Panel Discussion 1:
Mixed Methods in Assuring Content Validity

Fourth Annual

Patient-Reported Outcome (PRO) Consortium Workshop

April 24, 2013 ■ Silver Spring, MD

Co-sponsored by

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Disclaimer

The views and opinions expressed in the following PowerPoint slides are those of the individual presenters and should not be attributed to their respective companies, the Critical Path Institute, the PRO Consortium, or the ePRO Consortium.

These PowerPoint 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 Outline & Participants

- *Mixed Methods – FDA Perspective: Incorporating Mixed Methods to Enhance Content Validity in Drug-Development Tools*
  - **Moderator**: *Ashley Slagle, MS, PhD* – ORISE Fellow, Study Endpoints and Labeling Development (SEALD), Office of New Drugs (OND), CDER, FDA
  - **Presenter**: *James P. Stansbury, PhD, MPH* – Consumer Safety Officer, SEALD, OND, CDER, FDA
  - **Panelists**: *Laurie Burke, RPh, MPH* – Associate Director, OND, SEALD, CDER, FDA; *Lisa Kammerman, PhD* – Master Reviewer, Office of Biostatistics, CDER, FDA; *Scott Komo, DrPH* – Senior Statistical Reviewer, Office of Biostatistics, CDER, FDA; *Päivi Miskala, MSPH, PhD* – Study Endpoints Reviewer/Senior Clinical Analyst, SEALD, OND, CDER, FDA; *James Stansbury*

- *Mixed Methods – Industry and Academic Experience*
  - **Moderator**: *Josephine M. Norquist, MS* – Patient-Reported Outcomes Specialist, Department of Epidemiology, Merck Sharp & Dohme Corporation
  - **Presenters and Panelists**: *Joseph C. Cappelleri, PhD, MPH* – Senior Director, Biostatistics, Pfizer Inc.; *Ron D. Hays, PhD* – Professor, Department of Medicine, David Geffen School of Medicine, UCLA

Mixed Methods to Enhance Content Validity of Measures for Use in Drug-Development Trials

James P. Stansbury, PhD, MPH
Endpoints Reviewer

Study Endpoints and Labeling Development Staff
Office of New Drugs, CDER, FDA

Outline

- Introduction, terms, and scope of the discussion
- Why flexibility is useful—issues and solutions in instrument development and revision
- Conceptualizing alternative schemas

The goal is to…

- encourage development of the best possible evidence for content validity before moving to evaluation of other measurement properties
- offer suggestions that promote efficiency and add flexibility, not create new guidance
- open dialogue relevant to regulatory science, not restrict the path toward better measures for drug-development

Content Validity

"Evidence from qualitative research demonstrating that the instrument measures the concept of interest including evidence that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use. Testing other measurement properties will not replace or rectify problems with content validity."

- Attention to concept(s), domains, and items
- Recall period, scales, and item framing
- Perspectives from the target population

Establishing Content Validity: Review of Basics from FDA PRO Guidance

- Task that follows clear preliminary conceptualization—concept(s) and specific context(s) of use are appropriate

- Assessment of content validity requires evidence specific to the proposed context of use
  - If existing instrument is used for a new population or condition, additional evidence may be needed

- Content validity must be established before other evidence of construct validity, reliability or sensitivity to change can be interpreted
  - For older measures, content validity documentation is often unavailable

Advances with the PRO Guidance

- Well-Documented Qualitative studies to ensure content
  - Concept elicitation
  - Cognitive debriefing
- Strong emphasis on the patient perspective (for patient-reported outcome (PRO) instruments)

BUT challenges related to interpreting qualitative data may include…

- ambiguous meaning from discordant data
- challenges in targeting measures to populations
- difficulties balancing comprehensiveness and parsimony

Why Mixed Methods for Content Validation

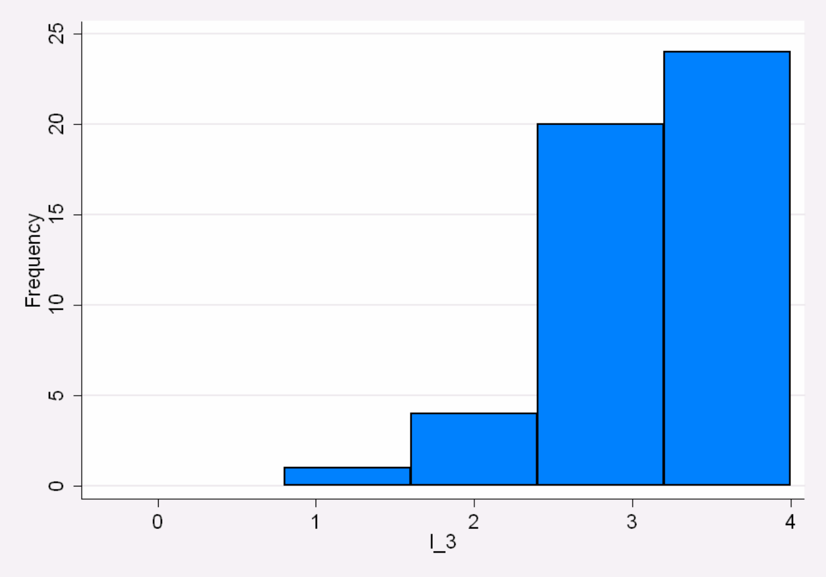
Intelligent, conscious integration of quantitative and qualitative data[†] in early instrument development or revision can help:

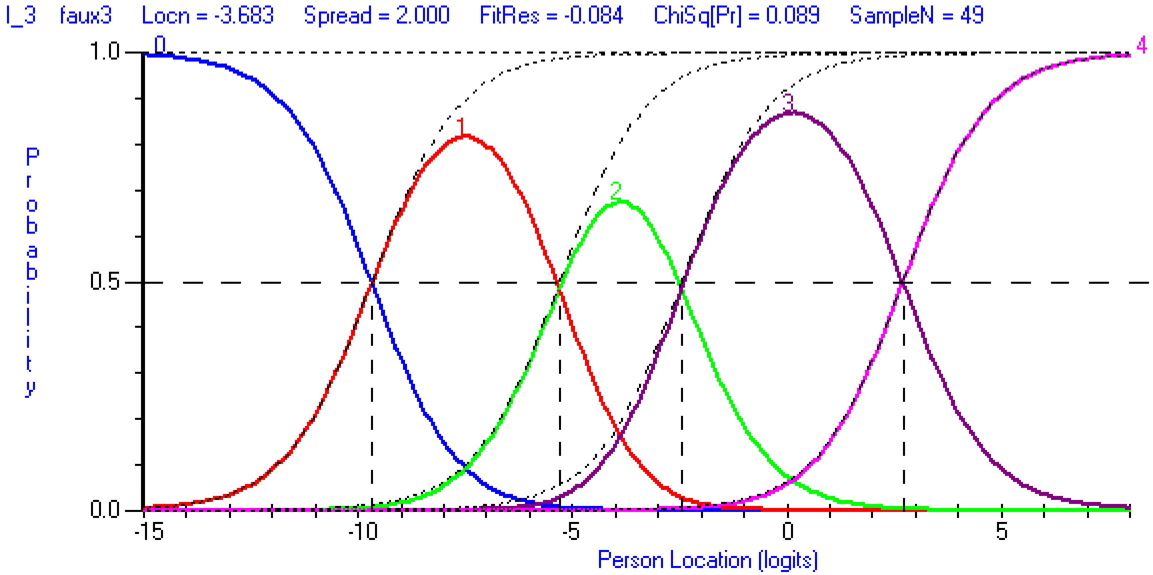
- aid item selection and flag item problems not always evident in qualitative interviewing
- gain an early "check" on measurement properties and glimpse of egregious problems using relatively small, well-targeted samples
- ensure a better match of the measure with population
- avoid gaps in measurement, and/or clusters of items.

