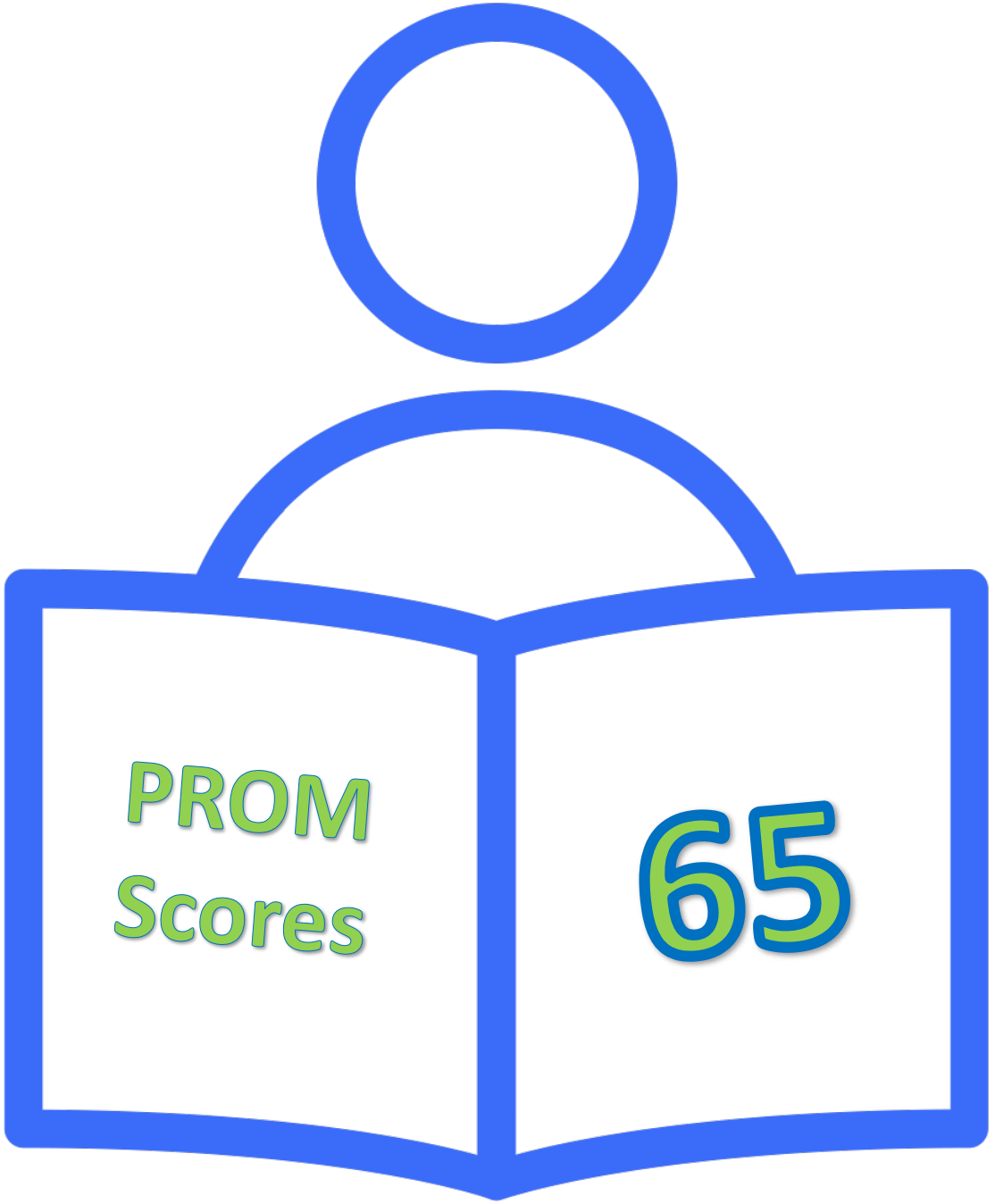
Will include tables of most probable items for PROMIS and Neuro-QOL measures

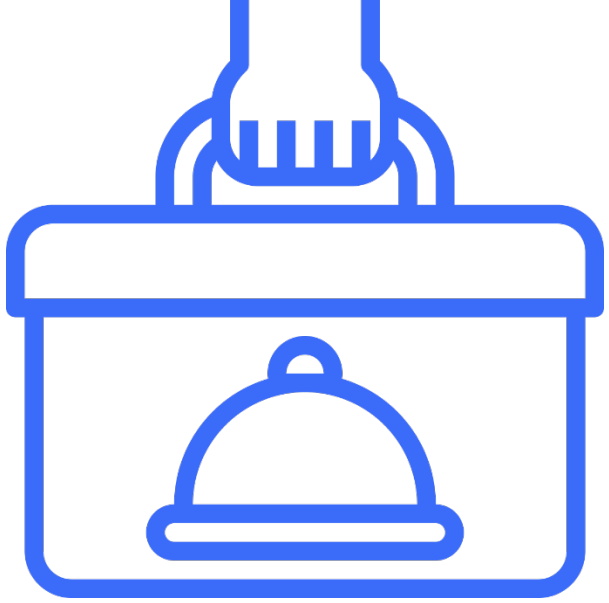
Will include sets of vignettes for Wave I PROMIS measures



Take Home Messages

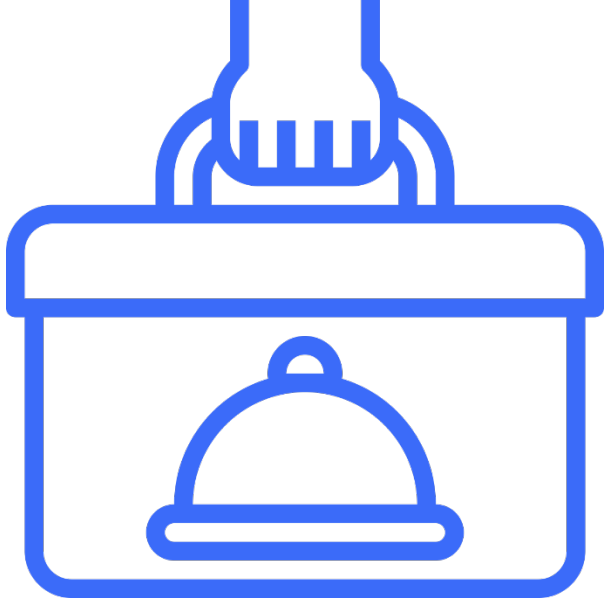
Stories are more interesting than numbers.





Take Home Messages

Idioscale Judgment is a unique method for patients to consider degrees of change.



Take Home Messages

Soon to have Shareware to support the use of Idioscale Judgment and Bookmarking

When?

Expected in early 2024

REFERENCES

Bookmarking (Non-exhaustive list, chosen to show variety)

Cohen ML, Harnish SM, Lanzi AM, Brello J, Victorson D, Kisala PA, Nandakumar R, Tulsky DS. Adapting a Patient-Reported Outcome Bookmarking Task to be Accessible to Adults With Cognitive and Language Disorders. J Speech Lang Hear Res. 2021 Nov 8;64(11):4403-4412. doi: 10.1044/2021_JSLHR-21-00071. Epub 2021 Oct 26. PMID: 34699261.

Cook KF, Cella D, Reeve BB. PRO-Bookmarking to Estimate Clinical Thresholds for Patient-reported Symptoms and Function. Med Care. 2019 May;57 Suppl 5 Suppl 1:S13-S17. doi: 10.1097/MLR.0000000000001087. PMID: 30985591.

