Guidebook

Preference Measurement in Economic Analysis

Patricia L. Sinnott, Vilija R. Joyce, Paul G. Barnett

March 2007

Preference Measurement in Economics Analysis. Guidebook.

Health Economics Resource Center (HERC)
VA Palo Alto Healthcare System
795 Willow Road (152 MPD)
Menlo Park, CA 94025
650-617-2630
650-716-2639 (fax)
herc@va.gov

Suggested citation: Sinnott PL, Joyce VR, Barnett PG. Preference Measurement in Economic Analysis. Guidebook. Menlo Park CA. VA Palo Alto, Health Economics Resource Center; 2007.

Acknowledgements: The authors wish to acknowledge the Forest Baker, Ph.D. who contributed early drafts of this manuscript. Funding for this guidebook was provided by the Cooperative Studies Program (CSP), a unit of the Veterans Health Administration (VHA) Clinical Science Research and Development Service.

Contents

Tables and Figures

1. Introduction and Overview

The continuing rise in the cost of healthcare and the wide range of treatment options available for most medical conditions suggests that cost-effectiveness analyses will be increasingly needed to resolve whether the benefits of treatment justify its cost (Gold, Siegel et al. 1996; Weinstein, Siegel et al. 1996). Because new treatments and interventions are rarely both more effective and less expensive than their predecessors, cost-effectiveness analysis has developed into a useful tool to guide health care practice and policy decisions. These analyses compare the costs of an intervention to natural units of benefit or health outcome, e.g. deaths averted or cases of disease identified. Of particular interest is a subset of cost-effectiveness analysis (also known as *cost-utility analysis* outside of the U.S.) which measures the benefit or health outcome in quality-of-life improvement, defined by the quality-adjusted life year or QALY.

This guide includes an overview of the concepts of preference measurement and quality-adjusted life years, a description of the most common techniques used for measuring preferences in economic evaluation, a summary of experience measuring preferences in the VA Cooperative Studies Program (CSP) (Spitzer, Dobson et al. 1981) clinical trials, and recommended criteria by which to select both the methods and measures to use in cost-effectiveness analysis (CEA). In Appendix 1 we have included sample questions from each of the multi-attribute health status classification systems included in the discussion of indirect methods. In Appendices 2 and 3, we have included links and references to various resources highlighted in the document.

2. Estimating quality-adjusted life years

The QALY is the most comprehensive measure of health outcome, simultaneously combining changes in morbidity and mortality into a single measure (Gold, Siegel et al. 1996). The QALY combines life expectancy with stated preferences or utilities for certain health states and allows comparison across treatments for heterogeneous conditions with various clinical effects (Torrance 1986; Brazier, Deverill et al. 1999; Guillemin 1999). In a study, QALYs are estimated by multiplying the average number of years subjects spend in each particular health state by a preference weight associated with that health state. These preference weights reflect the desirability or preference for that health state as estimated directly from the study subjects (direct measurement) or indirectly using a variety of measurement systems (indirect measurement). This preference weight is also known as the utility weight. These preference or utility weights are scaled from 0 – 1 where 0 represents death and 1 represents perfect health. Three kinds of information are needed to estimate the QALYs for a group or population:

- Descriptions of the various health states experienced during these lifetimes
- The duration of these health states
- An estimate of the preference or desirability of these health states for a given population.

Two methods and several classification systems are commonly used to estimate preference weights for a broad range of health states. The purpose of this guide is to help researchers choose the most appropriate method and system for a given study. Outpatient procedure codes

2.1 Definitions

The following definitions are critical to differentiating clinical health-related quality of life outcomes from economic health-related quality of life outcomes for use in CEA.

- *Quality of life* is a broad description of how a person feels and functions in isolation as well as within his or her political, economic and cultural environment

- *Health-related quality of life* is a narrower concept, including only those elements associated with biologic, physical, and emotional health, but not including politics, economics, or other environmental context (Torrance 1987; Torrance 1997)

- *Health states* are differentiated stages of a lifetime or disease progression which can be temporary or permanent.

- *Health status measures* are multi-question instruments which quantitatively measure the dimensions of health believed to be important to patients and influenced by disease, treatment, or natural changes in health, for example, changes associated with aging, pregnancy, or trauma (Brazier, Deverill et al. 1999). *Health status measures* are used to define individual profiles of health-related quality of life and to define and describe health states along various dimensions or domains of health, including perception of health, social function, psychological function, physical function, mobility, and pain. *Health status measures* can be generic, such as the SF-36 (Ware and Sherbourne 1992), or specific to a certain age, gender, population, or disorder, for example, the Spitzer Quality of Life (Spitzer, Dobson et al. 1981), the Oswestry Low Back Pain Questionnaire (Fairbank, Couper et al. 1980), and the Center for Epidemiologic Studies Depression Scale (Nordin, Alexandre et al. 2003). Health status measures alone are not used in economic analysis.

- *Preference-based health status measures* are instruments that define an individual's health state for use in economic analysis. In these measures, each possible health state is associated with an estimate of the value (preference or utility weight) that a surveyed sample of the general population (community sample) has attributed to these health states. These preference-based measures are used for estimation of QALYs (Torrance 1987; Gold, Siegel et al. 1996; Torrance 1997; Gold, Stevenson et al. 2002).

In this guide, we follow the convention of other researchers by using utilities to mean the expressed preferences or cardinal ranking of one health state when compared to another (Torrance 1987; Gold, Siegel et al. 1996).

2.2 Defining the preferences for the health states experienced during a trial

In a cost-effectiveness analysis done alongside a clinical trial, researchers elicit subject preferences for the health states they experience, using direct or indirect methods of elicitation. With direct methods, subjects in the trial directly score their preferences for the health states they experience. With indirect methods, subjects define the health states they experience by their

responses to surveys about various aspects of their health. These responses are aggregated into a single score and are linked through the proprietary scoring algorithms of the selected system to preference weights established by surveys of non-patient community samples. Using the same or similar surveys, population weights for indirect methods have been established through this survey methodology and the proprietary scoring algorithms. In both cases, direct and indirect, these preference weights are used to calculate QALYs.

2.3 Estimating the duration of each health state

The most accurate estimate of the duration of each health state an individual experiences would require constant and instantaneous measurements of health status, taken throughout a patient's life. Obviously, this continuous measurement is not feasible; therefore, in an economic evaluation, the duration of each health state is estimated using health status measurements taken at defined intervals. The more frequent the health status measurement, the more accurate the estimate of the duration of each health state (Gold, Siegel et al. 1996; Gold, Stevenson et al. 2002). In a study, the duration of each health state is usually assumed to last exactly one-half of the elapsed time between two measurement intervals as recommended by the U.S. Panel on Cost-Effectiveness in Health and Medicine (Gold, Siegel et al. 1996).

2.4 Estimation of a lifetime of QALYs

A lifetime of QALYs is estimated by multiplying the remaining number of years in each health state by the preference weight for that health state, and summing across the lifetime. Consider a patient with HIV who has undergone treatment with a medication. During the trial the patient experienced four different health states before suddenly dying:

Stage I	HIV+/asymptomatic
Stage II	HIV+/symptomatic
Stage III	AIDS
Stage IV	AIDS with cytomegalovirus

Using a direct method, the patient's preference for each health state was estimated at:

Stage I	.85
Stage II	.75
Stage III	.65
Stage IV	.40

The patient experienced Stage I for one-half year; Stage II for one year; Stage III for one year and Stage IV for 1.25 years. This patient lived 3.75 years under this treatment regimen. In this example, 2.325 QALYs were estimated for this patient during this time period. Table 1 demonstrates the calculation of a lifetime of QALYs for this patient.

Table 1: Total remaining lifetime of QALYs

Stage	Preference Weight (*pw*)	Duration (in years) of health state (*d*)	Total QALYs for this health state (*pw* * *d*)
I	.85	.5	.425
II	.75	1	.75
III	.65	1	.65
IV	.40	1.25	.5
Total		3.75	2.325

3. Common approaches to estimating preference weights for economic analysis alongside clinical trials

Preferences for health states have been estimated using off-the-shelf values from the literature, from patients using direct methods including the standard gamble (SG) and the time trade-off (TTO), and several rating scales including the visual analog scale (VAS) (Lenert, Cher et al. 1998). Preferences can also be estimated with indirect or multi-attribute methods, including the Health Utility Indexes (HUI II and III), the EuroQoL (EQ-5D), the Quality of Well-Being Scale (QWB), and the Short Form 6D (SF-6D).

On occasion, economic researchers will use a disease-specific or non-preference-based health status measure to determine outcomes, because other tools are not sensitive enough to the changes in health-related quality of life they expect to observe or need to quantify. In those cases, researchers gather disease-specific health related quality of life information from subjects in the study and estimate preference weights for these defined health states in subsequent studies. These methodologies are further described in Section 3.4.

3.1 Off-the-shelf utility or preference weights

If the health states of a study are known, and preference weights for these health states are available from the literature, then "off-the-shelf" utility or preference weights can be used to estimate QALYs. This method is usually restricted to modeling outcomes beyond the duration of a clinical trial.

3.1.1 Strengths and weaknesses of off-the-shelf utility or preference weights

Off-the-shelf utility weights provide an appealing method to estimate QALYs when direct or indirect methods are not feasible. Care must be taken when using these weights from the literature, however, as results are known to be significantly influenced by the elicitation procedures used in a study (Jansen, Stiggelbout et al. 2000; Lenert and Kaplan 2000). In particular, combining utility weights from several studies is not recommended because of this influence of the elicitation procedures.

3.2 Direct methods

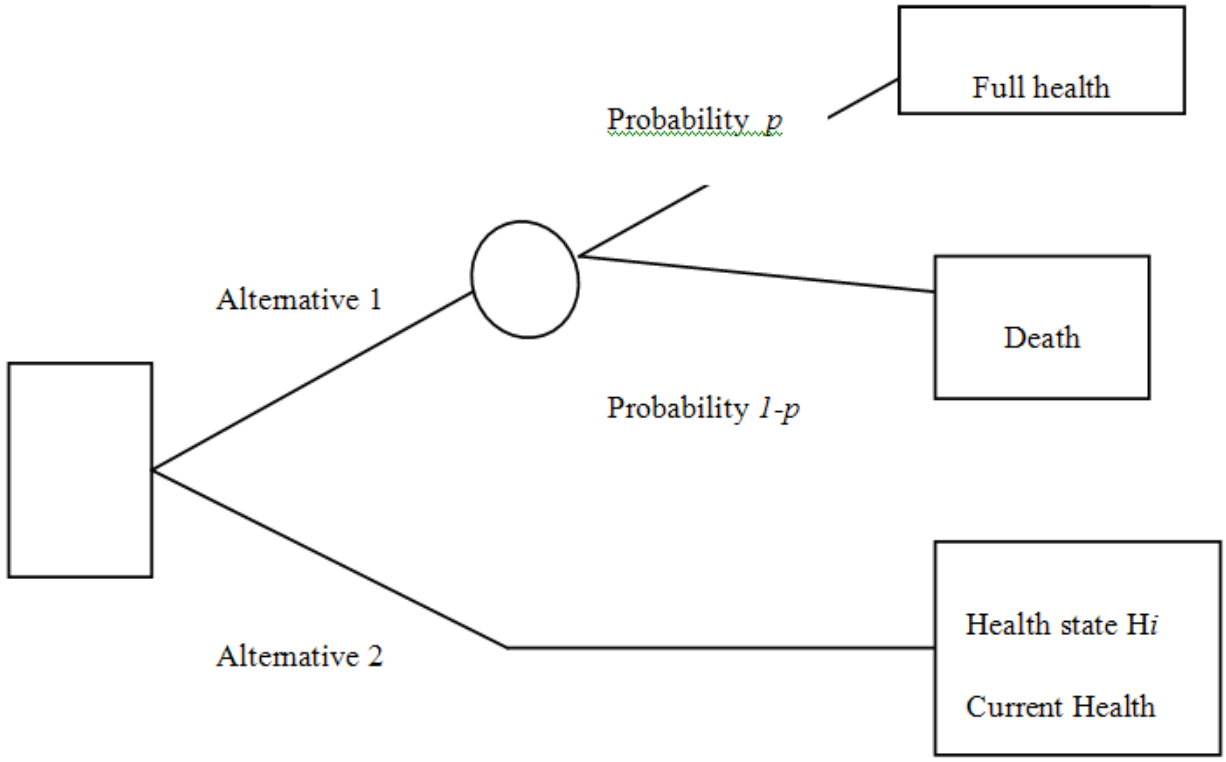
In direct methods, individuals are asked to rate the desirability of various health states. Individuals rank their preferences, making trade-offs between health states and alternatives. Individuals make judgments based on their own relative values for the various domains or characteristics of the health state experienced or described. For example, an individual who values being "medication-free" over mobility, will rank a health state differently from an individual who values mobility over freedom from medication.

