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
FDA
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.

Exploratory Quantitative Analysis
The SESIRnQ
<table>
<thead>
<tr>
<th>Item</th>
<th>Obs</th>
<th>Sign</th>
<th>item-test correlation</th>
<th>item-rest correlation</th>
<th>average inter-item covariance</th>
<th>alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>i_1</td>
<td>49</td>
<td>+</td>
<td>0.9099</td>
<td>0.8765</td>
<td>0.3869473</td>
<td>0.8292</td>
</tr>
<tr>
<td>i_3</td>
<td>49</td>
<td>+</td>
<td>0.8043</td>
<td>0.7225</td>
<td>0.3913265</td>
<td>0.8434</td>
</tr>
<tr>
<td>i_4</td>
<td>49</td>
<td>+</td>
<td>0.7235</td>
<td>0.6357</td>
<td>0.4318452</td>
<td>0.8594</td>
</tr>
<tr>
<td>i_5</td>
<td>49</td>
<td>+</td>
<td>0.7711</td>
<td>0.6776</td>
<td>0.4002551</td>
<td>0.8499</td>
</tr>
<tr>
<td>i_9</td>
<td>49</td>
<td>+</td>
<td>0.8806</td>
<td>0.8008</td>
<td>0.3244473</td>
<td>0.8249</td>
</tr>
<tr>
<td>i_2</td>
<td>49</td>
<td>+</td>
<td>0.7691</td>
<td>0.5818</td>
<td>0.346301</td>
<td>0.8925</td>
</tr>
<tr>
<td>Test scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3801871</td>
<td>0.8713</td>
</tr>
</tbody>
</table>

---

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

<table>
<thead>
<tr>
<th>Seq</th>
<th>Item Type</th>
<th>Location</th>
<th>SE</th>
<th>Residual</th>
<th>DF</th>
<th>ChiSq</th>
<th>DF</th>
<th>Prob</th>
<th>F-stat</th>
<th>DF1</th>
<th>DF2</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>i_1</td>
<td>Poly</td>
<td>-4.990</td>
<td>-0.872</td>
<td>37.00</td>
<td>5.810</td>
<td>4</td>
<td>0.213825</td>
<td>5.243</td>
<td>4</td>
<td>44</td>
<td>0.001529</td>
</tr>
<tr>
<td>2</td>
<td>i_3</td>
<td>Poly</td>
<td>-3.683</td>
<td>-0.084</td>
<td>37.00</td>
<td>5.595</td>
<td>4</td>
<td>0.231483</td>
<td>1.874</td>
<td>4</td>
<td>44</td>
<td>0.131905</td>
</tr>
<tr>
<td>3</td>
<td>i_4</td>
<td>Poly</td>
<td>0.674</td>
<td>0.660</td>
<td>37.00</td>
<td>4.619</td>
<td>4</td>
<td>0.328696</td>
<td>1.100</td>
<td>4</td>
<td>44</td>
<td>0.368765</td>
</tr>
<tr>
<td>4</td>
<td>i_5</td>
<td>Poly</td>
<td>1.009</td>
<td>0.829</td>
<td>37.00</td>
<td>6.317</td>
<td>4</td>
<td>0.176718</td>
<td>1.375</td>
<td>4</td>
<td>44</td>
<td>0.257918</td>
</tr>
<tr>
<td>5</td>
<td>i_9</td>
<td>Poly</td>
<td>4.699</td>
<td>-0.482</td>
<td>37.00</td>
<td>5.975</td>
<td>4</td>
<td>0.200999</td>
<td>3.890</td>
<td>4</td>
<td>44</td>
<td>0.008633</td>
</tr>
<tr>
<td>6</td>
<td>i_2</td>
<td>Poly</td>
<td>2.291</td>
<td>2.653</td>
<td>37.00</td>
<td>14.784</td>
<td>4</td>
<td>0.005171</td>
<td>3.145</td>
<td>4</td>
<td>44</td>
<td>0.023300</td>
</tr>
</tbody>
</table>
Person-Item Map

Scores in Logits—Rasch Measures

Direction Of Measure

Persons

Items

Item 9
Item 11
Item 4  Item 10
Item 8  Item 7
Item 3  Item 6  Item 2
Item 5  Item 1

Scores in Logits—Rasch Measures

Persons

Items

Item 9
Item 11
Item 4  Item 10
Item 8  Item 7
Item 3  Item 6  Item 2
Item 5  Item 1
Conceptualizing

Mixed Methods Approaches
Standard Approach

QUAL

Develop instrument

Quitative concept elicitation study (explore concept)

Cognitive debriefing (ensure item meaning and understanding)

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

Qual (qual)

