Guidance for Clinicians Wondering Whether HRQOL of an Individual Patient has Changed

Ron D. Hays, Ph.D.

October 24, 2023 (9:00-10:00 am)

PROMIS Health Organization Annual Meeting

Banff, Canada

Basic Premises and Terminology

- Statistically significant group mean difference may be trivial
  - MCID: Minimal or minimally clinically important difference
    - MID: Minimal or minimally important difference
    - MIC: Minimal or minimally important change
  - Meaningful change
    - An obviously important difference
- Individual change significant at p <.05 is always important
  - Coefficient of repeatability
  - Minimally detectable change
  - Smallest real difference
  - Smallest detectable change

Health-Related Quality of Life (HRQOL) Uses

- Research: Randomized clinical trials and observational studies
- Quality improvement
- Public reporting (e.g., CDC)
- Certification and recognition (NCQA)
- Value-based purchasing

Health-Related Quality of Life (HRQOL) Uses

- Research: Clinical trials and observational studies
- Quality improvement
- Public reporting
- Certification and recognition
- Value-based purchasing

Obstacles to Health Status Assessment in Ambulatory Settings

- How to fit it into brief patient encounters?
- When do I measure it?
- What do I do with the information?
- What difference does it make?

Wasson, Keller, Rubenstein et al. (1992), *Medical Care*

"Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations"

- HRQOL
  - is increasingly collected in a standardized fashion in routine clinical practice.
  - improves patient-clinician communication
  - may improve HRQOL
    - Velikova et al. (2004, J Clin Oncol)
    - Basch et al., (2017, JAMA)

Snyder, C.F., Aaronson, N. K., et al. (2012). *Quality of Life Research*, 21, 1305-1314.

Is Receiving Better Technical Quality of Care Bad for Health?

Change in SF-12 PCS regressed on process-of-care aggregate

Hypothesized positive effect, but regression coefficient was in the WRONG DIRECTION and not statistically significant:

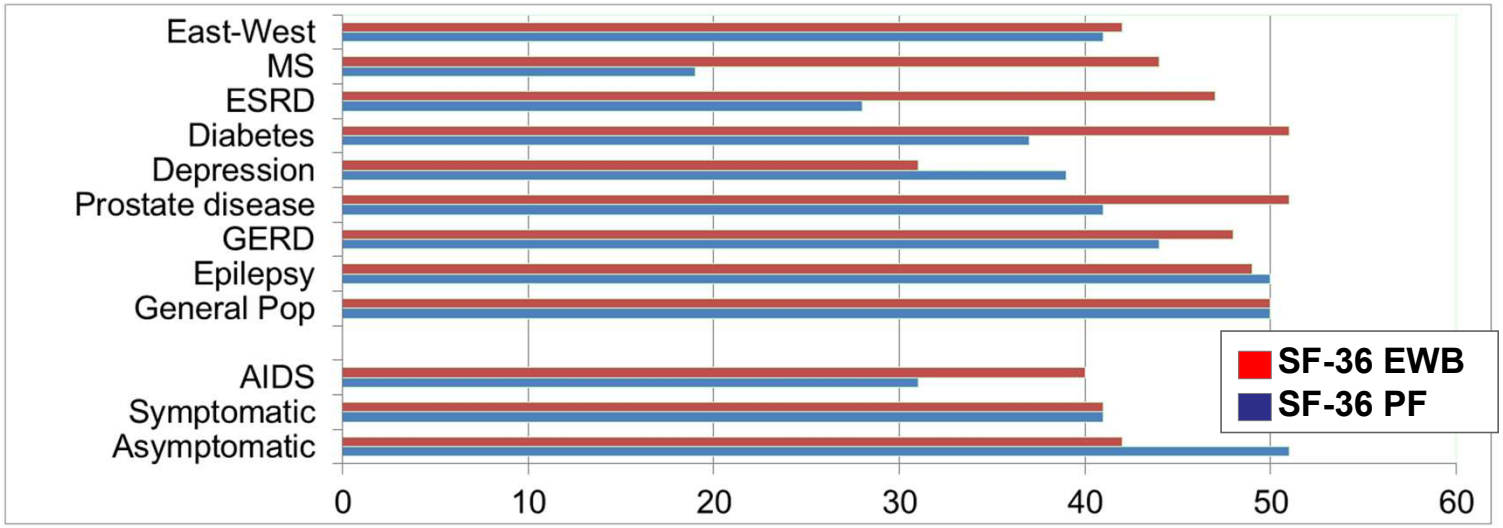
unstandardized beta = -1.41, p =.188

Kahn et al. (2007), *Health Services Research*, Article of Year

Individual Change in HRQOL

- Research: Clinical trials and observational studies
  - Supplement group mean differences
  - But is rarely reported.

Physical Functioning (PF) and Emotional Well-Being (EWB) at Baseline for 54 Patients at UCLA-Center for East West Medicine



MS = multiple sclerosis; ESRD = end-stage renal disease; GERD = gastroesophageal reflux disease.

Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K. (2005). Evaluating the statistical significance of health-related quality of life change in individual patients. *Evaluation and the Health Professions*, 28, 160-171.

Significant Mean improvement 6-weeks later in all but 1 SF-36 Scale

	Change	t-test	prob.
PF-10	1.7	2.38	.0208
RP-4	4.1	3.81	.0004
BP-2	3.6	2.59	.0125
GH-5	2.4	2.86	.0061
EN-4	5.1	4.33	.0001
SF-2	4.7	3.51	.0009
RE-3	1.5	0.96	.3400 ←
EWB-5	4.3	3.20	.0023
PCS	2.8	3.23	.0021
MCS	3.9	2.82	.0067

Change is in T-score metric.

Minimally important change (MIC) should not be used to identify responders to treatment

Underestimates the amount of change needed to be significant at the individual level due to larger measurement errors for individual change scores.

Quality of Life Research (2021) 30:2765–2772
https://doi.org/10.1007/s11136-021-02897-z

COMMENTARY

Between-group minimally important change versus individual treatment responders

Ron D. Hays[1] · John Devin Peipert[2]

Abu et al. (2020) used MIC threshold of **5** as cutoff to identify if patients changed on the Atrial Fibrillation Effect QualiTy-of-Life (AFEQT) Questionnaire

Table 2 Amount of change in atrial fibrillation effect on QualiTy-of-Life (AFEQT) scores needed for significant individual change (coefficient of repeatability)

	Overall score	Symptoms	Daily activities	Treatment concerns
Standard deviation	17.8	17.5	24.5	19.3
Internal consistency reliability	0.90*	0.95	0.94	0.90
Coefficient of repeatability	15.6	10.8	16.6	16.9

*Exact reliability not reported in Abu et al. [22] so we estimated this from prior work[23]

Abu HO, Saczynski JS, Mehawej J, Tisminetzky M, Kiefe CI, et al. 2020. *J Am Heart Assoc.* 9(18):e016651

What are the Minimum Clinically Important Differences in SF-36 Scores in Patients with Orthopaedic Oncological Conditions?

- 310 orthopedic oncology patients underwent musculoskeletal surgery
  - < 6 months (time 1) and 1-2 years (time 2) after surgery
- Distribution-based MIC "estimates" for PCS and MCS
  - Half SD (**5** for both)
  - Standard error of measurement (**6** for PCS and **5** for MCS)
- Anchor-based estimates for PCS and MCS
  - Compared to when you last completed the questionnaire, is your musculoskeletal condition *much better*, *somewhat better*, *about the same*, *somewhat worse*, or *much worse*?
  - **4** for improvement (PCS and MCS)

Ogura et al. (2020). *Clinical Orthopaedics and Related Research*, 478: 2148-2158.

What are the MCIDs for PROMIS, NDI, and ODI instruments among patients with spinal conditions?

- Median estimated MCID was **8** for PROMIS physical function
  - Hung et al. (2018, Clin Orthop Relat Res) estimated minimum detectable change
- Goldstein et al. (2015) also erroneously referred to significant individual change as the minimum clinically important change.