The standard gamble and the time trade-off methods are the direct methods used to estimate preference or utility weights for economic evaluation. Direct methods might be used with subjects in a clinical trial, or to establish preference weights for indirect multi-attribute classification systems. The advantages and disadvantages of direct methods are discussed below (Section 3.2.3).

3.2.1 Standard gamble (SG)

The SG asks participants to consider a choice between a gamble and continuation of life in the current health state. The gamble offered is the probability of perfect health (*p*) or certain death (*1-p*). The probabilities for the two options in the gamble are altered until the participant is indifferent between the gamble with the risk of immediate death, and continuation of life in his or her current health state (Torrance 1987; Gold, Siegel et al. 1996; Lenert, Cher et al. 1998). This gamble is diagramed in Figure 1 (Brazier, Deverill et al. 1999). The assumption is that in order to transition into perfect health, people living in poorer health will accept a higher risk of death than individuals living in good health.

Figure 1: Diagram of the standard gamble

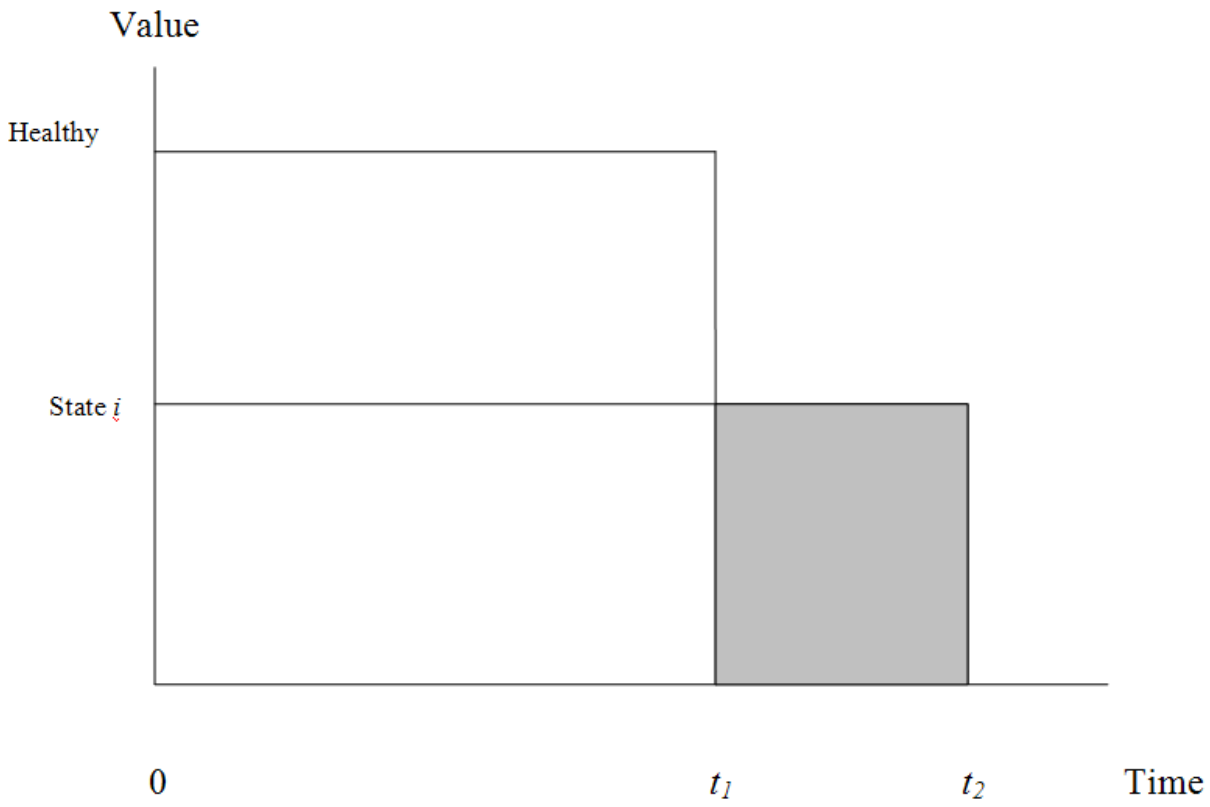


Different procedures are used to determine this indifference point. The simplest procedure asks participants to state how much risk of death would make the two options equally attractive. This can be done in an interview or on paper (Ross, Littenberg et al. 2003). Another set of procedures involves repeatedly presenting two options to patients, adjusting the risk (probability) of death between options until the participant is indifferent between the alternatives. This can be done using computer-based utility assessment software. The risk of immediate death can be adjusted using various methods to present alternative scenarios.

3.2.2 Time trade-off (TTO)

The participants using the TTO are asked to decide how much time in perfect health they would be willing to give up in order to escape their current health. In this case, both alternatives have certain outcomes (Torrance 1987; Green, Brazier et al. 2000). The idea is that in order to escape their current health condition, people living in poor health will accept a shorter life span in perfect health compared to people living in good health (see Figure 2) (Brazier, Deverill et al. 1999).

Figure 2: Diagram of the time trade-off



After finding the number of years in full health an individual is willing to forfeit in order to escape their current health state (or duration of life in perfect health that a participant finds equally attractive to living their full life in their current health) (t_2- t_1), the preference score is taken from the quotient of these two durations [perfect health duration / current health duration (t_1/ t_2); scaled 0-1]. The TTO is administered using the same techniques described above for the SG where alternative years forfeited are offered.

There are paper-and-pencil versions of both the SG and TTO available, but computer-based programs have greater consistency and reliability. See Appendix 2 for a list of available software for direct measurement of health utilities.

3.2.3 Strengths and weaknesses of direct methods

Direct methods to elicit preferences or utilities for health states are rooted in utility theory, providing confidence that the values are an accurate reflection of an individual's trade-off between morbidity and mortality. Direct elicitation is the preferred method if it is difficult to describe the health state (Brazier, Deverill et al. 1999). Direct methods are also preferred if the investigator is uncertain whether all domains that are important to the study are represented in an alternative method (Brazier, Deverill et al. 1999).

A disadvantage of direct methods is that they may not distinguish health-related quality of life from other factors. These other factors include the respondent's feelings about risk and individual valuation of health states. For example, some individuals may be unwilling to consider any risk (in the standard gamble) or give up any life (in the time trade off), no matter how poor the person's health state.

Direct methods may not always yield different scores for changes in health that are regarded as clinically significant. Brazier cites examples of this problem in clinical trials of erythropoietin and hip arthroplasty (Brazier, Deverill et al. 1999).

Direct methods are also complex to administer, burdening study participants and site staff. Studies that use direct elicitation have higher rates of missing data, including both individual items and complete responses. There is also concern that it may be unethical to ask a trial participant who is in frail health to respond to an instrument that requires the respondent to consider her own death.

Finally, direct measures, because of the unrelated variance (noise) associated with direct elicitation, will require an increased sample size compared to other methods in order to reach statistically significance of cost-effectiveness findings. A list of available software to measure utilities with direct methods is available in Appendix 2.

3.2.4 Visual analog scale (VAS)

While the VAS has often been used for direct measurement, concerns about its validity in economic analysis have been raised. Drummond and Brazier do not recommend using the VAS alone in economic evaluation because the method does not give the respondent a choice between two alternatives, and therefore, does not reflect the strength of preference necessary for economic evaluation (Brazier 2005; Drummond, Sculpher et al. 2005). There is also concern that rating scales are particularly subject to a variety of measurement biases. These include end-of-scale bias, where respondents avoid the extremes (0 or 100), and context bias, where respondents distribute responses over the scale or aggregate choices in certain areas of the scale, regardless of the differences in health states. For a further discussion, see (Brazier 2005; Drummond, Sculpher et al. 2005).

3.3 Multi-attribute or indirect methods

Indirect methods use multi-attribute health status classification systems to define preference weights for the various health states experienced by subjects in a trial. Using surveys of a sample of the population and direct methods (SG, TTO or VAS transformed to SG), developers of these systems have estimated preference or utility weights for each defined health state in their system. These surveys elicited the sample's preferences for various *individual attributes* of health. These attributes might include pain, mobility and self-care (see Appendix A for sample questions from each of the surveys discussed below). Preference scores for individual attributes of health have been transformed into a preference weight for each health state or combination of attributes in the system. These preference weights have been integrated into the scoring algorithms in the classification system.

In a cost-effectiveness analysis using an indirect method conducted alongside a clinical trial, subjects are surveyed with these multi-attribute systems. These surveys define the subject's overall health status along several domains or attributes of health. Each combination of findings defines a health state. These health states are then associated with the preference weights described above. In the cost-effectiveness analysis, the preference weights associated with each health state experienced by the subjects are used to calculate QALYs.

Preference-based multi-attribute classification systems all measure generic health status, but they vary by many factors including attributes, number of attribute levels, the description of the levels, the severity of the most severe level defined for each attribute, the number of health states defined, the communities from which the preference weights were estimated, and the theoretical approach to modeling preference data into the scoring formula of the classification system (Drummond, Sculpher et al. 2005). As a result, multi-attribute classification systems are not equally suited for all diseases or disorders.

3.3.1 Strengths and weaknesses of multi-attribute or indirect methods

Multi-attribute systems have less random error (better reliability) than direct elicitation (Brazier, Deverill et al. 1999) and minimize the unwanted variance that is attributable to factors other than the state of health, for example accommodation bias or the unwillingness to consider any risk as mentioned above.

One disadvantage of the multi-attribute systems is that they may not have sufficient response validity to capture the health effect of the intervention of interest. Additionally, the health state of interest or consequences of an intervention may not be adequately described by available multi-attribute systems. Both of these potential limitations must be considered when choosing an indirect or multi-attribute method. See Section 5 for further discussion. The five most commonly used multi-attribute health status classification systems are described below.

3.3.2 Health Utilities Index (HUI)

The HUI survey tools estimate the health utility of a patient's current health state by surveying the respondent in several domains of health. The preference weights for the health states in the HUI were defined by a large, Canadian, community sample in which participants rated hypothetical health states using the VAS. The VAS scores were transformed into SG scores, defining preference weights extrapolated from these health states based on multi-attribute utility theory. Health utility scores for individual health domains are also computed.

There are three versions of the HUI (HUI I, Mark2, and Mark3); the Mark2 and Mark3 have replaced the HUI I in most applications. The HUI Mark3 includes eight domains of health: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain (Horsman, Furlong et al. 2003). The HUI Mark2 consists of seven domains: sensation, mobility, emotion, cognition, self-care, pain, and fertility (Torrance, Furlong et al. 1995). The HUI is copyrighted and can be obtained, for a fee, from Health Utilities Inc (www.healthutilities.com). The HUI questionnaires consist of 15 or 40 questions, which can be used to derive both HUI Mark 2 and HUI Mark 3 comprehensive health utility scores (there is an additional question in each survey which is included for data collection purposes, but not included in the scoring algorithms). The HUI can

be self- or interviewer administered. Depending on the version, patients can be asked to consider their health over the past two, three, or four weeks. The HUI Mark3 has defined 972,000 health states.