Cook KF, Victorson DE, Cella D, Schalet BD, Miller D. Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. Qual Life Res. 2015 Mar;24(3):575-89. doi: 10.1007/s11136-014-0790-9. Epub 2014 Aug 23. PMID: 25148759.

Mann CM, Schanberg LE, Wang M, von Scheven E, Lucas N, Hernandez A, Ringold S, Reeve BB. Identifying clinically meaningful severity categories for PROMIS pediatric measures of anxiety, mobility, fatigue, and depressive symptoms in juvenile idiopathic arthritis and childhood-onset systemic lupus erythematosus. Qual Life Res. 2020 Sep;29(9):2573-2584. doi: 10.1007/s11136-020-02513-6. Epub 2020 May 14. PMID: 32410143.

Idioscale Judgment

Cook KF, Kallen MA, Coon CD, Victorson D, Miller DM. Idio Scale Judgment: evaluation of a new method for estimating responder thresholds. Qual Life Res. 2017 Nov;26(11):2961-2971. doi: 10.1007/s11136-017-1625-2. Epub 2017 Jun 17. PMID: 28624901.

Prediction Modeling

Terwee CB, Peipert JD, Chapman R, Lai JS, Terluin B, Cella D, Griffiths P, Mokkink LB. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res. 2021 Oct;30(10):2729-2754. doi: 10.1007/s11136-021-02925-y. Epub 2021 Jul 10. PMID: 34247326; PMCID: PMC8481206.

Reflections on the DIA Study Endpoints Community Working Group on Meaningful Change (Digital Technology)

Bill Byrom, PhD – Vice President, Product Intelligence and Positioning, and Principal, eCOA Science

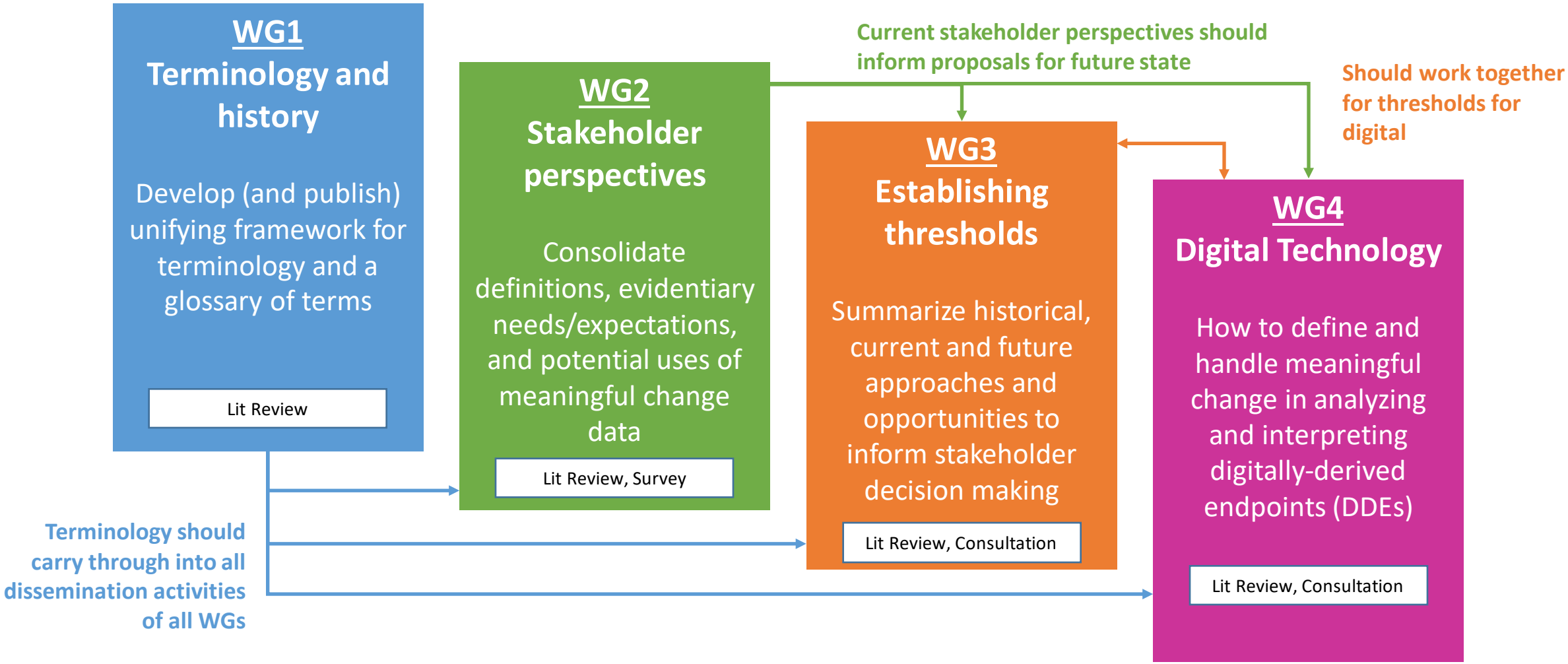
Signant Health, UK

@billbyrom

Learning objectives

- Understand the objectives and progress of the DIA Meaningful Change Working Group
  - Formed from: Study Endpoints, Statistics & Data Science, and Clinical Research Communities
- Explore challenges associated with meaningful change threshold estimation when considering endpoints derived from sensor data

Workstreams overview



Contributors

Leaders: Joan Buenconsejo, Helen Doll, Emuella Flood, Munish Mehra, Matt Reaney, Keith Wenzel

WG1: Terminology E Flood	WG2: Stakeholders B Campbell, M Reaney	WG3: Methods H Doll	WG4: Digital K Burrows, M McCarthy
Keri Brady	Veleka Allen	Rob Arbuckle	Peter Black
Nicole Clarke	Bill Byrom	Denise Bury	Joan Buenconsejo
Sonya Eremenco	Bob Campbell	Joe Cappelleri	Katie Burrows
Emuella Flood	Freda Cooner	Cheryl Coon	Joe Cappelleri
Bellinda King-Kallimanis	Stacie Hudgens	Helen Doll	Wen-Hung Chen
Jammbe Musoro	Carol Jamieson	Folke Folkvaljon	Charmaine Demanuele
Sandra Nolte	Hazel Lai	Lori McLeod	Pip Griffiths
Caroline Ward	Madhurima Majumder	Munish Mehra	Niklas Karlsson
	Nenad Medic	Lauren Nelson	Marie McCarthy
	Veronica Miller	Lindsay Petrenchik	Nikunj Patel
	Flo Mowlem	Sue Vallow	
	Kamila Novak	Diane Whalley	
	Dorothee Oberdhan		
	Matt Reaney		
	Vivian Shih		
	Glenn Vicary		
	William Wang		
	Jingjing Ye		
	Binglin Yue		