Clinically Meaningful Rehabilitation Outcomes of Low Vision Patients Served by Outpatient Clinical Centers

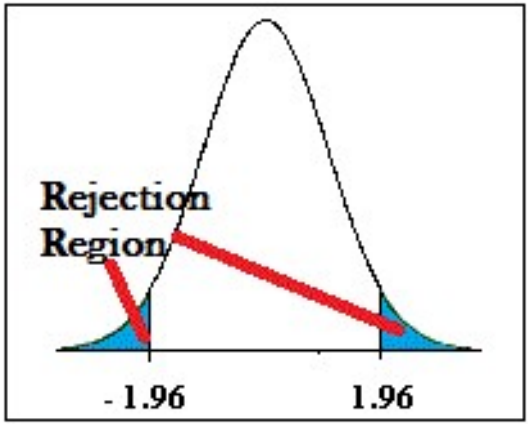
Judith E. Goldstein, OD; Mary Lou Jackson, MD; Sandra M. Fox, OD; James T. Deremeik, CLVT; Robert W. Massof, PhD; for the Low Vision Research Network Study Group

Identifying “Responders”: Reliable Change Index (RCI)

$$\frac{X_2 - X_1}{(\sqrt{2})(SEM)} \quad >= 1.96$$

$$SEM = SD_{bl} \times \sqrt{1 - r_{xx}}$$

SEM **= standard error of measurement**
SD_{bl} **= standard deviation at baseline**
r_{xx} **= reliability**



Amount of Change in Individual's Score for Statistical Significance (p <.05)

$$(\sqrt{2})\,(SD)\sqrt{(1 - r_{xx})}\,(1.96)$$

$$= 2.77 * SD * \sqrt{1 - r_{xx}} = 2.77 * SEM$$

"Coefficient of repeatability" (aka "minimally detectable change," "smallest real difference," and "smallest detectable change").

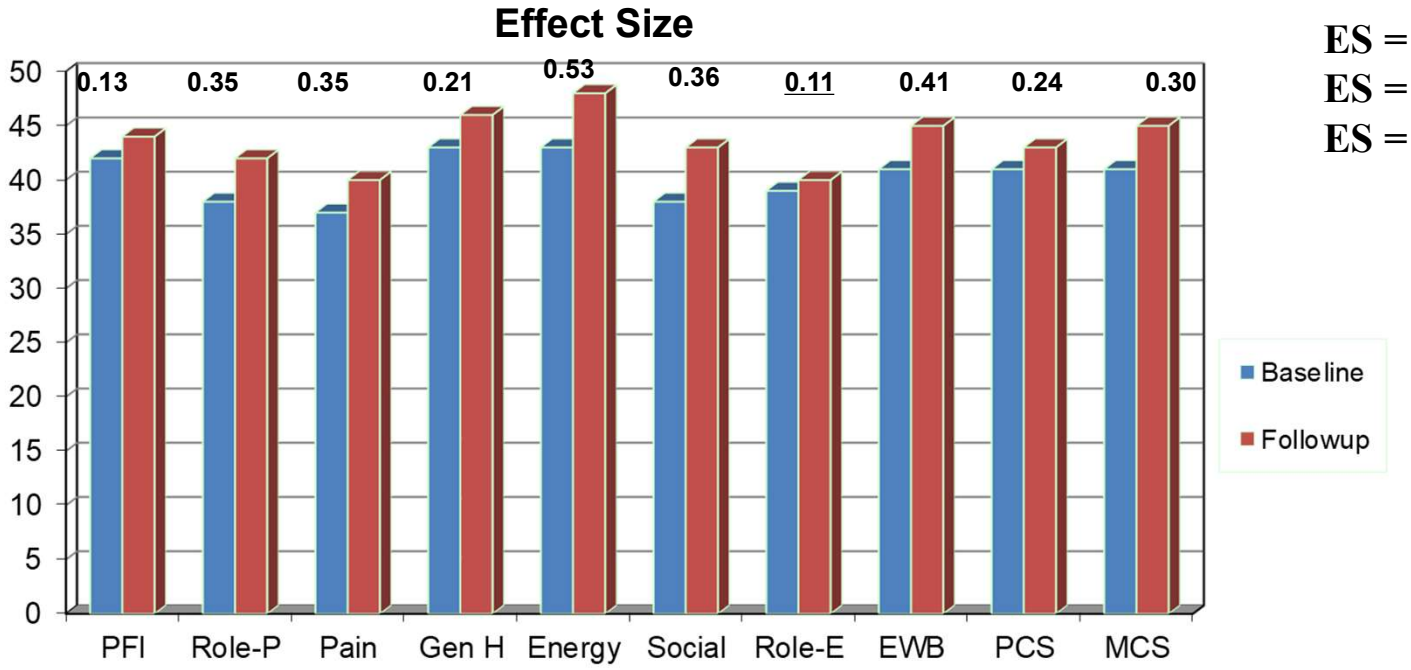
Note: SD = standard deviation and r_{xx} = reliability

