



Measurement in Clinical Trials: FDA COA Workshop Session 3

Content Validity

Jeremy Hobart MD, PhD

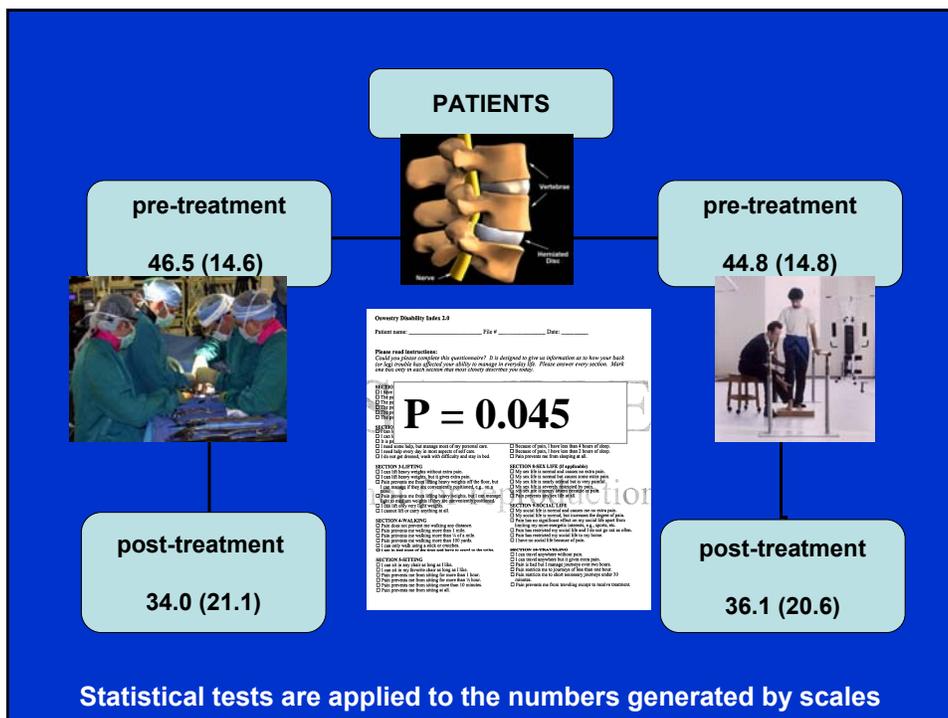
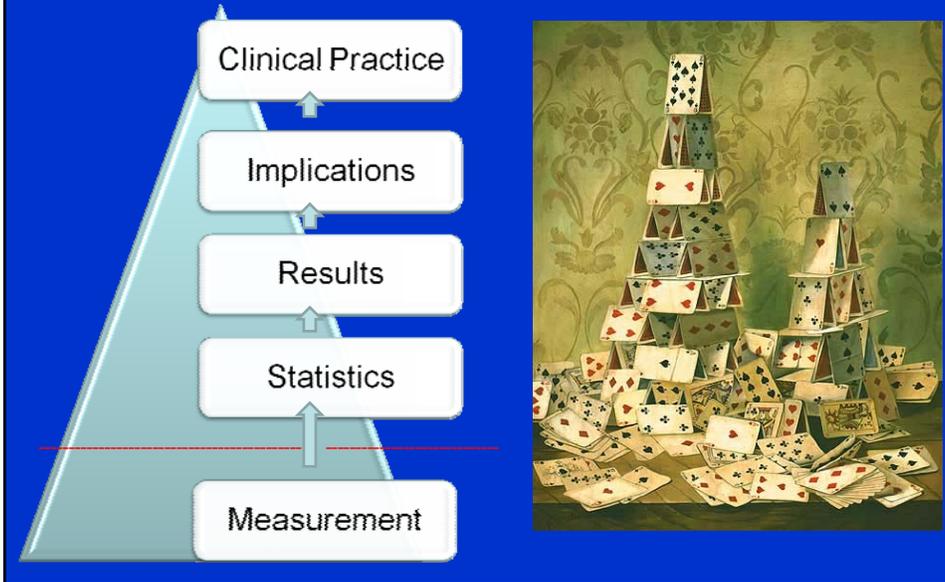
Professor of Clinical Neurology
Peninsula College of Medicine and Dentistry,
Plymouth, Devon, UK

Washington DC 19 October 2011

Overview

- **Measurement and measurement scales**
- **Validity: brief history & mantra**
- **Scale content: importance & guidance**
- **Current issues being debated**
- **Attempt to answer Bob Temple's questions**
 - **Back to the future**
 - What's really the problem?
 - An exemplar to aspire to
 - **Examples of statistical assistance for CV**
- **Conclusions**

Importance of measurement & “scales”



Randomised controlled trial to compare surgical stabilisation of the lumbar spine with an intensive rehabilitation programme for patients with chronic low back pain: the MRC spine stabilisation trial

Jeremy Fairbank, Helen Frost, James Wilson Macdonald, Ly-Mee Yu, Karen Barker, Ray Collins for the Spine Stabilisation Trial Group

Abstract
Objective: To assess stabilisation (spinal rehabilitation) for patients with chronic low back pain.
Design: Multicentre randomised controlled trial.
Setting: 12 secondary care centres in the UK.
Participants: 547 patients with low back pain of at least 6 weeks' duration.
Interventions: Surgical stabilisation (Laminectomy/Laminotomy) or intensive rehabilitation programme.
Main outcome measures: Disability (Oswestry Disability Index) at baseline and post-intervention (week 24).
Results: 170 patients were randomised to surgery and 377 to rehabilitation. The mean Oswestry Disability Index score at baseline was 44.5 (SD 14.0) in the surgical group and 44.5 (SD 14.0) in the rehabilitation group. This trial registered as a randomised controlled trial on ClinicalTrials.gov.
Conclusion: No significant difference was found between the two groups at week 24.

