Survey research: an overview of statistical and psychometric methods

Vanderbilt-Ingram Cancer Center Biostatistics Workshop
June 19, 2009

Irene D. Feurer, Ph.D.
Research Professor of Surgery and Biostatistics
Vanderbilt University Medical Center
irene.feurer@vanderbilt.edu
Workshop objectives

- Discuss types of surveys and their purposes
  - Surveys that are “scored”
  - Surveys requiring response summaries only

- Discuss review criteria for survey instruments
  - Scientific advisory committee of the Medical Outcomes Trust (Lohr et al. Clinical Therapeutics 1996;18:979-92.)
    - 8 criteria

- Focus on classical test (true-score) theory (not IRT)

- Examples
Examples

- **Summary response survey**
  - Vanderbilt Transplant Center Employment survey (pre and post-transplant forms)

- **Scored instruments**
  - SF-36®
  - Vanderbilt Transplant Center Patient Satisfaction Inventory
  - EQ-5D
  - “Memphis” symptom survey
  - QDQ, Vanderbilt Bill Wilkerson Center for Otolaryngology Quantitative Dizziness Questionnaire (GP Jacobson, et al., unpublished data)
Classical true-score theory: some assumptions

Error of measurement is unsystematic or random:

\[ X = T + E \]

- additivity of true score plus error of measurement

\[ \xi (X) = T \]
- expected value (population mean) of \( X \) is \( T \)

\[ \rho_{ET} = 0 \]
- error and true scores from a population are uncorrelated

\[ \rho_{E_1E_2} = 0 \]
- error scores of two different tests are uncorrelated

\[ \rho_{E_1T_2} = 0 \]
- error scores of one test are uncorrelated with error scores of another test
MOT scientific review criteria and attributes

- Conceptual and measurement model
- Reliability
- Validity
- Responsiveness
- Interpretability
- Respondent and administrative burden
- Alternative forms
- Cultural and language adaptations
MOT scientific review criteria and attributes

Conceptual and measurement model

- Underlying rationale and description of concepts measured
- Is reflected in survey’s scale of measurement
- Is reflected in subscale structure

1. What is the basis for combining items?
2. What descriptive statistics are appropriate for scales?
3. What evidence supports intended level of measurement (nominal, ordinal, interval, ratio scales)
4. Are the procedures for deriving scales scores from raw scores justified?
MOT scientific review criteria and attributes

Reliability

- Degree to which instrument is free of random (measurement) error
- Reproducibility of scores under “equivalent” circumstances

1. Test-retest (paired test statistics)
2. Inter-rater (kappa, ICC)
3. Equivalent or alternate forms ($r$)
4. Internal consistency (factor structure, coeff $\alpha$)
Single latent variable (factor/component) with five manifest variables
Scree plot of eigenvalues
13 core Vanderbilt Transplant Center
Patient Satisfaction Inventory (VTCPSI) items

The first component accounts for 56% of the covariance.
The second component accounts for only an additional 9%.

AHRQ 5 R03 HS013036.
### Summary parameters - VTCPSI

<table>
<thead>
<tr>
<th>Time point</th>
<th>N</th>
<th>13 items complete</th>
<th>n:p</th>
<th>KMO</th>
<th>Eigenval 1, 2 (%) var</th>
<th>Comp 1 (%)</th>
<th>Min load</th>
<th>Saturation</th>
<th>Coeff α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval</td>
<td>546</td>
<td>66%</td>
<td>28</td>
<td>.91</td>
<td>7.4, 1.2</td>
<td>57%</td>
<td>.53</td>
<td>.75</td>
<td>.92</td>
</tr>
<tr>
<td>Wait</td>
<td>514</td>
<td>82%</td>
<td>32</td>
<td>.92</td>
<td>7.4, 1.3</td>
<td>57%</td>
<td>.61</td>
<td>.75</td>
<td>.93</td>
</tr>
<tr>
<td>1 M</td>
<td>229</td>
<td>81%</td>
<td>14</td>
<td>.90</td>
<td>7.3, 1.0</td>
<td>56%</td>
<td>.56</td>
<td>.74</td>
<td>.92</td>
</tr>
<tr>
<td>3 M</td>
<td>177</td>
<td>89%</td>
<td>12</td>
<td>.91</td>
<td>7.0, 1.1</td>
<td>54%</td>
<td>.54</td>
<td>.73</td>
<td>.92</td>
</tr>
<tr>
<td>6 M</td>
<td>181</td>
<td>89%</td>
<td>12</td>
<td>.91</td>
<td>6.9, 1.3</td>
<td>53%</td>
<td>.54</td>
<td>.72</td>
<td>.91</td>
</tr>
<tr>
<td>Ann</td>
<td>1061</td>
<td>86%</td>
<td>70</td>
<td>.95</td>
<td>7.2, 1.2</td>
<td>56%</td>
<td>.51</td>
<td>.74</td>
<td>.92</td>
</tr>
<tr>
<td>Total</td>
<td>2708</td>
<td>81%</td>
<td>169</td>
<td>.94</td>
<td>7.2, 1.2</td>
<td>56%</td>
<td>.55</td>
<td>.74</td>
<td>.92</td>
</tr>
</tbody>
</table>
Vanderbilt Transplant Center Patient Satisfaction Inventory

