Literature Review to Determine Empirical Basis for Response Scale Selection in Patient-reported Outcome Measure Development

Katharine Gries, PharmD, PhD

Senior Research Associate, Outcomes Research, Evidera

SIXTH ANNUAL

PATIENT-REPORTED OUTCOME CONSORTIUM WORKSHOP

April 29–30, 2015 ■ Silver Spring, MD

CRITICAL PATH INSTITUTE
a decade of excellence

10 YEARS

Disclaimer

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

Session Objective

- Present results of a literature review project sponsored by the C-Path patient-reported outcome (PRO) Consortium
  - Review and summarize empirical evidence for common response scale types to enhance response scale selection for newly developed PRO measures

Literature Review to Determine Empirical Basis for Response Scale Selection in Patient-reported Outcome Measure Development

Prepared by:

- Margaret Vernon, PhD
- Katharine Gries, PharmD, PhD
- Shima Safikhani, MPH

Project Background and Rationale

- Response scale selection for patient-reported outcome (PRO) measures is a critical part of:
  - Content validity
  - Psychometric properties
  - Scoring
  - Interpretation of the score
  - Cross-cultural relevance and comparability
  - Ease of administration/operational usability
- Despite the importance of response scale selection for new measures, there seems to be little empirical basis for type of response scale selected
  - Response scale selection is often based on therapeutic convention or preference of the developer

Project Objectives

- Conduct a comprehensive review of the scientific literature:
  - Identify response scale option types
  - Review and summarize available empirical evidence for each type of scale by context of use
    - Psychometric properties and sensitivity/responsiveness
    - Therapeutic area
    - Study population
    - Format of PRO measure
    - Recall period of PRO measure
    - Acceptability from regulatory authorities
  - Enhance response scale selection for newly developed PRO measures

Specific Literature Search Objectives

- **Search 1**: Formal guidelines / review articles
  - Recommendations on the selection of response scale types
- **Search 2**: Direct and indirect empirical evidence for response scale types
  - Psychology and survey methodology literature

- Articles with direct or indirect empirical evidence for PRO response scale selection:
  - **Search 3**: based on recall period
  - **Search 4**: based on the format of the PRO measure or response scale
  - **Search 5**: based on scoring and psychometric property implications
  - **Search 6**: based on the intended study population
  - **Search 7**: based on therapeutic indications (of interest)

- **Search 8**: Response scales used in PRO measures achieving successful regulatory review by the United States (US) Food and Drug Administration (FDA) and European Medicines Agency (EMA) and included in the approved product label

Translatability Assessment and Electronic Feasibility

- Translatability and electronic implementation assessment of the identified response scale types
  - Feedback about format, words, or phrases which are structurally or culturally problematic when translated into different languages.
  - Implementation issues for electronic modes of data collection
    - Web-based, smart phone, tablet, interactive voice response (IVRS)

Summary of Literature Search

- Across the literature we identified and reviewed:
  - Almost 7,000 abstracts
  - Approximately 400 full text articles
- Final count: 196 unique references
- The FDA and EMA searches (search objective 8)
  - From 2006–2014
    - 36 unique PRO measures with claims in FDA labels
    - 37 unique PRO measures with claims in EMA labels

Common Response Scales Types Identified in the Literature Search

Response Scale Frequency in PRO Measure Literature

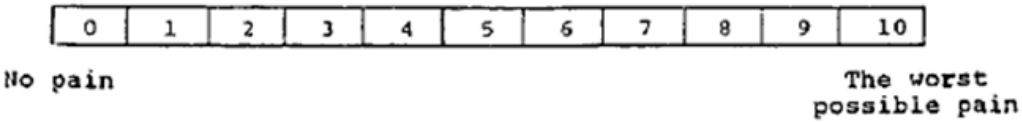
Response Scale Type	Frequency Count Across Searches 3–7
Visual Analog Scale (VAS)	68
Verbal Rating Scale (VRS)	59
Numeric Response Scale (NRS)	48
Faces	16
Other Graphical	13
Binary	5
Likert	3

Many studies included multiple scales, frequency counts do not represent unique studies.