Challenges for endpoints derived from wearable/sensor data

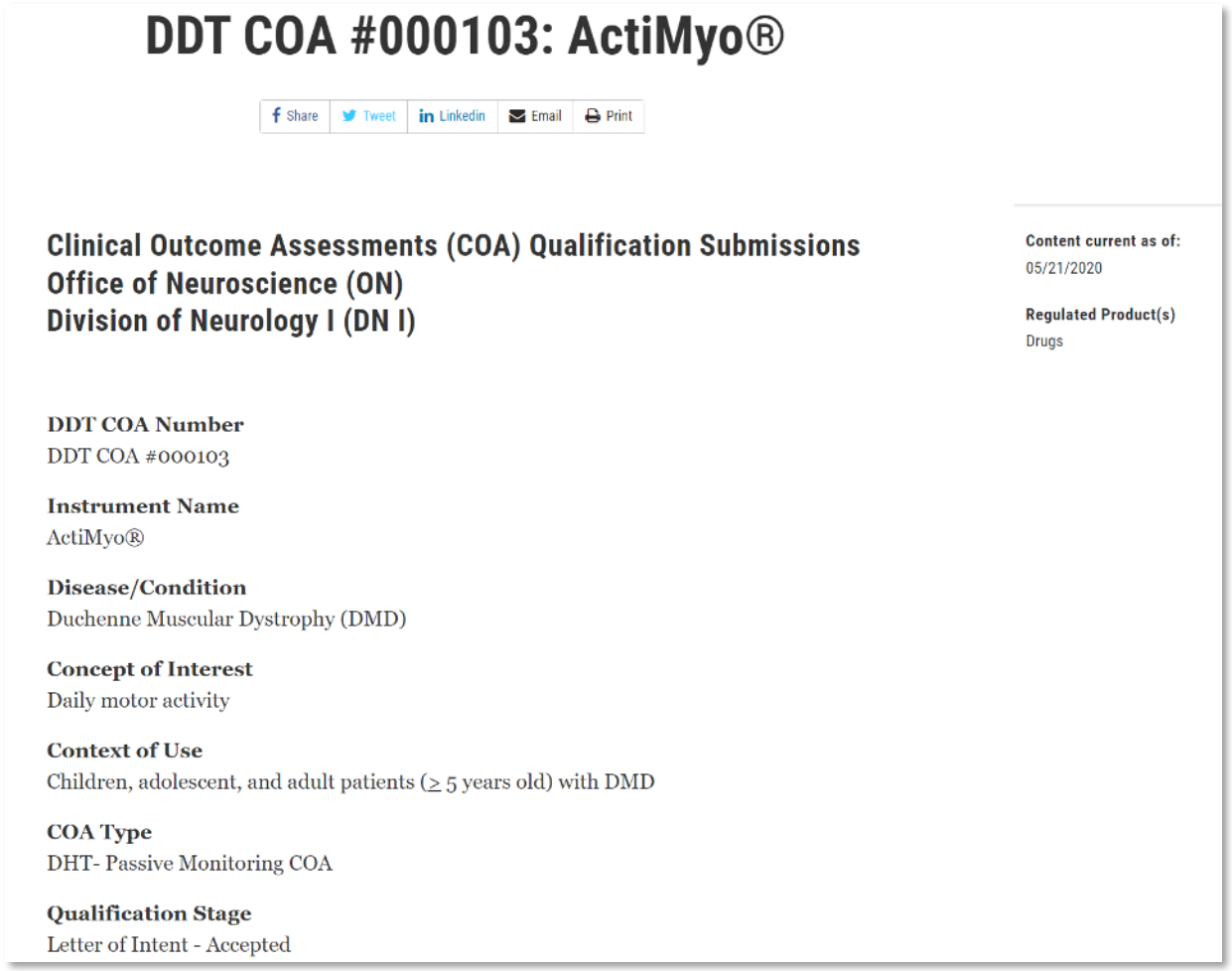
PRO CONSORTIUM
CRITICAL PATH INSTITUTE

Not new challenges, but ones that may be more pronounced with sensor-derived endpoints

1. Anchor selection – anchor scores may be less well correlated with digitally-derived measures, compared to when using the approach with patient-reported outcome measures (PROMs)
2. Endpoints may be more abstract and less easy to understand
3. Measurement comparability between sensors may impact meaningful change threshold definitions
4. Meaningful change may vary across the disease severity continuum
5. Current methodologies are imperfect

1. Anchor selection

Duchenne Muscular Dystrophy (DMD): stride length



DDT COA #000103: ActiMyo®

Share | Tweet | Linkedin | Email | Print

Clinical Outcome Assessments (COA) Qualification Submissions
Office of Neuroscience (ON)
Division of Neurology I (DN I)

Content current as of:
05/21/2020

Regulated Product(s)
Drugs

DDT COA Number
DDT COA #000103

Instrument Name
ActiMyo®

Disease/Condition
Duchenne Muscular Dystrophy (DMD)

Concept of Interest
Daily motor activity

Context of Use
Children, adolescent, and adult patients (≥ 5 years old) with DMD

COA Type
DHT- Passive Monitoring COA

Qualification Stage
Letter of Intent - Accepted

> "Your proposed minimally clinically important differences based on the standard deviation provides **only supportive information** as it does not directly convey the interpretation of meaningfulness (e.g., whether a 1.8 centimeter change in median stride length is a meaningful change to the patients)"

FDA Submission Decision and Recommendations
30-Aug-2018

- Typical performance outcome (PerfO) assessments in DMD
  - Six-minute walking test (6MWT)
  - Four stairs climbing test (4SCT)
  - North Star Ambulation Assessment (NSSA)

Example: Meaningful change in total steps per day in multiple sclerosis (MS)

OPEN ACCESS Freely available online

PLOS ONE

Clinical Importance of Steps Taken per Day among Persons with Multiple Sclerosis

Robert W. Motl[1]*, Lara A. Pilutti[1], Yvonne C. Learmonth[1], Myla D. Goldman[2], Ted Brown[3]

1 Department of Kinesiology and Community Health, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, 2 Department of Neurology, University of Virginia, Charlottesville, Virginia, United States of America, 3 MS Center at Evergreen, Evergreen Health, Evergreen, Washington, United States of America

Abstract

Background: The number of steps taken per day (steps/day) provides a reliable and valid outcome of free-living walking behavior in persons with multiple sclerosis (MS).

Objective: This study examined the clinical meaningfulness of steps/day using the minimal clinically important difference (MCID) value across stages representing the developing impact of MS.

Methods: This study was a secondary analysis of de-identified data from 15 investigations totaling 786 persons with MS and 157 healthy controls. All participants provided demographic information and wore an accelerometer or pedometer during the waking hours of a 7-day period. Those with MS further provided real-life, health, and clinical information and completed the Multiple Sclerosis Walking Scale-12 (MSWS-12) and Patient Determined Disease Steps (PDDS) scale. MCID estimates were based on regression analyses and analysis of variance for between group differences.

Results: The mean MCID from self-report scales that capture subtle changes in ambulation (1-point change in PDSS scores and 10-point change in MSWS-12 scores) was 779 steps/day (14% of mean score for MS sample); the mean MCID for clinical/health outcomes (MS type, duration, weight status) was 1,455 steps/day (26% of mean score for MS sample); real-life anchors (unemployment, divorce, assistive device use) resulted in a mean MCID of 2,580 steps/day (45% of mean score for MS sample); and the MCID for the cumulative impact of MS (MS vs. control) was 2,747 steps/day (48% of mean score for MS sample).

Conclusion: The change in motion sensor output of ~800 steps/day appears to represent a lower-bound estimate of clinically meaningful change in free-living walking behavior in interventions of MS.

Citation: Motl RW, Pilutti LA, Learmonth YC, Goldman MD, Brown T (2013) Clinical Importance of Steps Taken per Day among Persons with Multiple Sclerosis. PLoS ONE 8(9): e73247. doi:10.1371/journal.pone.0073247

Editor: Pablo Villoslada, Institute Biomedical Research August Pi Sunyer (IDIBAPS) – Hospital Clinic of Barcelona, Spain

Received April 1, 2013; **Accepted** July 18, 2013; **Published** September 4, 2013

Funding: This research was supported by an investigator-initiated grant from Acorda Therapeutics. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The funders did receive and read a copy of the paper and provided approval before submission.

