Preference Assessment

Henry Glick
HCMG 901
October 1, 2015

QALYs

- Economic outcome that combines preferences for both length of survival and its quality into a single measure
  - In many jurisdictions, QALYs represent recommended outcome of cost-effectiveness analysis
    [Gold et al., Cost-Effectiveness in Health and Medicine, summary recommendation IV.1, p. 308]

Not Congress

“The Patient-Centered Outcomes Research Institute . . . shall not develop or employ a dollars per quality adjusted life year (or similar measure that discounts the value of a life because of an individual’s disability) as a threshold to establish what type of health care is cost effective or recommended. The Secretary shall not utilize such an adjusted life year (or such a similar measure) as a threshold to determine coverage, reimbursement, or incentive programs under title XVIII”

The Patient Protection and Affordable Care Act
Question QALYs Answers

- How do we decide how much we should pay for:
  - Therapy that saves fully functional lives/life years
    VS
  - Therapy that saves less than fully functional lives/life years (e.g., a drug for heart failure that extends survival, but patients spend extra time with severe disease)
    VS
  - Therapy that doesn’t save lives/life years but improves patients’ functioning (e.g., patients with heart failure spend most of their remaining years in NYHA class I instead of NYHA class III)

QALY Scores

- QALY or preference scores generally range between 0 (death) and 1 (perfect health)
  - e.g., health state with preference score of 0.8 indicates that year in that state worth 0.8 years with perfect health
- Can be states worse than death with preference scores less than 0

Typology of Elicitation Methods

- Assesses or does not assess risk
- Scaling vs choice
- Preference for current health or preference for years of survival
- Direct vs indirect elicitation
- Whose preferences?
Incorporation of Risk Preference

• Measurement with risk theoretically appropriate
• Methods of assessment of QALYs differentiated by whether or not they incorporate preference for risk
  – Utilities when they do
  – Values when they don’t
• Refer to preference assessment, preference scores, or preferences when referring to generic assessment of QALYs

Scaling vs Choice

• Scaling
  – Rating scale, visual analog scale, feeling thermometer
• Choice
  – Standard gamble
  – Time trade-off

Risk and Choice

<table>
<thead>
<tr>
<th>Response Method</th>
<th>Question Framing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Certainty (values)</td>
</tr>
<tr>
<td>Scaling</td>
<td>Rating Scale</td>
</tr>
<tr>
<td></td>
<td>Category Scaling</td>
</tr>
<tr>
<td>Choices</td>
<td>Time trade-off</td>
</tr>
<tr>
<td></td>
<td>Paired comparison</td>
</tr>
</tbody>
</table>

From Drummond et al., Methods for Economic Evaluation of Health Care, p. 143…….
Current Health vs Years of Survival

- Can be measured as:
  - Series of valuations of current health
  - Explicit preference mapping for years of survival and their quality

Direct Elicitation vs Indirect Preference Assessment

- Direct elicitation: Direct rating of preference for health by respondent
  - Current health vs scenarios
- Indirect preference assessment: Uses instruments which have respondent directly rate functional status across a variety of domains and derives preference score from a scoring rule
- “Gold standard”: Although not necessarily feasible, QALYs constructed by use of direct elicitation, which incorporates risk, and accounts for duration of health states

Whose Preferences?

- Panel on Cost Effectiveness in Health and Medicine recommended reference case analysis that uses community preferences to value health
  - Empathy?
  - Trust those who already have disease?
- Also recommended sensitivity analysis that uses preferences of persons with condition
What are We Trying to Measure?

Utility Surface

- Gold standard version of this 3-dimensional surface would be measured with standard gambles
- As drawn, it is a simplification, because it assumes that you live all your years with a single morbidity level
- If people have a probability distribution over years of life and levels of morbidity, can compute expected utility, and translate it into healthy years
- Most of what is done in preference assessment uses alternative methods rather than gold standard

Outline

- Example of an approximate gold standard evaluation
  - Multiattribute utility functions
- Prescored health classification instruments
- Directly elicited preference scores
- Comparison of prescored instruments and direct elicitation
Objective

- Estimate QALYS related to shorter life expectancies and fully functional speech and those related to longer life expectancies with impaired speech
  - Do so using standard gambles and time trade-offs
Data Requirements

- Time preferences for survival (how people value life years)
- Tradeoff between types of speech and years of life
  - Hoarsened speech (normal)
  - Laryngeal or mechanically-assisted speech (impaired)
- Published data on survival rates

Step 1: Valuing Years of Life with Normal Speech

Evaluate utilities for different life expectancies with normal speech by use of standard gambles

\[
\begin{array}{c}
1.0 \\
p \\
1-p
\end{array}
\]

Choose \( Y \) so you are indifferent between \( Y \) with certainty and a 50/50 gamble for best and worst outcomes

Certainty Equivalent Standard Gamble

Choose \( Y \) so you are indifferent between \( Y \) with certainty and a 50/50 gamble for best and worst outcomes

\[
\begin{array}{c}
1.0 \\
0.5 \\
0.5
\end{array}
\]

\( ? Y ? ? \) (e.g., 7 years)

\( ? Y ? ? \) (e.g., 25 years)

\( ? Y ? ? \) (e.g., 0 years)
Probability Equivalent Standard Gamble
• Choose $p$ so you are indifferent between $Y$ with certainty and a $p/(1-p)$ gamble for best and worst outcomes

Gain Equivalent Standard Gamble
• Choose best outcome so you are indifferent between $Y$ with certainty and a 50/50 gamble for best and worst outcomes

Loss Equivalent Standard Gamble
• Choose worst outcome so you are indifferent between $Y$ with certainty and a 50/50 gamble for best and worst outcomes
Expected Utility

- Utility of gamble equals its expected value

\[ \text{Expected Utility} = \sum_{i} PR_i \times UT_i \]

where \(i\) equals possible outcomes, \(PR_i\) equals probability of \(i\)th outcome, and \(UT_i\) equals utility of \(i\)th outcome

Life-Year Utility Estimation

- Assume utility of living 0 years is 0 and utility of living 25 years is 1
- Assume that respondent reported she was indifferent between a 7 years with certainty and a 50/50 gamble of living 0 or 25 years
- What is utility of living 7 years?
- Suppose respondent had been given a second gamble in which she was asked for number of years that made her indifferent between \(y'\) years with certainty and a 50/50 gamble of living 0 or 7 years
  - If her response was 3 years, what is utility of 3 years?