Types of Response Scales

VRS

NRS



VAS (not to scale)

Types of Response Scales

Likert Scales

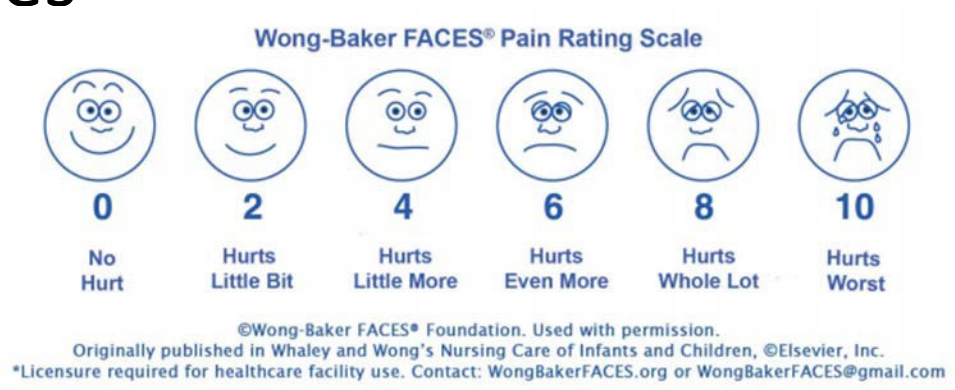
Please indicate how much you agree or disagree with each of these statements:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The president is doing a good job.	1	2	3	4	5
The Congress is doing a good job.	1	2	3	4	5
The Secretary of Defense is doing a good job.	1	2	3	4	5

Compared with 3 months ago, how would you rate your health in general now?

Much Better	Somewhat Better	About the Same	Somewhat Worse	Much Worse
❑	❑	❑	❑	❑

Faces Scales



Response Scale Selection for Use in PRO Measures

Response Scale Selection Based on Sensitivity/Responsiveness (n=13)

- Inconclusive to which response scale demonstrates greater responsiveness, however there was a slight trend in favor of NRS over VAS
- Dependent on the construct being measured
- 6- or 7-point VRS more responsive than 4- or 5-point VRS
- 6- or 7-point VRS may be as responsive as NRS or VAS

Response Scale Selection Based on Therapeutic Area

- Asthma (n=3)
  - VRS typically used
  - For ages 4–11 years old: fewer response options or VRS assisted by graphics
- Pain (n=62)
  - 11-point NRS likely the optimal response scale
    - Adult patients without cognitive impairment
- Fatigue (n=9)
  - VAS or NRS suitable
    - The NRS was more responsiveness than the VAS in exercise capacity study
- Oncology (n=14)
  - VAS or NRS typically used
    - 11-point NRS favored over VAS for ease of implementation, scoring, and good psychometric properties
  - Faces scales appropriate for pediatric populations

Response Scale Selection Based on Pediatric Population (n=31)

- Visual orientation of scales, emotions expressed in graphical scales, and word choice for verbal anchors can produce unexpected biases due to immaturity in abstract thinking skills of respondents
  - Child might choose a numerical response based on favorite number rather than representation of experience.
    - NRS and VAS suitable for children 7 and older
  - Faces scales appear to be well understood and produce reliable and valid estimates for children 4–12 years
  - VRS may be most suitable for older children (10–12 years and older) who are confident readers and have a higher level of reading comprehension
    - 5-point VRS may be appropriate for children as young as 7 years old
    - Likert scale appropriate for children 7 or 8 years of age and older

Response Scale Selection Based on Geriatric Populations (n=7)

- Evidence is somewhat inconsistent
  - VRS or NRS recommended over VAS or graphical scale
  - VAS preferred over 5-point VRS or a 4-faces scale
  - 7-point VRS more widely accepted and understood than a faces or NRS
  - NRS preferred by surgical patients over VAS or 6-point VRS
  - VAS and 5-point VRS yielded similar reliability and responsiveness, however the VRS was found to have superior interpretability
  - 4- and 5-point VRS demonstrated more stability than an 11-point NRS because of difficultly to complete as cognitive impairment increased

Translatability and Cultural Adaptation

- Choice of verbal anchors needs to be considered
- NRS
  - For the 0 end of the scale, common anchors include None, Not at all, Never, No pain, No [symptom], and Absent
    - *Absent* is particularly challenging to translate because it is not a simple term in the medical context and is not patient-friendly language
- VRS
  - Need to achieve similar intervals between the verbal descriptors to match the source language
  - Amounts of time "How much of the time…?" is difficult to translate
    - Simple frequency scales using terms such as Never, Rarely, Sometimes, Often, Always are recommended
- Faces Scales
  - There is little empirical evidence available to demonstrate cultural comparability across languages/cultures

Electronic Feasibility

- NRS
  - Easily implemented via Interactive voice response systems (IVRS) and other electronic modes
    - Formatting needs to be considered so that anchors are associated with intended number
- VAS
  - Impossible to implement in IVRS
  - Difficult to implement in other electronic modes due to need for specified length (100 mm)
- VRS
  - Number of response options and length of verbal descriptors should be carefully selected so as to lighten cognitive load (for IVRS)
  - Equidistant formatting on 1 screen for handheld/tablet devices

Discussion and Recommendations

VAS

VAS widely used in the PRO literature reviewed and generally psychometric properties and responsiveness are strong.

