Sources of Comparability Between Probability Sample Estimates and Nonprobability Web Sample Estimates

William Riley, National Cancer Institute
Ron Hays, University of California Los Angeles (UCLA)
Robert Kaplan, National Institutes of Health
David Cella, Northwestern University of Feinberg School of Medicine
William Thompson, Centers for Disease Control
Catherine Okoro, Centers for Disease Control
Satvinder Dhingra, Northrop Grumman and Centers for Disease Control
PROMIS Adult Self-Reported Health

PROMIS Profile Domains

Physical Health
- Physical Function
- Pain Intensity
- Pain Interference
- Fatigue
- Sleep Disturbance

Mental Health
- Depression
- Anxiety

Social Health
- Satisfaction with Participation in Social Roles

PROMIS Additional Domains

Pain Behavior
- Sleep-related Impairment
- Sexual Function

Anger
- Applied Cognition
- Alcohol Use, Consequences, & Expectancies
- Psychosocial Illness Impact

Satisfaction with Social Roles & Activities
- Ability to Participate in Social Roles & Activities
- Social Support
- Social Isolation
- Companionship

Global Health
Modern Psychometric Methodology

Item Response Theory and Computer Adaptive Testing

- Psychometric properties at the item level allow item banks that can be flexibly administered and further refined over time
- Improved efficiency (less respondent burden)
  - Automated administration, scoring and reporting
  - Tailored targeting of items
  - Item administration based on prior responses
- Improved precision
- Ability to crosswalk with existing instruments for comparability on the same metric
Fatigue Item Bank

- Chemotherapy trial: Items 1-10
- Rheumatoid Arthritis trial: CAT
- Parkinson’s Disease trial: Items 6-12
- Diabetes trial: Items 2, 4, 9, 13
- CHF trial: Items 1-5

Same metric, same meaning
Advancing Patient-Centered Outcomes

PROMIS: A Common Source of PROs

Clinical Practice
- Clinic
- Hospital

Clinical Research

Survey Research (CDC)

NIH

Industry
PROMIS Global Items (10)

1. In general, would you say your health is . . .
2. In general, would you say your quality of life is . . .
3. In general, how would you rate your physical health?
4. In general, how would you rate your mental health?
5. In general, how would you rate your satisfaction with social activities and relationships?
6. To what extent are you able to carry out your everyday physical activities?
7. How would you rate your pain on average?
8. How would you rate your fatigue on average?
9. In general, please rate how well you carry out usual social activities and roles.
10. How often have you been bothered by emotional problems?
Sample (n = 21,133)

- You Gov Polimetrix nonprobability Internet panel, augmented by clinical samples
- Age: 18-100 (mean = 53)
- 52% Female
- 9% Latino/Hispanic, 9% black, 2% other
- 3% < high school, 16% high school only
- 59% Married
- 39% Working full-time
### Physical Health

#### Item Parameters

<table>
<thead>
<tr>
<th>Item</th>
<th>A</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>b4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global03</td>
<td>2.31</td>
<td>-2.11</td>
<td>-0.89</td>
<td>0.29</td>
<td>1.54</td>
</tr>
<tr>
<td>Global06</td>
<td>2.99</td>
<td>-2.80</td>
<td>-1.78</td>
<td>-1.04</td>
<td>-0.40</td>
</tr>
<tr>
<td>Global07</td>
<td>1.74</td>
<td>-3.87</td>
<td>-1.81</td>
<td>-0.67</td>
<td>1.00</td>
</tr>
<tr>
<td>Global08</td>
<td>1.90</td>
<td>-3.24</td>
<td>-1.88</td>
<td>-0.36</td>
<td>1.17</td>
</tr>
</tbody>
</table>

3. In general, how would you rate your physical health?
6. To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries or moving a chair?
7. How would you rate your pain on average?
8. How would you rate your fatigue on average?

3: Poor, Fair, Good, Very Good, **Excellent**
6: Not at all, A Little, Moderately, Mostly, **Completely**
7: **No pain** (0) – Worse pain imaginable (10)
8: **None**, Mild, Moderate, Severe, Very Severe
Mental Health
Item Parameters

<table>
<thead>
<tr>
<th>Item</th>
<th>A</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>b4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global02</td>
<td>2.41</td>
<td>-2.45</td>
<td>-1.32</td>
<td>-0.19</td>
<td>1.07</td>
</tr>
<tr>
<td>Global04</td>
<td>3.67</td>
<td>-2.31</td>
<td>-1.26</td>
<td>-0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>Global05</td>
<td>2.98</td>
<td>-1.78</td>
<td>-0.90</td>
<td>-0.01</td>
<td>1.07</td>
</tr>
<tr>
<td>Global10</td>
<td>1.89</td>
<td>-2.82</td>
<td>-1.51</td>
<td>-0.25</td>
<td>0.99</td>
</tr>
</tbody>
</table>

2. In general, would you say your quality of life is ...?
4. In general, how would you rate your mental health, including your mood and your ability to think?
5. In general, how would you rate your satisfaction with social activities and relationships?
10. How often have you been bothered by emotional problems such as feeling anxious, depressed or irritable?