Articles

Cannabinoids for treatment of spasticity and other symptoms related to multiple sclerosis (CAMS study): multicentre randomised placebo-controlled trial

John Zujewski, Patrick Fox, Hilary Sanders, David Wright, Jane Whitty, Andrew Barton, Alan Thompson, on behalf of the UK MS Research Group

Abstract
Objective: To assess the efficacy and safety of nabiximols (a cannabis-based extract) for the treatment of spasticity and other symptoms in multiple sclerosis.
Design: Multicentre randomised placebo-controlled trial.
Setting: 12 secondary care centres in the UK.
Participants: 100 patients with multiple sclerosis.
Interventions: Nabiximols or placebo.
Main outcome measures: Spasticity (Modified Ashworth Scale) and other symptoms (Spasticity Symptom Severity Scale) at baseline and post-intervention (week 12).
Results: 50 patients were randomised to nabiximols and 50 to placebo. The mean Modified Ashworth Scale score at baseline was 3.5 (SD 1.5) in the nabiximols group and 3.5 (SD 1.5) in the placebo group. The mean Spasticity Symptom Severity Scale score at baseline was 12.5 (SD 4.5) in the nabiximols group and 12.5 (SD 4.5) in the placebo group. There was no significant difference between the two groups at week 12.
Conclusion: Nabiximols did not significantly improve spasticity or other symptoms compared with placebo.

Modafinil for fatigue in MS

A randomized placebo-controlled double-blind study

B. Stanekoff, MD, PhD*, E. Weusthuis, MD, PhD*, C. Confavreux, MD, G. Edan, MD, M. Debouveri, L. Rumbach, MD, T. Moreau, MD, PhD, J. Pellier, MD, PhD, C. Lubetki, MD, PhD, M. Clanet, J and French Modafinil Study Group†

Abstract—Objective: To assess whether modafinil, a wakefulness-promoting agent, is useful for fatigue in multiple sclerosis (MS). **Methods:** Patients with MS with stable disability, and a baseline score of 45 or more on the Modified Fatigue Impact Scale (MFIS), were eligible for the 6-week randomised, double-blind, placebo-controlled group study. The initial daily dose of modafinil was 200 mg for 1 week. Depending on tolerance, the dose was increased to 400 mg daily and remained unchanged between day 21 and day 35. The primary outcome was the change of MFIS score at day 35. **Results:** A total of 119 patients with MS were enrolled in the study. The mean MFIS score at baseline was 63 ± 9 in the placebo group and 63 ± 10 in the modafinil group. MFIS scores improved between day 0 and day 35 in both placebo-treated and modafinil-treated groups. No significant difference was detected between the two groups. There was no major safety concern. **Conclusion:** The improvement of fatigue in patients with multiple sclerosis treated with modafinil vs placebo according to the Fatigue Impact Scale.
 NEUROLOGY 2005;64:1139–1143

Surgical decompression for space-occupying cerebral infarction (the Hemispherectomy After Middle Cerebral Artery Infarction with Life-threatening Edema Trial [HAMLET]): a multicentre, open, randomised trial

Abstract
Objective: To assess the efficacy and safety of surgical decompression for space-occupying cerebral infarction.
Design: Multicentre randomised controlled trial.
Setting: 12 secondary care centres in the UK.
Participants: 100 patients with space-occupying cerebral infarction.
Interventions: Surgical decompression or medical management.
Main outcome measures: Mortality (Kaplan-Meier survival) and functional outcome (modified Rankin scale) at baseline and post-intervention (week 90).
Results: 50 patients were randomised to surgery and 50 to medical management. The mean modified Rankin scale score at baseline was 5.5 (SD 1.5) in the surgical group and 5.5 (SD 1.5) in the medical management group. The mean mortality at week 90 was 15% in the surgical group and 15% in the medical management group. There was no significant difference between the two groups at week 90.
Conclusion: Surgical decompression did not significantly improve mortality or functional outcome compared with medical management.

Fatigue is not associated with raised inflammatory markers in multiple sclerosis

Galvin Giovannoni, PhD; Alan J. Thompson, MD; David H. Miller, MD; and Edward J. Thompson, DSc

Abstract—Background: The pathogenesis of fatigue in patients with MS is poorly understood. **Objective:** To test the hypothesis that fatigue in MS is related to inflammatory disease activity as measured by systemic markers of inflammation. **Methods:** Fatigue as assessed by the Fatigue Questionnaire Scale (FQS) and Krupp's Fatigue Severity Scale (KFSS) was correlated with several inflammatory markers in 38 patients with MS (16 relapsing-remitting RRMS, 7 of whom had benign MS), 9 secondary progressive (SP), 13 primary progressive (PP). The markers included daily urinary neopterin excretion, a marker of interferon-γ-activated macrophage activity, and serum C-reactive protein (CRP) and soluble intercellular adhesion molecule-1 (sICAM-1) levels. Urinary neopterin excretion was measured daily for 2 weeks. **Results:** No correlation was found between urinary neopterin excretion, CRP, or sICAM-1 and the fatigue scores. However, patients with a raised serum CRP level had higher KFSS, but not FQS, scores than patients with normal CRP levels (KFSS, 50 ± 9 vs 41 ± 14, p = 0.05; FQS, 13 ± 4 vs 11 ± 5, p = NS). When assessed using the FQS, patients with RR and SP MS were more fatigued than patients with PP MS (RR, 12.8 [4 to 23] vs SP = 13.8 [3 to 18] vs PP = 9 [7 to 14], p = 0.02). The patients with benign MS were as fatigued as patients with nonbenign disease. **Conclusion:** The pathogenesis of fatigue in

Validity

- “the extent to which an instrument measures what (the thing) it purports to measure”
- “beguilingly simple”

A brief history of “validity”

1921 – Term first appears (15yrs post reliability)

1954/55/66/74/85 – APA/AERA/NCME: face, content, criterion, construct

1955 – Construct validity in psychological tests (Cronbach & Meehl)

1959 – Convergent and discriminant validation by the multi-method multi-trait matrix (Campbell & Fiske)

1960 – Explosion of types of validity (“over 100 types”)

1999 – Implosion of types of validity (APA “validity one concept”)

The mantra... “a building of evidence”

Content

Face

FUNCTIONAL ASSESSMENT OF MS (FAMS)					
• Please indicate how true each statement has been for you during the past 7 days					
During the past 7 days	Not at all	A little	Some-what	Quite a bit	Very much
1. Because of my physical condition, I have trouble meeting the needs of my family	0	1	2	3	4
2. I am able to work (include work in home)	0	1	2	3	4
3. I have trouble walking	0	1	2	3	4
4. I have to limit my social activity because of my condition	0	1	2	3	4
5. My legs are strong	0	1	2	3	4
6. I have trouble getting around in public places	0	1	2	3	4
7. I have to make plans around my condition	0	1	2	3	4

Correlations

Hypothesis testing

Group differences

Scale content and content validity

- Importance
- Methods: pre 2006
- A line in the sand
- Advances since 2006.....
-and their limitations
- Areas of current debate
- Issues for the future

Importance of scale content is obvious...