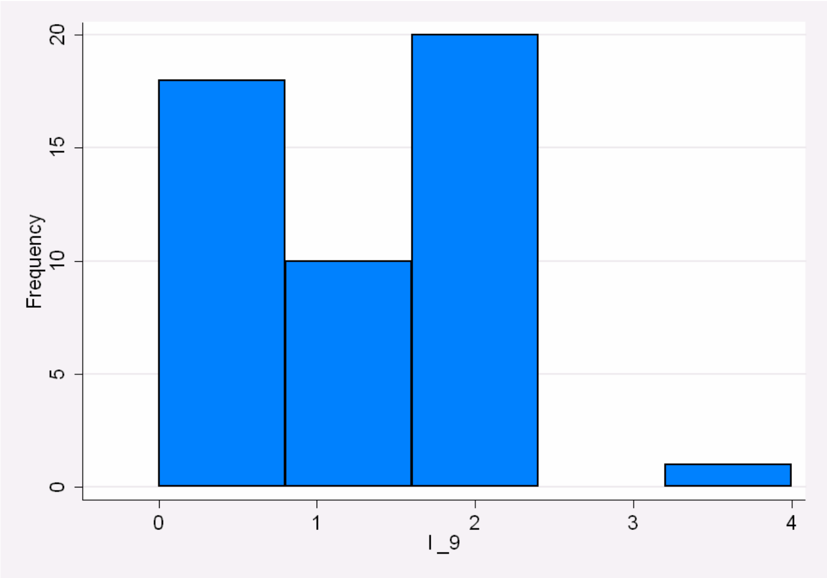
Creswell JW, Klassen AC, Plano Clark VI, Clegg Smith, K. (2011) *Best Practices for Mixed Methods Research in The Health Sciences*. Bethesda: NIH/OBSSR, under supervision of HI Meissner.

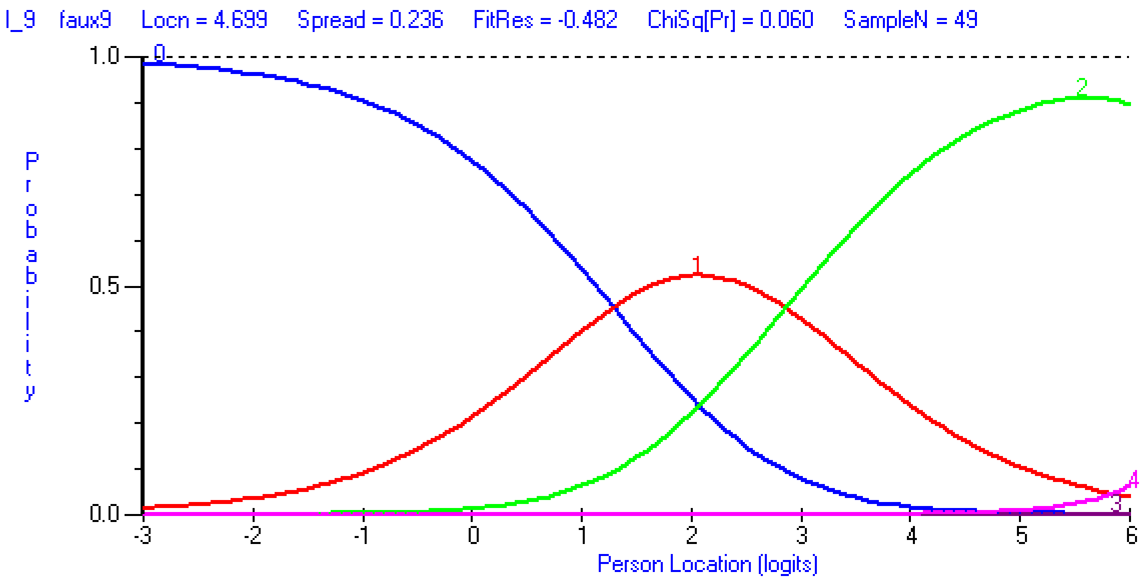
Exploratory Quantitative Analysis

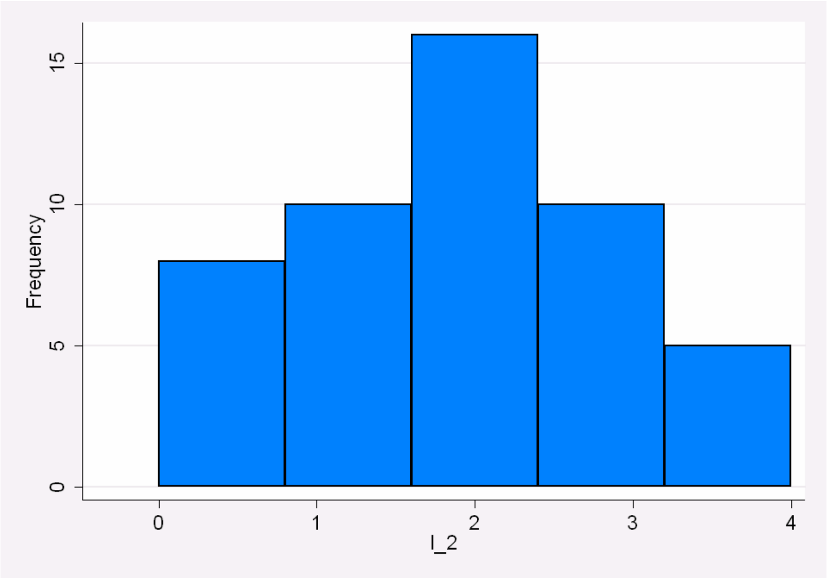
The SESIRnQ

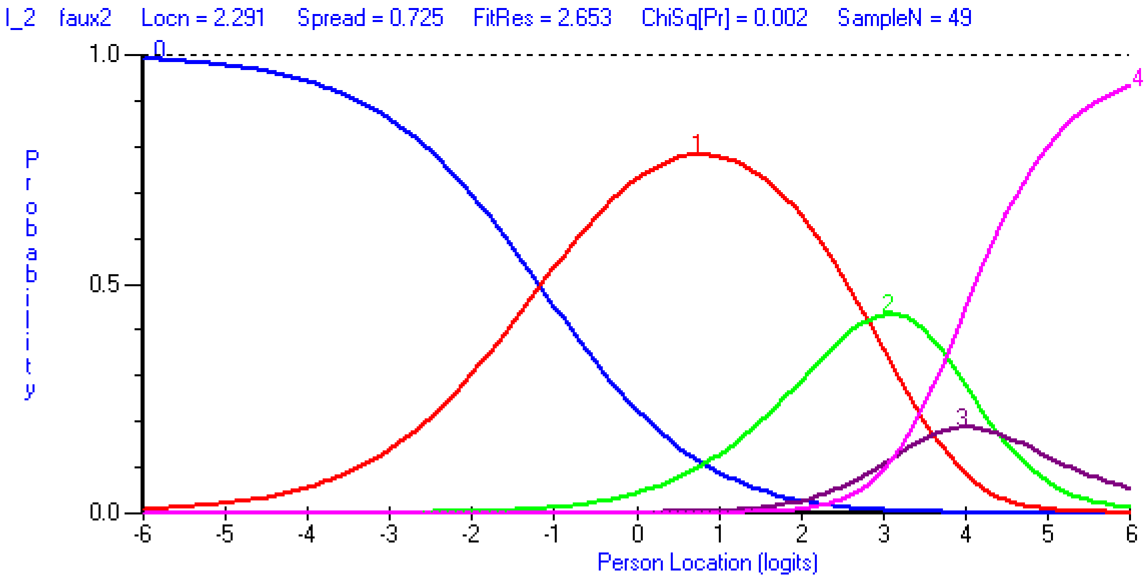

Frequency
0
5
10
15
20
25
0
1
2
3
4
I_3


I_3 faux3 Locn = -3.683 Spread = 2.000 FitRes = -0.084 ChiSq[Pr] = 0.089 SampleN = 49
Probability
1.0
0.5
0.0
0
1
2
3
4
-15
-10
-5
0
5
Person Location (logits)