Score interpretation

- n = 2708
- Mean = 34 ± 5
- Median = 36
- 25th %ile threshold = 31

Feurer et al. *Progress in Transplantation* 2007;17:121-128
Patient satisfaction with transplant healthcare delivery

25th %ile threshold
Two latent variables and five manifest variables

Factor A

V1
V2
V3
V4
V5

Factor B
SF-36® (v1) PCS and MCS

- Eight subscales
  - Physical function (PF)
  - Role physical (RP)
  - Bodily pain (BP)
  - General health (GH)
  - Vitality (VT)
  - Social functioning (SF)
  - Role emotional (RE)
  - Mental Health (MH)

- Two summary components
  - Physical (PCS)
  - Mental (MCS)

<table>
<thead>
<tr>
<th>Scale (Z scores)</th>
<th>PCS</th>
<th>MCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF</td>
<td>.42402</td>
<td>-.22999</td>
</tr>
<tr>
<td>RP</td>
<td>.35119</td>
<td>-.12329</td>
</tr>
<tr>
<td>BP</td>
<td>.31754</td>
<td>-.09731</td>
</tr>
<tr>
<td>GH</td>
<td>.24954</td>
<td>-.01571</td>
</tr>
<tr>
<td>VT</td>
<td>.02887</td>
<td>.23534</td>
</tr>
<tr>
<td>SF</td>
<td>-.00753</td>
<td>.26876</td>
</tr>
<tr>
<td>RE</td>
<td>-.19206</td>
<td>.43407</td>
</tr>
<tr>
<td>MH</td>
<td>-.22069</td>
<td>.48581</td>
</tr>
</tbody>
</table>
Simple structure
two latent variables and five manifest variables
Quantitative Dizziness Questionnaire
Item analysis (unpublished data)

- Starting p = 43
- Starting n:p = 4.5 (n ~ 200)
- Final p = 33
- Content expertise – drop 2 items
- Component ambiguity – drop 4 items
- Internal consistency reliability (coefficient $\alpha$) – drop 4 items
- 6 sales
  - Headache (5) $\alpha = .649$
  - Positional vertigo (6) $\alpha = .801$
  - Hydrops (5) $\alpha = .711$
  - Sup Canal Dehis (6) $\alpha = .863$
  - Multisensory (6) $\alpha = .791$
  - Psychophysiological (5) $\alpha = .836$
Unpublished data: Quantitative Dizziness Questionnaire (preliminary) headache scale

### Reliability Statistics

<table>
<thead>
<tr>
<th>Cronbach's Alpha</th>
<th>Cronbach's Alpha Based on Standardized Items</th>
<th>N of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>.618</td>
<td>.615</td>
<td>6</td>
</tr>
</tbody>
</table>

### Item-Total Statistics

<table>
<thead>
<tr>
<th></th>
<th>Scale Mean if Item Deleted</th>
<th>Scale Variance if Item Deleted</th>
<th>Corrected Item-Total Correlation</th>
<th>Squared Multiple Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>G3</td>
<td>8.17</td>
<td>19.747</td>
<td>.124</td>
<td>.099</td>
<td>.649</td>
</tr>
<tr>
<td>G5</td>
<td>9.11</td>
<td>17.361</td>
<td>.292</td>
<td>.088</td>
<td>.597</td>
</tr>
<tr>
<td>H1</td>
<td>9.21</td>
<td>15.460</td>
<td>.449</td>
<td>.405</td>
<td>.532</td>
</tr>
<tr>
<td>H2</td>
<td>9.60</td>
<td>16.221</td>
<td>.454</td>
<td>.439</td>
<td>.535</td>
</tr>
<tr>
<td>H6</td>
<td>9.71</td>
<td>15.918</td>
<td>.328</td>
<td>.151</td>
<td>.587</td>
</tr>
<tr>
<td>H7</td>
<td>9.47</td>
<td>15.798</td>
<td>.469</td>
<td>.240</td>
<td>.527</td>
</tr>
</tbody>
</table>
MOT scientific review criteria and attributes

Validity

- Degree to which a survey is actually measuring the intended concept
- A survey must be shown to be reliable before it can be considered valid

1. Content validity (related to “face validity”)
2. Criterion-related validity (reference to gold standard)
3. Construct validity (scores “behave as expected”)

HRQOL survey battery and assessment schedule
(Vanderbilt Transplant Center, effective January, 2002)