- May be appropriate for use in pediatric populations over the age of 7 or 8
  - BUT may not be as well understood in geriatric populations or populations with cognitive impairment
- Impossible to implement in an oral format
- Difficult to migrate to electronic visual formats due to requirements that the scale have a specific length
- Might be more difficult to implement (administer and score consistently) than other scale types

NRS

NRSs were widely utilized in PRO literature reviewed and generally psychometric properties and responsiveness are strong for this scale

- Preferred scale for pain measurement
- Pediatric Population: does not always produce comparable scores
  - NRS may not be advisable for use in younger age groups (4 to 11+ years)
- Suitable for translation and cross-cultural adaptation
- Easy to implement in various modes of administration including IVRS and electronic visual formats

VRS

VRSs were widely utilized in the PRO literature reviewed and generally psychometric properties and responsiveness are strong for this scale type

  - Scales with more response options (5 to 7) have better performance and responsiveness than scales with fewer response options (3 to 4)

- Selection of verbal anchors is important as acceptability and understanding is dependent on literacy levels of the sample
- Scale performance dependent on assumptions of equidistant intervals
- More difficult to translate and cross-cultural adaptation
- Typically used in multi-item scales
  - PRO measures to support FDA and EMA labeling claims

Faces and Other Graphical Scales

Less widely utilized in the PRO literature reviewed

- More research is needed regarding performance of these scales in specific contexts of use

- Faces scales work well in young pediatric samples, in adult samples, and in cross-cultural contexts
  - Pre-literate participants can more readily abstract their experience to the emotion expressed in the faces as compared to a verbal or numerical scale
  - More research is needed regarding cross-cultural comparability of facial cues presented in the faces response options
- Cannot be administered orally (IVRS) but could be administered via electronic visual modes

Discussion and Recommendations

- About 1/3 of articles identified were in the pain therapeutic area
  - Optimal response scale choice for pain may not always be generalizable to other contexts of use
  - Pain assessments typically consist of single-item questions which use either a VAS or an NRS response scale
    - Multi-item questionnaires also use these response scale types but more commonly use a VRS
  - Pain assessment may be different than other measurement concepts which require multiple items

- Response scale selection needs to take into account multiple factors
  - Target population
  - Study design
  - Concept of interest
  - Recall period
  - Data collection mode
  - Scale responsiveness

Discussion / Questions

Key References

- Example Scale Sources
  - Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. (1990) Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. *J Rheumatol.* 17(8):1022-1024.
  - Uebersax JS. Likert scales: dispelling the confusion. Statistical Methods for Rater Agreement website. 2006. Available at: http://john-uebersax.com/stat/likert.htm. Accessed: December 2, 2014
  - Vinik EJ, Hayes RP, Oglesby A, et al. The development and validation of the norfolk QOL-DN, a new measure of patients' perception of the effects of diabetes and diabetic neuropathy. Diab Tech Therapeutics *Diabetes technology & therapeutics*. 2005; 7(3): 497-508

Key References

- Therapeutic area
  - Liu, A.H., Zeiger, R., Sorkness, C., et al. (2007). Development and cross-sectional validation of the Childhood Asthma Control Test. J Allergy Clin Immunol. 119 (4):817-825.
  - Juniper, E. F., Guyatt, G. H., Feeny, D. H., Griffith, L. E., and Ferrie, P. J. (1997). Minimum skills required by children to complete health-related quality of life instruments for asthma: Comparison of measurement properties. Eur Respir J. 10 (10):2285-2294.
  - Jackson, D., Horn, S., Kersten, P., and Turner-Stokes, L. (2006). Development of a pictorial scale of pain intensity for patients with communication impairments: Initial validation in a general population. Clin Med. 6 (6):580-585.
  - Ritter, P. L., Gonzalez, V. M., Laurent, D. D. and Lorig, K. R. (2006). Measurement of pain using the visual numeric scale. Journal of Rheumatology. 33 (3):574-580.
  - Grant, S., Aitchison, T., Henderson, E., Christie, J., Zare, S., McMurray, J., and Dargie, H. (1999). A comparison of the reproducibility and the sensitivity to change of visual analogue scales, Borg scales, and Likert scales in normal subjects during submaximal exercise. Chest. 116 (5):1208-1217.
  - Hollen, P. J., Gralla, R. J., Kris, M. G., McCoy, S., Donaldson, G. W., and Moinpour, C. M. (2005). A comparison of visual analogue and numerical rating scale formats for the Lung Cancer Symptom Scale (LCSS): Does format affect patient ratings of symptoms and quality of life? Qual Life Res. 14 (3):837-847.