Frequency
0
5
10
15
20
0
1
2
3
4
I_9


I_9 faux9 Locn = 4.699 Spread = 0.236 FitRes = -0.482 ChiSq[Pr] = 0.060 SampleN = 49
Probability
1.0
0.5
0.0
0
1
2
3
4
-3
-2
-1
0
1
2
3
4
5
6
Person Location (logits)


Frequency
0
5
10
15
0
1
2
3
4
I_2


I_2 faux2 Locn = 2.291 Spread = 0.725 FitRes = 2.653 ChiSq[Pr] = 0.002 SampleN = 49
Probability
1.0
0.5
0.0
0
1
2
3
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Person Location (logits)

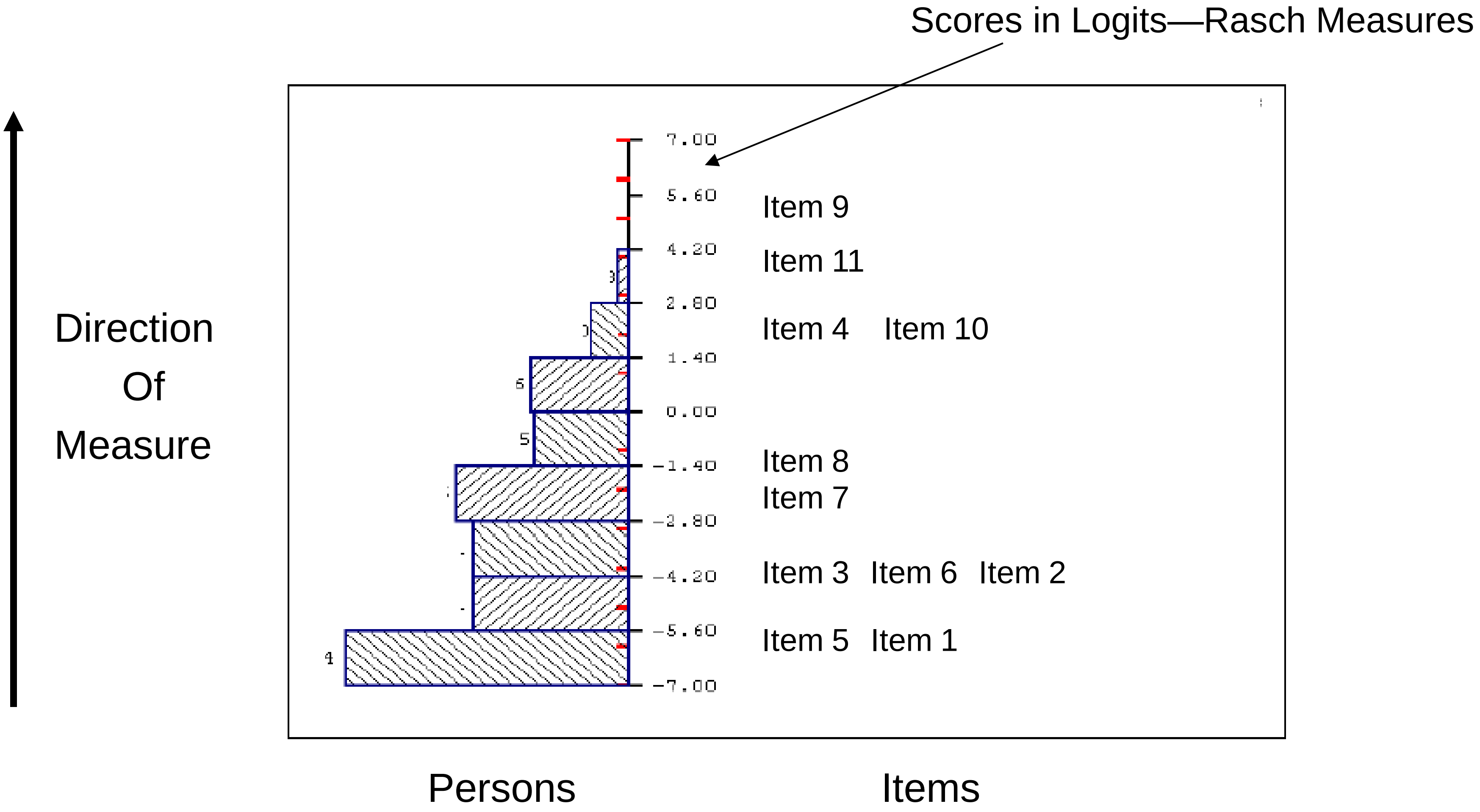
```
                                                average
                          item-test   item-rest   inter-item
Item         | Obs  Sign  correlation correlation covariance   alpha
i_1          |  49   +      0.9099      0.8765    .3869473    0.8292
i_3          |  49   +      0.8043      0.7225    .3913265    0.8434
i_4          |  49   +      0.7235      0.6357    .4318452    0.8594
i_5          |  49   +      0.7711      0.6776    .4002551    0.8499
i_9          |  49   +      0.8806      0.8008    .3244473    0.8249
i_2          |  49   +      0.7691      0.5818    .346301     0.8925
Test scale   |                                   .3801871     0.8713
```

```
RUMM2030        Project: FAUXDATA        Analysis: RATINGS
Title:  RATING SCALE ANALYSIS                         Date: 16 Apr 2013 10:45:41 PM
Display: INDIVIDUAL ITEM-FIT - Serial Order

Seq  Item Type  Location  SE    Residual   DF     ChiSq DF   Prob      F-stat  DF1  DF2   Prob

1    I_1  Poly   -4.990  0.422  -0.872  37.00   5.810  4  0.213825   5.243  4   44  0.001529
2    I_3  Poly   -3.683  0.313  -0.084  37.00   5.595  4  0.231483   1.874  4   44  0.131905
3    I_4  Poly    0.674  0.347   0.660  37.00   4.619  4  0.328696   1.100  4   44  0.368765
4    I_5  Poly    1.009  0.313   0.829  37.00   6.317  4  0.176718   1.375  4   44  0.257918
5    I_9  Poly    4.699  0.281  -0.482  37.00   5.975  4  0.200999   3.890  4   44  0.008633
6    I_2  Poly    2.291  0.171   2.653  37.00  14.784  4  0.005171   3.145  4   44  0.023300
```