Competing Interests: Acorda Therapeutics funded this study. RWM is a consultant for Acorda Therapeutics and Biogen Idec. LAP and YCL report no conflicts of interest. MDG has served as a consultant for Acorda Therapeutics and Novartis Pharmaceuticals, but does not report any current conflicts of interest. TB is a consultant for Acorda Therapeutics, Biogen Idec, Genzyme, Pfizer, and Teva. There are no patents, products in development or marked products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: robmotl@illinois.edu.

Introduction

There has been an ongoing debate regarding outcome measures in clinical research involving persons with multiple sclerosis (MS) [1], with increasing interest in approaches for objectively monitoring patients under real-world conditions [2]. This interest has highlighted the potential for the objective monitoring of free-living walking behavior using motion sensors such as accelerometers and pedometers in clinical research involving persons with neurologic diseases [3] including MS [2]. Such devices are worn around the waist or ankle during the waking hours of the day and over a representative sampling period (e.g., seven days). The motion sensors capture the total amount of walking undertaken in free-living conditions based on metrics such as steps taken per day (steps/day). The number of steps/day reflects a straight-forward metric of the overall amount of walking undertaken during one's everyday life, representing free-living walking behavior [2,3].

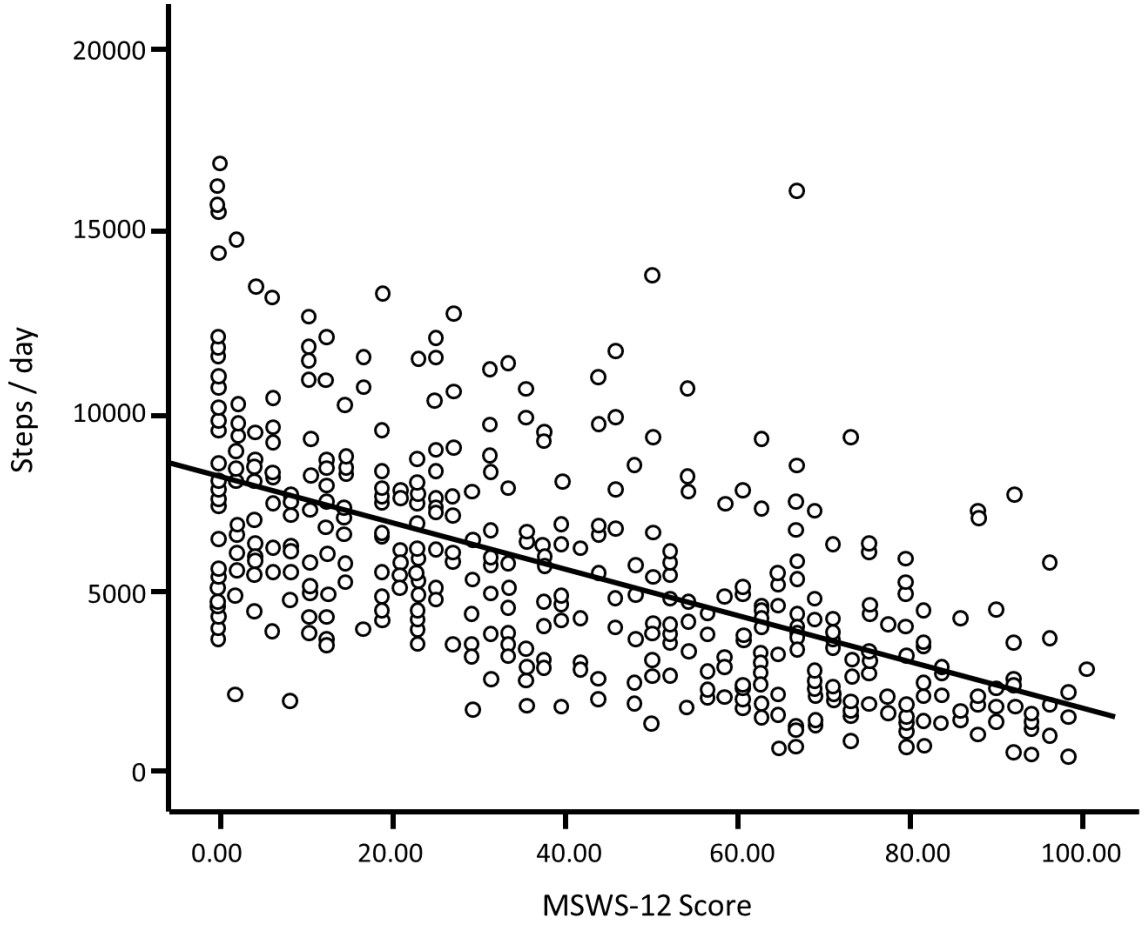
Accumulating data demonstrates that the number of steps/day provides a reliable and valid measure of free-living walking behavior in MS [4–7]. Steps/day has demonstrated acceptable test-retest reliability over a two-week time period in persons with MS [4], and as few as three days of data with an appropriate amount of wear time (i.e., 10 or more hours/day) yields a reliable estimate of usual ambulatory-based behavior [5]. Regarding validity, steps/day has correlated strongly with clinical (e.g., Expanded Disability Status Scale scores), performance (e.g., timed 25-foot walk and 6-minute walk), and patient-reported (e.g., 12-Item Multiple Sclerosis Walking Scale or MSWS-12 scores) measures of ambulation in persons with MS [6,7]. To date, there are no published data evaluating the clinical importance of differences in steps/day among those with MS (i.e., amount of difference in steps/day that actually reflects a difference) on a group level. Such minimal clinically important difference (MCID) values are necessary for designing and interpreting clinical trials

- 786 MS patients
- 157 healthy controls
- 3 – 7 days activity data (steps/day)
  - Yamax SW-200 pedometer

- Anchors
  - Multiple Sclerosis Walking Scale (MSWS)-12, a 12-item PRO measure assessing the impact of MS on walking-related activities
  - Patient-Determined Disease Steps (PDDS) scale

Example: Meaningful change in total steps per day in MS

Anchor measure 1: Multiple Sclerosis Walking Scale (MSWS-12)



- Between-group meaningful change threshold (MSWS-12) = 10
- Corresponds to 642 steps/day

Example: Meaningful change in total steps per day in MS

Anchor measure 2: Patient-determined disease steps scale (PDDS)