Individual T-Score Change Needed for p < .05 Significance at .90 and .95 Reliability

	p-value	Critical value	Critvalue * sq2	SEM	CR
Reliability =.90					
	.05	1.960	2.772	3.162	8.765
Reliability =.95					
	.05	1.960	2.772	2.236	6.198

CR = Coefficient of Repeatability (minimally detectable change, smallest real difference, smallest detectable change)

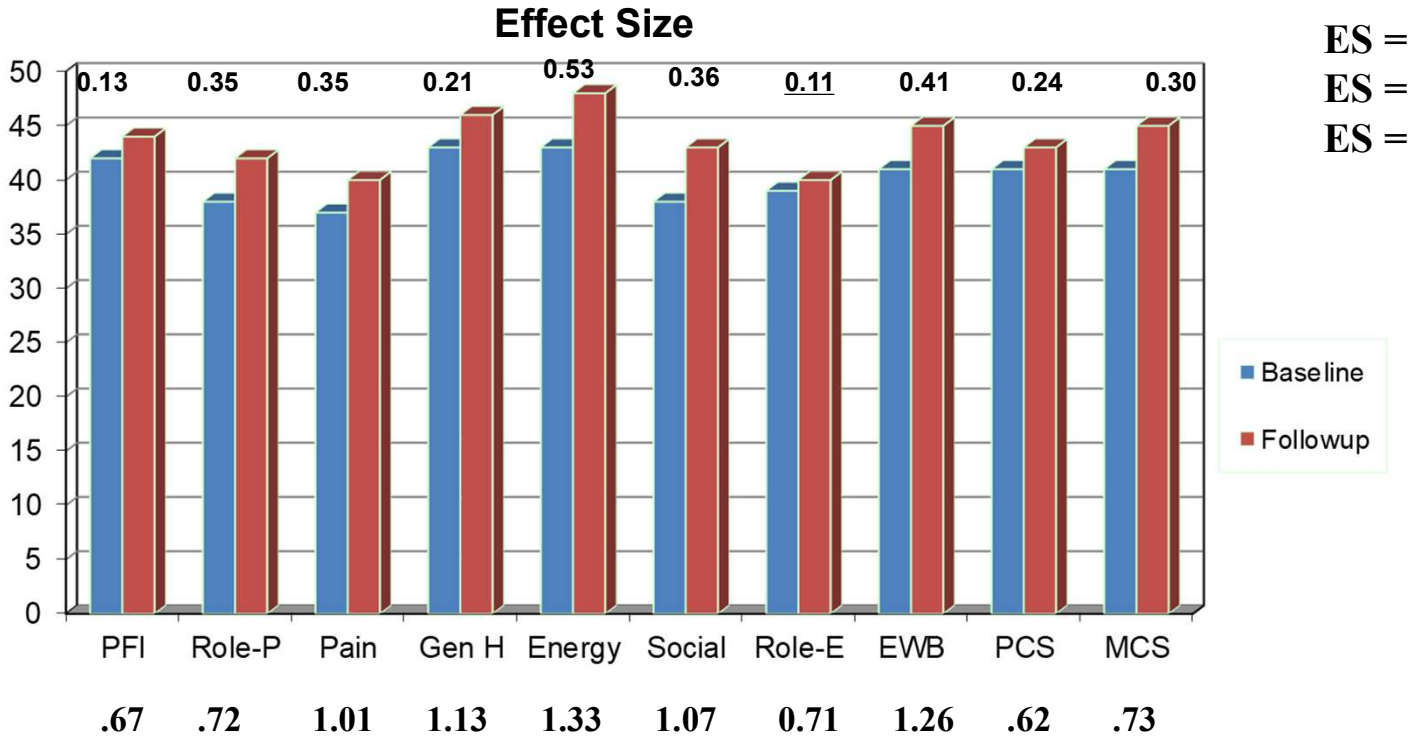
Effect Sizes for Mean SF-36 Score Changes