Person-Item Map



Scores in Logits—Rasch Measures

Direction Of Measure

Item 9

Item 11

Item 4 Item 10

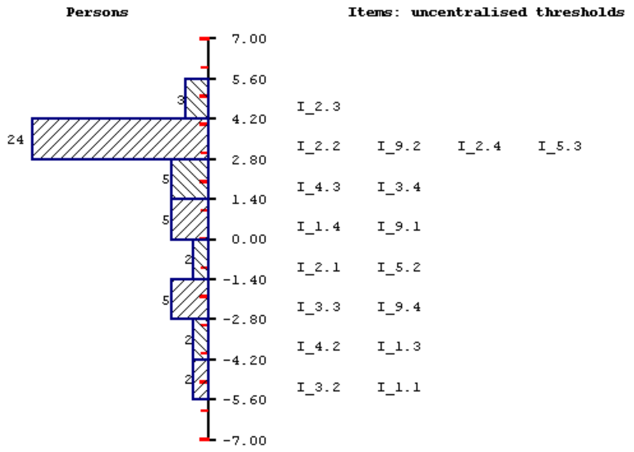
Item 8
Item 7

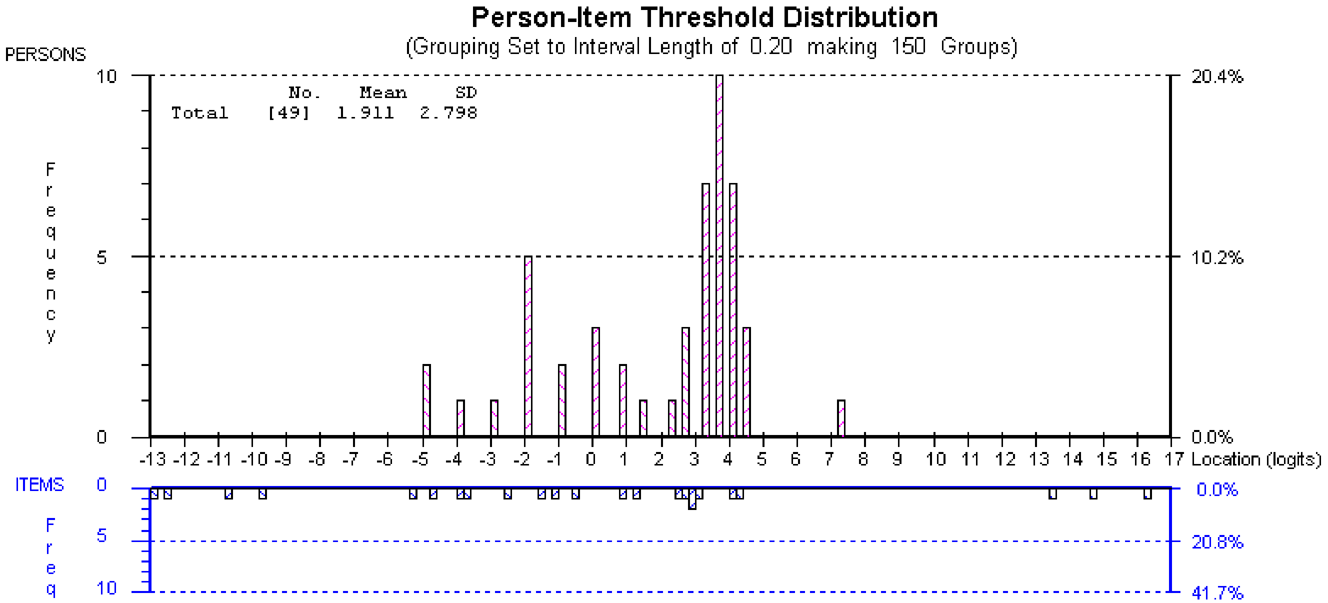
Item 3 Item 6 Item 2

Item 5 Item 1

Persons

Items




Person-Item Threshold Distribution

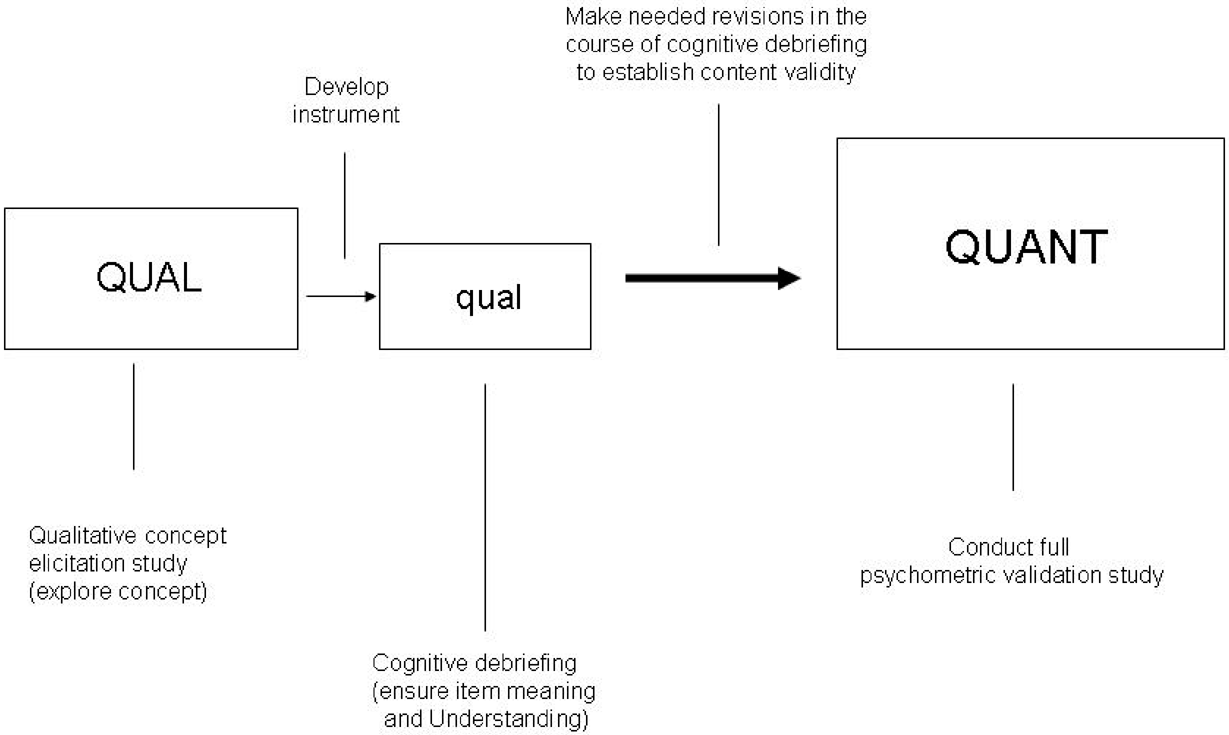
(Grouping Set to Interval Length of 0.20 making 150 Groups)

```
             No.    Mean     SD
Total      [49]   1.911   2.798
```

Conceptualizing

Mixed Methods Approaches

Standard Approach



Make needed revisions in the course of cognitive debriefing to establish content validity