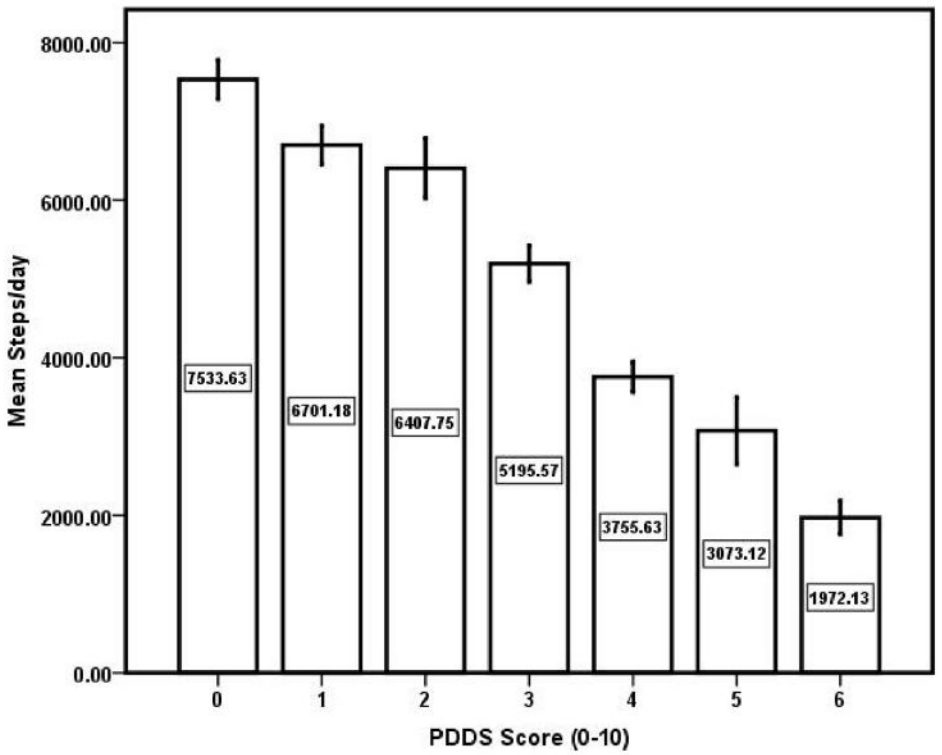


Figure 2. Bar graph of the association between Patient Determined Disease Steps (PDDS) scale scores and steps/day in persons with multiple sclerosis. The number within the bars represents the mean score for steps/day per level of the PDDS.
doi:10.1371/journal.pone.0073247.g002

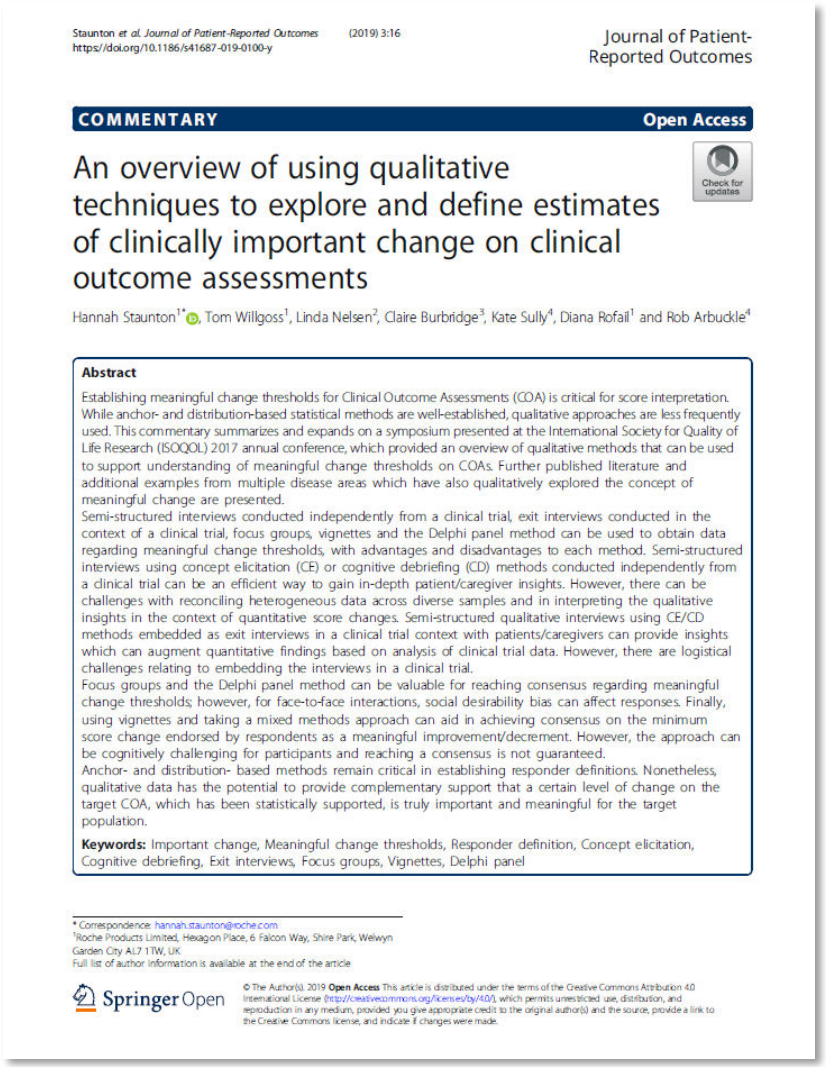
- Between-group meaningful change threshold (PDDS) = 1 point change
- Corresponds to 915 steps/day

- **Overall between-group meaningful change threshold = 779 steps**

(average of the threshold estimates from the 2 anchors: 642 and 915 steps per day)

2. Abstract endpoints

Qualitative methods



Staunton et al. Journal of Patient-Reported Outcomes (2019) 3:16
https://doi.org/10.1186/s41687-019-0100-y

Journal of Patient-Reported Outcomes

COMMENTARY Open Access

An overview of using qualitative techniques to explore and define estimates of clinically important change on clinical outcome assessments

Hannah Staunton[1*], Tom Willgoss[1], Linda Nelsen[2], Claire Burbridge[3], Kate Sully[4], Diana Rofail[1] and Rob Arbuckle[4]

Abstract

Establishing meaningful change thresholds for Clinical Outcome Assessments (COA) is critical for score interpretation. While anchor- and distribution-based statistical methods are well-established, qualitative approaches are less frequently used. This commentary summarizes and expands on a symposium presented at the International Society for Quality of Life Research (ISOQOL) 2017 annual conference, which provided an overview of qualitative methods that can be used to support understanding of meaningful change thresholds on COAs. Further published literature and additional examples from multiple disease areas which have also qualitatively explored the concept of meaningful change are presented.

Semi-structured interviews conducted independently from a clinical trial, exit interviews conducted in the context of a clinical trial, focus groups, vignettes and the Delphi panel method can be used to obtain data regarding meaningful change thresholds, with advantages and disadvantages to each method. Semi-structured interviews using concept elicitation (CE) or cognitive debriefing (CD) methods conducted independently from a clinical trial can be an efficient way to gain in-depth patient/caregiver insights. However, there can be challenges with reconciling heterogeneous data across diverse samples and in interpreting the qualitative insights in the context of quantitative score changes. Semi-structured qualitative interviews using CE/CD methods embedded as exit interviews in a clinical trial context with patients/caregivers can provide insights which can augment quantitative findings based on analysis of clinical trial data. However, there are logistical challenges relating to embedding the interviews in a clinical trial.

Focus groups and the Delphi panel method can be valuable for reaching consensus regarding meaningful change thresholds; however, for face-to-face interactions, social desirability bias can affect responses. Finally, using vignettes and taking a mixed methods approach can aid in achieving consensus on the minimum score change endorsed by respondents as a meaningful improvement/decrement. However, the approach can be cognitively challenging for participants and reaching a consensus is not guaranteed.

Anchor- and distribution- based methods remain critical in establishing responder definitions. Nonetheless, qualitative data has the potential to provide complementary support that a certain level of change on the target COA, which has been statistically supported, is truly important and meaningful for the target population.