These questions ask about limitations to your walking due to MS during the past two weeks.

For each statement, please circle the one number that best describes your degree of limitation.

Please answer all questions even if some seem rather similar to others, or seem irrelevant to you.

If you cannot walk at all, please tick this box:

In the past two weeks, how much has your MS ...

	Not at all	A little	Mod. amount	Quite a bit	Extremely
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
3. Limited your ability to climb up and down stairs?	1	2	3	4	5
4. Made standing when doing things more difficult?	1	2	3	4	5
5. Limited your balance when standing or walking?	1	2	3	4	5
6. Limited how far you are able to walk?	1	2	3	4	5
7. Increased the effort needed for you to walk?	1	2	3	4	5
8. Made it necessary for you to use support when walking outdoors (e.g. holding on to furniture, walls, a stick, etc.)?	1	2	3	4	5
9. Made it necessary for you to use support when walking indoors (e.g. using a stick, a frame, etc.)?	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Please check that you have circled ONE number for EACH question.

© 2000 Neurological Outcome Measures Unit

(better)



(worse)

..& irrespective of PRO, CRO, ObsRO

ADAS - Cognitive Behavior, page 9 of 9
Baseline Visit

13. **Number Cancellation**

Instructions for Example:

Please the patient from the top of the column and read the top of this page are the numbers. Throughout this page you will find these numbers mixed in with other numbers. As the you to begin here, "start at the beginning of the first row," and going across the by line, circle off each number that matches either of the two numbers at the top of the page. Please work as quickly as you can. (Throughout the course of 20 seconds.)

Instructions for Task:

Please the item from the top of the column and read the top of this page are the numbers. Throughout this page you will find these numbers mixed in with other numbers. As the you to begin here, "start at the beginning of the first row," and going across the by line, circle off the numbers that match the numbers at the top of the page. Please work as quickly as you can.

If the first cancellation done by the subject is incorrect, say, "These are the correct numbers to circle off," and point to the target numbers at the top of the page. If the subject becomes confused or stops writing during the test, repeat the standard instructions as needed throughout the test after 45 seconds.

13a. Number Cancellation: Number of targets hit

13b. Number Cancellation: Number of errors

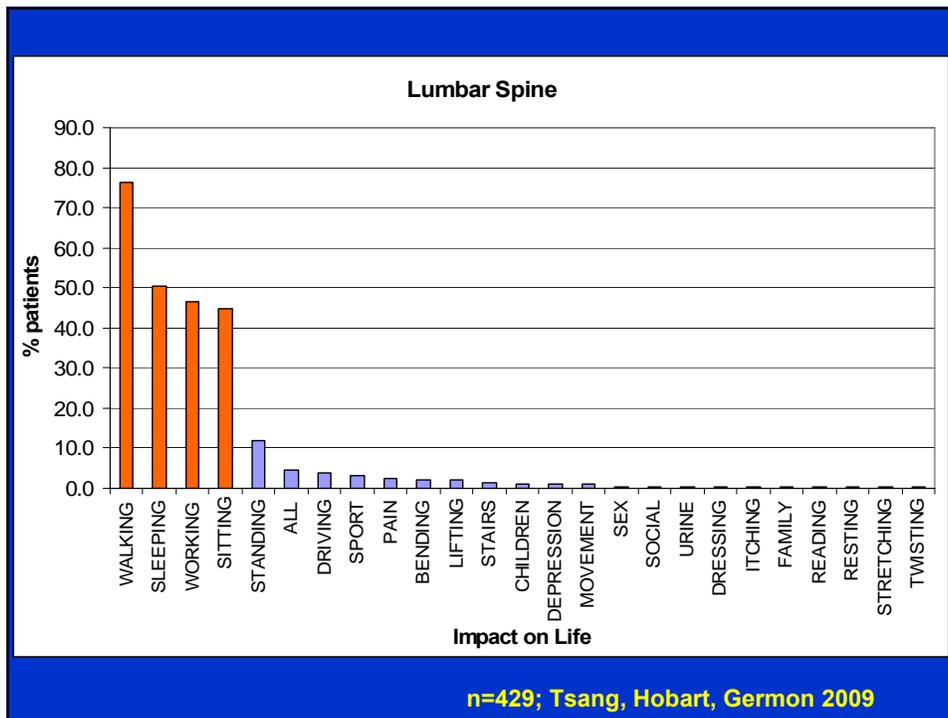
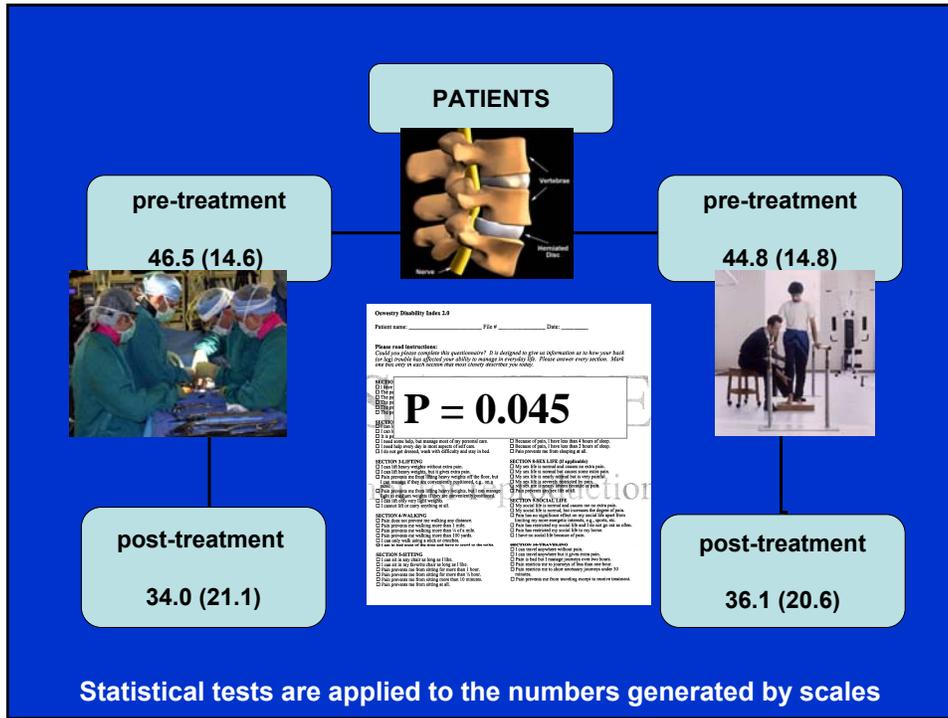
13c. Number Cancellation: Number of items cancelled of total

(better)



(worse)

...but can be difficult to demonstrate



Content validity: methods pre 2006

- Simplistic
- Representative of the domain of interest
- Item generation from qualitative work
- Limited (minimal?) guidance...
- ...the result was (is)

Imbalance in validity testing

- “Qualitative” tests of validity
 - Face validity
 - Content validity
- “Quantitative” tests of validity
 - Criterion validity
 - Construct validity
 - Convergent / discriminant
 - Group differences
 - Hypothesis testing
 - Etc.....