3.3.3 EuroQoL (EQ-5D)

The EQ-5D elicits a participant's description of his current health state by scoring items in five domains of health, placing the combined score in one of 243 possible health states. Each item assesses a different health domain, including mobility, self-care, usual activity, pain/discomfort, or anxiety/depression (Kind 1996; Kind, Dolan et al. 1998). Original preference weights were estimated for each health state based on a large, British, community sample in which participants rated hypothetical EQ-5D health states using the TTO (Dolan, Gudex et al. 1996). Preference weights from a US population sample are now available. A VAS is included in the EQ-5D questionnaire packet. The EQ-5D asks participants to consider their health during the day of the interview. Non-profit research groups can obtain the EQ-5D for free from the EuroQol Group (www.euroqol.org).

3.3.4 Quality of Well-Being Scale (QWB)

The original QWB is an interviewer-administered questionnaire with three functioning subscales (mobility, physical activity, and social activity) and additional items concerning symptoms or problems (Kaplan, Bush et al. 1976; Kaplan, Bush et al. 1979). More recently, a self-administered version has been developed, the Quality of Well-Being Scale, Self Administered (QWB-SA), which is a 76-item questionnaire with four subscales: symptom/problem complex, mobility, physical activity, and social activity (Kaplan, Sieber et al. 1997; Kaplan, Ganiats et al. 1998). The QWB-SA includes 1,215 health states. Preference weights were estimated from a survey of a sample of primary care patients in San Diego, CA, who rated hypothetical health states using the VAS. The QWB-SA asks participants to consider their health over the previous three days.

3.3.5 Short Form 6D (SF-6D)

The SF-6D (Brazier, Roberts et al. 2002) derives a participant's perceived health state from the SF-36, a popular health status measure (Ware and Sherbourne 1992; McHorney, Ware et al. 1993; Gandek, Ware et al. 1998), or the related SF-12 health status measure (Ware, Kosinski et al. 1996; Brazier, Roberts et al. 2002). When based on the SF-36, the SF-6D uses 10 items to create six health dimensions (physical functioning, role limitations, social functioning, pain, mental health, and vitality). When based on the SF-12, the SF-6D uses only seven items to create the same six health dimensions. The SF-6D's comprehensive preference weight estimate is based on a study of 836 members of the general public in Great Britain. The SF-6D derives preference weights for 18,000 health states. A one or four week recall period can be used with the SF-36 and SF-12. The SF-36 and SF-12 are copyrighted and can be obtained from www.sf-36.org. The scoring algorithm for the SF-6D is copyright-protected, but can be obtained from its developers at the University of Sheffield, free of charge for non-commercial use.

3.3.6 Summarized characteristics of multi-attribute health status classification systems

In the following tables, Tables 2, 3, 4 and 5, we have summarized key characteristics of four of the five most commonly used multi-attribute health status classification systems. We have also included the citations from which we extracted this information. In Table 2, we have listed key differentiating characteristics, including the domains and levels of domains, number of health states defined, and valuation techniques. You will see from this summary that comparison of these classification systems is complex. The systems include six to eight domains which are not constant across systems, and each domain has two to six different levels or classifications within the domain. The systems define 243 – 972,000 health states and are available in a variety of formats and languages. All the population samples from which the preference or utility weights were derived were from the general population (weights for the HUI2, which is not listed, were measured on a random sample of parents of school-aged children in Hamilton, Ontario, Canada) and these sample ranged in size from 504 to 3395 (Drummond, Sculpher et al. 2005). These populations were from the US, the United Kingdom, and Canada, although two of the systems have additionally elicited preference weights from several population samples beyond the original development sample. Two of the systems allow scores less than zero (i.e. worse than death) and the other two do not. Only one of the systems (HUI) claims to include interactions between levels and domains.

"Floor effects" refers to the limited ability of the system to differentiate between expected low-value or poor health states, while "ceiling effects" refers to the limited ability of the classification system to differentiate between high-value or good health states. Please note that for the HUI, the literature reports conflicting findings, e.g. that there both are (yes) and are not (no) ceiling effects. These conflicting findings might reflect the limited familiarity that non-patient survey respondents have with the conditions and symptoms being studied.

The terms "clinically important difference" or "minimally clinically important difference" refer to the smallest change in preference weight or utility score associated with a clinically meaningful change in health status. This table also includes available recall periods, score ranges, formats and languages.

Table 3 provides reported test-retest reliability – the ability to produce consistent results on different occasions given no evidence of change – for each classification system by medical condition (Bowling 2001). Correlations vary by instrument, by medical condition, and by the period of time between tests. A high correlation (0.85 or greater) indicates that an instrument has an acceptably low level of random measurement error. Note, however, that low correlations – particularly those with greater than two week intervals between tests – may reflect an actual change in health status, rather than poor reliability (McDowell and Newell 1996). This table is an update to Brazier's 1999 and 2005 multi-attribute instrument reviews (http://www.shef.ac.uk/scharr/sections/heds/discussion.html). We searched Medline for "Health Utilities Index" OR "HUI3", "Quality of Well-Being" OR "QWB", "EuroQol" OR "EQ-5D", and "Short-form 6D" OR "SF-6D" (July 24, 2006), and recorded clinical categories and conditions.

Table 4 provides a summary of publications by classification system, category, and numerically referenced conditions. Table 5 defines each referenced clinical condition. Medical conditions are

categorically organized - similar to the system used by the Center for the Evaluation of Value and Risk in Health (CEVR) at Tufts University-New England Medical Center (http://www.tufts-nemc.org/cearegistry/data/default.asp). Tables 4 and 5 are useful in determining which multi-attribute instruments have been used to measure preference-based quality of life in particular populations. For example, the EQ-5D is the most widely-used instrument to measure preference-based quality of life in the oncology literature.

Table 2: Summary of attributes in multi-attribute health status classification systems*

	HUI3	**QWB**	**EQ-5D**	**SF-6D**
	(Feeny, Furlong et al. 2002; HUInc 2004; Feeny 2005)	(Brazier 2005) (Sieber, Groessl et al. 2004) (Coons, Rao et al. 2000)	(EuroQoLGroup 2005; van Stel and Buskens 2006) (Brazier 2005) (Drummond, Sculpher et al. 2005) (Sieber, Groessl et al. 2004) (Conner-Spady and Suarez-Almazor 2003)	(MedicalOutcomesTrust 2006; van Stel and Buskens 2006) (Brazier 2005) (Drummond, Sculpher et al. 2005) (Conner-Spady and Suarez-Almazor 2003) (Brazier, Roberts et al. 2002)
Domains	Vision, hearing, speech, ambulation, dexterity, emotion, cognition, pain	Mobility, physical activity, social activity, symptom problem complex	Mobility, self-care, usual activity, pain/ discomfort, anxiety/ depression	Physical function, role limitation, social function, pain, mental health, vitality
Number of levels per domain	5-6	2-3	3	4-6
Number of health states	972,000	1170	243	18,000
Interactions	Yes	No	No	No
Valuation techniques	VAS transformed into SG	VAS	TTO	SG
Methods of extrapolation	Multi-attribute utility theory (MAUT)	Statistical	Statistical	Statistical
Size of sample from which preference weights were developed	504 (general population)	866 QWB, 430 QWB-SA (general	3395 (general population)	611 (general population)

* Health Utilities Index Mark III (HUI3); Quality of Well-Being (QWB); EuroQoL (EQ-5D); Short-Form 6D (SF-6D); visual analogue scale (VAS); standard gamble (SG); time trade-off (TTO)

		population)		
Origin of preference weights	Canada (Hamilton)	USA (San Diego)	UK	UK
Additional countries from which preference weights were developed	Austria, France, Japan, the Netherlands, Singapore, UK	-	Belgium, Denmark, Finland, Germany, Japan, New Zealand, Slovenia, Spain, US, Zimbabwe	-
Recall periods	1, 2 or 4 weeks, or usual health	3-6 days	today	1 or 4 weeks
Score ranges	-.36 – 1.00	.33 – 1.00 QWB .09 – 1.00 QWB-SA	-.59 – 1.00	.30 – 1.00
Formats available	Self, interviewer, proxy, web	Self, interviewer, proxy[†], phone	Self, proxy, phone	Self
Languages available	15 (7 in development)	9	> 70	> 50 (SF-36)
Floor effects	**No** (McDonough, Grove et al. 2005)[‡]	**No** (McDonough, Grove et al. 2005) (Andresen, Rothenberg et al. 1998)	**No** (Riazi, Cano et al. 2006) (McDonough, Grove et al. 2005)	**Yes** (Brazier, Roberts et al. 2004) **No** (McDonough, Grove et al. 2005)
Ceiling effects	**Yes** (Sung, Greenberg et al. 2003) (Feeny, Torrance et al. 1996) **No** (Naglie, Tomlinson et al. 2006) (McDonough, Grove et al. 2005)	**No** (Naglie, Tomlinson et al. 2006) (McDonough, Grove et al. 2005) (Andresen, Rothenberg et al. 1998) (Kaplan, Ganiats et al. 1998) (Fryback, Lawrence et	**Yes** (Hinz, Klaiberg et al. 2006) (Naglie, Tomlinson et al. 2006) (Schweikert, Hahmann et al. 2006) (Shaw, Johnson et al. 2005) (Wang, Kindig et al. 2005) (Brazier, Roberts et al.	**No** (Brazier, Roberts et al. 2004; McDonough, Grove et al. 2005)

† Sieber (2004) notes that the proxy method is not recommended. If used, results are to be interpreted with caution.

‡ McDonough, Grove et al. (2005) state "We did not find evidence of a floor effect for index values, but noted that each instrument (HUI3, EQ-5D, SF-6D, QWB) had large proportions of participants at the floor for either pain or physical function".

		al. 1997) (Andresen, Patrick et al. 1995)	2004) (Kaarlola, Pettila et al. 2004) (Oga, Nishimura et al. 2003) (Poissant, Mayo et al. 2003) (Konig, Ulshofer et al. 2002) (Wu, Jacobson et al. 2002) (Vitale, Levy et al. 2001) (Johnson and Pickard 2000) (Badia, Schiaffino et al. 1998) (Johnson and Coons 1998) **No** (Riazi, Cano et al. 2006) (McDonough, Grove et al. 2005) (Brazier, Walters et al. 1996)	
Clinically important difference (CID)[§]	0.06, 0.07 (Marra, Woolcott et al. 2005) 0.05[**] (Horsman, Furlong et al.	0.031 (Kupferberg, Kaplan et al. 2005) 0.03 (Kaplan 2005)	0.074[‡‡] (Walters and Brazier 2005) 0.07 (Lubetkin, 2005 #6)	0.041[§§] (Walters and Brazier 2005) 0.033[***] (Walters and Brazier 2003)

[§] Also known as minimum/minimal/minimally important difference (MID) and minimum/minimal/minimally clinically important difference (MCID).
[**] Clinically important difference in HUI3 single-attribute utility scores.
[‡‡] Mean estimate derived from 8 studies (range: 0.011 to 0.140).
[§§] Mean estimate derived from 8 studies (range: -0.011 to 0.097).
[***] Weighted mean estimate derived from 7 studies (range: 0.010 to 0.048).

	2003)	0.05	0.03
	0.03††	(Marra, Woolcott et al. 2005)	(Marra, Woolcott et al. 2005)
	(Horsman, Furlong et al. 2003)	0.033	
	(Drummond 2001)	(Sullivan, Lawrence et al. 2005)	
	(Grootendorst, Feeny et al. 2000)		

†† Clinically important difference in overall HUI3 scores.