Keywords: Important change, Meaningful change thresholds, Responder definition, Concept elicitation, Cognitive debriefing, Exit interviews, Focus groups, Vignettes, Delphi panel

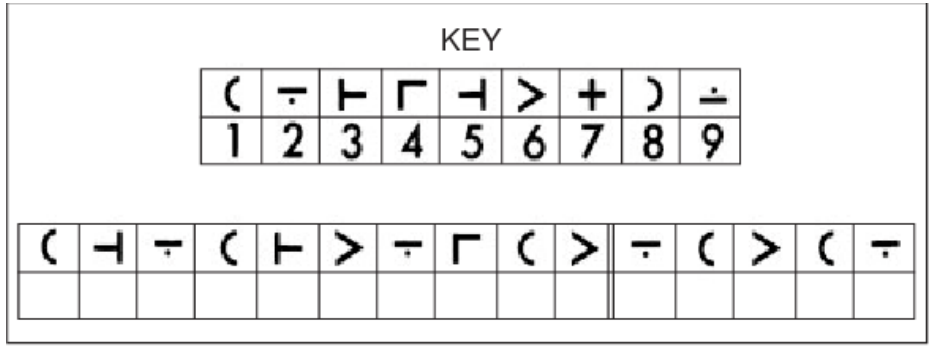
* Correspondence: hannah.staunton@roche.com
[1]Roche Products Limited, Hexagon Place, 6 Falcon Way, Shire Park, Welwyn Garden City AL7 1TW, UK
Full list of author information is available at the end of the article

Approaches

- Semi-structured interviews external to a clinical trial
  - Insights into what an important change might be
- Semi-structured exit interviews as part of a clinical trial
- Focus groups
- Vignettes (bookmarking / standard setting)
  - Use of hypothetical vignettes to reach consensus on thresholds for meaningful change
- Delphi Panel

ISPOR PerfO Assessment Task Force (currently active)

- **Indirect** vs **Direct** association with meaningful aspect of health
- Example:

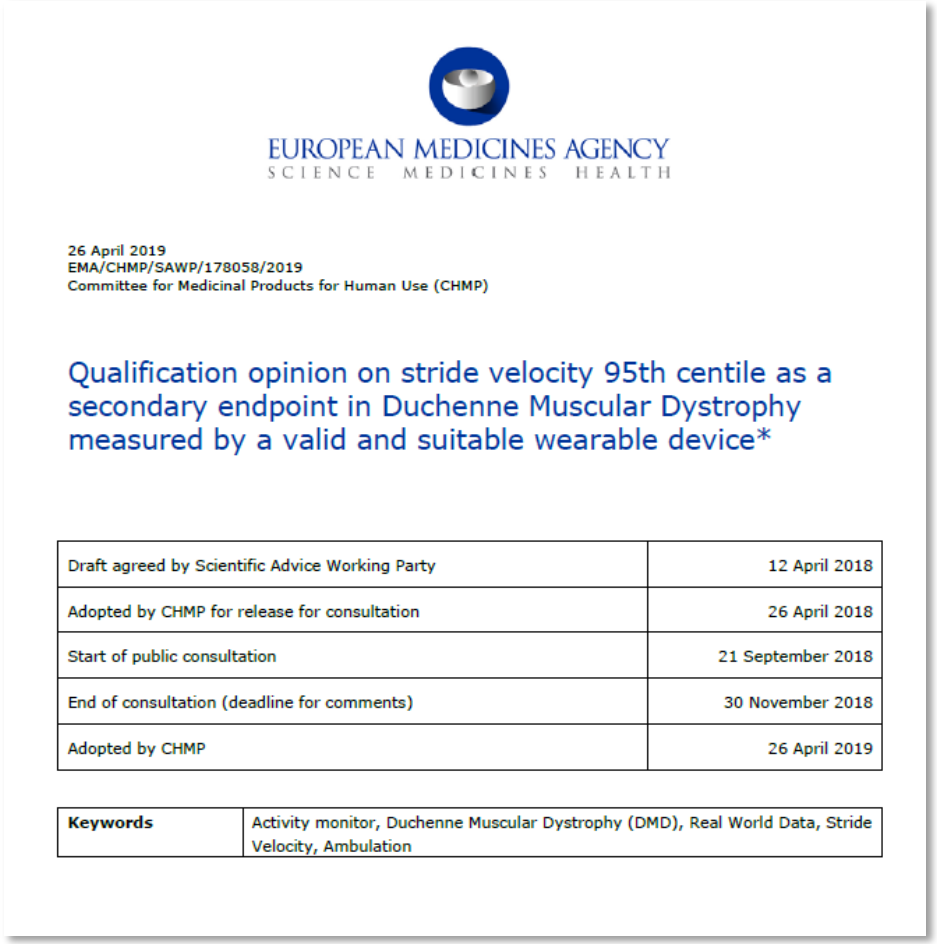


  - Symbol Digit Modalities Test (SDMT) to assess "processing speed" in multiple sclerosis (MS)
  - Not an activity performed in real life
  - May be difficult for patents and caregivers to have insight into the concept of processing speed

CONTRIBUTORS

- Chris J. Edgar, Cogstate Ltd.
- Elizabeth (Nicki) Bush, Janssen
- Heather R. Adams, University of Rochester
- Rachel Ballinger, ICON
- Bill Byrom, Signant Health
- Michelle Campbell, FDA
- Sonya Eremenco, Critical Path Institute
- Fiona McDougall, Genentech
- Elektra Papadopoulos, AbbVie
- Ashley F. Slagle, Aspen Consulting LLC
- Stephen Joel Coons, Critical Path Institute

DMD: Stride velocity 95th centile (SV95C)



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

26 April 2019
EMA/CHMP/SAWP/178058/2019
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion on stride velocity 95th centile as a secondary endpoint in Duchenne Muscular Dystrophy measured by a valid and suitable wearable device*

Draft agreed by Scientific Advice Working Party	12 April 2018
Adopted by CHMP for release for consultation	26 April 2018
Start of public consultation	21 September 2018
End of consultation (deadline for comments)	30 November 2018
Adopted by CHMP	26 April 2019

Keywords	Activity monitor, Duchenne Muscular Dystrophy (DMD), Real World Data, Stride Velocity, Ambulation

“CHMP considers that for ambulant Duchenne Muscular Dystrophy (DMD) patients 5 years of age and above:

Stride velocity 95th centile measured at the ankle (SV95C) is an **acceptable secondary endpoint** in pivotal or exploratory drug therapeutic studies for regulatory purposes when measured by a valid and suitable wearable device to quantify a patient’s ambulation ability directly and reliably in a continuous manner in a home environment and as an indicator of maximal performance”

EMA, Qualification Opinion, 26 April, 2019

Qualitative methodology challenges

- PROMs
  - Total score measures may be too abstract
  - Combine item-level estimates of meaningful change

- 95th centile of stride velocity may be very abstract for patient to understand
  - Stride velocity – the speed at which you can move your foot to place ahead of you when taking a step
  - 95th centile: the fastest times you do this
  - What makes a meaningful difference: 0.1 m/s, 0.05 m/s, etc?

3. Measurement comparability

Device agnostic thresholds?

This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

Measuring the Fitness of Fitness Trackers

Chelsea G. Bender, Jason C. Hoffstot
Brian T. Combs and Sara Hooshangi
Integrated Information, Science, and Technology
The George Washington University
Washington, DC, USA
Email: shoosh@gwu.edu

Justin Cappos
Computer Science and Engineering
New York University
New York, NY, USA
Email: jcappos@nyu.edu

***Abstract*—Data collected by fitness trackers could play an important role in improving the health and well-being of the individuals who wear them. Many insurance companies even offer monetary rewards to participants who meet certain steps or calorie goals. However, in order for it to be useful, the collected data must be accurate and also reflect real-world performance. While previous studies have compared step counts data in controlled laboratory environments for limited periods of time, few studies have been done to measure performance over longer periods of time, while the subject does real-world activities. There are also few direct comparisons of a range of health indicators on different fitness tracking devices. In this study, we compared step counts, calories burned, and miles travelled data collected by three pairs of fitness trackers over a 14-day time period in free-living conditions. Our work indicates that the number of steps reported by different devices worn simultaneously could vary as much as 26%. At the same time, the variations seen in distance travelled, based on the step count, followed the same trends. Little correlation was found between the number of calories burned and the variations seen in the step count across multiple devices. Our results demonstrate that the reporting of health indicators, such as calories burned and miles travelled, are heavily dependent on the device itself, as well as the manufacturer's proprietary algorithm to calculate or infer such data. As a result, it is difficult to use such measurements as an accurate predictor of health outcomes, or to develop a consistent criteria to rate the performance of such devices in head-to-head comparisons.**