ES = 0.20 Small
ES = 0.50 Medium
ES = 0.80 Large

PFI = Physical Functioning; Role-P = Role-Physical; Pain = Bodily Pain; Gen H=General Health; Energy = Energy/Fatigue; Social = Social Functioning; Role-E = Role-Emotional; EWB = Emotional Well-being; PCS = Physical Component Summary; MCS =Mental Component Summary.

Effect Sizes for Mean SF-36 Score Changes



ES = 0.20 Small
ES = 0.50 Medium
ES = 0.80 Large

PFI = Physical Functioning; Role-P = Role-Physical; Pain = Bodily Pain; Gen H=General Health; Energy = Energy/Fatigue; Social = Social Functioning; Role-E = Role-Emotional; EWB = Emotional Well-being; PCS = Physical Component Summary; MCS =Mental Component Summary.

7-31% of People in Sample Improved Significantly

	% Improving	% Declining	Improving - Declining
PF-10	13%	2%	+ 11%
RP-4	31%	2%	+ 29%
BP-2	22%	7%	+ 15%
GH-5	7%	0%	+ 7%
EN-4	9%	2%	+ 7%
SF-2	17%	4%	+ 13%
RE-3	15%	15%	0%
EWB-5	19%	4%	+ 15%
PCS	24%	7%	+ 17%
MCS	22%	11%	+ 11%

IRT Reliable Change Index (IRT)

$$\frac{X_2 - X_1}{SQRT(SEb^2 + SEf^2)}$$

SE_b^2 = Variance at baseline, SE_f^2 = Variance at follow-up

Jabrayilov et al. (2016). *Applied Psychological Measurement*

Different Change Categories in Observational Study of Chronic Low Back and Neck Pain Patients Getting Chiropractic Care (Baseline to 3 Months Later)

PROMIS Scale	**Worse**	**Same**	**Better**
Physical Function	3%	91%	6%

Significant according to IRT standard errors and two-tailed ($p < .05$) test.

Hays, Spritzer & Reise, (2023). *Psychometrika*

Expanding the Number of Change Categories Using Two- and One-Tailed Tests

	Definitely Worse	Probably Worse	Same	Probably Better	Definitely Better
Physical Function	3%	3%	84%	4%	6%

Note: *Definitely Worse* and *Definitely Better* groups defined as significant change according to item response theory standard errors and two-tailed test. *Probably Worse* and *Probably Better* groups defined as significant change according to one-tailed test.

Expanding the Number of Change Categories Using Two- and One-Tailed Tests

	Definitely Worse	Probably Worse	Same	Probably Better	Definitely Better
Physical Function	3%	3%	84%	4%	6%

Note: *Definitely Worse* and *Definitely Better* groups defined as significant change according to item response theory standard errors and two-tailed test. *Probably Worse* and *Probably Better* groups defined as significant change according to one-tailed test.