Utility for Years with Normal Speech

\[
\begin{align*}
\text{0-7 years: } & \quad U_i = \frac{0.5}{7} \times y_i \\
\text{7-25 years: } & \quad U_i = 0.5 + \frac{0.5}{18} \times (y_i - 7)
\end{align*}
\]
Trading Off Years With Impaired and Normal Speech

- McNeil et al. could have estimated utilities for morbid years by use of a probability-equivalent standard gamble of form:
  - Choose \( p \) so indifferent between 25 certain years with impaired speech and a \( p/(1-p) \) gamble for best and worst outcomes
- Instead used time trade-off method
  - How many years of normal speech (\( Y \)) are needed to be indifferent between 25 years of life with impaired speech and number (\( Y \)) needed with normal speech?

Impaired Speech Value Estimation

- Assume respondent reported she was indifferent between:

<table>
<thead>
<tr>
<th>Years w/ impaired speech</th>
<th>Years with normal speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 years</td>
<td>12.5 years</td>
</tr>
<tr>
<td>10 years</td>
<td>7 years</td>
</tr>
<tr>
<td>5 years</td>
<td>5 years (i.e. would not trade)</td>
</tr>
</tbody>
</table>

- What is value of living:
  - 10 years with impaired speech?
  - 25 years with impaired speech?
  - 5 years with impaired speech?

Value for Years with Impaired Speech

- \( 10M = 7H: \quad .5 + (.5/18) (7 -7)) = 0.5 \)
- \( 25M = 12.5H: \quad .5 + (.5/18) (12.5 -7)) = 0.653 \)
- \( 5M = 5H: \quad (.5 / 7) \times .5 = 0.357 \)
Utility for Years with Impaired Speech

Utility for Discounted Years

Preferences for Discounted Years

- When choosing between 1M lottery and 20-year 50k lottery (should be) choosing between present values
- Similarly, when choosing between 25 morbid years and Y healthy years also (should be) choosing between present values
- Time trade-off response expressed in undiscounted years (e.g., 12.5 vs 25), but (should be) reporting indifference between 17.94 discounted impaired years and 10.6 discounted healthy years
- Whether we use data in figure to calculate undiscounted or discounted years determines if results need further discounting
Summary of Time Trade-Off Method
1. Define outcomes
2. Select time periods for comparison
3. For each time period, determine number of years in better health state that is equivalent to given number of years in worse health state

Estimating QALYs
• So far, used standard gambles and time trade-offs to estimate health preferences for length and quality of survival
• Have not computed QALYs for radiation and surgery
• Common practice:
  – Calculate expected value of healthy years
    • Multiply probabilities of outcomes times value of outcomes (in healthy years)
    • Sum

Survival Probabilities / Discounted Healthy Years
• Assume a simplified survival curve in which patients live either 0, 5, or 25 years after either surgery or radiation
• Assume following survival probabilities and discounted life years

<table>
<thead>
<tr>
<th>Years</th>
<th>Radiation</th>
<th></th>
<th>Surgery</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prob</td>
<td>DHY</td>
<td>Prob</td>
<td>DHY</td>
</tr>
<tr>
<td>Die immediately</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>Survive 5 years</td>
<td>0.82</td>
<td>4.717</td>
<td>0.4</td>
<td>4.717</td>
</tr>
<tr>
<td>Survive 25 years</td>
<td>0.18</td>
<td>17.936</td>
<td>0.5</td>
<td>10.603</td>
</tr>
</tbody>
</table>

\* \( \frac{1}{1.03^3} + \frac{1}{1.03^4} + \frac{1}{1.03^5} + \frac{1}{1.03^6} + \frac{1}{1.03^7} = 4.717 \)
Decision Tree

- Surgery (7.188 discounted healthy years) preferred to Radiation (7.096 discounted health years)

Decision Tree, Rollback

- Surgery (7.188 discounted healthy years) preferred to Radiation (7.096 discounted health years)
Estimating QALYs (2)

- Common practice violates expected utility theory
  - Should be calculating expected value of utilities, not healthy years
- Correct approach calculates expected utility
  - Multiply probabilities of outcomes times value of outcomes (in utilities)
  - Sum
- Then translates expected utility into expected QALYs

Survival Probabilities / Utilities

- Assume same simplified survival curve in which patients live either 0, 5, or 25 years after either surgery or radiation
- Assume following survival probabilities and Utilities

<table>
<thead>
<tr>
<th>Years</th>
<th>Radiation</th>
<th>Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die immediately</td>
<td>0 0</td>
<td>0.1 0</td>
</tr>
<tr>
<td>Survive 5 years *</td>
<td>0.82 0.357</td>
<td>0.4 0.357</td>
</tr>
<tr>
<td>Survive 25 years</td>
<td>0.18 1</td>
<td>0.5 0.653</td>
</tr>
</tbody>
</table>

Source of Utilities
Decision Tree

- Radiation (0.473 utiles) preferred to Surgery (0.469 utiles)

Translating Expected Utilities to Discounted QALYs

- Already noted that utility for 7 or less undiscounted fully functional years (QALYs) equals:
  \[ U_i = \frac{0.5}{7} \times Y_i \]
- Can rearrange to solve for QALYs for utilities < 0.5:
  \[ Y_i = 2 \times 7 \times U_i \]
  - Radiation: \( 2 \times 7 \times 0.47274 = 6.618 \) undiscount QALYs
  - Surgery: \( 2 \times 7 \times 0.4693 = 6.570 \) undiscount QALYs
- After discounting:
  - 6.618 undiscounted QALYs = 6.097 disc QALYs
  - 6.570 undiscounted QALYs = 6.057 disc QALYs
- Compared to surgery, radiation yields 0.040 additional discounted QALYs
Source of Differences Between 2 Methods

<table>
<thead>
<tr>
<th></th>
<th>Expected years</th>
<th>Expected utility *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 years</td>
<td>0.4<em>4,717=1.887Y 0.4</em>.357</td>
<td>.1429u=1.999Y</td>
</tr>
<tr>
<td>25 years</td>
<td>0.5<em>10.60=5.301Y (0.5</em>.653)</td>
<td>.3265u=4.571Y</td>
</tr>
<tr>
<td>EU vs EY</td>
<td>5 years, +0.112; 25 years, -0.730; diff = -.618</td>
<td></td>
</tr>
<tr>
<td>Radiation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 years</td>
<td>0.82<em>4.72=3.868Y (0.82</em>.357)</td>
<td>.2927u=4.098Y</td>
</tr>
<tr>
<td>25 years</td>
<td>0.18<em>17.936=3.228Y (0.18</em>1)</td>
<td>.18u=2.520Y</td>
</tr>
<tr>
<td>EU vs EY</td>
<td>5 years, +0.230; 25 years, -0.708; diff = -.478</td>
<td></td>
</tr>
</tbody>
</table>