Table 3: Test-retest reliability* of multi-attribute health status classification systems by clinical condition and interval

Clinical Condition	Interval	HUI-3	QWB	EQ-5D	SF-6D
Ankylosing spondylitis (Haywood, Garratt et al. 2002)	2 weeks	-	-	0.83	-
Alzheimer's disease (Naglie, Tomlinson et al. 2006)	2 weeks	0.47	0.70	0.79	-
Breast hypertrophy – following reduction surgery (Thoma, Sprague et al. 2005)	~ 1 week	0.84	-	-	-
Breast hypertrophy – pre and post-operative (Kerrigan, Collins et al. 2000)	2-3 weeks	-	-	0.78	-
Burns (Anderson, Kaplan et al. 1989)	1 day	-	0.83	-	-
COPD (Anderson, Kaplan et al. 1989)					
	1 day	-	0.95	-	-
(Stavem 1999)	2 weeks	-	-	0.73	-
Dementia (Ankri, Beaufils et al. 2003)	3 days	-	-	0.74	-
Diabetes (Anderson, Kaplan et al. 1989)	1 day	-	0.94	-	-

* As measured by the intraclass correlation coefficient (ICC) unless noted otherwise.

Clinical Condition	Interval	HUI-3	QWB	EQ-5D	SF-6D
Epilepsy (Stavem, Bjornaes et al. 2001)	2 weeks	-	-	0.93	-
(Wiebe, Eliasziw et al. 2001)	12 weeks	0.71	-	-	-
Hip fractures – women at high risk of (Salkeld, Cameron et al. 2000)	3 weeks	-	-	0.61, 0.73, 0.88-	-
Hip fractures - recovering from (Jones, Feeny et al. 2005)	12 weeks	0.72	-	-	-
HIV/AIDS (Stavem, Froland et al. 2005)	2 weeks	-	-	0.78	0.94
Multiple sclerosis (Fisk, Brown et al. 2005)	2 weeks	0.87	-	0.81	0.83
Rheumatic disease (Luo, Chew et al. 2003)	1 week	0.75	-	0.64	-
Rheumatoid arthritis					
(Hurst, Jobanputra et al. 1994)	2 weeks	-	-	0.78	-
(Marra, Rashidi et al. 2005)	5 weeks	0.81	-	0.46	0.89
(Hurst, Jobanputra et al. 1994)	12 weeks	-	-	0.73	-
Stroke (Dorman, Slattery et al. 1998)	3 weeks	-	-	0.83	-
Miscellaneous					
Canadian general population (Boyle, Furlong et al. 1995)	4 weeks	0.77	-	-	-
	24 weeks	-	-		-
UK elderly women					

Clinical Condition	Interval	HUI-3	QWB	EQ-5D	SF-6D
(Brazier, Walters et al. 1996)				0.67†	

† Spearman rank correlation coefficient.

Table 4: Use of multi-attribute health status classification systems by clinical category and condition*

Clinical Category	HUI-3	QWB	EQ-5D	SF-6D†
1. Infectious and parasitic	1.1, 1.2	1.1	1.1	1.1
2. Neoplasms	2.1, 2.2, 2.6	2.1, 2.2, 2.3, 2.5, 2.6	2.1, 2.2, 2.3, 2.4, 2.5, 2.6	-
3. Endocrine, Nutritional, Metabolic, Immunity	3.1, 3.2, 3.3	3.1, 3.2	3.1, 3.2, 3.3	3.1, 3.2
4. Blood and Blood-Forming Organs	4.1	-	4.1	-
5. Mental	5.1, 5.2,	5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7	5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7	5.3
6. Nervous System and Sense Organs	6.1, 6.2, 6.3, 6.5	6.1, 6.2, 6.3, 6.4, 6.5	6.1, 6.2, 6.3, 6.4, 6.5	6.1, 6.2, 6.4
7. Circulatory System	7.1, 7.5, 7.6, 7.7, 7.10, 7.11	7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, 7.10, 7.12	7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, 7.8, 7.9, 7.10, 7.11, 7.12	7.6, 7.10, 7.12
8. Respiratory System	8.2, 8.3, 8.4, 8.6	8.1, 8.2, 8.3, 8.4, 8.5	8.2, 8.3, 8.4, 8.6	8.2, 8.4
9. Digestive System	9.2, 9.5, 9.6	-	9.2, 9.3, 9.4, 9.5	9.2, 9.5
10. Genitourinary System	10.2, 10.4	10.1, 10.2, 10.3, 10.4	10.1, 10.2, 10.3, 10.4	-
11. Pregnancy and Childbirth	-	11.1, 11.2, 11.3	11.1, 11.2, 11.3	-
12. Skin and Subcutaneous Tissue	12.3	-	12.1, 12.2, 12.3	-
13. Musculoskeletal System and Connective Tissue	13.2, 13.3, 13.4, 13.6, 13.7	13.2, 13.3, 13.4, 13.5, 13.6, 13.7	13.1, 13.2, 13.3, 13.4, 13.5, 13.6, 13.7, 13.8	13.1, 13.2, 13.4, 13.6, 13.8
14. Injury, Trauma, and Poisoning	14.2	14.1, 14.2	14.1, 14.2	-
15. Other	15.1	-	15.1	-

* A literature search by clinical category and instrument yielded too many citations to include in the text. However, a search by clinical category/condition (see Table 5) and instrument ("Health Utilities Index" OR "HUI3", "Quality of Well-Being" OR "QWB", "EuroQol" OR "EQ-5D", and "Short-form 6D" OR "SF-6D") would identify the studies on the topics noted here.

† Note that SF-6D preferences can be applied to any SF-36 dataset for purposes of economic evaluation. As of July 2006, the SF-36 has been documented in > 4,000 publications. Additional information available at http://www.sf-36.org/tools/sf36.shtml (accessed July 24, 2006).

Table 5: Breakdown of clinical conditions - key for table 4

1. Infectious and Parasitic Diseases
1.1 HIV/AIDS
1.2 Other (bacterial meningitis, herpes)

2. Neoplasms
2.1 Genitourinary cancers
2.2 Breast cancer
2.3 Lung cancer
2.4 Melanoma
2.5 Liver cancer
2.6 Other (colorectal cancer, hematopoietic stem cell transplantation, lymphoma, leukemia, chemotherapy)

3. Endocrine, Nutritional, and Metabolic Diseases and Immunity Disorders
3.1 Diabetes
3.2 Obesity
3.3 Other (hypopituitary, lupus)

4. Diseases of the Blood and Blood-Forming Organs
4.1 Hemophilia

5. Mental Disorders
5.1 Depression
5.2 Schizophrenia
5.3 Anxiety
5.4 Substance abuse
5.5 Bipolar disorder
5.6 Post-Traumatic Stress Disorder
5.7 Other (social phobia, panic disorder, attention-deficit hyperactivity disorder)

6. Diseases of the Nervous System and Sense Organs
6.1 Eyes
6.2 Ears and hearing
6.3 Dementia (Alzheimer's disease)
6.4 Multiple sclerosis
6.5 Other (ALS, fibromyalgia, migraine, ataxia, muscular dystrophy, Parkinson's disease, epilepsy, cerebral palsy)

7. Diseases of the Circulatory System
7.1 Angina
7.2 Cardiac Arrhythmia
7.3 Congestive Heart Failure
7.4 Hypertension
7.5 Myocardial Infarction
7.6 Stroke
7.7 Peripheral Vascular Disease (intermittent claudication)
7.8 Valvular heart Disease/Thromboembolism
7.9 Anticoagulation
7.10 Cardiac/Cerebral Investigations & Procedures (treatment of coronary artery disease)

Table 5: Breakdown of clinical conditions - key for table 4

8. Diseases of the Respiratory System
8.1 Acute Respiratory Failure
8.2 Chronic Respiratory Disease
8.3 Cystic Fibrosis
8.4 Obstructive Airways Disease/Sleep Apnea
8.5 Lung Transplant/End Stage Pulmonary Disease
8.6 Other (home mechanical ventilation, lung volume reduction surgery)

9. Diseases of the Digestive System
9.1 Gastrointestinal Bleeding
9.2 Hepatitis
9.3 Cirrhosis + Sequelae
9.4 Nausea, Vomiting and Bowel
9.5 Gastrointestinal Procedures and Surgeries
9.6 Other (dyspepsia)

10. Diseases of the Genitourinary System
10.1 Renal failure
10.2 Kidney dialysis
10.3 Kidney Transplant
10.4 Other Genito-Urinary (cystitis, urinary incontinence)

11. Pregnancy and Childbirth
11.1 Hysterectomy
11.2 Fertility
11.3 Other (neonatal circumcision, pain during pregnancy)

12. Diseases of the Skin and Subcutaneous Tissue (Non-Cancer)
12.1 Psoriasis
12.2 Acne
12.3 Other (shingles, dermatitis)

13. Diseases of the Musculoskeletal System and Connective Tissue
13.1 Upper Extremity
13.2 Hip
13.3 Knee and Foot
13.4 Spine
13.5 General Muscoloskeletal
13.6 Arthritis
13.7 Osteoporosis
13.8 Other (ankylosing spondylitis, acupuncture for low back pain, amputation)

14. Injury, Trauma, and Poisoning
14.1 Brain trauma injury (BTI)
14.2 Other

15. Other
15.1 Other (breast hypertrophy, mennorhagia, hydrocephalus)

3.4 Disease-specific health status measures in economic evaluation.

In some research, generic preference-based health status classification systems may not be sensitive or responsive enough to measure changes in the health related quality of life of interest for a particular economic evaluation. In a study of an intervention affecting sequelae of a disease or condition (e.g., urinary frequency or nausea and appetite), changes in those sequelae might be the outcome of interest but are not captured in a generic measure. For this reason, disease-specific health status measures might be collected during the trial and mapped or transformed into preference weights for estimation of QALYs for the economic evaluation. Note that, because these health status measures have not been developed for use in economic evaluation, a number of concerns about the methods of transformation have been expressed. For further description, see (Brazier, Deverill et al. 1999). One technique is to have the study participants define the health states relevant for the study with a disease-specific measure. Then, using a direct elicitation technique (SG or TTO) with a community sample, determine the preference weights for these health states. The preference weights from the community sample are then used in the estimation of the QALYs. A variant of this method is that used by Mary Goldstein et al. in the FLAIR project, which estimated preferences for health states based on the Katz index of activities of daily living (ADLs) and a population sample of older adults. For more information on this methodology see Goldstein (Goldstein, Michelson et al. 1993; Goldstein, Clarke et al. 1994) and Sims, Garber, et al.(Goldstein 2002; Sims, Garber et al. 2005). For more information on the several disease-specific measures see the Medical Outcomes Trust's descriptions of various instruments: http://www.outcomes-trust.org/instruments.htm.

3.5 Other measures

We have followed the recommendations of the U.K. Health Technology Program in excluding a number of measures from further consideration (Brazier, Deverill et al. 1999). We did not include the Rosser disability/distress scale or the 15D. Neither the Rosser nor 15D measures have received much use (Brazier and Roberts 2004). We have also excluded the visual analog scale (VAS), magnitude estimation, or person trade-off as possible measures to value health states. Because it is so simple, the visual analog scale has been widely used, but there is no justification for regarding the scale as a measure of preference or utility (see Section 3.2.4). Magnitude estimation and person trade-off are not widely used.

4. Summary of utility/preference measurement in VA clinical trials

Economic evaluations have been conducted alongside VA CSP clinical trials since 1986. In these studies, outcome and cost data collection protocols are developed in the planning phase of the trial, and the subsequent economic evaluations are integrated into the primary or secondary analysis plans. As of March 2007, economic evaluations have been included in 16 trials. Three are complete, 12 are ongoing, and one is currently organizational, pre-kickoff, or awaiting funding.

The following is a brief summary of VA experience estimating preference weights for use in cost effectiveness analysis alongside CSP trials. The information was provided by the principal

investigators, research assistants, and project coordinators for each study in early 2006. Several tables follow this section with more in-depth information.