Keywords—Fitness Trackers; Accuracy; Physical Activity; Free-living Conditions

I. Introduction

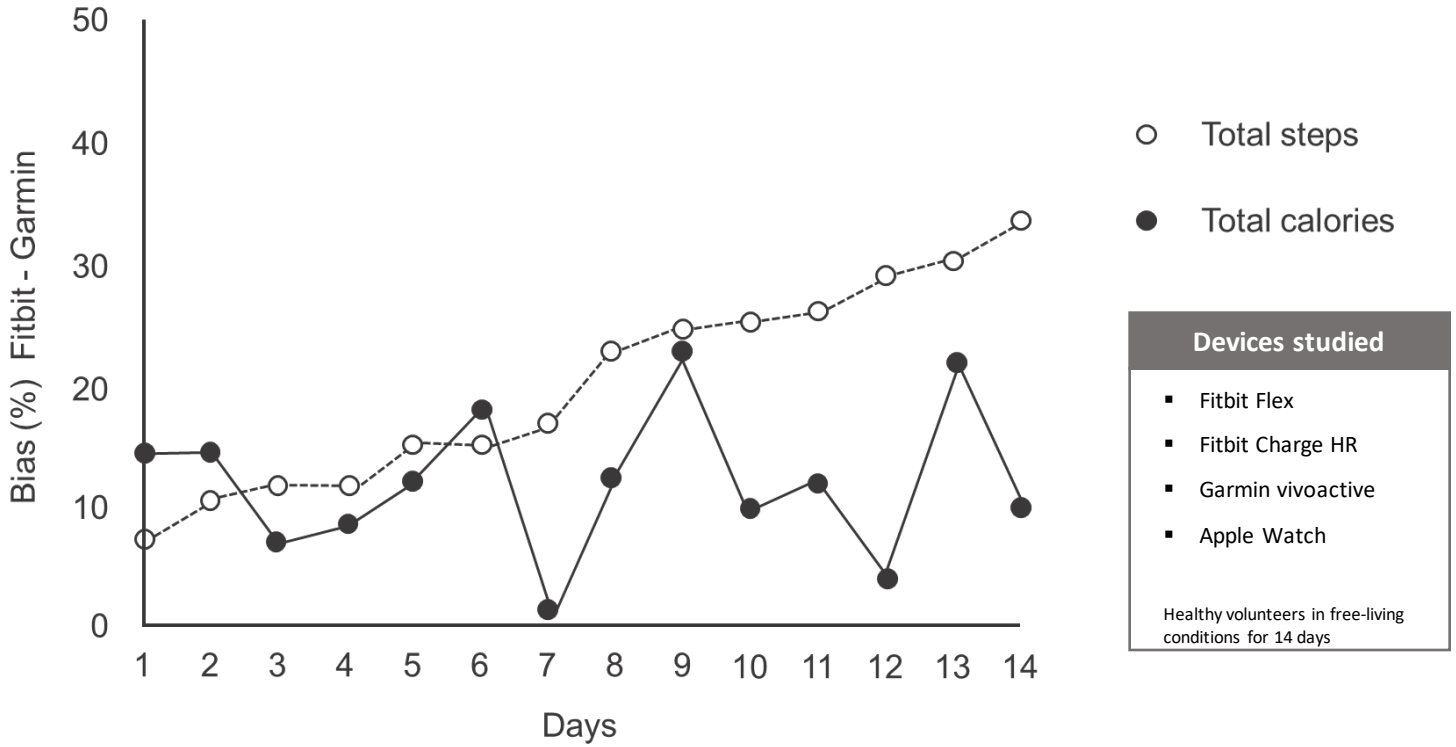
The past several years have seen an exponential growth in the market for personal wearable devices, with estimated sales of up to 126 million units anticipated by the end of 2019 [1]. Fitness tracking devices lead sales in this market, and continue to gain popularity as the correlation between an active lifestyle and the prevention of chronic diseases is demonstrated by research [2], [3]. These trackers give their users the ability to monitor and track key health markers, thus encouraging them to continue their healthy efforts.

As manufacturers try to improve the accuracy of these health measurements by adding functionality and introducing new devices into the market place at a rapid pace, the number of ways that this collected and stored data can be used also increases. Individuals can use data on their average daily/weekly physical activity to monitor their own health, or to identify key markers to report to their health providers. Public health researchers could use such data in aggregated form, in large-scale studies to monitor health related outcomes for different segments of the population. And, on a larger scale, programs sponsored by insurance companies can promote healthier lifestyles by offering incentivizing discounts on life and health insurance products based on the physical activity levels of consumers.

Such programs, however, rely on the ability of these devices to reliably generate accurate data. Data accuracy ultimately depends on two factors: the quality of the sensors embedded in the device, and the algorithm used to interpret the raw data. To this end, there has been a surge in the number of research studies testing the accuracy of wearable fitness devices as compared to research-grade accelerometers and multi-sensor devices [4]–[11]. Most of these studies have focused on a cross-sectional comparison of consumer-based products to research-grade gold standards only in a laboratory or a controlled real-world environment [4]–[7], [12]. Conducting experiments without the prescribed restrictions of a laboratory (i.e. under a free-living condition) is significantly more challenging, as the variations in speed, direction and intensity of physical activities are larger. This may be why only a few studies have measured the accuracy of trackers in free-living conditions [8]–[10] and most free-living studies have been short in duration, typically in the range of one or two days. Furthermore, the integrity of these results could also be compromised if the subjects under study (who are often volunteers) do not follow the experiment protocols.

In this work, we set out to compare parameters and experimental settings that have not been explored in previous work. We start by looking at other health indicators measured by these devices, such as calories burned or distance travelled. We designed a series of experiments to compare these along with the more commonly studied measure of step counts. While the step count provides a general sense of movement and physical activity, calories burned and the number of miles travelled could be better indicators of an individual's energy expenditure and, hence his/her physical fitness level. If the fitness trackers are to become an integral part of our health-

Bender CG, Hoffstot JC, Combs BT et al. Measuring the Fitness of Fitness Trackers. Sensors Applications Symposium 2017 IEEE, pp. 1-6, 2017, March.



> "Step count, distance travelled, and calories burned could vary significantly between devices used concurrently."