Develop instrument

QUAL

qual

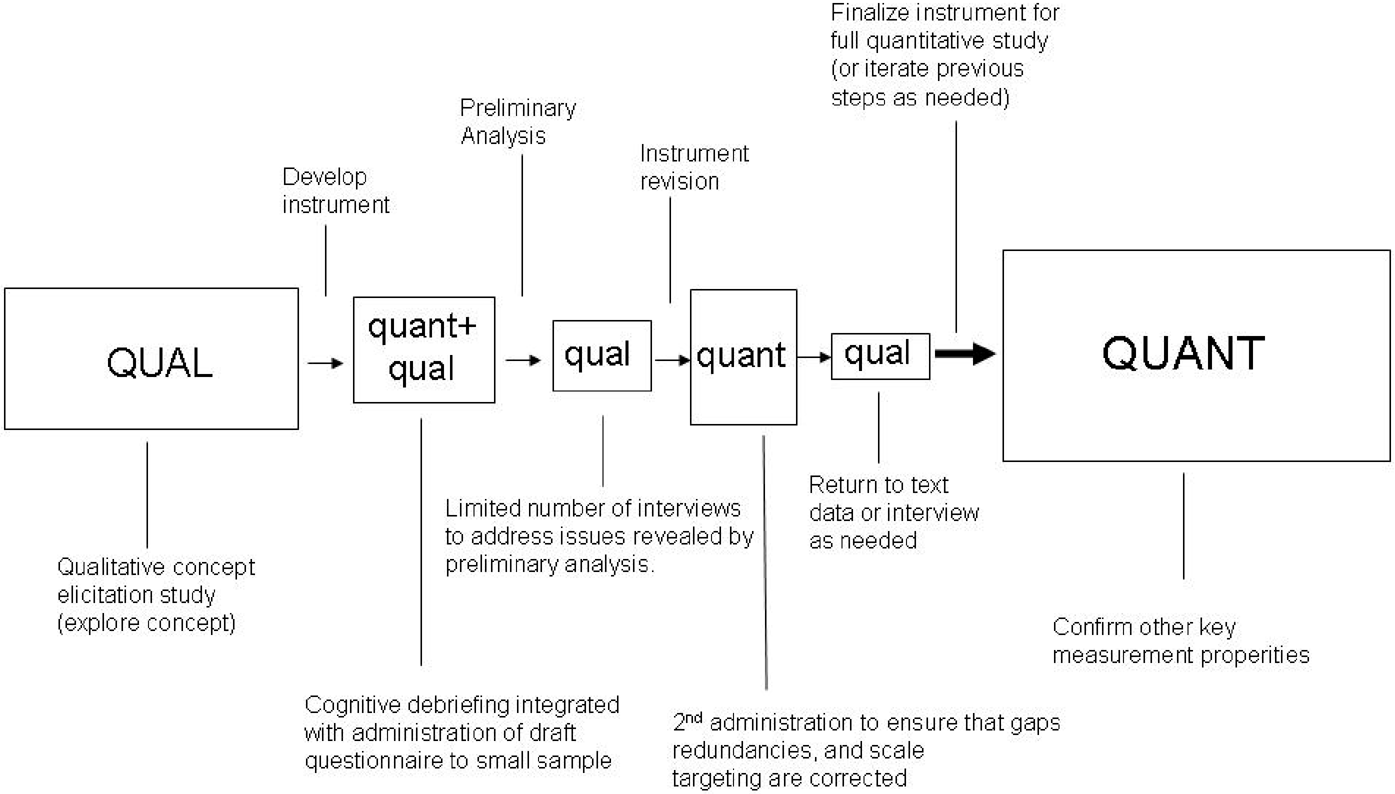
QUANT

Qualitative concept elicitation study (explore concept)

Cognitive debriefing (ensure item meaning and Understanding)

Conduct full psychometric validation study

An Iterative Approach



QUAL → quant+ qual → qual → quant → qual → QUANT

Develop instrument

Preliminary Analysis

Instrument revision

Finalize instrument for full quantitative study (or iterate previous steps as needed)

Qualitative concept elicitation study (explore concept)

Cognitive debriefing integrated with administration of draft questionnaire to small sample

Limited number of interviews to address issues revealed by preliminary analysis.

2nd administration to ensure that gaps redundancies, and scale targeting are corrected

Return to text data or interview as needed

Confirm other key measurement properities

Conclusions

- The traditional schema remains acceptable—but we see promise for greater flexibility
- Thoughtful integration of quantitative and qualitative methods may help…
  - ensure a better match of the measure with population
  - avoid gaps in measurement, and/or clusters of items
  - aid item selection and flag item problems not always evident in qualitative interviewing
  - gain an early “check” on measurement properties and glimpse of egregious problems using relatively small, well-targeted samples.

Item Calibration Stability (extent to which item difficulty parameter is stable relative to sample size)

Item Calibrations stable within	**Confidence**	**Minimum sample size range (best to poor targeting)**	**Size for most purposes**
± 1 logit	95%	16 – 36	30 (minimum for dichotomies)
± 1 logit	99%	27 – 61	50 (minimum for polytomies)
± ½ logit	95%	64 – 144	100
± ½ logit	99%	108 -- 243	150
Definitive or High Stakes	99%+ (Items)	250 -- 20*test length	250
Adverse Circumstances	Robust	450 upwards	500

“Small sample size? You can certainly perform useful exploratory work using Rasch analysis with a small sample. One of the foundational books in Rasch analysis, "Best Test Design" (Wright & Stone, 1979), is based on the analysis of a sample of 35 children and 18 items. The problem is not Rasch analysis, the problem is that a small sample is small for any type of definitive statistical analysis. There would be the same problem with any other type of statistical analysis.”

Linacre JM. 2012 [1994] Sample Size and Item Calibration Stability. *Rasch Measurement Transactions*. 7(4):328.

Mixed Methods Approach to Evaluating Content Validity: Review and Update

Joseph C. Cappelleri

Pfizer Inc

Fourth Annual
Patient-Reported Outcome (PRO) Consortium Workshop

April 24, 2013 ■ Silver Spring, MD

Co-sponsored by

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Instrument Development Process

Previous Process		Current Process
Scoping Stage	→	Scoping Stage
Qualitative Research Stage • Qualitative Interviews, no quantitative testing	→	Content Validity Stage • Mixed Methods – Qualitative Interviews & Quantitative Assessments
Quantitative Research Stage • Confirmatory Psychometric Analyses	→	Psychometric Analysis Stage • Confirmatory Psychometric Analyses

Mixed Methods

- Blends qualitative and quantitative methodologies into the assessment of content validity
- The approach is cyclical, iterative, and hypothesis-driven
- Anomalies that are detected should be explained, modifications to the instrument should be made, and further testing conducted

Historical Thread

- March 2012: Webinar on benefits of Rasch measurement model
- April 2012: C-Path panel on mixed methods approach to ensuring content validity
- June 2012: ISPOR Panel on classical test theory, item response theory, and Rasch measurement theory
- June 2012: Meeting at FDA
- October 2012: ISOQOL Panel & December 2012: Webinar -- Rasch modeling with small samples

Webinar on Benefits of Rasch Measurement Model (March 2012)

- Stacie Hudgens (presenter)
- Josephine Norquist, Denise Globe, Bryce Reeve (discussants)

- Rasch measurement model models the probability of a specific response based on item difficulty (severity) and person ability (ability)