Some curious scale development

- Top down rather than bottom up:
 - Item pool generation
 - Factor analysis to define scales
 - Statistical analysis and modification of scales
 - Scales named by their statistically driven content
- hence..... Scales with clinically curious item content

Please enter a tick in the app
Please give only one answer per row.

A "FATIGUE" SCALE

	Not at all	Slightly	Moderately	Extremely
1. I feel less alert.				
2. I feel that I am more isolated from social contact.				
3. I have to reduce my workload or responsibilities.				
4. I am more moody.				
5. I have difficulty paying attention for a long period.				
6. I feel like I cannot think clearly.				
7. I work less effectively (this applies to work both inside or outside of the home).				
8. I have to rely more on others to help me or do things for me.				
9. I have difficulty planning activities ahead of time.				
10. I am more clumsy and uncoordinated.				
11. I find that I am more forgetful.				
12. I am more irritable and more easily angered.				
13. I have to be careful about pacing my physical activities.				
14. I am less motivated to do anything that requires physical effort.				
15. I am less motivated to engage in social activities.				
16. My ability to travel outside my home is limited.				
17. I have trouble maintaining physical effort for long periods.				
18. I find it difficult to make decisions.				
19. I have few social contacts outside of my own home.				
20. Normal day-to-day events are stressful to me.				
21. I am less motivated to do anything that requires thinking.				
22. I avoid situations that are stressful to me.				
23. My muscles feel much weaker than they should.				
24. My physical discomfort is increased.				
25. I have difficulty dealing with anything new.				
26. I am less able to finish tasks that require thinking.				
27. I feel unable to meet the demands that people place on me.				
28. I am less able to provide financial support for myself and my family.				
29. I engage in less sexual activity.				
30. I find it difficult to organise my thoughts when I am doing things at home or at work.				
31. I am less able to complete tasks that require physical effort.				
32. I worry that about how I look to other people.				
33. I am less able to deal with emotional issues.				
34. I feel slowed down in my thinking.				
35. I find it hard to concentrate.				
36. I have difficulty participating fully in family activities.				
37. I have to limit my physical activities.				
38. I require more frequent or longer periods of rest.				
39. I am not able to provide as much emotional support to my family as I should.				
40. Minor difficulties seem like major difficulties.				

A “MOBILITY” SCALE

- Please indicate how true each statement has been for you during the past 7 days.

<i>During the past 7 days</i>	Not at all	A little	Some-what	Quite a bit	Very much
1. Because of my physical condition, I have trouble meeting the needs of my family	0	1	2	3	4
2. I am able to work (include work in home)	0	1	2	3	4
3. I have trouble walking	0	1	2	3	4
4. I have to limit my social activity because of mv condition	0	1	2	3	4
5. My legs are strong	0	1	2	3	4
6. I have trouble getting around in public places	0	1	2	3	4
7. I have to make plans around my condition	0	1	2	3	4

Then: A line in the sand.....

Guidance for Industry
 Patient-Reported Outcome Measures:
 Use in Medical Product Development
 to Support Labeling Claims

U.S. Department of Health and Human Services
 Food and Drug Administration
 Center for Drug Evaluation and Research (CDER)
 Center for Biologics Evaluation and Research (CBER)
 Center for Devices and Radiological Health (CDRH)
 December 2009
 Clinical/Medical

....and content validity suddenly becomes a hot topic

- Stimulated debate
- Exposed limitations
- Methodological development
- 2 examples: position paper; guidance
- Transition period with shifting sands

(NB: value of FDA 2009 doc under-appreciated)

Qual Life Res
DOI 10.1007/s11136-011-9990-8

REVIEW

Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting

Susan Magasi · Gery Ryan · Dennis Revicki ·
William Lenderking · Ron D. Hays · Meryl Brod ·
Claire Snyder · Maarten Boers · David Cella

Accepted: 3 August 2011
© Springer Science+Business Media B.V. 2011

Recommendations of Magasi *et al.*

- The adoption of a consensus definition of content validity;
- The development of content validity guidelines;
- Generalizability be assessed by empirical research;
- The use of generic measures as the foundation for PRO assessment in clinical trials.



Content Validity—Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1—Eliciting Concepts for a New PRO Instrument

Donald L. Patrick, PhD, MSPH^{1,*}, Laurie B. Burke, RPh, MPH², Chad J. Gwaltney, PhD³, Nancy Kline Leidy, PhD⁴, Mona L. Martin, RN, MPA⁵, Elizabeth Molsen, RN⁶, Lena Ring, PhD⁷

¹Department of Health Services, University of Washington, Seattle, Silver Spring, MD, USA; ²Department of Community Health, Bro Corporation, Bethesda, MD, USA; ³Health Research Associates, Laurensville, NJ, USA; ⁴Health Economics & Outcomes Research, Uppsala University, Uppsala, Sweden



Content Validity—Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part II—Assessing Respondent Understanding

Donald L. Patrick, PhD, MSPH^{1,*}, Laurie B. Burke, RPh, MPH², Chad J. Gwaltney, PhD³, Nancy Kline Leidy, PhD⁴, Mona L. Martin, RN, MPA⁵, Elizabeth Molsen⁶, Lena Ring, PhD⁷

¹Department of Health Services, University of Washington, Seattle, WA, USA; ²Office of New Drugs, Center for Drug Evaluation Research, Food and Drug Administration, Silver Spring, MD, USA; ³Department of Community Health, Brown University, Providence, RI, USA, and PRO Consulting, Pittsburgh, PA, USA; ⁴United Biosource Corporation, Bethesda, MD, USA; ⁵Health Research Associates, Inc., Seattle, WA, USA; ⁶International Society for Pharmacoeconomics and Outcomes Research, Laurensville, NJ, USA; ⁷Health Economics & Outcomes Research Division, AstraZeneca, Södertälje, Sweden, and Pharmaceutical Outcomes Research, Department of Pharmacy, Uppsala University, Uppsala, Sweden



