


**Preference Assessment**

Henry Glick

EPI 550

April 15, 2020



---

---

---

---

---


---

---

---

**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 [Neumann et al., Cost-Effectiveness in Health and Medicine. Second edition, summary recommendation 7.1, p. 375]



---

---

---

---

---

---


---

---

**But 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 in NYHA class III)

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/Preference 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 fully functional/"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

Response Method	Question Framing	
	Certainty (values)	Risk (utilities)
Scaling	Rating Scale	
	Category Scaling	--
	Visual Analog	
Choices	Time trade-off	Standard gamble
	Paired comparison	

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



---

---

---

---

---

---

---

---

### Years of Quality-Adjusted Survival

- Gold standard preference assessment directly measures preferences for level of morbidity and its duration
  - Sankey's review of 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



---

---

---

---

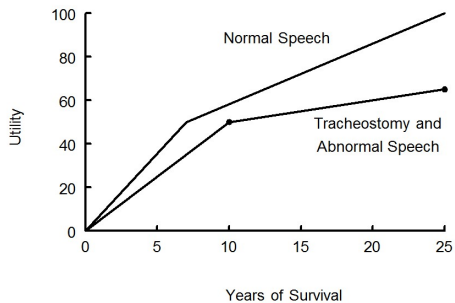
---

---

---

---

### Sankey's Preference Mapping of Quality and Quantity of Survival



---

---

---

---

---

---

---

---

### Preferences for Current Health

- Because Sankey already addressed development of a preference mapping and because most preference assessment is based on a series of valuations of current health, in following discussion review methods for latter type of assessment



---

---

---

---

---

---

---

---

### Direct Elicitation vs Indirect Preference Assessment

- Direct elicitation: Direct rating of preference for health by respondent
  - Can be used to assess current health or to generate preference mapping
- Indirect preference assessment: Uses instruments which have respondent directly rate functional status across a variety of domains but 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 recommends a reference case analysis that uses community preferences to value health [Neumann et al., Cost-Effectiveness in Health and Medicine. Second edition, summary recommendation 7.4, p. 375]
  - Empathy
  - Trust those who already have disease?
- Also recommend a sensitivity analysis that uses preferences of persons with condition



---

---

---

---

---

---

---

---

### Outline

- Prescored health classification instruments
  - EQ-5D
  - HUI2
  - HUI3
  - SF-6D
- Direct elicitation
  - Standard Gambles
  - Time trade-off
  - Rating scale
- Comparison of Methods



---

---

---

---

---

---

---

---

### 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



---

---

---

---

---

---

---

---

### Direct Elicitation

- Second dominant approach for estimating preference scores directly elicits participants' preferences for their current health
- Methods include:
  - Standard gamble
  - Time trade-off
  - Rating scales
- When administered to study participants, these methods yield measures of patient preference



---

---

---

---

---

---

---

---

### Scenarios

- A third approach describes disease scenarios to members of general public and directly elicits preferences for these scenarios
- We do not discuss this method below. Rather, in what follows, we describe prescored instruments and direct elicitation methods.



---

---

---

---

---

---

---

---

### Prescored Health Classification Instruments



---

---

---

---

---

---

---

---

### Prescored Instruments

- A number of prescored instruments currently available for measurement of preference scores for current health
  - EuroQol instrument (EQ-5D)
  - 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 minutes 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-3L Levels of Function

- In original instrument, each domain defined by 3 levels of function from good to poor
- 3-levels 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<sup>5</sup>) health states



---

---

---

---

---

---

---

---

### EQ-5D-5L Levels of Function

- More recently, 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<sup>5</sup>) health states



---

---

---

---

---

---

---

---

### EQ-5D “Tariffs”/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
- Shaw et al. developed a 3-level US scoring rule from responses from 3773 respondents from a multistage probability sample of noninstitutionalized English- and Spanish-speaking adults, aged 18 and older  
(Shaw JW, et al. US valuation of the EQ-5D health states. Developing and testing of the D1 valuation model. Med Care. 2005;43:203-20.)
- 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)