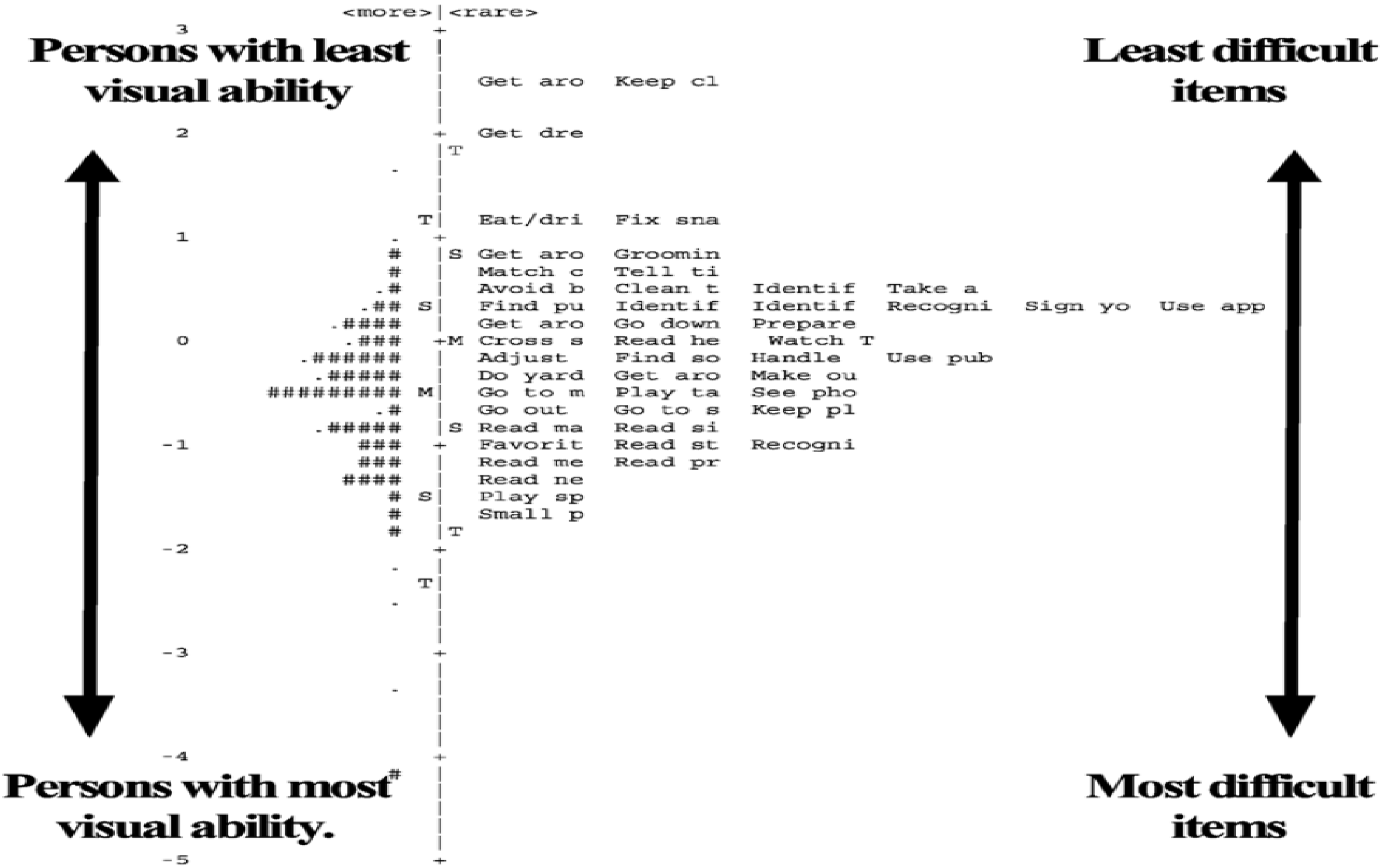
- The higher a person’s level on the underlying construct, the more likely they are to endorse more severe symptom severity (a higher score represents more symptom severity)

Rasch Person-Item Map

- A way to visualize the patient distribution relative to the item distribution

- Can assess the following
  - Presence of items at the ceiling
  - Gaps in the item distribution
  - Redundancy of items in the distribution

Webinar on Benefits of Rasch Measurement Model (March 2012): Stelmark et al. *Journal of Rehabilitation Research & Development* 2004 41(2):233-242.



Persons with least visual ability

Least difficult items

```
            <more>|<rare>
   3              +
                  |
                  |   Get aro   Keep cl
                  |
   2              +   Get dre
                  |T
              .   |
                  |
                  |
                T |   Eat/dri   Fix sna
   1          .   +
              #   |S  Get aro   Groomin
              #   |   Match c   Tell ti
             .#   |   Avoid b   Clean t   Identif   Take a
            .## S |   Find pu   Identif   Identif   Recogni   Sign yo   Use app
          .####   |   Get aro   Go down   Prepare
   0       .###   +M  Cross s   Read he    Watch T
        .#######  |   Adjust    Find so   Handle    Use pub
         .######  |   Do yard   Get aro   Make ou
     ########## M |   Go to m   Play ta   See pho
             .#   |   Go out    Go to s   Keep pl
         .######  |S  Read ma   Read si
  -1        ###   +   Favorit   Read st   Recogni
            ###   |   Read me   Read pr
           ####   |   Read ne
              # S |   Play sp
              #   |   Small p
              #   |T
  -2              +
              .   |
                T |
              .   |
                  |
  -3              +
                  |
              .   |
                  |
  -4              +
              #   |
                  |
  -5              +
```

Persons with most visual ability.

Most difficult items

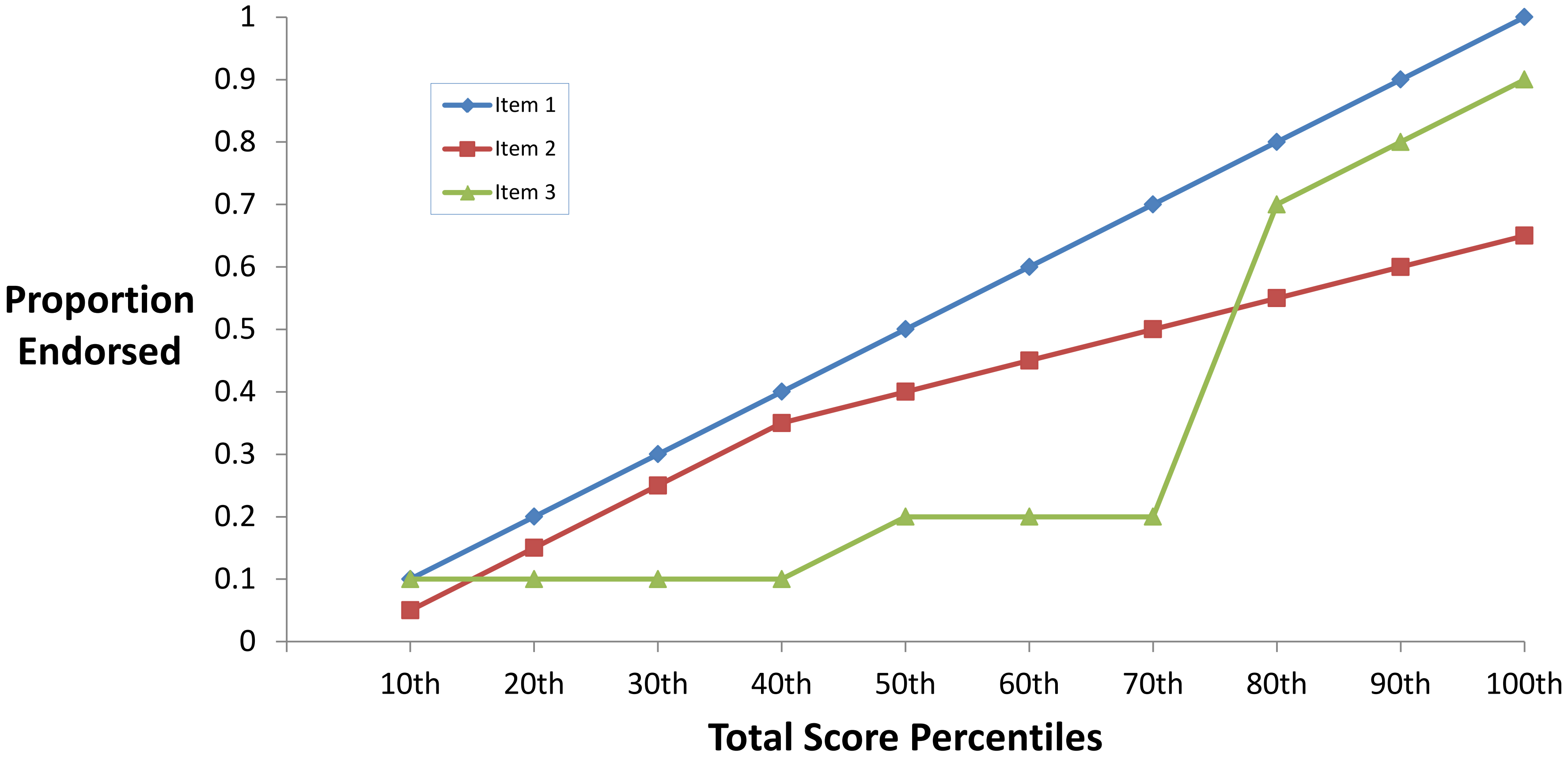
C-Path Panel: Mixed Methods Approach to Ensuring Content Validity (April 2012)