5 good practice steps:

1. Determine the context of use
2. Develop protocol for qualitative concept elicitation & analysis
3. Conduct the CE interviews & focus groups
4. Analyze qualitative data
5. Document concept development & elicitation methods & results



5 good practice steps:

1. Create draft instrument from findings of concept elicitation
2. Design cognitive interview process to document content validity for the planned context of use
3. Conduct cognitive interviews
4. Revise PRO instrument accordingly
5. Document cognitive interview results for evaluation of content validity

Table 1 – Five steps to elicit concepts for new patient-reported outcome instruments and document content validity consistent with good research practices. Steps to develop an instrument, evaluate the new measure through cognitive interviewing, and document that aspect of content validity are addressed in Part 2 of the Task Force report [2].

1. Determine the context of use (medical product labeling)
 - Understand the disease or condition in the target population
 - Develop an endpoint model for the context of use
 - Consider the target population – cultural/language groups
 - Consider preliminary issues related to instrument content and structure
 - Consider the theoretical and qualitative methodologic approach
 - Develop an hypothesized conceptual framework
2. Develop the research protocol for qualitative concept elicitation and analysis
 - Define the target sample characteristics
 - Select the data collection method – focus groups, individual interviews, both
 - Determine the setting and location for data collection
 - Develop the interview guide—draft, pilot, revise
3. Conduct the concept elicitation interviews and focus groups
 - Obtain institutional review board approval
 - Recruit and train sites
 - Recruit participants; monitor sample characteristics to assure diversity of participation from the target population
 - Select and train interviewers
 - Conduct interviews—implement quality control measures
 - Record or videotape interviews
 - Transcribe and clean transcripts
4. Analyze the qualitative data
 - Analyze qualitative data according to theoretical approach used
 - Establish preliminary coding framework; update as data are coded
 - Establish coding procedures and train coders
 - Organize data using a qualitative research software program
 - Assess saturation
 - Interpret results
5. Document concept development and elicitation methodology and results
 - Provide target claims and any other context for use
 - Describe target population
 - Provide hypothesized and revised disease model and any input from content experts
 - Provide endpoint model
 - Provide conceptual framework and revisions made from preliminary to revised
 - Provide study methods via protocols and guides
 - Provide summary of results, including evidence of saturation
 - Provide transcripts of interviews and focus group
 - Track origin and derivation of concepts captured in the patient-reported outcome instrument
 - Summarize qualitative data
 - Provide key references

1. Develop items based on findings from concept elicitation
 - Develop criteria for item selection according to purpose of instrument and conceptual framework
 - Select recall period and modes of administration
 - Draft instructions
 - Determine wording of each new question
 - Match each new item to response scale
 - Review items against item criteria
 - Select items for cognitive interviews
 - Determine readability
 - Determine order and sequence
 - Format the actual instrument for cognitive interviewing
2. Design cognitive interview process for the planned context of use
 - Identify population
 - Design cognitive interview process
 - Develop protocol and cognitive interview guide
3. Conduct cognitive interviews
 - Train interviewers
 - Train subject to think aloud
 - Use verbal probes
 - Monitor interview quality
 - Record and transcribe
 - Prepare result summaries
4. Make decisions to revise the patient-reported outcome instrument
 - Employ an iterative process
 - Reduce ambiguity in item language
 - Assess saturation
 - Balance respondent input with principles of item construction and decisions on conceptual framework
5. Document cognitive interview results for evaluation of content validity
 - Complete item tracking matrix including final item, final response scale, any preliminary domain assignment, description of intent of item, and patient quotes supporting item intent

Fig. 1 – Five good practices in using cognitive interviews to evaluate patient understanding of a new patient-report outcome instrument.

Some specific debates

- Definitions of content validity
- The role of statistical tests
- Precedence:
 - Qualitative / quantitative ?
 - Patients or experts?
- Which qualitative method ?
 - Phenom.; grounded theory; critical social theory
- Which quantitative method (CTT, IRT, Rasch mt) ?
- How much validation is required ?
- What happens when important items don't "work" ?
- Disease specific or domain specific ?

Proposed definition (Magasi *et al.*)

tion as a starting point: “Content validity is the extent to which a scale or questionnaire represents the most relevant and important aspects of a concept in the context of a given measurement application”. A

- Reminiscent of SS Stevens 1946 defn measurement
 - “assignment of numbers according to rule”
- Not clear how this definition of content validity might be tested and, more importantly, falsified

Using quantitative (statistical) methods in content validity testing (Magasi *et al.*)

While qualitative methods help define the boundaries of a concept and the initial item content, quantitative analyses provide insight into content validity of multi-item and multi-dimensional PRO measures. Quantitative methods can be used to explore and confirm the dimensionality and structure of multi-item scales; evaluate item bias across demographic groups; and examine relationships among health concepts. Results of quantitative psychometric analyses confirm and extend the qualitative research findings. To ensure an instrument's validity, developers would benefit from avoiding a false dichotomization of qualitative and quantitative methods and adopting iterative mixed methods approaches. Confirmation of content validity is dependent on the accumulation of research evidence. However, once sufficient evidence from multiple sources is demonstrated, it is reasonable to conclude that there is enough information on content validity of the targeted PRO. Thus, quantitative evidence is a critical part of the iterative process in developing content valid PRO measures.

In the PRO instrument development process, exploratory and confirmatory factor analysis can evaluate items for fit within a hypothesized domain by demonstrating that items with a specified domain scale load onto the factors [15]. Items that cross-load on multiple factors may be removed from an instrument. Factor analysis allows us to understand the internal structure of a PRO measure and to evaluate the consistency of the factor structure across different samples. Factor analysis also leads to the development of summary scores from multi-domain measures. Confirmatory factor analysis (CFA) within the context of content validity allows for the evaluation of specific

hypotheses about factor structures and content and can be used to hierarchically test for invariance in factor structure across groups [16, 17].