Key References

- Study Population
  - Cohen, L. L., Lemanek, K., Blount, R. L., Dahlquist, L. M., Lim, C. S., Palermo, T. M., McKenna, K. D. and Weiss, K. E. (2008). Evidence-based assessment of pediatric pain. Journal of Pediatric Psychology. 33 (9):939-955.
  - McGrath, P. A. (1987). An assessment of children's pain: A review of behavioral, physiological and direct scaling techniques. Pain. 31 (2):147-176.
  - von Baeyer, C. L. (2006). Children's self-reports of pain intensity: Scale selection, limitations and interpretation. Pain Research and Management. 11 (3):157-162.
  - Dahlquist, L. M. (1990). Obtaining child reports in health care settings. Through the eyes of the child: Obtaining self-reports from children and adolescents. Needham Heights, MA, US, Allyn & Bacon: 395-439.
  - Matza, L. S., Patrick, D. L., Riley, A. W., Alexander, J. J., Rajmil, L., Pleil, A. M. and Bullinger, M. (2013). Pediatric patient-reported outcome instruments for research to support medical product labeling: Report of the ISPOR PRO good research practices for the assessment of children and adolescents task force. Value in Health. 16 (4):461-479.
  - Matza, L.S., Swensen, A.R., Flood, E.M., et al. (2004). Assessment of health-related quality of life in children: a review of conceptual, methodological, and regulatory issues. Value Health. 7 (1):79-92.
  - Schofield, P., Clarke, A., Faulkner, M., Ryan, T., Dunham, M. and Howarth, A. (2006). Assessment of pain in older people. Where are we now and what needs to be done? International Journal on Disability and Human Development. 5 (1):3-8.
  - Castle, N. G., and Engberg, J. (2004). Response formats and satisfaction surveys for elders. Gerontologist. 44 (3):358-367.
  - Jones, K. R., Vojir, C. P., Hutt, E. and Fink, R. (2007). Determining mild, moderate, and severe pain equivalency across pain-intensity tools in nursing home residents. Journal of Rehabilitation Research and Development. 44 (2):305-314.
  - Gagliese, L., Weizblit, N., Ellis, W., and Chan, V. W. (2005). The measurement of postoperative pain: A comparison of intensity scales in younger and older surgical patients. Pain. 117 (3):412-420.
  - Hagell, P., Knutsson, I. (2013). Single-item assessment of perceived health in neurological disorders: verbal response categories vs visual analog scale. Abstract presented at: ISOQOL 20th Annual Conference.

Backup Slide

Search Criteria: Empirical Evidence For PRO Measure Response Scale Selection

Type	Search Terms
Response scale terms	'response scale':ab,ti OR 'response scales':ab,ti OR likert:ab,ti OR 'likert scale'/exp OR 'visual analog scale':ab,ti OR 'visual analog scales':ab,ti OR 'visual analogue scale':ab,ti OR 'visual analog scale'/exp OR 'numerical rating scale':ab,ti OR 'numerical rating scales':ab,ti OR 'verbal rating scale':ab,ti OR 'verbal rating scales':ab,ti OR 'competence scale':ab,ti OR 'competence scales':ab,ti OR 'frequency scale':ab,ti OR 'frequency scales':ab,ti OR 'extent scale':ab,ti OR 'extent scales':ab,ti OR 'comparison scale':ab,ti OR 'comparison scales':ab,ti OR 'performance scale':ab,ti OR 'performance scales':ab,ti OR 'developmental scale':ab,ti OR 'developmental scales':ab,ti OR 'qualitative scale':ab,ti OR 'qualitative scales':ab,ti OR 'agreement scale':ab,ti OR 'agreement scales':ab,ti OR 'categorical scale':ab,ti OR 'categorical scales':ab,ti
PRO terms	'patient satisfaction'/exp OR (patient* NEAR/2 satisfaction):ab,ti OR (patient* NEAR/2 reported):ab,ti OR 'self report'/exp OR (self NEAR/1 report*):ab,ti OR 'patient preference'/exp OR (patient* NEAR/2 preference*):ab,ti OR (patient* NEAR/1 assess*):ab,ti OR 'self evaluation':ab,ti OR 'self evaluations':ab,ti OR (patient* NEAR/2 rating):ab,ti OR (patient* NEAR/2 rated):ab,ti OR 'self-completed':ab,ti OR 'self- administered':ab,ti OR (self NEAR/1 assessment*):ab,ti OR 'self-rated':ab,ti OR 'patient based outcome':ab,ti OR 'self evaluation'/exp OR experience*:ab,ti

Literature Review to Determine Empirical Basis for Response Scale Selection in Patient-reported Outcome Measure Development

Prepared by:

- Margaret Vernon, PhD
- Katharine Gries, PharmD, PhD
- Shima Safikhani, MPH