- J. Jason Lundy (organizer)
- Joseph C. Cappelleri, Jeremy Hobart, Ron D. Hays (presenters)
- James P. Stansbury (FDA response)

- Descriptive merits of classical test theory
  - Item difficulty, item-scale correlations (discrimination), curves
  - Reliability
  - Analogies made to item response theory

- Benefits of item response theory
  - Item fit
  - Response theory ordering
  - Targeting and precision

Classical Test Measurement: Item Curves



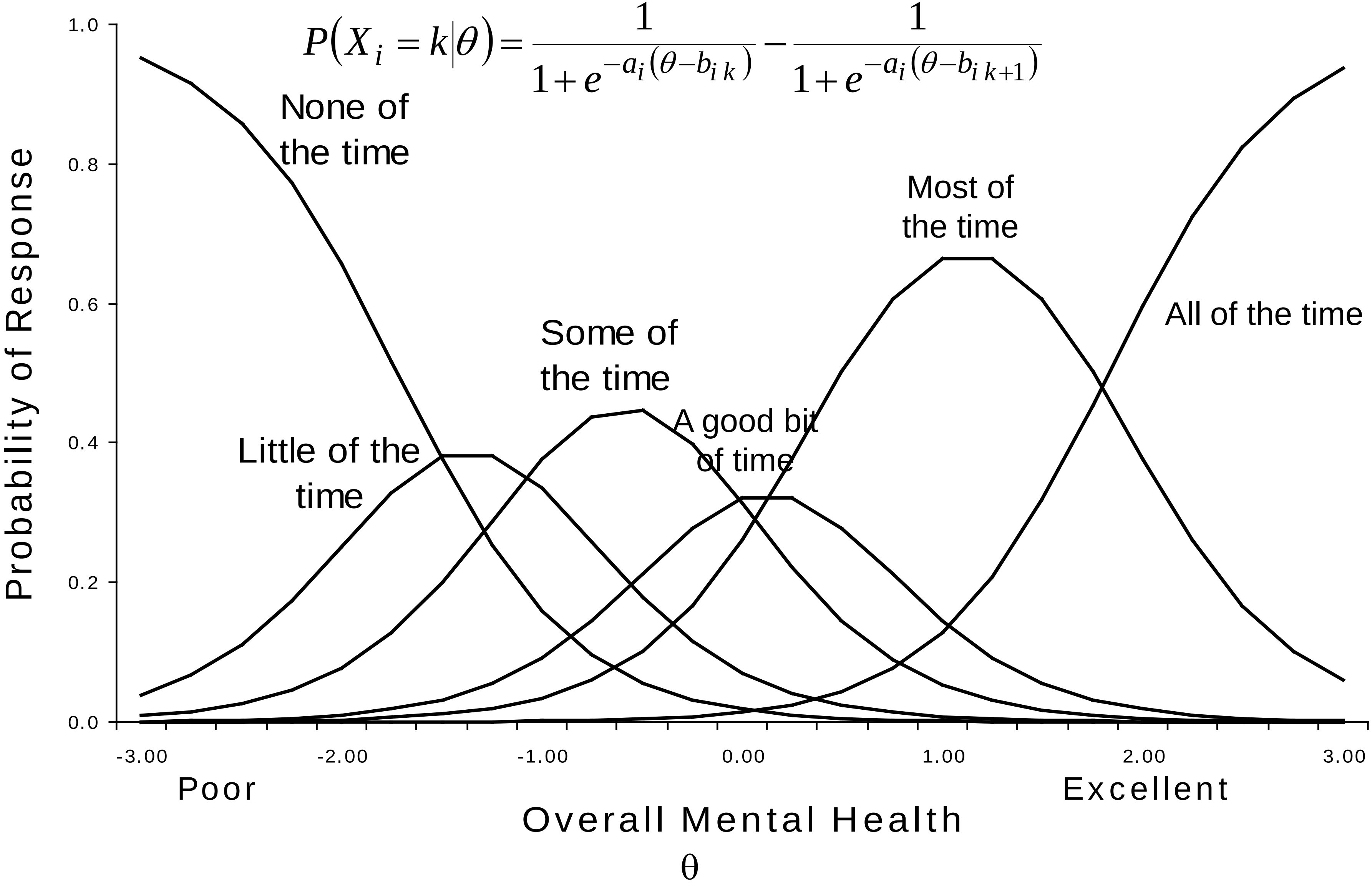
Item 1: Equally good at discriminating across the continuum of the attribute

Item 2: Discriminates better at the lower end than at the upper end of the attribute

Item 3: Discriminates better at upper end, especially between 70th and 80th percentiles

Item Response Theory - Graded Response Model: Item Characteristic Curve for One Item

Did you have a lot of energy?



$$P\left(X_i = k \middle| \theta\right) = \frac{1}{1+e^{-a_i(\theta-b_{i\,k})}} - \frac{1}{1+e^{-a_i(\theta-b_{i\,k+1})}}$$

ISPOR Panel on Classical Test Theory, Item Response Theory, and Rasch Measurement Theory (June 2012)

- Jennifer Petrillo (moderator)
- Stefan Cano, Lori McLeod, Cheryl Coon (presenters)

- Baseline data from trial of 240 patients with visual impairment due to diabetic macular edema

- Evaluation of 25 items on Visual Functioning Questionnaire (assumed unidimensional for this exercise)

- Comparisons between classical test theory, item response theory (graded response model), and Rasch measurement theory

Empirical Lessons (ISPOR Panel, June 2012)

- In this theoretical exercise, each methodology provided complementary information with the potential to optimize instrument composition and scoring

- Some similarities among the three approaches as to what items to keep and what items to modify or delete

- Why are the recommendations different?
- How would your recommendations affect the development of an instrument?
- As a PRO instrument developer or sponsor, why should I use YOUR method over the others?

Meeting at FDA (June 2012)

- Laurie Burke, June Cai, Stefan Cano, Joe Cappelleri, Wen-Hung Chen, Cheryl Coons, Stephen Joel Coons, Sheri Fehnel, Jeremy Hobart, Stacie Hudgens, Lisa Kammerman, Dianne Kennedy, Bob Massof, Paive Miskala, Elektra Papadopoulos, Donald Patrick, Elisabeth Piault-Louis, James Stansbury, Jessica Voqui, Marc Walton

- Quantitative Content Validity Testing
  - Explore evaluation of item content using analyses
  - Assist as a "guide to sensible thinking" in early instrument development
  - Make decisions about whether to go forward with full psychometric testing
  - Or iterate with continued qualitative research
  - Mitigate risk related to Phase 3 signal detection and interpretation

Meeting at FDA (June 2012)

What is the range of item responses relative to the sample (distribution of item responses/endorsement)? How does the sample utilize the categories across the range of responses? What are the frequencies of endorsement of individual items?
Does the instrument measure across the full range of the population (scale to sample targeting)? What is the distribution of the total scores? Are there ceiling or floor effects?
Are the response options used by patients in an informative fashion and as intended? Does a higher response option mean more of a problem than a lower response option? Do the intervals have meaning?
Does the item order reflect the clinically hypothesized item order, if relevant? Does item order reflect the importance/bother ratings from the patients?