2, 4, 5: Poor, Fair, Good, Very Good, Excellent
10: Never, Rarely, Sometimes, Often, Always
Physical and Mental Health  
\( (r = 0.63) \)

- Physical (\( \alpha = 0.81 \))
  - \( r = 0.82 \) with EQ-5D
  - \( r = -0.75 \) (pain impact), -0.73 (fatigue), 0.71 (physical functioning), & -0.67 (pain behavior)

- Mental (\( \alpha = 0.86 \))
  - \( r = 0.61 \) with EQ-5D
  - \( r = -0.71 \) (depression), - 0.65 (anxiety), & 0.60 (satisfaction with discretionary social activities)
### General Health Item Comparisons

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
<th>vs. MEPS</th>
<th>vs. NHANES</th>
<th>vs. BRFSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Health (5-Excellent, 4-Very Good, 3-Good, 2-Fair, 1-Poor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004 MEPS</td>
<td>20777</td>
<td>3.56</td>
<td>0.012</td>
<td>3.54</td>
<td>3.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001-2002 NHANES</td>
<td>6873</td>
<td>3.50</td>
<td>0.017</td>
<td>3.47</td>
<td>3.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 BRFSS</td>
<td>352036</td>
<td>3.52</td>
<td>0.004</td>
<td>3.52</td>
<td>3.52</td>
<td>3.53</td>
<td></td>
</tr>
<tr>
<td>PROMIS: General population (Unweighted)</td>
<td>11794</td>
<td>3.50</td>
<td>0.009</td>
<td>3.48</td>
<td>3.52</td>
<td>-0.085, -0.035</td>
<td>-0.032, 0.032</td>
</tr>
<tr>
<td>PROMIS: General population (weighted)</td>
<td>218022053</td>
<td>3.42</td>
<td>0.016</td>
<td>3.39</td>
<td>3.45</td>
<td>-0.175, -0.108</td>
<td>-0.120, -0.043</td>
</tr>
<tr>
<td>PROMIS: General population sub-sample</td>
<td>2196</td>
<td>3.53</td>
<td>0.020</td>
<td>3.49</td>
<td>3.57</td>
<td>-0.068, 0.008</td>
<td>-0.013, 0.073</td>
</tr>
</tbody>
</table>

Note: equivalencies are printed in boldface type.

---

Liu et al. Representativeness of PROMIS Internet Panel, J Clin Epi 2010;63: 1169-78
PROMIS GLOBAL AMONG DIFFERENT SAMPLING METHODOLOGIES
Sampling Methodologies

- **National Health Interview Survey (2010 - a probability sample of households)**
  - The National Health Interview Survey is a cross-sectional household interview survey.
  - Data are collected through a personal household interview conducted by interviewers employed and trained by the U.S. Bureau of the Census according to procedures specified by the NCHS.
  - Sampling and interviewing are continuous throughout each year.
  - The sampling plan is redesigned after every decennial U.S. census.

- **Health Styles (2010 - a non-probability mail panel)**
  - Porter Novelli contracted with Synovate
  - The sampling and data collection are conducted by Synovate, Inc.
  - The Synovate, Inc. consumer mail panel contains approximately 200,000 potential respondents.

- **Health Styles (2012 - a probability Internet panel)**
  - Porter Novelli contracted with Knowledge Networks
  - Knowledge Networks maintains an online panel of 50,000 representative of the entire U.S. population.
  - Panel members randomly recruited by probability-based sampling (using both random-digit dial and address-based sampling methods).
  - If needed, households provided with a laptop computer and access to the Internet.

- **PHSB (2013 - a non-probability sample constructed to be representative of the U.S. population)**
  - YouGov utilizes different modes of recruitment continuously over time ensuring hard-to-reach populations adequately represented in survey samples
  - Sampling targets set based on gender, age, race, and education of this group using information from the 2010 American Community Survey
  - Weighted the matched set of survey respondents to known characteristics in the U.S. using propensity score weighting
## Population Characteristics