4.1 Selection of preference measurement tools in CSP trials

Cost-effectiveness analyses, including estimation of QALYs, have been conducted in conjunction with VA clinical trials since 1986. Before 2001, researchers either transformed general or disease-specific health status measures, or used direct elicitation techniques to estimate utility or preference weights for economic evaluations. The decision to use a particular method was based on its use in prior studies of the same condition or the lack of sensitivity of other methods to detect changes in the outcome of interest. After 2001, studies have primarily employed multi-attribute health status classification systems, sometimes alongside direct elicitation techniques. Economists' describe their reasons for choosing a particular method to include:

1) Literature supports use of a particular tool with a particular disease or condition

2) Patient burden (simplicity for patient)

3) Tool does (or does not) include questions specific to our outcome of interest (e.g. dexterity or activities of daily living)

4) Community sample used to establish preference weights in the classification system most closely resembles study population (e.g. US community sample and US study population; European community sample and European study population)

5) Interest in comparing sensitivity of instruments (when two or more tools are used alongside a trial)

6) Feasibility

4.2 Strengths and weaknesses or difficulties administering the tools

Staff report that, in general, patients have difficulty with direct elicitation techniques, both with the concepts in the SG and TTO and in understanding the survey questions. Additionally, staff indicate that patients express frustration with the burden of many questionnaires and frequently repeated measurement: staff have observed patients skipping questions or repeating keystrokes in order to finish quickly. Staff have also found the software difficult to set up and maintain. Significantly, in one study, computer pentablets/laptops had to be repaired or replaced much sooner than expected and prior to the end of the study.

Staff and patient experiences suggest that minimal time is necessary to administer and complete the multi-attribute surveys. Except in one study (CSP530) where patients are critically ill and proxies have been used to complete surveys, the reported completion rate is very high. Please see Tables 6, 7 and 8 for more detail on VA experience estimating preference weights alongside CSP trials.

Table 6. Results from PI survey of VA CSP experience with preference measurement – background

Start Year	Study number and name	Economist	Study population (February 2007)	Status of study	Preference measure-ment tools used*	Paper or compu-ter (pro-gram used)	Frequency of collection	Reasons for selecting this tool	Other HRQoL collected
1986	246 A Randomized Study Of Prostatic Surgery For Moderately Symptomatic Benign Prostatic Hyperplasia In Elderly Men	Denise Hynes	561 elderly male veterans with moderately symptomatic benign prostatic hyperplasia	Completed	Urinary bother score (transformed)†	Paper	Unknown	Tool selected by previous economist – reasons unknown.	-
1994	6 Effectiveness of Geriatric Evaluation and Management (GEM) Units and Geriatric Evaluation and Management Clinic (GEMC) Follow-Up	Ciaran Phibbs	1388 hospitalized, frail patients > age 65, following stabilization of acute illness	Completed	Activities of Daily Living (ADL) (transformed)‡	Paper	-	-	Instrumental Activities of daily living (IADL), SF-36

* **SG**: standard gamble; **TTO**: time trade-off; **QWB**: Quality of Well-Being; **VAS**: visual analog scale; **EQ-5D**: EuroQoL; **HUI**: Health Utilities Index

† Transformation method: Investigator method of mapping to utilities (unknown)

‡ Transformation method: (Goldstein 2002; Sims, Garber et al. 2005)

Start Year	Study number and name	Economist	Study population (February 2007)	Status of study	Preference measure-ment tools used*	Paper or compu-ter (pro-gram used)	Frequency of collection	Reasons for selecting this tool	Other HRQoL collected
1997	430 Reducing the Efficacy-Effectiveness Gap in Bipolar Disorder	Henry Glick	330 patients with bipolar disorder	Completed	EQ-5D, SF-36 transformed to SF-6D	Paper Paper	-	-	-
1998	424 Clinical Outcomes, Revascularization and Aggressive Drug Evaluation (COURAGE)	Paul Barnett	2287 patients with chronic angina pectoris (Canadian Cardiovascular Society (CCS) Class I-III), uncomplicated MI, or asymptomatic (or "silent") myocardial ischemia	Primary analysis & publications	SG	Compu-ter (U-titer pentablet)	Baseline, 3mo, 6mo, 12mo intervals	Needed measure more sensitive to outcomes than currently available multi-attribute utility systems.	SF-36, Seattle Angina Quest. (SAQ)

Start Year	Study number and name	Economist	Study population (February 2007)	Status of study	Preference measure-ment tools used*	Paper or compu-ter (pro-gram used)	Frequency of collection	Reasons for selecting this tool	Other HRQoL collected
1998	456 Tension Free Inguinal Hernia Repair: Comparison Of Open And Laparoscopic Surgical Techniques	Denise Hynes	2,164 men with inguinal hernia	Primary analysis & publications	SF-36 (transformed)§	Paper - interview	Baseline (pre-op), 3mo, 6mo, 12mo, 24mo (post-op)	Prior studies' methods. Add to existing body of SF-36 data. HUI discussed, but decided SF-36 was sufficient. Why? CE was *not* the primary outcome.	-
2001	512 A Tri-National Randomized Controlled Trial to Determine the Optimal Management of Patients with HIV infection - First and Second-Line Active Anti-Retroviral Therapy has Failed	Wei Yu, Doug Owens	368 total (288 VA) patients with advanced HIV disease who have failed conventional HAART	Ongoing – patient follow-up	SG, TTO, VAS, EQ-5D, HUI	Compu-ter (U-titer), Compu-ter (U-titer), Paper, Paper, Paper (all self)	Baseline, 1.5mo, 3mo intervals (all tools)	International trial – each country requested preferred tool with community preferences that closely reflect own study population.	MOS-HIV
2003	474 Radial	Todd	614 VA	Ongoing –	HUI	Paper	Baseline	Includes dexterity	Seattle

§ Transformation method: (Nichol, Sengupta et al. 2001) (HUI2-derived utilities)

Start Year	Study number and name	Economist	Study population (February 2007)	Status of study	Preference measure-ment tools used*	Paper or compu-ter (pro-gram used)	Frequency of collection	Reasons for selecting this tool	Other HRQoL collected
	Artery vs. Saphenous Vein Grafts in Coronary Artery Bypass Surgery	Wagner	patients with coronary artery disease who have agreed to undergo coronary artery bypass surgery	patient accrual			(pre-op), 10days (DC hospital), 12mo, 60mo	and mobility questions. Choice of questions to cover outcomes.	Angina Quest. (SAQ), Arm and leg function, Dynamometer
2003	481 The Home International Normalized Ratio (Kaplan, Anderson et al.) Monitor Study (THINRS)	Ciaran Phibbs	2923 VA patients on warfarin with either atrial fibrillation or a mechanical heart valve	Ongoing - patient follow-up	HUI	Paper	3mo intervals	Includes utilities with valid North American weights. Simple.	-
2003	530 Intensive vs. Conventional Renal Support in Acute Renal Failure	Mark Smith	1023 critically ill patients with acute renal failure	Ongoing - patient accrual	HUI	Paper – interview	2 and 12 mo post-enrollment	Simplicity. Previous use. Lit review.	-
2004	519 Integrating Clinical Practice	Mark Smith	672 smokers undergoing	Ongoing – patient	QWB	Paper – self	Baseline,	Previous smoking	Smoking Cessation

Start Year	Study number and name	Economist	Study population (February 2007)	Status of study	Preference measure-ment tools used*	Paper or compu-ter (pro-gram used)	Frequency of collection	Reasons for selecting this tool	Other HRQoL collected
	Guidelines for Smoking Cessation into Mental Health Care for Veterans with Post Traumatic Stress Disorder		mental health care for PTSD	accrual			18mo, Final visit	studies. Lit review.	QOL scale = (SF-36 + 5 smoking cessation questions)
2005	535 Anabolic Steroid Therapy on Pressure Ulcer Healing in Persons with Spinal Cord Injury	Doug Bradham	407 registered/ 100 randomized spinal cord injury patients with a chronic Stage III or IV pressure ulcer of the pelvic region	Ongoing – patient accrual	HUI	Paper	-	Literature review. Reliability. Most valid across cultures. QWB doesn't quantify as well.	SF-36

Start Year	Study number and name	Economist	Study population (February 2007)	Status of study	Preference measure-ment tools used*	Paper or compu-ter (pro-gram used)	Frequency of collection	Reasons for selecting this tool	Other HRQoL collected
2005	553 Chemotherapy after Prostatectomy (CAP) For High Risk Prostate Carcinoma: A Phase III Randomized Study	Wei Yu	22 patients that underwent prostatectomy (CAP) for high-risk prostate cancer	Ongoing – patient accrual	VAS, EQ-5D	Paper, Paper	3mo, 6mo, 9mo, 12mo, 6mo intervals	EQ-5D replaced lengthy QWB. EQ-5D's simplicity reduced patient burden	FACT-P, Subjective Significance Quest. (SSQ)
2006	555 Impact of Long-Acting Injectible Risperidone on Cost-Effectiveness of Treatment for Veterans with Schizophrenia	Doug Leslie, Paul Barnett	77 veterans with a primary diagnosis of schizophrenia who had at least one psychiatric hospitalization in the past year	Ongoing – patient accrual	QWB	Paper	1.5mo, 3mo intervals	Pyne 2003 MedCare study…found QWB best for schizophrenia. Kay 1987 Schizo Bulletin-PANSS (positive and negative syndrome scale)	Heinrichs-Carpenter Quality of Life Scale, SF-36, Lehman Quality of Life Interview, PANSS

Start Year	Study number and name	Economist	Study population (February 2007)	Status of study	Preference measure-ment tools used*	Paper or compu-ter (pro-gram used)	Frequency of collection	Reasons for selecting this tool	Other HRQoL collected
2006	558 Robotic Assisted Upper-Limb Neuroreha-bilitation in Stroke Patients	Todd Wagner	158 (target) chronic stroke patients with moderate to severe upper extremity impairment	Ongoing – patient accrual	HUI, VAS	-	6mo intervals	Improvement in mobility and dexterity best addressed by HUI.	Fugl-Meyer score of neurological impairment
2006	560 Bronchitis and Emphysema Advice and Training to Reduce Hospitalization (BREATH)	Todd Wagner	24 patients with COPD	Ongoing – patient accrual	EQ-5D	-	-	-	SF-12
2007	551 Rheumatoid arthritis: Comparison of active therapies in patients with active disease despite methotrexate therapy	Ciaran Phibbs	600 (target) rheumatoid arthritis patients with active disease despite treatment with MTX (Disease Activity Score with 28 joints (DAS28) of ≥ to 4.4 units)	Pre-kickoff	VAS, EQ-5D, HUI	Paper, Paper, Paper	Baseline, 6mo, 12mo	Previous use by investigators and in rheumatoid arthritis literature. Using more than one instrument to compare sensitivity of instruments.	-

Table 7: Results from PI and site staff survey of VA CSP experience with preference measurement – problems identified[*]

Study number	Economist	Site Staff (BIO, RA, SC, Other)[†]	Preference measurement tools used[‡]	Problems Identified
246	Denise Hynes	Other	Urinary bother score (transformed)[§]	Other: - First CSP study to incorporate QoL and costs, but primitive. - Predates QoL standardized instruments; standards of CEA have since changed. - UBS developed by Dr. Rutgers – scaled to calculate bother score. Methodology: translated from different disease areas. - Created time trade-off scenarios; however, they didn't work - instruments too complex for patients
6	Ciaran Phibbs	-	Activities of Daily Living (ADL) (transformed)[**]	-
430	Henry Glick	-	EQ-5D, SF-36 transformed to SF-6D	-

[*] CSP studies 535, 553, 551, 555, 558, and 560 were excluded. Their study status was pre-patient accrual at the time of this survey (early 2006).

[†] **BIO**: biostatistician, data manager; **RA**: research coordinator, research assistant, research associate; **SC**: site coordinator, national clinical coordinator.