ISOQOL Panel (October 2012) and Webinar (December 2012): Rasch Modeling with Small Samples

- Heather Gelhorn, Kathy Wyrwich, Wen-Hung Chen, William Lenderking, Ying Jin, Dennis Revicki
- 768 subjects from the PROMIS pain behavior item bank were used to generate subsets of small samples for the Rasch modeling
- 10 items selected as an unidimensional subset
- Samples of 30, 50, 100, and 250 were randomly drawn 10 times each from the total sample
- Rasch analysis was conducted for each of the random samples, as well as the full sample

Conclusions

- Based on the results from larger samples, the conclusion would be totally the opposite

- Contradictory results were primarily due to the less robust estimation of the threshold parameters caused by the sparse data when sample size was small

- Results of this study suggest that Rasch modeling on small sample size is not recommended

Questions Posed to David Andrich and Mike Linacre in Separate E-mails to Them

- You have recommended at least 10 observations per category are needed for polytomous items.

- Is that at least 10 observations per category for each item?

- Therefore, for nine items with five categories each (four thresholds), assuming a rating scale model (the same set of five categories per item), what is the minimum sample size needed?

- Would it be 50 individuals (i.e., 10 observations times five categories)?

  Or would it be 50 times 9 items = 450 individuals?

  Or would it be something else?

Response from David

- David: Some number of the order of 450.

Dialogue with Mike

- Mike: If each individual responds to all 9 items, then the person sample for the "10 for each category" criterion could be as small as 5 x 10 = 50. OK?

- Joe: I ask the same question to David Andrich and he said "Some number of the order of 450.“

- Mike: David and I are answering different questions.

  David's answer: "450 is a robust sample size assuming somewhat adverse conditions."

  My answer: "50 is the minimum possible sample size assuming perfect conditions."

Dialogue with Mike (continued)

Mike: The table on http://www.rasch.org/rmt/rmt74m.htm is a useful guide (now updated to match David's recommendation).

Rules *Intended* for Dichotomous Items

Item Calibrations stable within	Confidence	Minimum sample size range (best to poor targeting)	Size for most purposes
± 1 logit	95%	16 -- 36	30 (minimum for dichotomies)
± 1 logit	99%	27 -- 61	50 (minimum for polytomies)
± ½ logit	95%	64 -- 144	100
± ½ logit	99%	108 -- 243	150
Definitive or High Stakes	99%+ (Items)	250 -- 20*test length	250
Adverse Circumstances	Robust	450 upwards	500

Dialogue with Mike (continued)

- Joe: It would also be helpful if you would define the terms "adverse conditions" and "perfect conditions" in terms of sample size estimation.

- Mike: Perfect fit: item mean-squares in the range 0.8 - 1.2

  Ordinary fit: item mean-squares in the range 0.5 - 1.5

  Adverse fit: item mean-squares in the range 0.0 - 2.0

  Glaring misfit: several item mean-squares > 2.0, maybe with zero or negative point-biserials

These are rough guidelines with grey areas and they often must be adjusted for reality, see "Reasonable Mean-square fit values"

http://www.rasch.org/rmt/rmt83b.htm

Dialogue with Mike (continued)

- Joe: Have you had an opportunity yet to critique the attached webinar presentation?

- Mike: Rasch is not concerned with content validity (in the title of the webinar). Rasch cannot know what is and what is not included in the content area. Rasch is concerned about construct validity.

- Mike: Does the item hierarchy make sense? If it does, we have success! If it does not, then the instrument (not Rasch) is in trouble.

Dialogue with Mike (continued)

- Mike: A quick scan of the 46 slides suggests that the authors are too rigid in their application of Rasch methodology. According I agree with their conclusion:
  - "Results of this study suggest that Rasch modeling on small sample size is not recommended [in the way that the authors apply it]"
- Mike: My conclusion would be "Results of this suggest that Rasch modeling should be applied in a different, more substantive, more communicative and less statistical way."
- Mike: This webinar does not display even one model or empirical ICC. Pictures are much, much better at communicating information than tables of numbers.

Potential Use of Online Panels as part of a Mixed Methods Approach to Evaluating Content Validity

RON D. HAYS

UCLA DEPARTMENT OF MEDICINE

4TH ANNUAL PATIENT-REPORTED OUTCOME (PRO) CONSORTIUM WORKSHOP

April 24, 2013 ■ Silver Spring, MD

Co-sponsored by

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Patient-Reported Outcomes Measurement Information System (PROMIS®)

- Multiple modes of survey data collection
  - Mail, face-to-face or telephone interviews
  - Web-based surveys
- PROMIS internet panel of about 12k
- > 1 million members who regularly participate in online surveys

Liu, H. et al. (2010). Representativeness of the Patient-Reported Outcomes Measurement Information System internet panel. J Clinical Epidemiology, 63, 1169-1178.

Sample-matching Methodology

- Non-probability based recruitment of panel
- Target subset with selected characteristics
  - n = 11,796 overall
  - Subgroups with lower response rates oversampled
- PROMIS targets ("Quota sampling")
  - 50% female
  - 20% 18-29, 30-44, 45-59, 60-74 and 75+
  - 12.5% black, 12.5% Hispanic
  - 10% < high school graduate

PROMIS Internet Sample versus Census

	PROMIS Sample	2000 Census
% Female	55%	52%
% Hispanic	13%	11%
% Black	10%	11%
% < High school	3%	20%
% High school/GED	19%	29%
% > High school	78%	51%
Mean age	50	45

Analytic Weights (Post-Stratification Adjustment)

- Compensate for nonresponse and non-coverage
- Weight sample to have same distribution on demographic variables
    - gender x age x race/ethnicity, education, marital status, and income
- Iterative proportional fitting or raking

PROMIS Internet Sample (Weighted) versus Census

	PROMIS Sample	2000 Census
% Female	52%	52%
% Hispanic	11%	11%
% Black	11%	11%
% < High school	20%	20%
% High school/GED	29%	29%
% > High school	51%	51%
Mean age	45	45

In general, how would you rate your health? (5 = excellent; 4 = very good; 3 = good; 2 = fair; 1 = poor)

Sample	Mean (1-5 possible score)
PROMIS	3.53
2004 MEPS	3.56
2001-2002 NHANES	3.50
2005 BRFSS	3.52

Other Internet Panel Examples

- NIH Toolbox (R. Gershon)
  - Delve, Inc databases assembled using online self-enrollment, enrollment through events hosted by the company, and random telephone calls from market research representatives
- PROMIS Valuation Study (B. Craig)
  - Each of 7 vendors recruited 1000 respondents by sending members an e-mail invitation containing payment information and a member-specific hyperlink

Testing and Marketing Companies

- Psychological Corporation national standardization sample for RAND-36
- 800 18-89 years respondents from U.S. general population
- Stratified sample to represent on selected demographic variables
  - 100 males and 100 females in each of four age groups (18-24, 25-44, 45-64, 65+), stratified by race/ethnicity and level of education

Pros and Cons of Internet Panels

- PROs
  - Relatively inexpensive
  - Faster
  - Able to get to low incidence subgroups
- CONs
  - Highly educated respondents
  - May differ from target on unmeasured characteristics
  - Data integrity (False answers, duplicates)

Discussion

Q & A