* Translation between EU and years simplified because sum of EU is not greater than 0.5

Generalization of Results
- Technique used to assess preferences for laryngeal cancer describes two curves in a three dimensional space (years of life with normal quality [X] speech, years of life with impaired speech [Y], utility [Z])
- If there are several levels of speech (or quality is continuous), method could be used to assess levels of speech quality and a utility surface could be estimated

Multiattribute Utility Theory
- McNeil et al. developed two utility curves, one for normal speech and one for impaired speech
  - Defined utility function for impaired speech by use of utility function for normal speech
- More generally, when developing a utility function for two (or more) attributes:
  - First develop conditional utility functions for each attribute, where each conditional utility function ranges from 0 to 1
    - e.g., in a blood bank, separate utility functions for wastage and shortage
  - Second develop scaling constants that allow combination of conditional utility functions
Prescored Health Classification Instruments

Preferences for Duration of Morbidity

- Gold standard preference assessment directly measures preferences for level of morbidity and its duration
  - McNeil article provides an example
- Most QALY estimates ignore preferences for duration
  - QALYs usually calculated by multiplying duration of a given level of morbidity times a preference score for that level of morbidity
  - A second-best solution that allows direct assessment of preferences for current health by participants in prospective studies
- Measured by use of prescored instruments OR via direct elicitation

Prescored Health State Classification Instruments

- One of two dominant approaches for QALY measurement uses prescored health state classification instruments (indirect utility assessment)
- Participants' report their functional status across a variety of domains
- Preference scores derived from scoring rules that have usually been developed by use of samples from general public
- Prescored instruments considered to satisfy “community preferences” recommendation of Panel on Cost-Effectiveness
Prescored Instruments

- A number of prescored instruments currently available for measurement of preference scores for current health
  - EuroQol instrument (EQ-5D), 3 and 5 level
  - Health Utilities Index Mark 2 (HUI2)
  - Health Utilities Index Mark 3 (HUI3)
  - SF-6D
  - Quality of Well-Being Scale (QWB)
  - 15D
  - Disability and Distress Index (DDI)
- Most ask participants or proxies to report on health status of patient

EQ-5D, HUI2, HUI3, and SF-6D

- EQ-5D, HUI2, HUI3, and SF-6D are four of most commonly used prescored preference assessment instruments
- All four share features of ease of use
  - e.g., high completion rates and ability to be filled out in 5 min or less
- All have been used to assess preferences for wide variety of diseases

EuroQol instrument

- EuroQol instrument made up of two parts:
  - Health state classification instrument (EQ-5D) and its attendant scoring rule
  - 100-point visual analog scale
    - A form of direct elicitation
EQ-5D Domains

- EQ-5D health state classification instrument has 5 domains
  - Mobility
  - Self-care
  - Usual activities
  - Pain/discomfort
  - Anxiety/depression

EQ-5D Levels of Function

- In original instrument, each domain defined by 3 levels of function from good to poor
  - 3-level generally worded:
    - "I have no problems..."
    - "I have some problems..."
    - "I am unable....."
  - 3 levels for each of 5 domains used to define 243 (3^5) health states

EQ-5D Levels of Function (2)

- In recently released instrument, each domain defined by 5 levels of function from good to poor
  - 5-level generally worded:
    - "I have no problems..."
    - "I have slight problems..."
    - "I have moderate problems..."
    - "I have severe problems..."
    - "I am unable to.../ I have extreme problems..."
  - 5 levels for each of 5 domains used to define 3125 (5^5) health states
EQ-5D Scoring Rule(s)

- Principal 3 level scoring rule developed by Dolan by use of time trade-off responses from a representative sample of 2997 noninstitutionalized adults from England, Scotland, and Wales
- 3 level scoring rules exist for at least 10 additional countries (Szende, Oppe, Devlin eds. EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Springer, 2010)

### Scoring 3 Level EuroQol (Dolan)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility</td>
<td>0.000</td>
<td>0.069</td>
<td>0.314</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Care</td>
<td>0.314</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual Activities</td>
<td>0.000</td>
<td>0.104</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain/Discomfort</td>
<td>0.000</td>
<td>0.123</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety/Depression</td>
<td>0.000</td>
<td>0.071</td>
<td>0.236</td>
</tr>
</tbody>
</table>

### Dolan Scoring Rule

- Scoring formula:
  - If all domains are level 1: 1.000
  - If at least one domain has a score of 2 and no domains have a score of 3 (i.e., worst functioning): 0.929 – sum of scores
  - If one or more domains have a score of 3: 0.65 – sum of scores
Sampling of Other 3 Level Scoring Rules

<table>
<thead>
<tr>
<th></th>
<th>Dolan</th>
<th>Denmark</th>
<th>Netherlands</th>
<th>Spain</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>-.069</td>
<td>-.053</td>
<td>-.036</td>
<td>-.106</td>
</tr>
<tr>
<td>M3</td>
<td>-.314</td>
<td>-.411</td>
<td>-.161</td>
<td>-.430</td>
</tr>
<tr>
<td>S2</td>
<td>-.104</td>
<td>-.053</td>
<td>-.082</td>
<td>-.134</td>
</tr>
<tr>
<td>S3</td>
<td>-.214</td>
<td>-.192</td>
<td>-.152</td>
<td>-.309</td>
</tr>
<tr>
<td>U2</td>
<td>-.036</td>
<td>-.048</td>
<td>-.032</td>
<td>-.071</td>
</tr>
<tr>
<td>U3</td>
<td>-.094</td>
<td>-.144</td>
<td>-.057</td>
<td>-.195</td>
</tr>
<tr>
<td>P2</td>
<td>-.123</td>
<td>-.062</td>
<td>-.086</td>
<td>-.089</td>
</tr>
<tr>
<td>P3</td>
<td>-.386</td>
<td>-.396</td>
<td>-.329</td>
<td>-.261</td>
</tr>
<tr>
<td>A2</td>
<td>-.071</td>
<td>-.068</td>
<td>-.124</td>
<td>-.062</td>
</tr>
<tr>
<td>A3</td>
<td>-.236</td>
<td>-.367</td>
<td>-.325</td>
<td>-.144</td>
</tr>
<tr>
<td>Any 2 or 3</td>
<td>-.071</td>
<td>-.114</td>
<td>-.071</td>
<td>-.024</td>
</tr>
<tr>
<td>Any 3</td>
<td>-.279</td>
<td>--</td>
<td>-.234</td>
<td>-.291</td>
</tr>
</tbody>
</table>