[‡] **SG**: standard gamble; **TTO**: time trade-off; **QWB**: Quality of Well-Being; **VAS**: visual analog scale; **EQ-5D**: EuroQoL; **HUI**: Health Utilities Index

[§] Transformation method: Investigator method of mapping to utilities (unknown)

[**] Transformation method: (Goldstein 2002; Sims, Garber et al. 2005) (SG-derived utilities)

Study number	Economist	Site Staff (BIO, RA, SC, Other)†	Preference measurement tools used‡	Problems Identified
424	Paul Barnett	RA SC1, SC2	SG	Economist: - Pentablets did not last as long as was budgeted. Not durable. RA: - SG gives everyone the most trouble. - It is the hardest for elderly cardiac patients who can't relate to going blind/other questions. - Out of 48 sites, only 5 are using the pentablet/online SG. - Most often, coordinators will use paper forms and never submit the SG. SC1: - (When the study) first started, the average person thought it was a horrible questionnaire. Hated it. - Poor wording. Questions impossible to answer. - After some time, patients would just select "yes, yes, yes" or "no, no, no". SC2: - Pentablets no longer hold charge.
456	Denise Hynes	SC	SF-36 (transformed)††	SC: - Dependent on how well the patients could read. - Questions somewhat redundant.

†† Transformation method: (Nichol, Sengupta et al. 2001) (HUI2-derived utilities)

Study number	Economist	Site Staff (BIO, RA, SC, Other)†	Preference measurement tools used‡	Problems Identified
512	Wei Yu	RA	SG,	Economist:
	Doug Owens	SC1, SC2	TTO,	- Laptop lifespan short ~ 3 yrs. Need hardware/software support…lots of problems with software. Lengthy set-up time at beginning of study.
			VAS,	RA:
			EQ-5D,	- Software problems – either not recording data or overwriting data. - Laptop CMOS battery problem (date/time not saved). Affects all laptops. - Data sent via floppy – periodically faulty floppies, data security issues. - Lengthy patient accrual period – numerous coordinator switches at sites. New coordinators require training. Laptops sometimes go missing for a period of time. One laptop stolen.
			HUI	SC1: - VAS throws pts for a loop. They don't understand that the "square is them". - Some (patients) are computer savvy, but some are not comfortable with taking the (SG/TTO) survey at all.
				SC2: - HUI Q16 tends to throw them off - vertical instead of horizontal layout - often gets skipped. - Some have learned to "game" the (TTO/SG) survey; others struggle (suspects literacy issues). Long timers in study have got the survey "down to 4 key strokes" - trying to get through it as quickly as possible - accuracy/validity is in question. - Has patient with retinitis - blindness TTO question - doesn't know how to answer - throws pts off. - Multiple surveys are a problem. Patients often don't understand the difference between them. - Patients tired of TTO/SG survey "sick of it" - refusing - insisting that their health hasn't changed. Upset about having to give up this or that. Again, specific only to computer-based survey (not VAS, EQ-5D, HUI) - pts are "irritated" and "emotive".

Study number	Economist	Site Staff (BIO, RA, SC, Other)†	Preference measurement tools used‡	Problems Identified
474	Todd Wagner	BIO SC	HUI	BIO: - Patients often forgot to complete the survey or they refused (most HUIs administered by mail). SC: - Most patients are aged 50+ - they don't understand the questions. Unsure of how to answer questions. Patients have never described this to a doctor. - Pressure from spouse to record a different answer.
481	Ciaran Phibbs	SC	HUI	Economist: - Some patients do not receive a follow-up visit; therefore, they are not completing the HUI. SC: - (Patients) complain more and more every time they have to fill it out.
530	Mark Smith	SC1, SC2	HUI	SC1: - HUI is conducted via phone - often need to repeat questions. - Some very ill patients don't understand the HUI. - Some patients are uncomfortable with the HUI. SC2: - Patients stumble on question #19 (getting around) - skip pattern is confusing, irritating, and frustrating to patients.

Study number	Economist	Site Staff (BIO, RA, SC, Other)†	Preference measurement tools used‡	Problems Identified
519	Mark Smith	SC	QWB	Economist: - Patients don't know some terms on the form. - Patients do not understand that they can choose from multiple answers (despite explanation provided on the form and given by the coordinators). SC: - Compared to other forms, need to spend some extra time explaining QWB. - Patients find QWB more complicated.

Table 8: Results from site staff survey of VA CSP experience with preference measurement – administration and implementation issues

			Patient Experience *Estimated average…*		Site Staff Experience *Estimated average number of minutes…*						
Study number [*]	**Preference measure-ment tools used** [†]	**Site Staff (BIO, RA, SC, Other)** [‡]	**Minutes to completion**	**Complexity** [§]	**Discussing survey with patient/ prepping to take survey**	**Transmitting survey to coordinating center**	**Responding to coordina-ting center queries**	**Surveys Received** [**] ---------- **Surveys Expected** [††]	**% Surveys Received of Expected**	**Surveys Complete** [‡‡] ---------- **Surveys Received**	**% Surveys Complete of Received**
246	Urinary bother score (trans-formed) [§§]	-	-	-	-	-	-	-	-	-	-

[*] CSP studies 535, 553, 551, 555, 558, and 560 were excluded. Their study status was pre-patient accrual at the time of this survey (early 2006).

[†] **SG**: standard gamble; **TTO**: time trade-off; **QWB**: Quality of Well-Being; **VAS**: visual analog scale; **EQ-5D**: EuroQoL; **HUI**: Health Utilities Index

[‡] **BIO**: biostatistician, data manager; **RA**: research coordinator, research assistant, research associate; **SC**: site coordinator, national clinical coordinator.

[§] **Complexity**: very easy, easy, medium, difficult, or very difficult

[**] **Surveys Received**: Number of surveys received from patients (includes surveys with missing responses)

[††] **Surveys Expected**: Number of surveys expected if all patients had completed all surveys at all time points (excluding patients who have died, dropped out, or who have been lost to follow-up)

[‡‡] **Surveys Complete**: Number of surveys that do *not* include missing responses.

[§§] Transformation method: Investigator method of mapping to utilities (unknown)

		Patient Experience *Estimated average…*			Site Staff Experience *Estimated average number of minutes…*						
Study number *	**Preference measure-ment tools used**†	**Site Staff (BIO, RA, SC, Other)** ‡	**Minutes to completion**	**Complexity** §	**Discussing survey with patient/ prepping to take survey**	**Transmitting survey to coordinating center**	**Responding to coordina-ting center queries**	**Surveys Received**** ---------- **Surveys Expected**††	**% Surveys Received of Expected**	**Surveys Complete** ‡‡ **-** ---------- **Surveys Received**	**% Surveys Complete of Received**
6	Activities of Daily Living (ADL) (trans-formed)***	-	-	-	-		-	-	-	-	-
430	EQ-5D, SF-36 transformed to SF-6D	-	-	-	-		-	-	-	-	-
424	SG	BIO	-	-	-	-	-	9820 / 16742	59%	-	-
		SC1	20	Difficult	3	3	0	-	-	-	-
		SC2	5	Difficult	2	1	0	-	-	-	-

*** Transformation method: (Goldstein 2002; Sims, Garber et al. 2005)

			Patient Experience *Estimated average…*		Site Staff Experience *Estimated average number of minutes…*						
Study number [*]	**Preference measure-ment tools used** [†]	**Site Staff (BIO, RA, SC, Other)** [‡]	**Minutes to completion**	**Complexity** [§]	**Discussing survey with patient/ prepping to take survey**	**Transmitting survey to coordinating center**	**Responding to coordina-ting center queries**	**Surveys Received** [**] ---------- **Surveys Expected** [††]	**% Surveys Received of Expected**	**Surveys Complete** [‡‡] ---------- **Surveys Received**	**% Surveys Complete of Received**
456	SF-36 (trans-formed) [†††]	BIO	-	-	-	-	-	5207 / 5888	88%	5207 / 5207	100%
		SC	4	Medium	2	1	2	-	-	-	-
512	SG & TTO,	RA	6	-	-	-	-	2481 / 2884 [‡‡‡]	86%	2351 / 2481	95%
		SC1	8	Difficult	5	5	2	-	-	-	-
		SC2	14	Difficult to Very Difficult	10	5	5	-	-	-	-
	EQ-5D & VAS,	RA	-	-	-	-	-	2693 / 2884	93%	2594 / 2693	96%
		SC1	5	Medium	3	5	0	-	-	-	-
		SC2	3	Very Easy	2	5	0	-	-	-	-

[†††] Transformation method: (Nichol, Sengupta et al. 2001) (HUI2-derived utilities)

[‡‡‡] Of SG & TTO surveys *not* received (2884-2481 = 403), 31 (8%) were marked "too ill", 71 (18%) "refused", 221 (55%) had "missed their visit", and 80 (20%) were marked as "other". "Other" refers to remote follow-up patients (with only minimal data collection), coordinators who forgot or failed to administer the survey, and patients who were running late and did not complete all instruments.

			Patient Experience *Estimated average…*		Site Staff Experience *Estimated average number of minutes…*						
Study number [*]	**Preference measure-ment tools used** [†]	**Site Staff (BIO, RA, SC, Other)** [‡]	**Minutes to completion**	**Complexity** [§]	**Discussing survey with patient/ prepping to take survey**	**Transmitting survey to coordinating center**	**Responding to coordina-ting center queries**	**Surveys Received** [**] ---------- **Surveys Expected** [††]	**% Surveys Received of Expected**	**Surveys Complete** [‡‡] ---------- **Surveys Received**	**% Surveys Complete of Received**
	HUI	RA	-	-	-	-	-	2695/2884	93%	2592/2695	96%
		SC1	7	Medium	3	5	0	-	-	-	-
		SC2	9	Easy	2	5	0	-	-	-	-
474	HUI	BIO	-	-	-	-	-	1080/1087	99%	1078/1080	~100%
		SC1	15	Difficult to Very Difficult	10	5	15	-	-	-	-
		SC2	5	Very Easy	2	1	0	-	-	-	-
481	HUI	BIO	-	-	-	-	-	5754/6216	93%	5655/5754	98%
		SC1	15	Easy	5	5	4	-	-	-	-
		SC2	10	Easy	1	1	0	-	-	-	-
		SC3	10	Easy	2	1	1	-	-	-	-

			Patient Experience *Estimated average…*		Site Staff Experience *Estimated average number of minutes…*						
Study number [*]	**Preference measure-ment tools used** [†]	**Site Staff (BIO, RA, SC, Other)** [‡]	**Minutes to completion**	**Complexity** [§]	**Discussing survey with patient/ prepping to take survey**	**Transmitting survey to coordinating center**	**Responding to coordina-ting center queries**	**Surveys Received** [**] ---------- **Surveys Expected** [††]	**% Surveys Received of Expected**	**Surveys Complete** [‡‡] ---------- **Surveys Received**	**% Surveys Complete of Received**
530	HUI	BIO	-	-	-	-	-	325/354	92%	229/325	70%
		SC1	38	Medium	5	5	0	-	-	-	-
		SC2	13	Easy	7	9	0	-	-	-	-
519	QWB	BIO	-	-	-	-	-	382/382	100%	305/380	80%
		SC1	8	Medium	9	1	0	-	-	-	-
		SC2	10	Easy to Medium	1	3	2	-	-	-	-

5. Selecting a preference assessment method and measure

This section describes the selection of a preference assessment method (off-the-shelf, direct or indirect) and classification system for a clinical trial. The best method and measure will depend on the population in which it is used, the complexity of the intervention, the health states to be measured, and the time frame and resources of the study.

We briefly review the guidance provided in other publications on selecting a preference assessment method and measure. We then describe steps to gather information about the available alternatives and then define a process for selecting a method, considering the alternatives in ascending order of complexity. At each step, we offer criteria for considering whether a more complex method is needed. We then provide recommendations for choosing a measurement system and then summarize recommendations for planning a clinical trial.

5.1 Recommendations from others

The U.S. Panel on Cost-Effectiveness in Health and Medicine has made several recommendations for selecting a preference assessment method (Gold, Siegel et al. 1996). The Panel recommends that the measure include domains important for the problem under consideration and reflect the impact of morbidity on productivity and leisure. The Panel recommends that community utility values be employed in order to adopt the societal perspective, and that these values be drawn from a representative, community-based U.S. sample (for U.S. studies). The panel acknowledged that this is not always possible, and that values of individual patients may be used "as an approximation" (p. 106). The Panel noted that off-the-shelf utility values, determined in another study, might be suitable in certain situations. They also acknowledged that in some studies, a generic measure may be inadequate to measure the health changes of interest for the study. They recommended that in those cases, a disease-specific measure might be used, with preference or utility weights estimated after the study.