<b>Mobility</b>	
1. I have no problems walking about	0.000
2. I have some problems walking about	0.069
3. I am confined to bed	0.314
<b>Self-Care</b>	
1. I have no problems with self-care	0.000
2. I have some problems washing and dressing myself	0.104
3. I am unable to wash or dress myself	0.214
<b>Usual Activities</b>	
1. I have no problems with performing my usual activities	0.000
2. I have some problems with performing my usual activities	0.036
3. I am unable to perform my usual activities	0.094
<b>Pain/Discomfort</b>	
1. I have no pain or discomfort	0.000
2. I have some pain or discomfort	0.123
3. I have extreme pain or discomfort	0.386
<b>Anxiety/Depression</b>	
1. I am not anxious or depressed	0.000
2. I am moderately anxious or depressed	0.071
3. I am extremely anxious or depressed	0.236




---

---

---

---

---

---

---

---

---

---

---

---

### Dolan Scoring for EuroQol

- 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




---

---

---

---

---

---

---

---

---

---

---

---

### US Scoring Rule (Shaw)

	Mean Rule
M2	.146
M3	.558
S2	.175
S3	.471
U2	.140
U3	.374
P2	.173
P3	.537
A2	.156
A3	.450
# Non-1s	-.140
(#2s (0 to 4)) <sup>2</sup>	.011
(#3s (0 to 4))	-.122
(#3s (0 to 4)) <sup>2</sup>	-.015




---

---

---

---

---

---

---

---

---

---

---

---

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

- Directly elicited scores have recently been published in a number of countries, e.g.,

Canada	Japan	Thailand
China	Korea	Uruguay
England	Malaysia	US (2)
France	Netherlands	Vietnam
Germany	Poland	
Indonesia	Spain	
Ireland	Taiwan	




---

---

---

---

---

---

---

---

---

---

### Selected EQ-5D-5L Tariffs

Domain	Japan	Netherlands	Uruguay	US
MO2	-0.0639	-0.035	-0.0140	-0.096
MO3	-0.1126	-0.057	-0.0322	-0.122
MO4	-0.1790	-0.166	-0.1077	-0.237
MO5	-0.2429	-0.203	-0.2987	-0.322
SC2	-0.0436	-0.038	-0.0256	-0.089
SC3	-0.0767	-0.061	-0.0609	-0.107
SC4	-0.1243	-0.168	-0.1169	-0.220
SC5	-0.1597	-0.168	-0.2734	-0.261
UA2	-0.0504	-0.039	-0.0424	-0.068
UA3	-0.0911	-0.087	-0.0455	-0.101
UA4	-0.1479	-0.192	-0.1183	-0.255
UA5	-0.1748	-0.192	-0.2315	-0.255
PD2	-0.0445	-0.066	-0.0171	-0.060
PD3	-0.0682	-0.092	-0.0607	-0.098
PD4	-0.1314	-0.360	-0.1870	-0.318
PD5	-0.1912	-0.415	-0.2705	-0.414
AD2	-0.0718	-0.070	-0.0095	-0.057
AD3	-0.1105	-0.145	-0.0435	-0.123
AD4	-0.1682	-0.356	-0.1043	-0.299
AD5	-0.1960	-0.421	-0.1771	-0.321
Con	0.9391	0.953	0.9874	1




---

---

---

---

---

---

---

---

---

---

### (Very Different) Canadian EQ-5D-5L Tariffs

Domain *	Tariff
MO	-0.0389
SC	-0.0458
UA	-0.0195
PD	-0.0444
AD	-0.0376
MO45	-0.0510
SC45	-0.0584
UA45	-0.1103
PD45	-0.1409
AD45	-0.1277
N45 <sup>2</sup>	0.0085
Cons	1.1351

- Mo, SC, UA, PD, AD = domain level (1-5); MO45, SC45, UA45, PD45, AD45 = 0/1 variable representing level 4/5 function; N45<sup>2</sup> = square of number of level 4/5 domains