US Scoring Rule

<table>
<thead>
<tr>
<th></th>
<th>Mean Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>.146</td>
</tr>
<tr>
<td>M3</td>
<td>.558</td>
</tr>
<tr>
<td>S2</td>
<td>.175</td>
</tr>
<tr>
<td>S3</td>
<td>.417</td>
</tr>
<tr>
<td>U2</td>
<td>.140</td>
</tr>
<tr>
<td>U3</td>
<td>.374</td>
</tr>
<tr>
<td>P2</td>
<td>.173</td>
</tr>
<tr>
<td>P3</td>
<td>.537</td>
</tr>
<tr>
<td>A2</td>
<td>.156</td>
</tr>
<tr>
<td>A3</td>
<td>.450</td>
</tr>
<tr>
<td># Non-1s</td>
<td>-.140</td>
</tr>
<tr>
<td>(#2s (0 to 4))²</td>
<td>0.011</td>
</tr>
<tr>
<td>(#3s (0 to 4))</td>
<td>-12.2</td>
</tr>
<tr>
<td>(#3s (0 to 4))²</td>
<td>-0.015</td>
</tr>
</tbody>
</table>

Selected US Scores

<table>
<thead>
<tr>
<th>State</th>
<th>Mean Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>11112</td>
<td>.844</td>
</tr>
<tr>
<td>11212</td>
<td>.833</td>
</tr>
<tr>
<td>13122</td>
<td>.489</td>
</tr>
<tr>
<td>11231</td>
<td>.463</td>
</tr>
<tr>
<td>13123</td>
<td>.323</td>
</tr>
<tr>
<td>11233</td>
<td>.29</td>
</tr>
<tr>
<td>13131</td>
<td>.269</td>
</tr>
<tr>
<td>33211</td>
<td>.248</td>
</tr>
<tr>
<td>32313</td>
<td>.167</td>
</tr>
<tr>
<td>33223</td>
<td>.061</td>
</tr>
<tr>
<td>33332</td>
<td>-.035</td>
</tr>
<tr>
<td>33333</td>
<td>-.152</td>
</tr>
</tbody>
</table>
Mapping 5L Domains To 3L

<table>
<thead>
<tr>
<th>SL→3L</th>
<th>Mobility</th>
<th>Self-Care</th>
<th>Activity</th>
<th>Pain</th>
<th>Anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>1→1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1→2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1→3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2→1</td>
<td>0.18</td>
<td>0.17</td>
<td>0.20</td>
<td>0.20</td>
<td>0.21</td>
</tr>
<tr>
<td>2→2</td>
<td>0.82</td>
<td>0.83</td>
<td>0.80</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>2→3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3→1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3→2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3→3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4→1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4→2</td>
<td>0.93</td>
<td>0.76</td>
<td>0.67</td>
<td>0.60</td>
<td>0.51</td>
</tr>
<tr>
<td>4→3</td>
<td>0.07</td>
<td>0.24</td>
<td>0.33</td>
<td>0.40</td>
<td>0.49</td>
</tr>
<tr>
<td>5→1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5→2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5→3</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

EQ-5D Scoring Rule(s), 5 Level

- Direct elicitation scoring rules currently under development
- Currently available scores generated from crosswalk between 5L and 3L
  - Respondents responded to both 3 and 5 level instruments
  - Scores derived for each 3 level response
  - Regression developed to predict 5-level score from 3-level score

Selected State Scores, 5L

<table>
<thead>
<tr>
<th>State</th>
<th>US</th>
<th>UK</th>
<th>Japan</th>
<th>Thailand</th>
<th>Zimbabwe</th>
</tr>
</thead>
<tbody>
<tr>
<td>11211</td>
<td>0.888</td>
<td>0.906</td>
<td>0.843</td>
<td>0.791</td>
<td>0.666</td>
</tr>
<tr>
<td>12211</td>
<td>0.828</td>
<td>0.806</td>
<td>0.773</td>
<td>0.657</td>
<td>0.789</td>
</tr>
<tr>
<td>21212</td>
<td>0.816</td>
<td>0.778</td>
<td>0.702</td>
<td>0.627</td>
<td>0.783</td>
</tr>
<tr>
<td>22222</td>
<td>0.678</td>
<td>0.592</td>
<td>0.592</td>
<td>0.467</td>
<td>0.653</td>
</tr>
<tr>
<td>31111</td>
<td>0.816</td>
<td>0.763</td>
<td>0.728</td>
<td>0.576</td>
<td>0.767</td>
</tr>
<tr>
<td>31221</td>
<td>0.796</td>
<td>0.723</td>
<td>0.674</td>
<td>0.572</td>
<td>0.756</td>
</tr>
<tr>
<td>31212</td>
<td>0.716</td>
<td>0.609</td>
<td>0.614</td>
<td>0.493</td>
<td>0.677</td>
</tr>
<tr>
<td>32342</td>
<td>0.579</td>
<td>0.441</td>
<td>0.525</td>
<td>0.355</td>
<td>0.591</td>
</tr>
<tr>
<td>12344</td>
<td>0.471</td>
<td>0.230</td>
<td>0.547</td>
<td>0.345</td>
<td>0.511</td>
</tr>
<tr>
<td>12345</td>
<td>0.370</td>
<td>0.063</td>
<td>0.522</td>
<td>0.263</td>
<td>0.447</td>
</tr>
<tr>
<td>33333</td>
<td>0.597</td>
<td>0.516</td>
<td>0.532</td>
<td>0.393</td>
<td>0.596</td>
</tr>
<tr>
<td>43434</td>
<td>0.437</td>
<td>0.215</td>
<td>0.454</td>
<td>0.218</td>
<td>0.493</td>
</tr>
<tr>
<td>44444</td>
<td>0.323</td>
<td>0.036</td>
<td>0.397</td>
<td>0.110</td>
<td>0.365</td>
</tr>
<tr>
<td>45454</td>
<td>0.129</td>
<td>-0.244</td>
<td>0.292</td>
<td>-0.084</td>
<td>0.119</td>
</tr>
<tr>
<td>55555</td>
<td>-0.109</td>
<td>-0.594</td>
<td>-0.111</td>
<td>-0.452</td>
<td>-0.145</td>
</tr>
</tbody>
</table>
Pediatric EQ-5D-Y *