The U.K. National Health Service Health Technology Assessment Program commissioned a review of health status measures, including preference assessment methods (Brazier, Deverill et al. 1999). This review represents one of the most comprehensive sources of information on preference assessment, and provides a list of criteria to be used to evaluate preference assessment measures (see Section 3). This review recommended that standard gamble or time trade-off methods be the ultimate source of all health state utility values and that multi-attribute classification system be used in all economic evaluations conducted alongside clinical trials. The HUI and EQ-5D were recommended, while the QWB and SF-6D were not. The QWB was not recommended because at that time, a self-administered version of the survey was not available. The SF-6D was not included because it was a very new method under evaluation and results were not available at the time of the 1999 Brazier report. A more recent comparison of preference-based measures of health by Brazier (Brazier 2005) http://www.shef.ac.uk/content/1/c6/01/87/47/0505FT.pdf includes both of these classification systems in its list of the five (including HUI2 and 3) most frequently used preference-based health status measures.

A task force of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) http://www.ispor.org/ has also made recommendations (Ramsey, Willke et al. 2005). This task force observed that EQ-5D, HUI, and QWB are being used in many clinical trials. It reported that direct elicitation by standard gamble or time trade-off is often difficult and that these direct measures may not be responsive to changes in health status. As a result, this task force recommended that the decision about whether to use a direct elicitation method should be decided on a case-by-case basis.

A key text on cost-effectiveness analysis by Drummond et al. makes several suggestions for selecting a measure for indirect measurement of utilities (Drummond, Sculpher et al. 2005). The authors recommend that the researcher select an established instrument that is not too burdensome. The instrument should measure attributes that are likely to be important in the patient population, and it should be proven to be responsive in that population. These authors suggest that pilot testing may be needed to select an instrument. Some studies may select several instruments, selecting one as primary. Drummond and co-authors suggest that relevance of the community that assigned the utility values not be used as a criterion for selecting a measure as research concludes that different communities assign very similar utility values (Drummond, Sculpher et al. 2005). Brazier, however, disagrees and recommends selecting a measure based on the similarity of the community to the study population (Brazier 2005).

5.2 Review the literature

The first step in selecting the method and measure to estimate preference weights is to complete a review of previous studies of preference measurement in a similar population. The literature review should be comprehensive.

5.2.1 Identify the relevant literature

Researchers should identify relevant studies through a search of the literature and a search of relevant web resources. A catalog of preference weights, sorted by disease area and including citations, is available from the web site of the Center for the Evaluation of Value and Risk in Health at Tufts University-New England Medical Center (http://www.tufts-nemc.org/cearegistry/data/default.asp). This catalogue is maintained by a group led by Peter Neumann and covers the period between 1976 and 2001. A related database of cost-effectiveness ratios that may be helpful in locating preference weights for particular medical conditions is also available from this Tufts website.

A search for the relevant literature should include a search of all appropriate databases including the National Library of Medicine PubMed database, the Science Citation Index Expanded (which can search forward for articles that cite a previous work), and the Cumulative Index to Nursing and Allied Health Literature (CINAHL). Because “quality-of-life” is such a broad topic, Brazier et al. (p. 11) suggest searching for a broad combination of quality of life headings, and then limiting the search to studies that also include headings for cost and economics (Brazier, Deverill et al. 1999).

Table 9 provides some examples of search strategies. Because authors and publications refer to measures using different names, the search should include all possible variants and

abbreviations. Note, however, that the Health Utilities Index is abbreviated as "HUI," and an unrestricted search using this abbreviation will retrieve papers by authors with this common surname, or words that contain these three letters. The search strategy must use limits to avoid retrieving a large number of unrelated papers. A similar strategy is needed to find studies that abbreviate time trade-off as TTO. Studies that report information on more than one measure are especially valuable, as they allow a comparison of the performance of different measures in the same population.

Table 9 gives the number of citations found using each search method, during a search conducted in July 2006. It is also possible to search for papers on one of the preference assessment measures listed in Table 9, restricted to a condition of interest. For example, a PubMed search restricted to standard gamble and diabetes identified seven studies, and a search restricted to SF-6D and diabetes identified six studies.

Table 9: Search strategies for identifying preference-based quality of life literature

Measure	Medline Search Request	Number of citations found in PubMed July 2006
Standard Gamble	"Standard Gamble"	406
Time Trade Off	"Time Trade Off" OR (TTO AND ("Quality of Life" OR QALY OR "Quality Adjusted Life Year" OR Preference OR preferences OR utility OR utilities)	380
EQ-5D	EuroQoL OR "EQ 5D" OR EQ5D	869
QWB	"Quality of Well Being" OR QWB	197
HUI	Health Utilities Index OR "Health Utility Index" OR (HUI NOT HUI[Author] AND ("Quality of Life" OR QALY OR "Quality Adjusted Life Year" OR Preference OR preferences OR utility OR utilities)	348
SF-6D	SF-6D or SF6D	59

5.2.2 Validity, reliability, construct validity, and responsiveness

The literature review should focus on the psychometric properties of these methods and measures, including validity, reliability, construct validity and responsiveness of the system in the population with the medical condition of interest. *Validity* shows the extent to which an instrument measures what was intended (Hays, Anderson et al. 1998). *Descriptive validity* is the ability of the measure to accurately represent the health state of interest. Note that small

variations in the description of a health state can result in large changes in the community valuations of the health state (Brazier, Deverill et al. 1999).

Content validity is the extent to which the measure covers the range of outcomes that is appropriate to the population under study (Hays, Anderson et al. 1998). It indicates whether the domains are relevant to the intent (Scientific Advisory Committee of the Medical Outcomes Trust 2002). Since it is not feasible for the measure to be comprehensive, content validity evaluates whether anything important has been omitted (Brazier, Deverill et al. 1999). The views of patients, clinicians, and the community are assessed to determine this. Evaluations of content validity are often subjective and formal assessment is rare (McDowell and Newell 1996). A related concept is *face validity*, which reflects whether the items in an instrument are logical and appropriate to the health state of the individuals being assessed (Brazier, Deverill et al. 1999).

Reliability is the extent to which a measure yields the same score when the underlying health state remains unchanged (Hays, Anderson et al. 1998). It is an indication of how free the instrument is from random error. Reliability consists of internal consistency and reproducibility.

Responsiveness is the ability of the measure to detect differences in health that are clinically significant (Hays, Anderson et al. 1998). This can be assessed by comparing groups that are known to have a difference in quality of life. It is commonly assumed that the measure that discriminates the largest effect is the best, but economists are also interested in an accurate measure of the value of the difference (Brazier, Deverill et al. 1999).

5.3 Selection of a preference assessment method

We propose that, based on the literature review, the analyst find the simplest, least costly, and most feasible method that will answer the economic hypotheses of the trial, recognizing that there may be a trade-off between validity and expense. The analyst should consider each possible method starting with the simplest and working to the more complex (burdensome and expensive). The methods include:

- Using off-the-shelf utility values
- Assessing patients with an indirect method that includes a generic multi-attribute health status classification system for which preference weights have been established, or a disease-specific quality of life measure for which preference weights will need to be developed, and
- Direct elicitation of preferences from the patients in the study

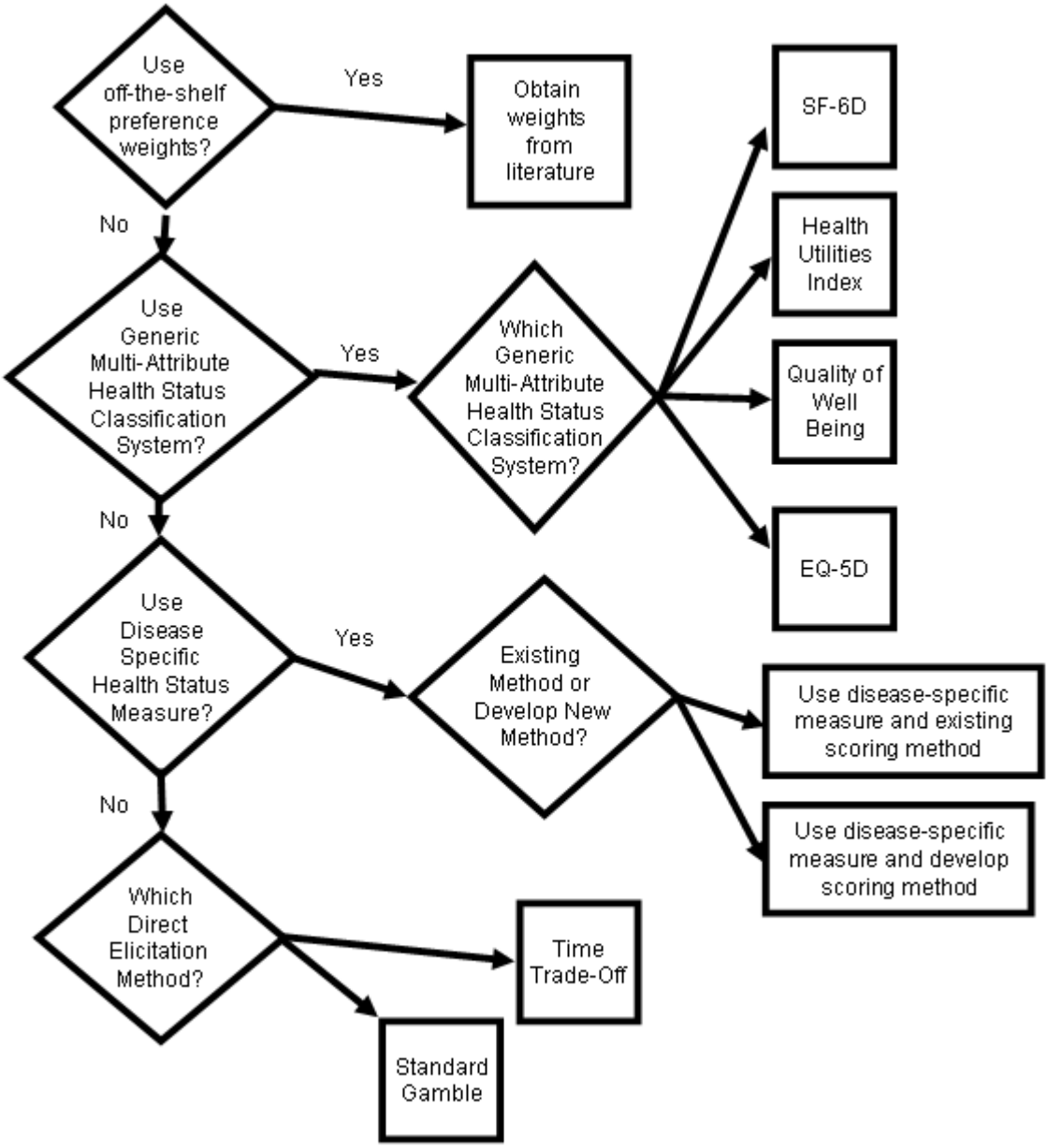
This process is summarized in the flow chart (Figure 3). We describe criteria for judging whether the method is adequate, or whether a more complex method is needed.

The method should capture the most important effects of the intervention. If the intervention is intended to prevent mortality or major disability, an off-the-shelf preference weight or a multi-attribute method may be sufficient. These methods would be able to represent the consequences of a disabling health event. If the intervention has more complex effects in multiple health

domains, however, an off-the-shelf method will likely be insufficient. Consider an intervention that reduces symptoms of a disease, but produces uncomfortable side-effects, or an inconvenient treatment regime. The quality of life and associated health states that might result from this intervention may not be very well represented in a generic multi-attribute measure. In such a case, it may be possible to use a disease-specific measure to characterize the health state, and conduct a separate study outside of the trial to assign values to these health states. This approach would be used if the disease-specific measure had greater discriminant validity than the generic multi-attribute measure (see Section 5.4.8).