---

---

---

---

---

---

---

---

---

---

### Valuing State 23245

Country	Equation	Score
Canada	$1.1351 + 4 \cdot 0.0085 - 2 \cdot 0.0389 - 3 \cdot 0.0458 - 2 \cdot 0.0195 - 4 \cdot 0.0444 - 5 \cdot 0.0376 - 0.1409 - 0.1277$	0.281
Japan	$0.9391 - 0.0639 - 0.0767 - 0.0504 - 0.1314 - 0.1960$	0.421
Netherlands	$0.953 - 0.035 - 0.061 - 0.039 - 0.360 - 0.421$	0.037
UK	$1 - 0.9675 \cdot (0.051 + 0.076 + 0.051 + 0.276 + 0.301)$	0.267
Uruguay	$0.9874 - 0.0140 - 0.0609 - 0.424 - 0.1870 - 0.1771$	0.506
US	$1 - .096 - .107 - .068 - .318 - .321$	0.090




---

---

---

---

---

---

---

---

---

---

---

---

### Pediatric EQ-5D-Y \*

- “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)

\* Wille et al. Development of the EQ-5D-Y: a child-friendly version of the EQ-5D. Qual Life Res. 2010;19:875-86.

\* Ravens-Sieberer et al. Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study. Qual Life Res (2010) 19:887-897




---

---

---

---

---

---

---

---

---

---

---

---

### 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”




---

---

---

---

---

---

---

---

---

---

---

---

### Pediatric EQ-5D-Y Tariffs

- At least one published scoring rule for children's preferences
  - Craig BM, et al. Valuation of child health-related quality of life in the United States. Health Economics. 2016; 25: 768-77.
- 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



---

---

---

---

---

---

---

---

### Scoring HUI2

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	0.84
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




---

---

---

---

---

---

---

---

---

---

---

---

### Scoring HUI2

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...	0.97
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 HUI2

- 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
	$(1.06 \text{ score}) - 0.06$	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

Description		
1. Able to see well enough to read ordinary newspaper and recognize a friend on the other side of the street, without glasses or contact lenses	1.00	
2. Able to see well enough to read ordinary newspaper and recognize a friend on the other side of the street, but with glasses	0.98	
3. Able to read ordinary newspaper with or without glasses but unable to recognize a friend on the other side of the street, even with glasses	0.89	
4. Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newspaper, even with glasses	0.84	
5. Unable to read ordinary newspaper and unable to recognize a friend on the other side of the street, even with glasses	0.75	
6. Unable to see at all	0.61	




---

---

---

---

---

---

---

---

---

---

### Scoring HUI3

• Scoring formula:

$$U^* = 1.371(w_1 * w_2 * w_3 * w_4 * w_5 * w_6 * w_7 * w_8) - 0.371$$

Domain	Level	Score
Vision	2	0.98
Hearing	1	1.00
Speech	2	0.94
Ambulation	3	0.86
Dexterity	2	0.95
Emotion	3	0.85
Cognition	2	0.92
Pain	2	0.96

$$\prod (1.371 \text{ score}) - 0.371 = 0.404$$




---

---

---

---

---

---

---

---

---

---

Estimated Rates of Health Problems of the General Adult U.S. Population (Luo et al. Medical Care. 2005;43:1080)

Instrument/ Domain	Level					
	1	2	3	4	5	6
EQ-5D (n=3977)						
Mobility	81.14	18.66	0.21	NA	NA	NA
Self-care	95.93	3.81	0.26	NA	NA	NA
Usual activities	84.56	13.59	1.84	NA	NA	NA
Pain/discomfort	59.15	37.10	3.75	NA	NA	NA
Anxiety/depression	73.71	23.89	2.39	NA	NA	NA
HUI2 (n=3889)						
Sensation	38.88	48.24	11.27	1.61	NA	NA
Mobility	86.12	9.45	3.72	0.71	0	NA
Emotion	66.96	29.94	2.17	0.57	0.37	NA
Cognition	65.29	33.13	1.51	0.06	NA	NA
Self-care	96.50	2.93	0.21	0.36	NA	NA
Pain	42.54	45.14	8.25	3.06	1.01	NA
HUI3 (n=3907)						
Vision	42.30	54.13	1.15	1.83	0.50	0.09
Hearing	92.83	0.91	1.91	2.56	0.30	1.49
Speech	92.69	5.00	1.81	0.32	0.18	NA
Ambulation	86.09	9.48	2.38	1.33	0.44	0.27
Dexterity	91.69	6.38	0.90	0.94	0.05	0.04
Emotion	72.09	23.29	3.42	0.96	0.24	NA
Cognition	65.29	4.12	20.72	7.77	2.03	0.06
Pain	45.45	36.42	12.10	4.44	1.59	NA