Individual Change in HRQOL

- Quality improvement
  - Individual patient monitoring

Use of HRQOL Measures in Clinical Practice

- IDEAL
  - Identify/elicit the problem
  - Discuss/co-create planned actions
  - Enact action(s)
  - Action(s)
  - Learn about the effects

Reliability Target for Individual Assessment

- 0.90 or above
  - SE = SD (1- reliability)$^{1/2}$
  - Reliability = 1 – (SE/10)^2
- Reliability = 0.90 when <u>SE = 3.2</u>
  - 95% CI = true score +/- 1.96 x SE

(observed score = true score if = mean)

PROMIS CAT Report

Computerized Adaptive Test (CAT) Report

Date: 01-Nov-10

Your age: 50

Your gender: Male

Computerized Adaptive Tests: Depression, Physical Function

Your score on the Depression CAT is 70. The average score is 50.

Your score indicates that your level of Depression is higher (worse) than:

- **98 percent of people in the general population**
- **96 percent of people age 45-54**
- **98 percent of males**

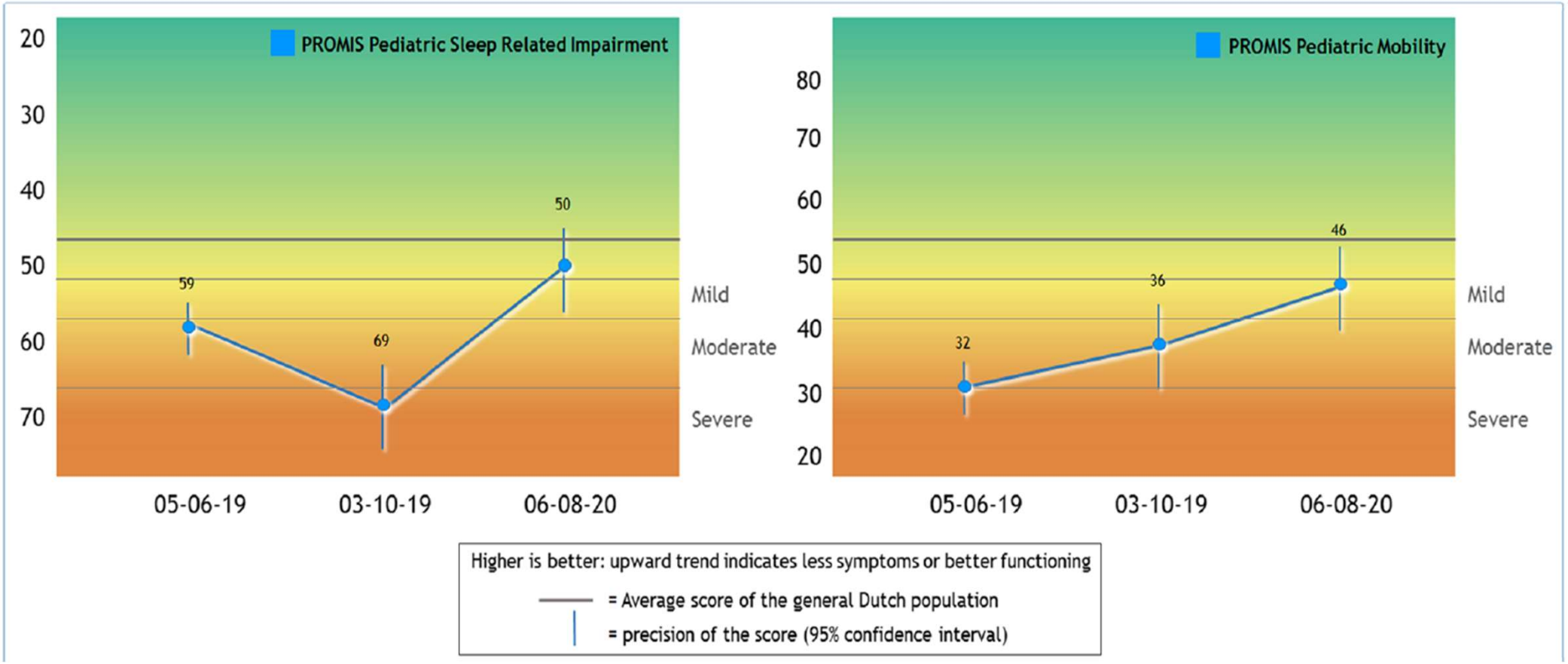
Your score on the Physical Function CAT is 33. The average score is 50.