Structural equation modeling (SEM) provides the framework from which CFA can be used to evaluate the extent to which concepts being measured across groups of people have the same characteristics and boundaries in relation to other concepts. Once measurement equivalence is established, SEM can be used to evaluate structural (e.g., group) equivalence. SEM is also used to confirm hypothesized factor structures of PRO measures [18, 19]. SEM is used to evaluate the factorial validity of PRO measures by confirming specified relationships between the PROs and antecedents and consequences of interest. SEM allows researchers to assess multiple domains simultaneously and examine the longitudinal relationships between clinical and PRO end points. Finally, SEM can be used to cross-validate PRO measures across subgroups (i.e., gender, language versions, etc.). SEM allows for the evaluation of complex relationships between clinical and PROs [20]. For example, these models have been used to examine the longitudinal relationships between treatment-related impact on hemoglobin in patients with chemotherapy-induced anemia and the effect of changes in hemoglobin on changes in patient-reported fatigue [21]. SEMs can also be used to evaluate and confirm PRO end point models.

Measurement invariance can also be evaluated using differential item functioning (DIF). DIF examines the relationship among item responses, levels of a concept being measured, and subgroup membership [22, 23]. For any given level of a concept, the probability of endorsing a specific item response should be independent of group membership. There are two kinds of DIF. Uniform DIF is consistent across the range of the concept being measured, and non-uniform DIF varies depending on the concept level. DIF testing can be done with ordinal logistic regression. DIF is identified as a significant effect of subgroup membership on item score after controlling for the level of the concept. The concept level is approximated by summing across items or estimating IRT scores.

IRT information curves show the investigator where an item bank or instrument is not covering the continuum of severity or impairment [22]. This information is useful in targeting when additional development work and domain content coverage is needed.

Using quantitative (statistical) methods in content validity testing (Patrick *et al.*)

In addition to the qualitative work, quantitative evaluation of items, such as assessment of how well items address the entire continuum of patient experience of the concept is useful and desirable, regardless of if the concept is a symptom, behavior, or feeling. Rasch analysis or item-response theory methods can be used to evaluate item information curves and what part of the response continuum items address [26,27]. The use of quantitative data in the absence of prior knowledge, frameworks, and qualitative considerations can lead to a theoretical instrumentation producing scores with unknown meaning. Similarly, the use of qualitative data alone to substantiate an instrument may be rhetorically convincing, but scientifically incomplete.

BUT....

Empirical issues requiring explicit study

These debates likely to continue until we have
explicit objective methods of testing /
disproving "validity"

Why ?.....

A scale is an hypothesis....

These questions ask about limitations to your walking due to MS during the past two weeks.

- For each statement, please circle the one number that best describes your degree of limitation.
- Please answer all questions even if some seem rather similar to others, or seem irrelevant to you.
- If you cannot walk at all, please tick this box:

In the past two weeks, how much have your MS ...

	Not at all	A little	Mod. amount	Quite a bit	Extremely
--	------------	----------	-------------	-------------	-----------

MSWS-12 is an hypothesis of how walking ability could be measured

when walking outdoors (e.g. using a stick, a frame, etc)?

	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Please check that you have circled ONE number for EACH question.

© 2000 Neurological Outcome Measures Unit

(better)



(worse)

A scale is an hypothesis....

These questions ask about limitations to your walking due to MS during the past two weeks.

- For each statement, please circle the one number that best describes your degree of limitation.
- Please answer all questions even if some seem rather similar to others, or seem irrelevant to you.
- If you cannot walk at all, please tick this box:

In the past two weeks, how much have your MS ...

	Not at all	A little	Mod. amount	Quite a bit	Extremely
--	------------	----------	-------------	-------------	-----------

MSWS-12 is an hypothesis of how walking ability could be measured

when walking outdoors (e.g. using a stick, a frame, etc)?

	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Please check that you have circled ONE number for EACH question.

© 2000 Neurological Outcome Measures Unit

(better)



(worse)

..wrestling with two major uncertainties

These questions ask about limitations to your walking due to MS during the past two weeks.

For each statement, please circle the one number that best describes your degree of limitation.

Please answer **all** questions even if some seem rather similar to others, or seem irrelevant to you.

If you cannot walk at all, please tick this box

In the past two weeks, how much have your MS ...

	Not at all	A little	Mod-erately	Quite a bit	Extremely
11. Limited your ability to walk?	1	2	3	4	5

What is the definition of the variable ?
How is it best articulated with “words” ?

10. Slowed down your walking?

	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Please check that you have circled ONE number for EACH question.

© 2005 Neurological Outcome Resource Unit

(better)



(worse)

clinical implications of uncertainty

These questions ask about limitations to your walking due to MS during the past two weeks.

For each statement, please circle the one number that best describes your degree of limitation.

Please answer **all** questions even if some seem rather similar to others, or seem irrelevant to you.

If you cannot walk at all, please tick this box

In the past two weeks, how much have your MS ...

	Not at all	A little	Mod-erately	Quite a bit	Extremely
11. Limited your ability to walk?	1	2	3	4	5

Scale construction is....

an iterative on-going process of hypothesis generation, testing, and revision requiring all the help we can get from available methods

(better)



(worse)

An issue raised 30 yrs ago

Perceptual and Motor Skills, 1982, 55, 415-426.

TESTING CONSTRUCT THEORIES

A. JACKSON STENNER AND MALBERT SMITH, III
NTS Research Corporation¹

Journal of Educational Measurement¹
Volume 20, No. 4 Winter 1983

TOWARD A THEORY OF CONSTRUCT DEFINITION

A. JACKSON STENNER
ComputerLand
MALBERT SMITH III
ComputerLand
DONALD S. BURDICK
Duke University

The real problem....

- Validity methods currently used don't answer the question:
- *"the extent to which an instrument measures what it purports to measure"*
- Circumstantial evidence only
- Root of problem: Data-driven not theory-driven

A Solution for the future

- 1) Theory-driven measurement
- 2) Formal methods to test theories

Developing a Construct Theory

- What is the '*something*' that causes variation?
- Construct definition seeks to test theories about this '*something*', thereby specifying the meaning of a construct

An exemplar

- Stenner, Smith & Burdick (1983)
- 25+ years of work in education

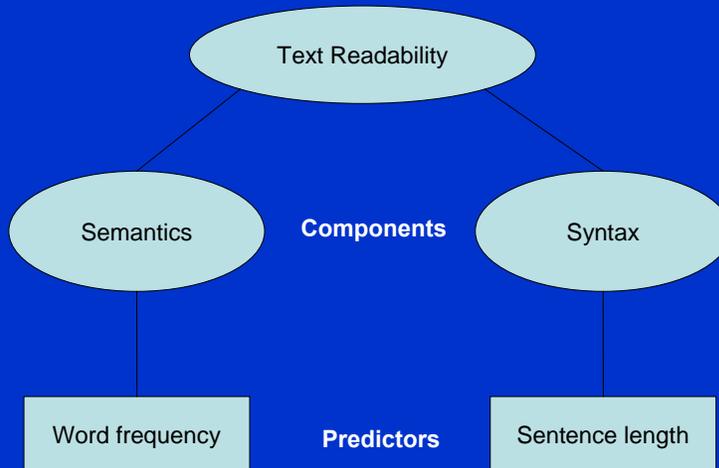


The Lexile Scale for reading ability

Which intrinsic features of text make one passage more difficult for a person to comprehend than another?