4. Severity-dependent thresholds

Variable meaningful change thresholds

SAGE-Hindawi Access to Research
International Journal of Inflammation
Volume 2011, Article ID 231926, 6 pages
doi:10.4061/2011/231926

Review Article

Definition of Nonresponse to Analgesic Treatment of Arthritic Pain: An Analytical Literature Review of the Smallest Detectable Difference, the Minimal Detectable Change, and the Minimal Clinically Important Difference on the Pain Visual Analog Scale

Melissa E. Stauffer,[1] Stephanie D. Taylor,[2] Douglas J. Watson,[2] Paul M. Peloso,[2] and Alan Morrison[1]

[1] *Scribco Pharmaceutical Writing, P.O. Box 1525, Blue Bell, PA 19422, USA*
[2] *Departments of Global Health Outcomes, Epidemiology, and Clinical Development, Merck & Co., Inc., One Merck Drive, Whitehouse Station, NJ 08889, USA*

Correspondence should be addressed to Melissa E. Stauffer, melstauff@charter.net

Received 25 January 2011; Accepted 8 March 2011

Academic Editor: Bernhard Rintelen

Our objective was to develop a working definition of nonresponse to analgesic treatment of arthritis, focusing on the measurement of pain on the 0–100 mm pain visual analog scale (VAS). We reviewed the literature to assess the smallest detectable difference (SDD), the minimal detectable change (MDC), and the minimal clinically important difference (MCID). The SDD for improvement reported in three studies of rheumatoid arthritis was 18.6, 19.0, and 20.0. The median MDC was 25.4 for 7 studies of osteoarthritis and 5 studies of rheumatoid arthritis (calculated for a reliability coefficient of 0.85). The MCID increased with increasing baseline pain score. For baseline VAS tertiles defined by scores of 30–49, 50–65, and >65, the MCID for improvement was, respectively, 7–11 units, 19–27 units, and 29–37 units. Nonresponse can thus be defined in terms of the MDC for low baseline pain scores and in terms of the MCID for high baseline scores.

1. Introduction

Nonsteroidal anti-inflammatory drugs (NSAIDs) are the first-line treatment for osteoarthritis [1] and a cornerstone of pharmacologic management of other arthritic and rheumatologic illnesses [2]. The dichotomous classification of patients into responders and nonresponders to NSAIDs began in the 1970s [3]. Walker et al. provided evidence of the validity of the responder/nonresponder thesis in rheumatoid arthritis and osteoarthritis [4].

Both the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) have developed response criteria for rheumatoid arthritis. The ACR criteria are based on a core set of measures with 7 components [5]. EULAR response criteria are based on the multiitem Disease Activity Score 28 (DAS28) [6]. A composite index of patient-reported pain, physical function, and global assessment appears to be as good as the entire ACR core set or the DAS28 for determining response to treatment in clinical trials of rheumatoid arthritis [7].

The most recent definition of response to NSAID treatment in osteoarthritis was determined by a joint task force comprising members of the Outcome Measures in Rheumatology (OMERACT) committee and the Osteoarthritis Research Society International (OARSI) [8]. In the OMERACT-OARSI guidelines, response is a binary, composite endpoint based on three patient-reported core outcome measures: pain, physical function, and the patient global assessment. Of these three, pain is considered the primary outcome measure of interest [9].

In clinical trials, definitions of the effectiveness of analgesic therapies for arthritic pain are focused on response to treatment. In clinical practice, however, decisions to change the analgesic dose or drug for an individual patient

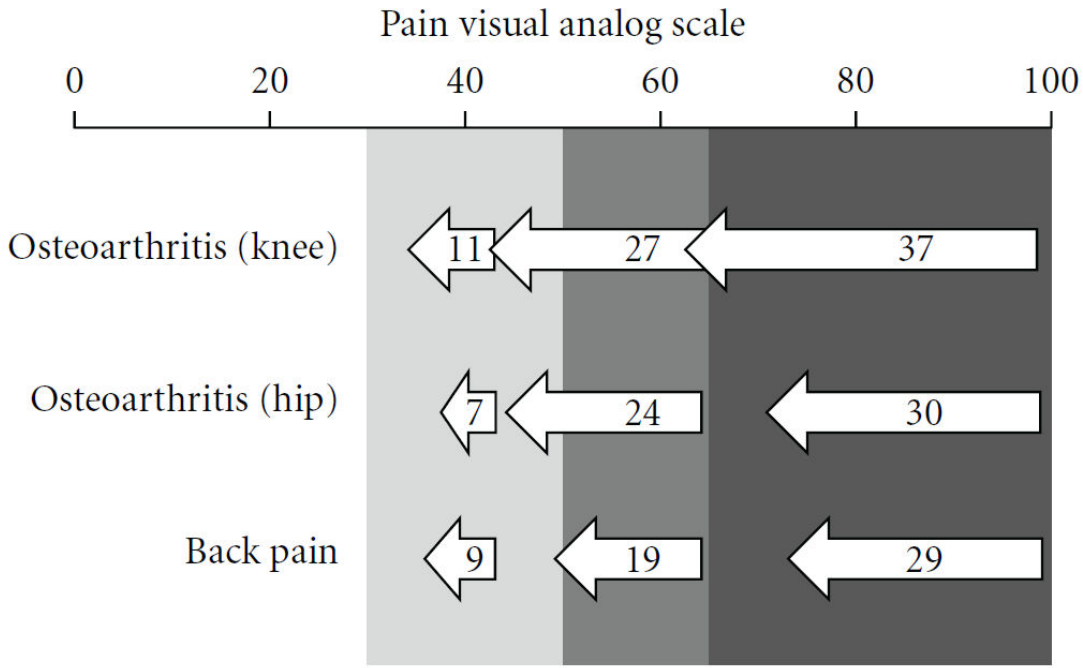


FIGURE 2: Minimal clinically important difference by tertile of baseline VAS score. The light, medium, and dark gray sections of the VAS (30–49, 50–65, and >65) correspond to the rounded tertiles of baseline pain VAS scores reported by Tubach et al. [30]. Arrows depict the MCID for improvement reported by Tubach et al. (osteoarthritis of knee and hip) [30] and Hägg et al. (back pain) [18].

5. Current methodologies imperfect

Imperfect methods

1. Anchors: How robust are the meaningful change thresholds for anchor measures?
2. Distribution methods: distributional assumptions fail to represent patient views.
   - Note: German Institute for Quality and Efficiency in Healthcare (IQWiG) 15% threshold guidance
3. Qualitative methods: abstractness of measures can make these approaches challenging.

Triangulation recommended to mitigate limitations in individual approaches, where they occur.

DIA Next Steps: Survey response request

The DIA Meaningful Change Working Group Survey

Objectives:

- Evaluate understanding and definitions of meaningful change among different stakeholders in the healthcare arena
  - Regulators, payers, healthcare professionals, sponsors, patients, and caregivers.
- DIA is now circulating the survey to various communities to ask members to complete it
  - The survey will take about 15 minutes for you to complete. Your responses are anonymous.
- Our working group would be very grateful if you would be willing to complete it.

www.surveymonkey.com/r/983NJFJ

Conclusion

- Meaningful within-person change thresholds are essential to enable robust inferences from trial data
- The challenges in identifying meaningful change thresholds may be more pronounced when using sensor data and digitally-derived endpoints
- Triangulation of methods is essential to mitigate shortcomings of individual approaches
- More research and development into novel methods is encouraged

Panel Discussion and Q&A

Moderator

- *Rebecca M. Speck, PhD, MPH* – Clinical Outcome Assessment Scientist, Eli Lilly and Company

Presenters

- *Elizabeth (Nicki) Bush, MHS* – Senior Director, Endpoints and Measurement Strategy, Janssen Pharmaceutical Companies of Johnson & Johnson
- *Devin Peipert, PhD* – Assistant Professor of Medical Social Sciences, Northwestern University
- *Karon Cook, PhD* – Research Professor (Retired), Feinberg School of Medicine, Northwestern University
- *Bill Byrom, PhD* – Vice President, Product Intelligence and Positioning, and Principal, eCOA Science, Signant Health, UK

Additional Panelists

- *Selena Daniels, PharmD, PhD* – Clinical Outcome Assessment Team Leader, Division of Clinical Outcome Assessment, Office of Drug Evaluation Sciences, Office of New Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration
- *Monica Morell, PhD* – Patient-Focused Statistical Support Reviewer, Division of Biometrics III, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Thank you!
Day 1 Wrap Up