Your score indicates that your level of Physical Function is higher (better) than:

- **6 percent of people in the general population**
- **9 percent of people age 45-54**
- **5 percent of males**

---

van Muilekom et al. (2021)



Likely Change

- Donaldson, 2008, *QLR*
  - Suggested relaxing .05 p-value because it misclassifies patients who feel they have changed.
- Peipert et al. 2023, *QLR*
  - .32 p-value corresponded more closely than .05 to mean change for those who were *a little* or *lot better* (worse).

T-Score Change Needed for Statistical Significance

	p-value	Critical value	Critvalue * sq2	SEM	CR
Reliability =.90					
	.05	1.960	2.772	3.162	8.765
	.10	1.650	2.333	3.162	7.379
	.32	0.994	1.406	3.162	4.445
	.50	0.674	0.953	3.162	3.014
Reliability=.95					
	.05	1.960	2.772	2.236	6.198
	.10	1.650	2.333	2.236	5.218
	.32	0.994	1.406	2.236	3.143
	.50	0.674	0.953	2.236	2.131

Meaningful individual change

- *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims* (2009)
- *Incorporating Clinical Outcome Assessments into Endpoints for Regulatory Decision-Making* (December 6, 2019, patient-focused drug development guidance public workshop)

- Focuses on <u>average change</u> for patients who improved or got worse.

"Another anchor-based approach to defining responders makes use of patient ratings of change administered at different periods of time or upon exit from a clinical trial. These numerical ratings range from *worse* to *the same* and *better*. The difference in the PRO score for persons who rate their condition *the same* and *better* or *worse* can be used to define responders to treatment" (p. 25).

Meaningful Change on the ISS

- Impact Stratification Score (ISS) assessed at baseline and 6 weeks
  - 750 active-duty U.S. military with low back pain treated with usual care alone or usual care plus chiropractic
- *ISS reliability = 0.92, SD = 8.5*
  - *Coefficient of repeatability = 8 for p<.05*
- Compared to your first visit, your low back pain is:
  - *1: A little better*, *Moderately better*, *Much better*, or *Completely gone*
  - *2: Moderately better*, *Much better*, or *Completely gone*
- *Optimal cut-point = 6 (for #1 above) and 8 (for #2 above)*

Hays & Peipert, 2021, *Quality of Life Research*

PROMIS

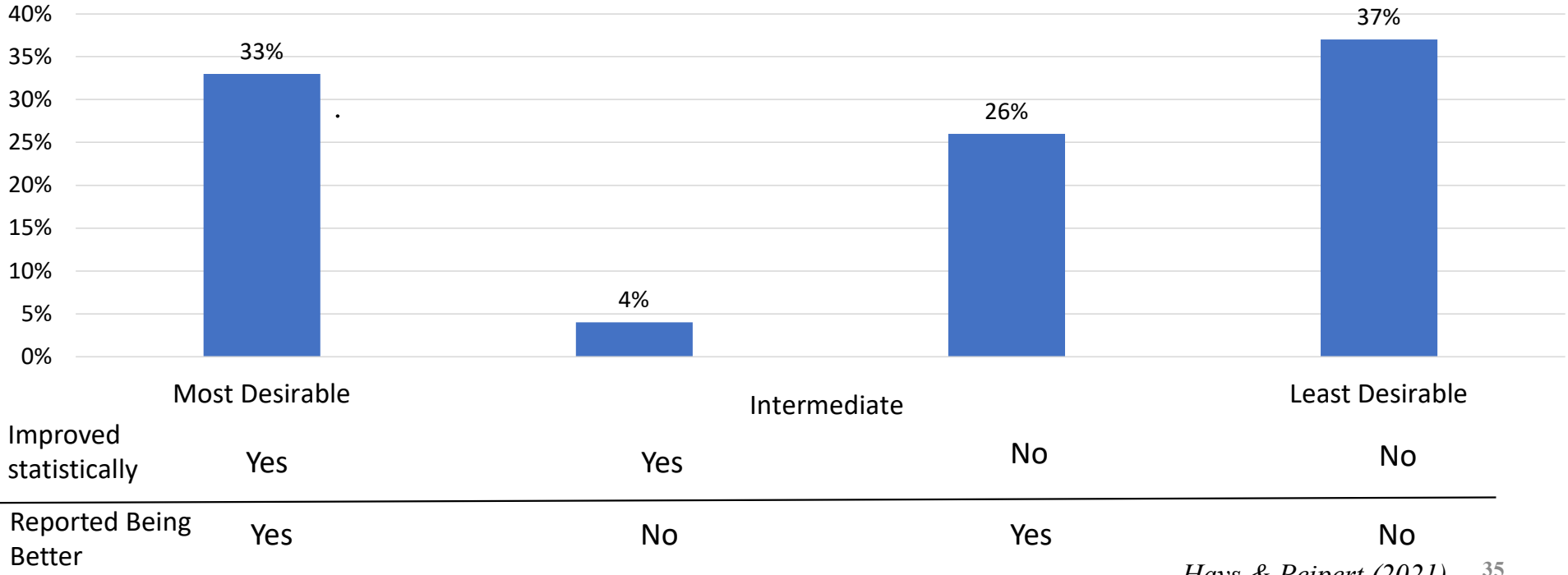
HealthMeasures
TRANSFORMING HOW HEALTH IS MEASURED