<table>
<thead>
<tr>
<th></th>
<th>PHSB 2013 (n=3,500)</th>
<th>HealthStyles 2012 (n=3,503)</th>
<th>HealthStyles 2010 (n=4,184)</th>
<th>NHIS 2010 (n=27,157)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unweighted</td>
<td>Weighted</td>
<td>Unweighted</td>
<td>Weighted</td>
</tr>
<tr>
<td><strong>Race or Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>2,635</td>
<td>75.3</td>
<td>67.4</td>
<td>2,641</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>326</td>
<td>9.3</td>
<td>11.4</td>
<td>334</td>
</tr>
<tr>
<td>Hispanic</td>
<td>311</td>
<td>8.9</td>
<td>14.1</td>
<td>116</td>
</tr>
<tr>
<td>Other</td>
<td>228</td>
<td>6.5</td>
<td>7.2</td>
<td>412</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1,968</td>
<td>56.2</td>
<td>52.0</td>
<td>1,770</td>
</tr>
<tr>
<td>Male</td>
<td>1,532</td>
<td>43.8</td>
<td>48.0</td>
<td>1,733</td>
</tr>
<tr>
<td><strong>Age of Respondent</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>283</td>
<td>8.1</td>
<td>13.1</td>
<td>317</td>
</tr>
<tr>
<td>25-34</td>
<td>565</td>
<td>16.1</td>
<td>15.7</td>
<td>418</td>
</tr>
<tr>
<td>35-44</td>
<td>786</td>
<td>22.5</td>
<td>18.9</td>
<td>518</td>
</tr>
<tr>
<td>45-54</td>
<td>533</td>
<td>15.2</td>
<td>16.8</td>
<td>718</td>
</tr>
<tr>
<td>55-64</td>
<td>751</td>
<td>21.5</td>
<td>18.2</td>
<td>706</td>
</tr>
<tr>
<td>65+</td>
<td>582</td>
<td>16.6</td>
<td>17.3</td>
<td>826</td>
</tr>
</tbody>
</table>
T-Scores
PROMIS Global Physical Health

Physical Health G03

Physical Function G06

Pain G07

Fatigue G08

NHIS 2010 | HS 2010 MAIL | HS 2012 INTERNET | PHSB 2013 INTERNET
PROMIS Global Mental Health

Quality of Life G02

Mental Health G04

Social Discretionary G05

Emotional Problems G10

- NHIS 2010
- HS 2010 MAIL
- HS 2012 INTERNET
- PHSB 2013 INTERNET
Discussion

• Comparable results from probability and nonprobability samples

• NHIS difference potentially due to mode of administration, not probability vs. nonprobability sampling
  – Interviewer interaction produces a 0.2 – 0.5 SD increase in HRQL responses compared to mail/Internet responses
    Hays et al., Values in Health, 2009, 12: 1035-1039

• Limitations of comparing probability and nonprobability sampling
  – Other survey confounds
  – If different, was it the sampling approach?
  – If the same, is it generalizable?
Pros and Cons of Internet Panels

• Pros
  – Rapid and efficient data collection
  – Ability to recruit targeted samples
  – Ease of replication

• Cons
  – Selection biases
  – Difficulty ensuring integrity, security of the data
  – Higher rates of loss to follow-up
  – Considerable variability among Internet opt-in panels
But Are Nonprobability Samples Really That Different?

- Little practical distinction between opting in vs. opting out (Gotway Crawford, 2013)
  - Growing limits on RDD (no landline, call screening) introduce a form of selection bias
  - Weighting by the inverse of the selection probability (design-based) only works if the nonresponse rate is small (Rivers, 2013)
    - 10% nonresponse rate upper limit (Cochran, 1977)
    - 30% nonresponse rate with flawed results (Lohr, 2010)
    - What are we to expect of >95% nonresponse rates?
  - Modeling based inferences required for self-selection vs. nonresponse are mathematically indistinguishable (Rivers, 2013)

From JSSAM (2013) comments on the Summary Report of the AAPOR Task Force on Nonprobability Sampling
Weighting Adjustments for Opt-In Internet Panels

– Sample weighting
  • Cell weighting – adjust the sample distributions so they conform to the population distributions on a cell by cell basis
    – Assumes missing at random
    – Less stable aggregated estimates with large numbers of cells
  • Raking – iterative matching of cell counts to marginal distributions of the grouping variable
  • Propensity score weighting – correct selection bias in internet surveys but limited if only a few variables can be used to generate propensity scores
One Approach for Addressing the Problem of Estimate Precision from Non Probability Samples: Bayesian Credibility Intervals

If you have a uniform(0,1) prior (the usual diffuse prior), then the posterior is:

$$\beta(np+1,n(1-p)+1)$$

where $p$ is the sample fraction

Roshwalb et al. (2012) Towards the use of Bayesian credibility intervals in online survey results. NY: Ipsos Public Affairs
Conclusion

• Nonprobability Internet Panel Surveys are here to stay
• Focus on improving these methodologies and addressing their limitations
• Probability samples will remain the standard by which nonprobability samples are compared, and data from them remain critical for modeling and weighting nonprobability samples