QUANT

Conduct full psychometric validation study
An Iterative Approach

QUAL

quant+ qual

Preliminary Analysis

Instrument revision

quant

Limited number of interviews to address issues revealed by preliminary analysis

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

qual

Return to text data or interview as needed

quant

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

qual

Confirm other key measurement properties

QUANT

Cognitive debriefing integrated with administration of draft questionnaire to small sample

Qualitative concept elicitation study (explore concept)
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)

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

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

<table>
<thead>
<tr>
<th>Previous Process</th>
<th>Current Process</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scoping Stage</strong></td>
<td>Scoping Stage</td>
</tr>
<tr>
<td><strong>Qualitative Research Stage</strong></td>
<td>Content Validity Stage</td>
</tr>
<tr>
<td>• Qualitative Interviews, no quantitative testing</td>
<td>• Mixed Methods – Qualitative Interviews &amp; Quantitative Assessments</td>
</tr>
<tr>
<td><strong>Quantitative Research Stage</strong></td>
<td>Psychometric Analysis Stage</td>
</tr>
<tr>
<td>• Confirmatory Psychometric Analyses</td>
<td>• Confirmatory Psychometric Analyses</td>
</tr>
</tbody>
</table>
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

Persons with least visual ability

- Get ar o
- Keep cl
- Get dre
- T
- Eat/dri
- Fix sna
- S Get ar o
- Groomin
- Match c
- Tell ti
- Avoid b
- Clean t
- Identif
- Take a
- S Find pu
- Identif
- Identif
- Recogni
- Sign yo
- Use app
- M Cross s
- Read he
- Watch T
- Adjust
- Find so
- Handle
- Use pub
- Do yard
- Get ar o
- Make ou
- Go to m
- Play ta
- See pho
- Go out
- Go to s
- Keep pl
- S Read ma
- Read si
- Favorit
- Read st
- Recogni
- Read me
- Read pr
- S Play sp
- # Small p

Persons with most visual ability.

Least difficult items

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
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?

\[ P(X_i = k | \theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{ik+1})}} \]

- None of the time
- Little of the time
- Some of the time
- Most of the time
- All of the time

Overall Mental Health

Probability of Response

θ
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)


- 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
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>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?</td>
<td></td>
</tr>
<tr>
<td>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?</td>
<td></td>
</tr>
<tr>
<td>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?</td>
<td></td>
</tr>
<tr>
<td>Does the item order reflect the clinically hypothesized item order, if relevant? Does item order reflect the importance/bother ratings from the patients?</td>
<td></td>
</tr>
</tbody>
</table>
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 $5 \times 10 = 50$. 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](http://www.rasch.org/rmt/rmt74m.htm) is a useful guide (now updated to match David's recommendation).
### Rules *Intended* for Dichotomous Items

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

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

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
<table>
<thead>
<tr>
<th></th>
<th>PROMIS Sample</th>
<th>2000 Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Female</td>
<td>55%</td>
<td>52%</td>
</tr>
<tr>
<td>% Hispanic</td>
<td>13%</td>
<td>11%</td>
</tr>
<tr>
<td>% Black</td>
<td>10%</td>
<td>11%</td>
</tr>
<tr>
<td>% &lt; High school</td>
<td>3%</td>
<td>20%</td>
</tr>
<tr>
<td>% High school/GED</td>
<td>19%</td>
<td>29%</td>
</tr>
<tr>
<td>% &gt; High school</td>
<td>78%</td>
<td>51%</td>
</tr>
<tr>
<td>Mean age</td>
<td>50</td>
<td>45</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>PROMIS Sample</th>
<th>2000 Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Female</td>
<td>52%</td>
<td>52%</td>
</tr>
<tr>
<td>% Hispanic</td>
<td>11%</td>
<td>11%</td>
</tr>
<tr>
<td>% Black</td>
<td>11%</td>
<td>11%</td>
</tr>
<tr>
<td>% &lt; High school</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>% High school/GED</td>
<td>29%</td>
<td>29%</td>
</tr>
<tr>
<td>% &gt; High school</td>
<td>51%</td>
<td>51%</td>
</tr>
<tr>
<td>Mean age</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>
In general, how would you rate your health? (5 = excellent; 4 = very good; 3 = good; 2 = fair; 1 = poor)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean (1-5 possible score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROMIS</td>
<td>3.53</td>
</tr>
<tr>
<td>2004 MEPS</td>
<td>3.56</td>
</tr>
<tr>
<td>2001-2002 NHANES</td>
<td>3.50</td>
</tr>
<tr>
<td>2005 BRFSS</td>
<td>3.52</td>
</tr>
</tbody>
</table>
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