T-Score Change Needed for Statistical Significance

	p-value	CR	Meaningful		MIC
Reliability =.90					
	.05	8.765	8	Moderately better +	
	.10	7.379			
	.32	4.445			
	.50	3.014		A little better	3
Reliability=.95					
	.05	6.198	6	A little better +	
	.10	5.218			
	.32	3.143		A little better	3
	.50	2.131			

6-Week Change in Impact Stratification Score in 750 active-duty U.S. military with low back pain treated with usual care alone or usual care plus chiropractic

- **59**% reported retrospectively that they were a *little better*, *moderately better*, *much better,* or their pain was *completely gone* post-baseline.
- **37**% had a statistically significant ($p < .05$) improvement.



Hays & Peipert (2021)

Another Consideration in Assessing Individual Change: Where One Ends Up

A primary care physician might be interested in whether a patient ends up within the normal blood pressure range following initiation of high blood pressure medicine.

A rehabilitation clinician might want to know if a patient with impaired physical functioning at the beginning of treatment ends up functioning as well as other people with a similar condition.

Person Fit

Person misfit may be suggestive of response carelessness or cognitive errors due to survey items being difficult to comprehend….

Person fit using the standardized Z(L) fit index.

Large negative Z(L) values indicate unlikely response patterns given the model.

On PROMIS physical functioning item bank someone reported

- A *little difficulty* being out of bed most of the day.

&

- Able to run 5 miles *without any difficulty*

Reise (1990); Hays, Calderón et al. (2018)

Multiple Observations Are Ideal

- But the number recommended is large
  - 10+ per subject
    - Borckardt et al. (2008), *American Psychologist*
  - 15 per subject
    - Moinpour et al. (2017),*Quality of Life Research*
- See critique of RCI
  - "When (Not) to Rely on the Reliable Change Index"
    - https://osf.io/3kthg

Recommendations for Your Patients

- Evaluate the statistical significance of individual change at multiple p-values (.05, .10, .32).
- Combine statistical significance with the patient's perception of change
  - Definitely improved statistically
    - Patient felt they improved (retrospectively)
    - Patient felt they did not improve
  - Probably improved statistically
    - Patient felt they improved
    - Patient felt they did not improve
  - Stayed the same statistically
    - Patient felt they improved
    - Patient felt they did not improve
    - Patient felt they got worse
  - Probably got worse
    - Patient felt they got worse
    - Patient felt they did not get worse
  - Definitely got worse
    - Patient felt they got worse
    - Patient felt they did not get worse


Thanks Eh !
Ron Hays