- Recently proposed “child-friendly” version of EQ-5D
  - Children’s preference scores currently unavailable
- Same 5 (renamed) domains:
  - Mobility
  - Looking after myself (Self-care)
  - Doing usual activities (Usual activities)
  - Having pain or discomfort (Pain/discomfort)
  - Feeling worried, sad or unhappy (Anxiety/depression)


Pediatric EQ-5D-Y (2)

- Main difference in question wording occurs in most severe level of each domain
  - “a lot of problems walking about” vs “confined to bed”
  - “a lot of problems washing and dressing” vs “unable to wash or dress”
  - “a lot of problems doing my usual activities” vs “unable to perform my usual activities”
  - “a lot of pain and discomfort” vs extreme pain and discomfort
  - “very worried, sad or unhappy” vs “extremely anxious or depressed”
- Currently no value set but at least 3 studies have used adult scoring rules

HUI2

- 7 domain instrument with varying numbers of levels depending on domain
- Domains and number of levels include:
  - Sensory with 4 levels
  - Mobility with 5
  - Emotion with 5
  - Cognition with 4
  - Self-care with 4
  - Pain with 5
  - Fertility with 3
- Multiple levels of seven domains can be used to define 24,000 health states
HUI2 Scoring Rule

- HUI2 has 2 multiplicative scoring rules derived from responses of 293 parents of school children drawn from general population in Canada
  - Because rules were initially developed to evaluate a therapy for childhood cancer
- Focus on utility scoring rule developed by use of standard gambles
- At least one other scoring rule has been proposed

### Sensation
1. Able to see, hear, and speak normally for age (1.00)
2. Requires equipment to see or hear or speak (0.95)
3. Sees, hears, or speaks with limitations even with equipment (0.86)
4. Blind, deaf, or mute (0.61)

### Mobility
1. Able to walk, bend, lift, jump and run normally for age (1.00)
2. Walks, bends, lifts jumps or runs with some limitations (0.97)
3. Requires mechanical equipment (0.94)
4. Requires the help of another person to walk or get around (0.73)
5. Unable to control or use arms and legs (0.58)

### Emotion
1. Generally happy and free from worry (1.00)
2. Occasionally fretful, angry, irritable, anxious, depressed (0.93)
3. Often fretful, angry, irritable, anxious, depressed (0.81)
4. Almost always fretful, angry, irritable, anxious, depressed (0.70)
5. Extremely fretful, angry, irritable, or depressed (0.53)

### Cognition
1. Learns and remembers normally for age (1.00)
2. Learns and remembers more slowly than normal for age (0.95)
3. Learns and remembers very slowly (0.88)
4. Unable to learn and remember (0.65)

### Self-Care
1. Eats, bathes, dresses and uses the toilet normally for age (1.00)
2. Eats, bathes, dresses or uses the toilet independently but... (0.87)
3. Requires mechanical equipment to eat, bathe, dress (0.91)
4. Requires the help of another person to eat, bathe, dress (0.80)
Scoring HUI2

Pain
1. Free of pain and discomfort 1.00
2. Occasional pain 0.97
3. Frequent pain. Discomfort relieved by oral medicines 0.85
4. Frequent pain. Discomfort requires prescription narcotics 0.64
5. Severe pain 0.38

Fertility
1. Able to have children with a fertile spouse 1.00
2. Difficulty in having children with a fertile spouse 0.97
3. Unable to have children with a fertile spouse 0.88

Scoring formula:
1.06 \( (w_1 \times w_2 \times w_3 \times w_4 \times w_5 \times w_6 \times w_7) \) - 0.06

Domain Level Score
Sensory 2 0.95
Mobility 3 0.84
Emotional 2 0.93
Cognitive 3 0.88
Self-care 2 0.97
Pain 4 0.64
Fertility 2 0.97

\( \prod \) 0.393

(1.06 score) - 0.06 0.357

HUI3

- HUI3 has 8 domains each with 5 or 6 levels depending on domain. Domains and number of levels include:
  - Vision, 6 levels
  - Hearing, 6
  - Speech, 5
  - Ambulation, 6
  - Dexterity, 6
  - Emotion, 5
  - Cognition, 6
  - Pain, 5

- Levels of domains can be used to define 972,000 health states
HUI3 Scoring Rule

- As with HUI2, HUI3 has two multiplicative scoring rules.
- For HUI3, derived from responses from random sample of 256 adults drawn from general population in Hamilton, Ontario.
- Also, as with HUI2, focus on utility scoring rule developed by use of standard gambles.

Scoring HUI3

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses</td>
<td>1.00</td>
</tr>
<tr>
<td>2. Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses</td>
<td>0.98</td>
</tr>
<tr>
<td>3. Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses</td>
<td>0.89</td>
</tr>
<tr>
<td>4. Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses</td>
<td>0.84</td>
</tr>
<tr>
<td>5. Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses</td>
<td>0.75</td>
</tr>
<tr>
<td>6. Unable to see at all</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Scoring formula:

\[ U^* = 1.371 \left( w_1 \times w_2 \times w_3 \times w_4 \times w_5 \times w_6 \times w_7 \times w_8 \right) - 0.371 \]

<table>
<thead>
<tr>
<th>Domain</th>
<th>Level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision</td>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>Hearing</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>Speech</td>
<td>2</td>
<td>0.94</td>
</tr>
<tr>
<td>Ambulation</td>
<td>3</td>
<td>0.86</td>
</tr>
<tr>
<td>Dexterity</td>
<td>2</td>
<td>0.95</td>
</tr>
<tr>
<td>Emotion</td>
<td>3</td>
<td>0.85</td>
</tr>
<tr>
<td>Cognition</td>
<td>2</td>
<td>0.92</td>
</tr>
<tr>
<td>Pain</td>
<td>2</td>
<td>0.96</td>
</tr>
</tbody>
</table>

\[ \prod \]

\[ (1.371 \text{ score}) - 0.371 = 0.404 \]