---

---

---

---

---

---

---

---

---

---



### SF-6D

- 6 domain instrument – derived from SF-12 and SF-36 – with varying numbers of levels depending on domain
- Domains 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
  - Craig BM, Pickard S, Stolk E, Brazier JE. US valuation of the SF-6D. Med Decis Making. 2013;33:793-803




---

---

---

---

---

---

---

---

---

---

---

---

### Comparison of Prescored Instruments

	EQ-5D	HUI2	HUI3	SF-6D *
# scores>0.9 (N)	1	27	14	38
# scores<0.0 (N)	84	63	643k	0
Average score†	0.137	0.286	-0.101	0.612
Lowest score	-0.594	-0.025	-0.359	0.345

\* Based on SF-12 version

† Assumes equal weighting of states




---

---

---

---

---

---

---

---

---

---

---

---

### Use of Multiple Instruments in Same Populations \*

Instru-ment	Weighted mean (SD) †	Instru-ment	Weighted mean (SD) †	# res-ponses	# samples	% signi-ficant
EQ5D	0.762 (0.313)	HUI2	0.802 (0.241)	15,123	27	59
EQ5D	0.763 (0.314)	HUI3	0.709(0.288)	19,311	53	55
EQ5D	0.729 (0.335)	SF6D	0.722 (0.184)	12,529	30	63
HUI2	0.797 (0.231)	HUI3	0.717 (0.285)	17,921	39	77
HUI2	0.767 (0.215)	SF6D	0.707 (0.169)	12,101	34	76
HUI3	0.672 (0.286)	SF6D	0.714 (0.169)	15,074	37	65

\* Included 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 of 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
- Sankey's laryngeal cancer example illustrated use of these method to assess preferences for duration of morbidity (i.e., curves he drew)
- Three most common methods for doing so are:
  - Standard gamble (SG)
  - Time trade-off (TTO)
  - Rating scale (RS)



---

---

---

---

---

---

---

---

### 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



---

---

---

---

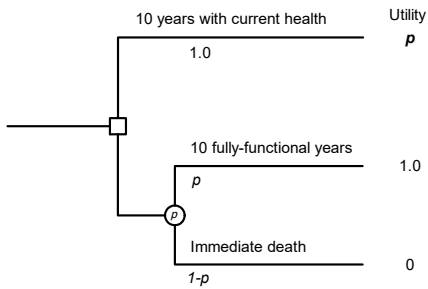
---

---

---

---

### Depiction of Probability-Equivalent 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 ( $p * 1 = p$ )
  - $(1 - p)$  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
- As Sankey noted last class, 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 (II)

- As Sankey noted last class, rating scales neither satisfy axioms of expected utility theory, nor require that participants choose between health outcomes
- If 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



---

---

---

---

---

---

---

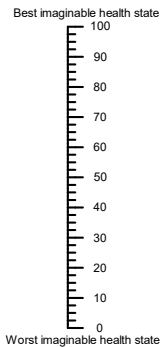
---

### Rating Scale Example

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked by 100, and the worst state you can imagine is marked by 0.

Your own health state today

We would like you to indicate on this scale how good or bad your health is today. Please do this by drawing a line from the box to whichever point on the scale that indicates how good or bad your health is currently.