Many years and ~50 variables later...

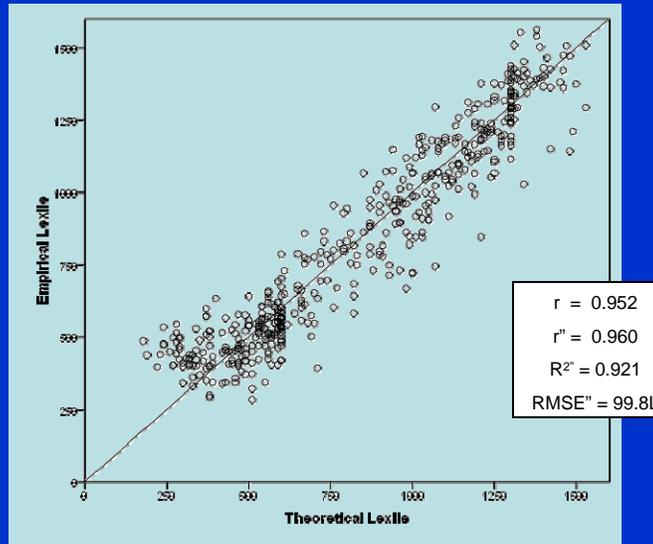


Construct specification equation: links theory to observation

$$\text{Text difficulty} = (a \times \text{LMSL}) - (b \times \text{MLWF}) - c$$



Plot of Theoretical Text Complexity versus Empirical Text Complexity for 475 articles



Yes, but applicability to health measurement?

- Can these principles be applied to health
- For example, measurement of upper limb functioning
- Stage 1 develop a construct theory

A Construct Theory

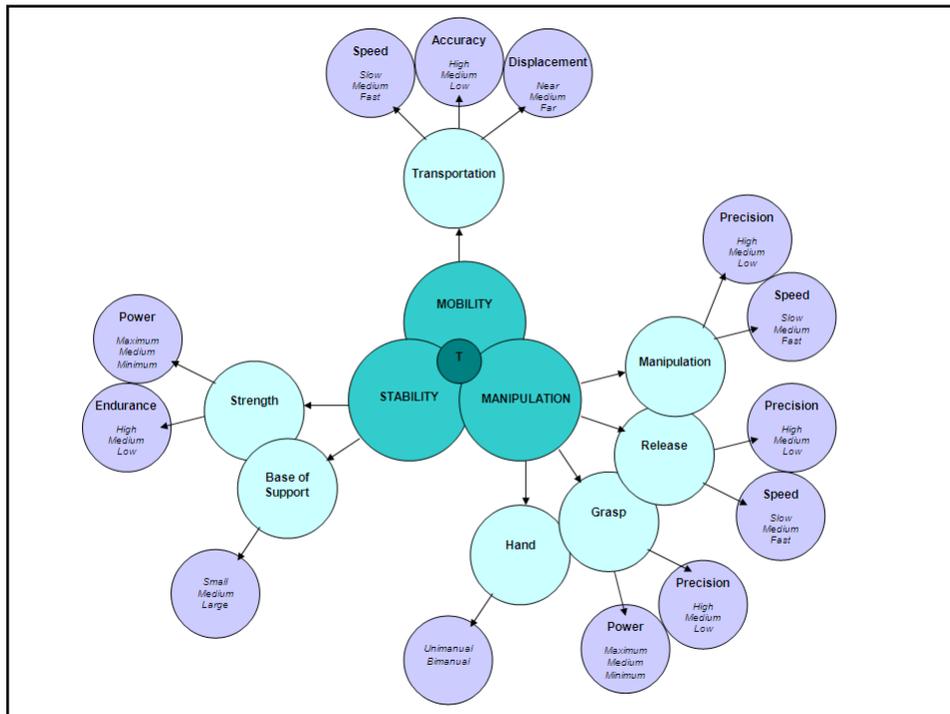
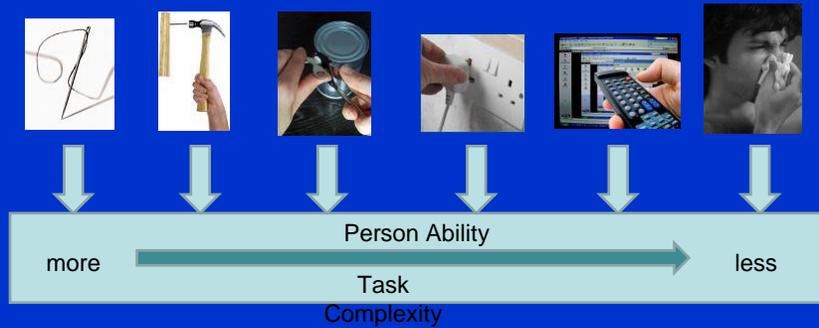
- Identify motor components of tasks that characterise upper limb functioning
- Examine items and identify characteristics that account for variance in task difficulties
- Devise method to test theory against observation

Selected Items from an ULF Scale (easiest to most difficult)

Code	Statement	Location	Rank Order	
IB_33	Take off specs	3.21	1	1
IB_32	Put on specs	3.192	2	2
IB_4	Blow nose	2.798	3	3
IB112	Crumple paper	2.48	8	4
IB133	Turn on TV	1.994	16	5
IB114	Stick stamp	1.579	21	6
IB147	Turn door handle	0.939	38	7
IB_58	Butter slice of bread	0.56	56	8
IB_54	Unwrap sweet	0.546	58	9
IB154	Insert plug	0.042	79	10
IB_65	Open crisps	0.025	81	11
IB113	Wrap up gift	-0.564	108	12
IB_67	Peel fruit	-0.611	110	13
IB_77	Open tin with ringpull	-1.019	127	14
IB163	Use light weights	-1.552	144	15
IB_41	Put in earrings	-1.611	146	16
IB158	Hammer nail	-1.982	155	17
IB_34	Fasten necklace	-2.581	158	18
IB104	Thread needle	-2.767	162	19
IB156	Place object on shelf	-2.925	163	20
IB164	Use heavy weights	-3.403	164	21

Upper Limb Function

Which intrinsic features of upper limb tasks make one task more difficult for a person to perform than another?