<table>
<thead>
<tr>
<th>Instrument/Domain</th>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D (n= 3977)</td>
<td>Mobility</td>
<td>81.14</td>
<td>18.66</td>
<td>0.21</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Self-care</td>
<td>95.93</td>
<td>3.81</td>
<td>0.26</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Usual activities</td>
<td>84.56</td>
<td>13.59</td>
<td>1.84</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Pain/discomfort</td>
<td>59.15</td>
<td>37.10</td>
<td>3.75</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Anxiety/depression</td>
<td>73.71</td>
<td>23.89</td>
<td>2.08</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>HUI2 (n=3889)</td>
<td>Sensation</td>
<td>38.88</td>
<td>48.24</td>
<td>11.27</td>
<td>1.61</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Mobility</td>
<td>86.12</td>
<td>9.45</td>
<td>3.72</td>
<td>0.71</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Emotion</td>
<td>66.96</td>
<td>29.94</td>
<td>2.17</td>
<td>0.57</td>
<td>0.37</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Cognition</td>
<td>65.29</td>
<td>33.13</td>
<td>1.51</td>
<td>0.06</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Self-care</td>
<td>96.50</td>
<td>2.93</td>
<td>0.21</td>
<td>0.36</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Pain</td>
<td>42.54</td>
<td>45.14</td>
<td>8.25</td>
<td>3.06</td>
<td>1.01</td>
<td>NA</td>
</tr>
<tr>
<td>HUI3 (n=3907)</td>
<td>Vision</td>
<td>42.30</td>
<td>54.13</td>
<td>1.15</td>
<td>1.83</td>
<td>0.50</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Hearing</td>
<td>92.83</td>
<td>0.91</td>
<td>1.91</td>
<td>2.56</td>
<td>0.30</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>Speech</td>
<td>92.69</td>
<td>5.00</td>
<td>1.81</td>
<td>0.32</td>
<td>0.18</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Ambulation</td>
<td>86.09</td>
<td>9.48</td>
<td>2.38</td>
<td>1.33</td>
<td>0.44</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Dexterity</td>
<td>91.69</td>
<td>6.38</td>
<td>0.90</td>
<td>0.94</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Emotion</td>
<td>72.09</td>
<td>23.29</td>
<td>3.42</td>
<td>0.96</td>
<td>0.24</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Cognition</td>
<td>65.29</td>
<td>4.12</td>
<td>20.72</td>
<td>7.77</td>
<td>2.03</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Pain</td>
<td>45.45</td>
<td>36.42</td>
<td>12.10</td>
<td>4.44</td>
<td>1.59</td>
<td>NA</td>
</tr>
</tbody>
</table>

### SF-6D

- 6 domain instrument – derived from SF-12 or SF-36 – with varying numbers of levels depending on domain
- Domains and number of levels include:
  - Physical functioning
  - Role limitations
  - Social Functioning
  - Pain
  - Mental health
  - Vitality
- Multiple levels of 6 domains used to define either 7500 health states (SF-12 version) or 18,000 states (SF-36 version)

### SF-6D Scoring Rule

- Additive scoring rules derived by Brazier and colleagues from a valuation survey that elicited standard gamble preference scores from 611 members of UK general population
  - Separate rules for SF-12 and SF-36 versions
- Several country-specific scoring rules have also been published
Comparison of Prescored Instruments

<table>
<thead>
<tr>
<th></th>
<th>EQ-5D</th>
<th>HUI2</th>
<th>HUI3</th>
<th>SF-6D *</th>
</tr>
</thead>
<tbody>
<tr>
<td># scores&gt;0.9 (N)</td>
<td>1</td>
<td>27</td>
<td>14</td>
<td>38</td>
</tr>
<tr>
<td># scores&lt;0.0 (N)</td>
<td>84</td>
<td>63</td>
<td>643k</td>
<td>0</td>
</tr>
<tr>
<td>Average score†</td>
<td>0.137</td>
<td>0.286</td>
<td>-0.101</td>
<td>0.612</td>
</tr>
<tr>
<td>Lowest score</td>
<td>-0.594</td>
<td>-0.025</td>
<td>-0.359</td>
<td>0.345</td>
</tr>
</tbody>
</table>

* Based on SF-12 version  
† Assumes equal weighting of states

Use of Multiple Instruments in Same Populations *

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Weighted mean (SD) †</th>
<th>Instrument</th>
<th>Weighted mean (SD) †</th>
<th># responses</th>
<th>% significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ5D</td>
<td>0.762 (0.313)</td>
<td>HUI2</td>
<td>0.802 (0.241)</td>
<td>15,123</td>
<td>59</td>
</tr>
<tr>
<td>EQ5D</td>
<td>0.763 (0.314)</td>
<td>HUI3</td>
<td>0.709 (0.288)</td>
<td>19,311</td>
<td>55</td>
</tr>
<tr>
<td>EQ5D</td>
<td>0.729 (0.335)</td>
<td>SF6D</td>
<td>0.722 (0.184)</td>
<td>12,529</td>
<td>63</td>
</tr>
<tr>
<td>HUI2</td>
<td>0.797 (0.231)</td>
<td>HUI3</td>
<td>0.717 (0.285)</td>
<td>17,921</td>
<td>77</td>
</tr>
<tr>
<td>HUI2</td>
<td>0.767 (0.215)</td>
<td>SF6D</td>
<td>0.707 (0.169)</td>
<td>12,101</td>
<td>76</td>
</tr>
<tr>
<td>HUI3</td>
<td>0.672 (0.286)</td>
<td>SF6D</td>
<td>0.714 (0.169)</td>
<td>15,074</td>
<td>65</td>
</tr>
</tbody>
</table>

* Includes studies that assessed at least 3 of 4 instruments OR 2 prescored instruments and at least 2 direct assessment methods  
† Weights based on number of respondents in each sample

Conclusions: Multiple Instruments

- 37 studies; 71 samples of respondents; between 12,101 and 19,311 responses for each pair of instruments  
- Weighted average preference scores appear highest for HUI2 followed by EQ-5D, SF-6D, and HUI3  
- All instruments yielded statistically significantly different preference scores in more than 50% of samples in which they were compared  
- Weighted standard deviations appear smallest for SF-6D and largest for HUI3  
  - All else equal (no sure thing), SF-6D would allow enrollment of smaller sample sizes while providing equivalent power to detect differences
Minimally/Clinically Important Difference (MID/CID)

- MIDs reported in literature
  - EQ-5D, 0.03-0.05
  - HUI2 and HUI3, 0.01 to 0.04
  - SF-6D, 0.033

- Idea underlying MID: There exists a single boundary between changes in health that are and are not important, independent of both health endowment and cost of preventing decrement / improving health

MID Not an Economic Concept

- Unwillingness to play Russian roulette for any finite amount of money at same time as engaging in other risky behaviors thought to be explained by reference to health endowment
- Willingness to pay out-of-pocket for pain reliever for simple headache suggests any increment in health quality can be important if its cost is small enough
- Alternative economic definition of minimally important difference (??):
  - Any difference we are willing to pay to modify
  - Under this definition, 0.005 increment would be important if cost of treatment was $1

Relative Responsiveness

- 4 studies suggest equivalent responsiveness between HUI3 and EQ-5D, but 3 indicate HUI3 more responsive
- 3 suggest equivalent responsiveness between HUI3 and HUI2, but 3 indicate HUI3 more responsive, while one indicates reverse
- Most evidence for SF-6D indicates equivalence with other three instruments; few studies reporting differences tend to balance out
Superiority?