---

---

---

---

---

---

---

---

### Use of Multiple Methods in Same Populations \*

Instru- ment	Weighted mean (SD) †	Instru- ment	Weighted mean (SD) †	# res- ponses	# samples	% signi- ficant
SG	0.864 (0.178)	TTO	0.832 (0.207)	6815	67	49
SG	0.862 (0.180)	RS	0.719 (0.183)	7158	78	77
TTO	0.826 (0.210)	RS	0.719 (0.182)	7176	73	67

\* Includes 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 appeared 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 \*

Instru- ment	Weighted mean (SD) †	Me- thod	Weighted mean (SD) †	# res- ponses	# samples	% signi- ficant
EQ-5D	0.733 (0.224)	SG	0.834 (0.222)	1059	16	38
EQ-5D	0.731 (0.222)	TTO	0.793 (0.257)	1227	15	42
EQ-5D	0.732 (0.226)	RS	0.708 (0.200)	1420	22	23
HUI2	0.750 (0.164)	SG	0.892 (0.170)	257	7	71
HUI2	0.848 (0.162)	TTO	0.807 (0.198)	107	3	67
HUI2	0.750 (0.164)	RS	0.739 (0.173)	257	7	43

\* 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) \*

Instru- ment	Weighted mean (SD) †	Me- thod	Weighted mean (SD) †	# res- ponses	# samples	% signi- ficant
HUI3	0.701 (0.251)	SG	0.836 (0.215)	1020	17	41
HUI3	0.643 (0.245)	TTO	0.785 (0.247)	1188	16	56
HUI3	0.652 (0.250)	RS	0.710 (0.188)	1381	23	30
SF-6D	0.678 (0.169)	SG	0.878 (0.200)	296	6	100
SF-6D	0.681 (0.172)	TTO	0.732 (0.306)	355	3	67
SF-6D	0.671 (0.157)	RS	0.695 (0.209)	505	7	14

\* 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




---

---

---

---

---

---

---

---

**Constructing QALYs by Use of Preference Scores**

Hypothetical responses to the HUI2 measured quarterly for 2 years\*

Month	SE	MO	EM	CO	SC	PN	FE	Score
0	1	2	2	1	1	1	1	0.896
3	1	2	2	1	1	1	1	0.896
6	1	3	3	1	2	3	1	0.535
9	1	3	3	1	1	2	1	0.640
12	1	2	3	1	1	2	1	0.748
15	1	2	2	1	1	2	1	0.868
18	1	2	2	1	1	1	1	0.896
21	1	2	2	1	1	1	1	0.896
24	1	2	2	1	1	1	1	0.896

\*SE: sensory; MO: mobility; EM: emotion; CO: cognition; SC: self-care; PN: pain; and FE: fertility




---

---

---

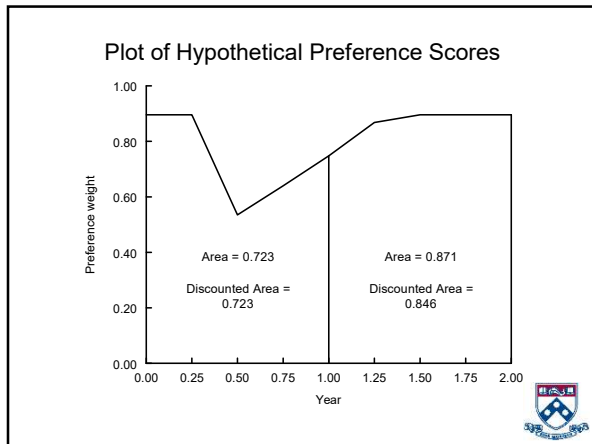
---

---

---

---

---




---

---

---

---

---

---

---

---

### Frequency of Elicitation

- Preferences usually measured for all study participants at prescheduled intervals, e.g., baseline and semi-annually thereafter
- Other designs that yield unbiased results include assessment at random intervals or random assignment to assessment intervals
- One design that will lead to biased results is purposely sampling when a clinical outcome occurs, such as onset of myocardial infarction during follow-up

---

---

---

---

---

---

---

---

### Frequency of Elicitation (II)

- Frequency of elicitation depends on beliefs about how rapidly preferences are expected to change, likely duration of changes, length of follow-up, and resources available for data collection
- For studies that last several years, routinely recommend assessing preferences at least twice a year
  - In a recent long-term clinical trial, we expected an initial rapid change and recommended quarterly assessment during first year of follow-up
  - Thereafter we measured preferences semi-annually

---

---

---

---

---

---

---

---



### Short-term "Hellish" Experiences

- Suppose you have cavity and 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