<table>
<thead>
<tr>
<th>Survey Instrument</th>
<th>(# items)</th>
<th>Eval</th>
<th>Listed q 6 M</th>
<th>6 Hr Pre</th>
<th>1 M Post</th>
<th>3 M Post</th>
<th>6 M Post</th>
<th>Ann Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-36®</td>
<td>(36)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Beck Anxiety Inventory</td>
<td>(21)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CES-D Depression Scale</td>
<td>(20)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Satisfaction Inventory©</td>
<td>(16)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Overall Health (VAS)</td>
<td>(1)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Employment (20 pre, 17 post)</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ-5D</td>
<td>(5)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Symptom survey (pending)</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Functional Performance</td>
<td>(1)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

PAIS discontinued January, 2002
EQ-5D added July, 2006

Listed Transplant

VANDERBILT

idf 6/19/2009 19
Criterion-related validity of EQ-5D preference weights in 285 liver transplant candidates and recipients theoretically-targeted associations

<table>
<thead>
<tr>
<th>Dimension</th>
<th>PCS</th>
<th>MCS</th>
<th>PF</th>
<th>RP</th>
<th>BP</th>
<th>GH</th>
<th>VT</th>
<th>SF</th>
<th>RE</th>
<th>MH</th>
<th>BAI</th>
<th>CES-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D HUI</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Care</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual Activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety/Dep</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAS</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dimensions and HUI of EQ-5D should have a strong correlation with scales/subscales from the “gold-standard” that measures similar constructs/domains (Spearman \( r \geq 0.5 \))

Criterion-Related Validity
(with correction for “attenuation”)

- Classic test theory model
- Establishes the criterion-related validity of a target test ($T_x$) after adjusting for the known reliability of the target test and/or the gold standard ($T_y$)

$$r_{T_{EQ} T_{SF-36}} = r_{T_x T_y} = \frac{r_{xy}}{(r_{xx'} * r_{yy'})^{1/2}}$$

$r_{T_x T_y} =$ adjusted validity coefficient
$r_{xy} =$ unadjusted correlation between $T_x$ and $T_y$
$r_{xx'}$ and $r_{yy'} =$ reliability coefficients
(e.g. test-retest reliability from literature)
EQ-5D US population-based preference weights by period in liver transplant patients

* p < 0.05 vs. pre-transplant

EQ-5D US preference weights provide a valid metric for cost-utility analysis in the liver transplant setting.

MOT scientific review criteria and attributes

Responsiveness

- Instrument’s ability to detect change over time
  1. longitudinal data
  2. effect sizes
  3. clinically relevant change
  4. “signal to noise ratio”
Effect of time on physical HRQOL by transplant organ group

Fixed effects: time p=0.001, time x organ p <0.001
Effect of time on mental HRQOL by transplant organ group

Fixed effects: time p=0.038, time x organ p <0.001
Anxiety symptoms by measurement period

![Box plot showing Anxiety symptoms by measurement period](image)

- **Eval**
- **1-6M post**
- **Wait list**
- **1-3Y post**
- **4-6Y post**
- **> 6Y post**

- Moderate to severe symptoms
Depression symptoms by measurement period

Moderate to severe symptoms
MOT scientific review criteria and attributes

Respondent and administrative burden

- Strike a balance between construct coverage and cost in terms of time
- Related to survey completion rates
Original patient report survey battery redundancy of PAIS and SF-36 scales

<table>
<thead>
<tr>
<th></th>
<th>Voc</th>
<th>Dom</th>
<th>Sex</th>
<th>Fam</th>
<th>Soc</th>
<th>Psy</th>
<th>HC</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCS</td>
<td>0.20</td>
<td>0.24</td>
<td>0.08</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
<td>0.65*</td>
</tr>
<tr>
<td>MCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.72*</td>
</tr>
</tbody>
</table>

Entries are standardized regression coefficients (β)

* P < 0.001

### HRQOL survey battery and assessment schedule

(Vanderbilt Transplant Center, effective January, 2002)

<table>
<thead>
<tr>
<th>Survey Instrument</th>
<th>(# items)</th>
<th>Eval</th>
<th>Listed q 6 M</th>
<th>6 Hr Pre</th>
<th>1 M Post</th>
<th>3 M Post</th>
<th>6 M Post</th>
<th>Ann Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-36®</td>
<td>(36)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Beck Anxiety Inventory</td>
<td>(21)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CES-D Depression Scale</td>
<td>(20)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Satisfaction Inventory©</td>
<td>(16)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Overall Health (VAS)</td>
<td>(1)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Employment (20 pre, 17 post)</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>(5)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Symptom survey (pending)</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Functional Performance</td>
<td>(1)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

PAIS discontinued January, 2002
EQ-5D added July, 2006
MOT scientific review criteria and attributes

Alternate forms
Cultural and language adaptations
Interpretability

- Ability to administer instrument in another format
- Validated alternative language forms
- Can qualitative meaning be assigned to quantitative scores?
Additional considerations

- Availability of suitable norms
- “Rules” for handling missing item data
Developing and refining a survey battery
striking a balance

- Validated set of surveys (literature, context, etc.)
- Should you develop a survey “from scratch”
- Response burden
- Breadth of constructs
- Relevance to the clinical setting or research questions
- Redundancy
- Ongoing process