- Most studies that evaluated correlations between preference scores found them to be correlated
  - Correlations greater than 0.66 for all instruments in 2 large studies (1 healthy population; 1 diseased)
- Most that evaluated correlations between preference scores and convergent validity criteria found them to be correlated
- Most studies that evaluated responsiveness concluded that all instruments were responsive
- Most studies concluded there is little evidence that one instrument superior to another

But Which is Measuring QALYs?

- (By now) large number of authors have concluded that while all four instruments appear to be measuring quality of life, constructs being measured not identical and preference scores differ
  - "The index scores are not interchangeable in the calculation of longitudinal-based QALYs" (Conner-Spady, 2003)
  - "...results underscore the lack of interchangeability among different preference-based measures" (Feeny, 2012)

Withholding Judgment

- Given instruments should all be measuring same construct and lack of evidence of superiority of 1 instrument over another, disagreement in scores problematic
- (Continuing) widespread direct comparison of instruments not providing answer about when 1 instrument better than another
  - In part because correlation between instruments’ scores and convergent validity criteria and relative responsiveness not sufficient selection criterion
  - Having higher correlations with convergent validity criteria or being more responsive needn’t translate into being a better instrument
Directly Elicited Preference Scores

Direct Elicitation from Participants
- Second common approach for assessing QALYs directly elicits preferences from study participants
- Laryngeal cancer example illustrated use of SG and TTO to assess preferences for duration of morbidity
- Standard gamble (SG), time trade-off (TTO), and rating scale (RS) also can be used to develop scores like those from prescored instruments

Probability-Equivalent Standard Gamble
- Most common SG method for eliciting preference for current health
- Select certain life expectancy with current health (e.g., 10 years); identify best and worst outcomes: same number of years fully functional (e.g., 10 fully functional years) vs immediate death
- Offer subject choice between 10 certain years with current health and a 1-p/p chance for 0 and 10 fully functional years
- Participant asked to identify p such that she is indifferent between certain current health and gamble
**Utility of Gamble**
- Preference or utility score for current health equals probability that makes respondent indifferent between certain amount and gamble
- By indicating indifference, respondent indicates utility of certain outcome identical to expected utility of gamble
- By setting utility of worst outcome to 0 and utility of best outcome to 1, expected utility of gamble equals \( p \) times utility of best outcome \( (p \times 1 = p) \)
  - \( (1 - p) \) drops out because utility of worst outcome is set to 0

**Time Trade-Off**
- Step 1: Select life expectancy for current health (e.g., 10 years) and conduct time-trade-off
- Step 2: Offer 10 years with current health or willingness to live for some shorter amount of time with full functioning
- Step 3: If willing to trade-off, how many out of 10 years would you give up so that you’d have full functioning for remainder? For example, would you give up 3 years and choose 7 years with full functioning rather than 10 years with current health? If not, what number of years with full functioning would be equal to 10 years of current health?
  - Suppose answer was 7 healthy years?
Time Trade-Off (2)

- Step 4: Preference / value score equals number of healthy years divided by 10 years with current health
  - 7 / 10 = 0.7 = Preference score for year with current health
- Unlike standard gambles, TTOs do not satisfy axioms of expected utility theory
  - Because not measured with risk
- Like standard gambles, do require participants to choose between health outcomes

Rating Scale

- Rating scale – also referred to as visual analog scale or feeling thermometer – asks participants to rate how good or bad their current health is on a 0–1 or 0–100 scale
  - 0 often represents worst imaginable health or death
  - 1 often represents “best imaginable health” or “full health”
- Rating scales can vary in presentation in terms of length of line, whether drawn vertically or horizontally, and whether intervals marked out with numbers
- Some have argued that having intervals marked out with numbers can induce memory effects and clustering

Rating Scale (2)

- Rating scales neither satisfy axioms of expected utility theory, nor require that participants choose between health outcomes
- If a rating scale ranges between 0 and 1, point on line selected by participant represents preference score; if scale ranges between 0 and 100, point on line divided by 100 represents score
Use of Multiple Methods in Same Populations *

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Weighted mean (SD) †</th>
<th>Instrument</th>
<th>Weighted mean (SD) †</th>
<th># responses</th>
<th># samples</th>
<th>% significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>0.864 (0.178)</td>
<td>TTO</td>
<td>0.832 (0.207)</td>
<td>6815</td>
<td>67</td>
<td>49</td>
</tr>
<tr>
<td>SG</td>
<td>0.862 (0.180)</td>
<td>RS</td>
<td>0.719 (0.183)</td>
<td>7158</td>
<td>78</td>
<td>77</td>
</tr>
<tr>
<td>TTO</td>
<td>0.826 (0.210)</td>
<td>RS</td>
<td>0.719 (0.182)</td>
<td>7176</td>
<td>73</td>
<td>67</td>
</tr>
</tbody>
</table>

* Included studies that assessed all 3 direct assessment methods OR 2 direct assessment methods and at least 2 prescored instruments  † Weights based on number of respondents in each sample

Conclusions: Multiple Instruments

- 37 studies; 84 samples of respondents; between 6815 and 7176 responses for each pair of instruments
- Weighted average mean scores confirm suggestion in literature that difference between SG and TTO responses (~0.03) smaller than difference between SG and RS (~0.14) and TTO and RS (~0.11)
- But all 3 methods yielded significantly different preferences scores in 49% or more of samples in which they were compared.
Tengs and Lin Meta-Analyses

- Meta-analyses of responses from patients, caregivers, providers, and members of community who rated current health or disease scenarios for HIV or stroke
- TTOs appear to yield highest preference scores
- SG scores appear 0.1 lower than TTOs (p=0.16 for HIV and p=0.08 for stroke)
- RS scores -0.02 less than SG scores when rating HIV (RS vs SG, NS; RS vs TTO, p = 0.001)
- RS scores -0.11 less than SG scores when rating stroke (RS vs SG, p-value not reported; RS vs TTO, p=0.006)