Some examples to address Bob Temple's questions

Items are used as indicators of variables

These questions ask about limitations to your walking due to MS during the past two weeks.

For each statement, please circle the one number that best describes your degree of limitation.

Please answer all questions even if some seem rather similar to others, or seem irrelevant to you.

If you cannot walk at all, please tick this box:

In the past two weeks, how much has your MS ...

	Not at all	A little	Mod. anxiety	Quite a bit	Extremely
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
3. Limited your ability to climb up and down stairs?	1	2	3	4	5
4. Made standing when doing things more difficult?	1	2	3	4	5
5. Limited your balance when standing or walking?	1	2	3	4	5
6. Limited how far you are able to walk?	1	2	3	4	5
7. Increased the effort needed for you to walk?	1	2	3	4	5
8. Made it necessary for you to use support when walking outdoors (e.g. holding on to furniture, walls, a stick, etc.)?	1	2	3	4	5
9. Made it necessary for you to use support when walking indoors (e.g. using a stick, a frame, etc.)?	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Please check that you have circled ONE number for EACH question.

© 2000 Neurological Outcome Measures Unit

(better)



(worse)

Rasch analysis

• These questions ask about limitations to your walking due to MS during the past two weeks.

• For each statement, please circle the one number that best describes your degree of limitation.

• Please answer all questions even if some seem rather similar to others, or seem irrelevant to you.

• If you cannot walk at all, please tick this box:

In the past two weeks, how much has your MS ...	Not at all	A little	Moderately	Quite a bit	Extremely
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
3. Limited your ability to climb up and down stairs?	1	2	3	4	5
4. Made standing when doing things more difficult?	1	2	3	4	5
5. Limited your balance when standing or walking?	1	2	3	4	5
6. Limited how far you are able to walk?	1	2	3	4	5
7. Increased the effort needed for you to walk?	1	2	3	4	5
8. Made it necessary for you to use support when walking indoors (e.g. holding on to furniture, using a stick, etc)?	1	2	3	4	5
9. Made it necessary for you to use support when walking outdoors (e.g. using a stick, a frame, etc)?	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Please check that you have circled ONE number for EACH question

© 2000 Neurological Outcome Measures Unit

Rasch analysis of the DASH Cano, Barrett, Zajicek, Hobart Multiple Sclerosis 2011

+ resulting correspondence

Evidence from fatigue in multiple sclerosis for a theory-driven paradigm in scale development & evaluation

Jeremy Hobart¹, Stefan Cano¹, Rachel Baron¹, Alan Thompson²,
Steven Schwid^{3*}, John Zajicek¹ and David Andrich⁴



• These questions ask about limitations to your walking due to MS during the past two weeks.
 • For each statement, please circle the one number that best describes your degree of limitation.
 • Please answer **all** questions even if some seem rather similar to others, or seem irrelevant to you.
 • If you cannot walk at all, please tick this box:

In the past two weeks, how much have your MS...	Not at all	A little	Moderately	Quite a bit	Extremely
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
3. Limited your ability to climb up and down stairs?	1	2	3	4	5
4. Made standing when doing things more difficult?	1	2	3	4	5
5. Limited your balance when standing or walking?	1	2	3	4	5
6. Limited how far you are able to walk?	1	2	3	4	5
7. Increased the effort needed for you to walk?	1	2	3	4	5
8. Made it necessary for you to use support when walking indoors (e.g. holding on to furniture, using a stick, etc)?	1	2	3	4	5
9. Made it necessary for you to use support when walking outdoors (e.g. using a stick, a frame, etc)?	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Please check that you have circled ONE number for EACH question.
© 2006 International Multiple Sclerosis Society

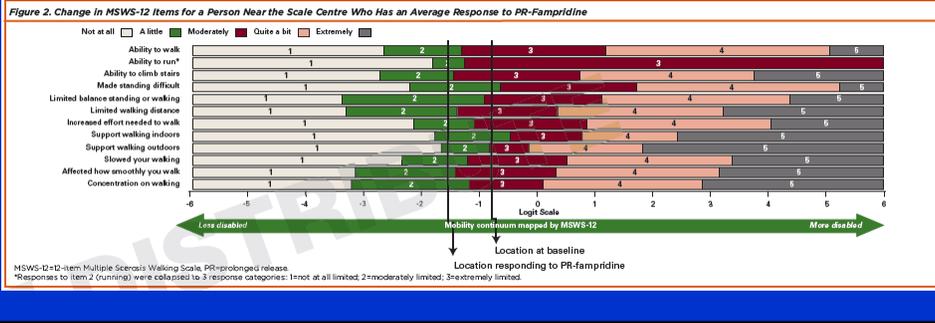


Figure 3: The Clinical Impact of PR-Fampridine for a Person Near the Middle of the MSWS-12 Range Who Has an Average Responder Change

MSWS-12 Item	Response	
	Off PR-Fampridine	On PR-Fampridine
1. Ability to walk	Moderately	A little
2. Ability to run	Extremely	Moderately
3. Ability to climb stairs	Moderately	A little
4. Made standing difficult	A little	A little
5. Limited balance standing or walking	Moderately	A little
6. Limited walking distance	Moderately	A little
7. Increased effort needed to walk	Moderately	A little
8. Support walking INDOORS	A little	A little
9. Support walking OUTDOORS	Moderately	A little
10. Slowed your walking	Moderately	A little
11. Affected how smoothly you walk	Moderately	A little
12. Concentrate on walking	Moderately	A little
		10 items change

MSWS-12=12-item Multiple Sclerosis Walking Scale.

Conclusions

- Not easy. Shifting sands. I empathise
- Scale content determines what is measured (validity)
- Importance of validity: can't settle for weak science
- FDA guidance: line in sand and stimulus to field
- Guidance, esp. Patrick et al. important
- Do advances tell us what we are measuring ?
- Need definitions & testable theories for "constructs"
- *[Also need for empirical work into what to measure]*
- Change paradigm: A scale is just an hypothesis
- Some things may not be "measurable"