---

---

---

---

---

---

---

---

APPENDIX  
Implementation Issues,  
Direct Elicitation



---

---

---

---

---

---

---

---

Open- Versus Closed-Ended Questions

- Standard gambles and time trade-offs can be administered by use of a single open-ended question,
  - “Which  $p$  makes you indifferent?” or “How many years with full function make you indifferent?”
- They are more commonly administered by use of a series of close-ended questions
  - e.g., “Would you rather live with your current health for 10 years or would you choose a gamble in which you have a 90% chance of living 10 fully functional years and a 10% chance of dying immediately.”
  - Probabilities are changed and question repeated until respondent reports she is indifferent between options



---

---

---

---

---

---

---

---

Search Procedures

- When offered as series of close-ended questions, questions can:
  - Ping pong from high to low to high
  - Offer probabilities or years of healthy survival in steps from maximum to minimum (titration down)
  - Offer them from minimum to maximum (titration up)
  - Be posed by use of interval division search strategies (bisecting search routines)



---

---

---

---

---

---

---

---

### Effects of Search Procedures

- Lenert et al. have reported that different search procedures can have strong and persistent effects on reported preference scores for both standard gambles and time trade-offs
  - Supports findings of an earlier study by Percy and Llewellyn-Thomas
- Hammerschmidt et al., on other hand, did not see significant differences between results of mailed questionnaire standard gambles that used top-down versus bottom-up search procedures
  - Supports earlier findings by Tsevat et al.




---

---

---

---

---

---

---

---

### Time Horizon

- Preferences for highly confining health states appear to be a decreasing function of time, whereas preferences for inconvenient health states appear to be an increasing function of time

Torrance et al., 1972




---

---

---

---

---

---

---

---

### Effects of Different Time Horizons

- Most investigators who have empirically assessed effect of time horizon have found that longer time horizons, associated with smaller preference scores
  - Finding holds for standard gambles, time trade-offs, and rating scales

Morbid Years	Healthy Years	TTO Weight
25	12.5	0.5
10	7	0.7
5	5	1.0

McNeil et al. 1981




---

---

---

---

---

---

---

---

### What Time Horizons Have Investigators Used?

- Out of 35 studies that asked patients to rate their current health by use of standard gamble, time trade-off, and rating scale:
  - 10 used time horizons  $\leq 15$  years
  - 11 used horizons of 20–60 years
  - 13 used life expectancy as time horizon
- Unclear how much variability of results in literature arises because of use of different time horizons



---

---

---

---

---

---

---

---

### Methods of Administration

- Standard gambles and time trade-offs most commonly administered by use computer followed by of interview
- U-Titer, U-Maker, and iMPACT, and custom-developed software
- Interviews often use aids such as chance boards, decision wheels, and pie charts
- All three methods can be self-completed by participants



---

---

---

---

---

---

---

---

### Telephone Surveys

- van Wijck et al. have reported that telephone interviews (preceded by a mailed survey) yield standard gamble and time trade-off results that are similar to those obtained by face-to-face interview



---

---

---

---

---

---

---

---



### Mailed Surveys

- Good evidence of feasibility of use of rating scales in mailed, self-completed surveys [63,64]
- Evidence for feasibility of mailed, self-completed standard gambles appears more mixed
  - Ross et al. and Littenberg et al. reported a one-page paper standard gamble is a reliable measure of patient preference and is suitable for use in mailed surveys
  - Hammerschmidt et al., have reported substantial feasibility problems for mailed, self-completed standard gambles



---

---

---

---

---

---

---

---

### General Practicality

- Green et al. report substantial evidence supporting all 3 methods' practicality in terms of completion and response
  - Discount claims that standard gambles are too complex or not intuitively obvious to participants
  - Do note rating scales may be "slightly better in terms of response rate and cost"
- Also note standard gambles and time trade-off methods may "result in a larger number of refusals, missing values, and inconsistent responses" than do rating scales



---

---

---

---

---

---

---

---

### General Practicality (II)

- Woloshin et al. more recently raised concerns about quality of results from standard gambles and time trade-offs among less numerate participants
- Green et al. report that all three methods have acceptable levels of reliability, although they found some evidence that time trade-off may have slightly better test-retest performance



---

---

---

---

---

---

---

---