Comparison Of Prescored Instruments And Direct Elicitation

Prescored Instruments vs Direct Elicitation *

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Weighted mean (SD) †</th>
<th>Method</th>
<th>Weighted mean (SD) †</th>
<th># responses</th>
<th># samples</th>
<th>% significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D</td>
<td>0.733 (0.224)</td>
<td>SG</td>
<td>0.834 (0.222)</td>
<td>1059</td>
<td>16</td>
<td>38</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>0.731 (0.222)</td>
<td>TTO</td>
<td>0.793 (0.257)</td>
<td>1227</td>
<td>15</td>
<td>42</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>0.732 (0.226)</td>
<td>RS</td>
<td>0.708 (0.200)</td>
<td>1420</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>HUI2</td>
<td>0.750 (0.164)</td>
<td>SG</td>
<td>0.862 (0.170)</td>
<td>257</td>
<td>7</td>
<td>71</td>
</tr>
<tr>
<td>HUI2</td>
<td>0.848 (0.162)</td>
<td>TTO</td>
<td>0.807 (0.198)</td>
<td>107</td>
<td>3</td>
<td>67</td>
</tr>
<tr>
<td>HUI2</td>
<td>0.750 (0.164)</td>
<td>RS</td>
<td>0.739 (0.173)</td>
<td>257</td>
<td>7</td>
<td>43</td>
</tr>
</tbody>
</table>

* Included studies that assessed at least 2 prescored instruments and at least 2 direct assessment methods
† Weights based on number of respondents in each sample
Prescored Instruments vs Direct Elicitation (2) *

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Weighted mean (SD) †</th>
<th>Method</th>
<th>Weighted mean (SD) †</th>
<th># responses</th>
<th># samples</th>
<th>% significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI3</td>
<td>0.701 (0.251)</td>
<td>SG</td>
<td>0.836 (0.215)</td>
<td>1020</td>
<td>17</td>
<td>41</td>
</tr>
<tr>
<td>HUI3</td>
<td>0.643 (0.245)</td>
<td>TTO</td>
<td>0.785 (0.247)</td>
<td>1188</td>
<td>16</td>
<td>56</td>
</tr>
<tr>
<td>HUI3</td>
<td>0.652 (0.250)</td>
<td>RS</td>
<td>0.710 (0.188)</td>
<td>1381</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td>SF-6D</td>
<td>0.678 (0.169)</td>
<td>SG</td>
<td>0.878 (0.200)</td>
<td>296</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>SF-6D</td>
<td>0.681 (0.172)</td>
<td>TTO</td>
<td>0.732 (0.306)</td>
<td>355</td>
<td>3</td>
<td>67</td>
</tr>
<tr>
<td>SF-6D</td>
<td>0.671 (0.157)</td>
<td>RS</td>
<td>0.695 (0.209)</td>
<td>505</td>
<td>7</td>
<td>14</td>
</tr>
</tbody>
</table>

* Included studies that assessed at least 2 prescored instruments and at least 2 direct assessment methods
† Weights based on number of respondents in each sample

Conclusions: Prescored vs Direct Assessment

- 11 studies; 29 samples of respondents; between 107 and 1420 responses for each pair of instruments
- 4 prescored instruments appear most similar to RS
  - Weighted mean differences: 0.024, 0.011, -0.058, and -0.024 for RS vs EQ-5D, HUI2, HUI3, and SF-6D
  - Significant differences in only 23%, 43%, 30%, and 14% of samples for RS vs EQ-5D, HUI2, HUI3, and SF-6D (but small sample sizes)
- SG and TTO both had scores generally substantially larger than EQ-5D, HUI3, and SF-6D scores

What to Make of These Findings?

- General recommendation for use of preferences from general public in economic evaluations
  - One rationale for use of prescored instruments
- Some evidence that patients’ ratings of own health are higher than general public’s ratings of scenarios that mirror patients’ health
  - Evidence not conclusive
- Appears that RS – which some consider least preferred method for direct elicitation of preferences – no worse at reproducing results of prescored instruments than other direct elicitation methods
  - May be better
Preference Scores for Selected Health States *

<table>
<thead>
<tr>
<th>Health State</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergic Rhinitis</td>
<td>0.853</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>0.834</td>
</tr>
<tr>
<td>Migraine</td>
<td>0.806</td>
</tr>
<tr>
<td>Essential hypertension</td>
<td>0.789</td>
</tr>
<tr>
<td>Depression disorders</td>
<td>0.732</td>
</tr>
<tr>
<td>Blindness / low vision</td>
<td>0.694</td>
</tr>
<tr>
<td>CVA</td>
<td>0.650</td>
</tr>
<tr>
<td>Heart failure</td>
<td>0.636</td>
</tr>
<tr>
<td>Senility w/ psychosis</td>
<td>0.545</td>
</tr>
<tr>
<td>Death</td>
<td>0.0</td>
</tr>
</tbody>
</table>


Literature-Based Sources

- Tufts Medical Center, Institute for Clinical Research and Health Policy Studies, Center for the Evaluation of Value and Risk in Health, CEA Registry
  www.cearegistry.org
  OR
  https://research.tufts-nemc.org/cear/

Constructing QALYs by Use of Preference Scores

<table>
<thead>
<tr>
<th>Month</th>
<th>SE</th>
<th>MO</th>
<th>EM</th>
<th>CO</th>
<th>SC</th>
<th>PN</th>
<th>FE</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.896</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.896</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0.535</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.640</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.748</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.868</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.896</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.896</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.896</td>
</tr>
</tbody>
</table>

*SE: sensory; MO: mobility; EM: emotion; CO: cognition; SC: self-care; PN: pain; and FE: fertility
Short-term "Hellish" Experiences

- Suppose you have a cavity and the dentist plans to drill tooth for 10 minutes (.000019013 years)
- Suppose you rate drilling minutes as having a preference score of 0 (i.e., one loses 0.000019013 QALYS by having one's teeth drilled)
- If intervention to relieve pain costs $5, cost/QALY saved equals $263,000 ($5/0.000019013)
- Should we recommend against such interventions?
  - If not, what needs to be changed in our calculation?

Choice Between Instruments/Methods

- None of evidence presented runs counter to recommendation to measure both general public's and patients' preferences
- But review has not led to strong conclusions about best methods for measurement
Choice Between Instruments/Methods (II)

• Complicated to get preferences exactly right
  – Human preferences so variable and having so many determinants
  – All measurement techniques flawed
• Many of “recommendations” from commentators seem based on theories that ignore complexity and flaws
  – It would be easy to recommend sensitivity analysis for preference scores, but strategy is costly
• Conclusion: Not clear that strong recommendations about adoption of specific methods or instruments are supportable