In some studies, however, possible health states may involve complex combinations of symptoms, disability, and side effects difficult to describe to community raters. In such cases, a direct method, such as the standard gamble or time trade-off, might be used with trial participants who are well-versed in their current health.

Figure 3: Algorithm for selecting a preference measure for a clinical trial



5.3.1 Off-the-shelf utility or preference weights

Off-the-shelf preference weights from the literature can be applied to known or hypothetical health states. This method is usually restricted to modeling outcomes beyond the duration of a clinical trial. If off-the-shelf weights are not available in the literature or the health states expected during the trial are not known, other methods will be necessary to assess preferences. See cautions in Section 3.1.1 about the influence of elicitation procedures on health state valuations and the precaution against combining values from several studies (Brazier and Deverill 1999; Brazier, Deverill et al. 1999; Chapman, Stone et al. 2000; Jansen, Stiggelbout et al. 2000; Lenert and Kaplan 2000).

5.3.2 Indirect methods

We agree with the ISPOR recommendations (see Section 5.1) that the multi-attribute health status classification systems, including HUI, EQ-5D, QWB, and SF-6D provide the most feasible methods of determining preference weights in the context of a clinical trial, if the measure is responsive in the population of interest. These indirect methods are easier to administer than direct elicitation and they use community values of preferences as recommended by the U.S. Panel on Cost-Effectiveness in Health and Medicine (Gold, Siegel et al. 1996). We include use of disease-specific measures in this category. We address selection of a classification system in the next section.

5.3.3 Direct elicitation of utilities

If off-the-shelf and indirect methods are not appropriate for your study, then utilities may be elicited directly from trial participants using the standard gamble or the time trade-off. The standard gamble is considered the most theoretically valid technique to measure utilities (Torrance 1986). The time trade-off, however, was designed to be less challenging than the standard gamble, so its advantage is some reduction in burden, at the expense of theoretical validity. The literature review should address whether one of the methods has more desirable psychometric properties in the population of interest.

When a direct method is used, it is desirable to gather a disease-specific health status measure at the same time. If the participant fails to complete the direct method, the missing value may be imputed based on response to the disease-specific measure. A logical extension is to combine the measures and develop a preference weight scoring system for the disease-specific measure. This would reduce the contribution of non-health factors to the variance in outcomes.

Direct elicitation in a clinical trial, has participants or study subjects assign utility or preference weights to health states. This, however, is inconsistent with U.S. guidelines for cost-effectiveness analysis, which call for community values. Community values often differ from the values assigned by affected individuals. It is hypothesized that individuals who experience the health state likely accommodate to that state and will assign higher values than members of the community (Slevin, Stubbs et al. 1990; Gold, Siegel et al. 1996). These guidelines however, acknowledge that direct elicitation may be necessary in some cases.

5.4 Criteria for evaluating multi-attribute health status classification systems for preference weight estimation

If the literature review and published psychometric information suggest that the indirect method would be the best method for determining preference weights for a particular study, the researcher will need to choose between several multi-attribute health status classification systems. Researchers will also need to consider the feasibility of using the system, the availability of alternative forms and translations, theoretical justification, and the appropriateness of the population sampled to estimate the preference weights integrated into the specific classification system.

5.4.1 Content and face validity

Both Brazier (2005) and Drummond (2005) suggest that while it will matter which system is chosen, all of the multi-attribute systems described above have demonstrated reliability, feasibility, validity, and responsiveness. Brazier suggests there is no reason to believe these preference-based measures of health status are less stable than non-preference-based and notes that all are moderately correlated at .5 - .6. Both experts suggest that content validity is the primary criteria for choosing between these well-known preference-based health status classification systems. Because these systems vary in the dimensions or attributes they include, the levels and number of levels included in each dimension, the description of the levels, the severity of the most severe level described, as well as the theoretical approach to modeling health status survey results and the community or population surveyed, researchers should carefully confirm that any system accurately depicts the health states likely to be experienced in the study. Drummond suggests a researcher should review the following questions (Drummond, Sculpher et al. 2005):

- Is the instrument seen as credible (is it an established instrument that has demonstrated feasibility, reliability, validity and responsiveness in a number of studies)?
- Does the system cover the attributes and levels of attributes important to the study?
- Has the system been used with a similar population and shown to be responsive?
- Will the system likely be responsive with this population?
- Are there floor or ceiling effects which might affect the sensitivity of this system to changes in this population or this intervention?
- Is a health status worse than death important for this study; does the system include states worse than death?
- Does the planned audience for this study have a preference for a classification system?
- Is the system based on sound theory?
- Is the patient burden acceptable?
- Is the overall cost of licensing and administration feasible?

5.4.2 Response validity

A second criterion is the response validity of the measure. The review of the literature is necessary to ascertain that the measure can detect important differences in health states that are expected to be affected by the intervention being tested. For example, if there are side-effects from treatment, then the measure should reflect the effect of side effects on preference weights.

5.4.3 Feasibility

A third criterion is feasibility. A measure is more feasible if it takes less time to complete and if there are fewer missing responses and missing items within a response. Feasibility can be quantified by the response rate, that is, the percentage of scheduled assessments that are actually completed. It can also be quantified by the rate of incompleteness, that is, the number of responses that have missing items. The EQ-5D is quite a bit shorter, thus easier to administer than the HUI or QWB. The researcher must trade off the potential for greater responsiveness of the latter instruments with the greater burden they impose on study respondents and research staff.

5.4.4 Respondent burden

The respondent burden has been defined as the demands placed on those who respond to the instrument. Administrative burden has been defined as the demand placed on those who administer it (Scientific Advisory Committee of the Medical Outcomes Trust 2002). Subject and administrative burden are linked to feasibility. More burdensome instruments will be less likely to be properly administered and complete. Planners must trade off burden with response validity.

5.4.5 Alternative forms and translations

The performance of an instrument may be affected by the mode of administration. Modes include self-report, interviewer administered, and computer-assisted methods (Scientific Advisory Committee of the Medical Outcomes Trust 2002).

Instruments not only need to be translated, they also need to be adapted the culture of the respondent (Scientific Advisory Committee of the Medical Outcomes Trust 2002).

If the study participant cannot undertake complex cognitive tasks, the preference measure may need to be completed by a proxy. Not all measures can be as easily completed by a proxy.

5.4.6 Theoretical justification

Utility assessment is based on expected utility theory. Some researchers feel that measures consistent with the assumptions of utility theory should be favored. Theoretical justification is an argument typically adopted by economists. Psychologists rely more on psychometric properties in the selection of preference assessment measures.

Since it is based on the classic theory that is the underpinning of utility assessment, the standard gamble is often regarded as having the best theoretical justification for assigning values to health states. Multi-attribute health status classification systems, such as the SF-6D, use preference weights that were developed by community responses to the standard gamble. Some researchers, for example Brazier et al believe that theoretical justification is a weak criterion for selecting a measure. They point out ways in which raters do not follow the axioms of utility theory, and argue that this weakens the theoretical justification for the standard gamble (Brazier, Deverill et al. 1999).

5.4.7 Appropriateness of the population sampled to obtain preference weights

There is debate in the field about the translatability of responses from a community sample to a study population in a different country. The U.S. Panel on Cost-Effectiveness in Health and Medicine recommended that preference weights come from a sample that represents the U.S. population (Gold, Siegel et al. 1996). The U.K. National Institute for Clinical Excellence specifies that preference measures should be based on values established in a U.K. population. However, Drummond and co-authors provide evidence that the different communities assign very similar values to a particular health state, and that this concern should not affect the choice of a preference assessment measure (Drummond, Sculpher et al. 2005). Brazier (2005), on the other hand, suggests that the literature indicates that responses do vary by community as well as other socioeconomic and demographic differences.

5.4.8 Disease-specific health status classification systems

If generic preference-based health status classification systems have not been tested in the population of interest, or if generic measures do not have sufficient response validity for the condition of interest, a researcher might select a disease-specific measure to elicit the descriptions of health states experienced during a trial. Although disease-specific quality of life measures are not designed to yield preference weights, they have been used to estimate preferences in some conditions. The disease-specific health status measure is applied in the trial population, and the results characterize the health states that are common in that population. A separate survey of community members is conducted to elicit preference weights for these health states (Brazier, Deverill et al. 1999). Drummond suggests transforming health status measures to preference weights by anchoring the poorest health state in the medical condition, to the 0 to 1 scale for preference weights (Drummond, Sculpher et al. 2005).

As a disease-specific measure is tailored for the disease of interest, this method may provide greater response validity than is possible with a generic preference-based health status measure. Disease-specific measures also impose a smaller response burden than direct elicitation. This method was used in CSP 006 – Effectiveness of Geriatric Evaluation and Management (GEM) Units – which conducted a survey of older adults to assign preference weights to responses about Activities of Daily Life (Bravata, Nelson et al. 2005).

A significant disadvantage of this method is that it requires additional resources, as a community study must be conducted. It is also possible that more than one disease-specific instrument will be needed, for example, to capture the side effects of treatment as well as symptoms of disease. If there is a very large number of different potential health states to characterize, valuation may be prohibitively complex, and direct elicitation would be more straightforward.

5.5 Summary of recommendations for planning a clinical trial

In summary, we recommend that once planners have confirmed that an economic evaluation is appropriate for the clinical trial, the researcher should next identify the expected health states for the study population. The researcher should then determine whether these health states will develop during the term of the trial, and then evaluate existing literature on preference assessment measures that have been used in a similar population. The evaluation should allow

selection of the least burdensome measure that meets psychometric and other criteria. The measure should be valid, reliable, and responsive among patients with the disease to be studied. Researchers are encouraged to use more than one preference assessment measure in order to compare the performance of the measures and to add to the literature on the psychometric properties of preference assessment measures. The choice of measures should be justified in the study protocol and the implementation of the measure should be done to enhance the measure's feasibility, reliability and validity (Stalmeier, Goldstein et al. 2001).

5.6 Reporting results of preference measurement

Even though preference and utility measurement are widely used in health care economic analyses, the psychometric properties of these measures are not well known for most patient populations. For this reason, CSP researchers are encouraged to contribute to the literature on preference assessment. Researchers should fully describe their methods, following the recommendations of the consensus report from the Utilities Interest Group of the Society for Medical Decision Making (Siegel, Weinstein et al. 1996; Stalmeier, Goldstein et al. 2001). Such thorough reporting will contribute to the improved understanding of how measures perform in specific populations.

References

Anderson, J. P., Kaplan, R. M., Berry, C. C., et al. Interday reliability of function assessment for a health status measure. The Quality of Well-Being scale. Med Care 1989; 27(11): 1076-83.

Andresen, E. M., Patrick, D. L., Carter, W. B., et al. Comparing the performance of health status measures for healthy older adults. J Am Geriatr Soc 1995; 43(9): 1030-4.

Andresen, E. M., Rothenberg, B. M. and Kaplan, R. M. Performance of a self-administered mailed version of the Quality of Well-Being (QWB-SA) questionnaire among older adults. Med Care 1998; 36(9): 1349-60.

Ankri, J., Beaufils, B., Novella, J. L., et al. Use of the EQ-5D among patients suffering from dementia. J Clin Epidemiol 2003; 56(11): 1055-63.

Badia, X., Schiaffino, A., Alonso, J., et al. Using the EuroQoI 5-D in the Catalan general population: feasibility and construct validity. Qual Life Res 1998; 7(4): 311-22.

Bayoumi, A. M. (2003). "Prospec." 1.1. from http://individual.utoronto.ca/bayoumi/prospec/.

Bowling, A. (2001). Measuring Disease. Buckingham, PA, Open University Press.

Boyle, M. H., Furlong, W., Feeny, D., et al. Reliability of the Health Utilities Index--Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res 1995; 4(3): 249-57.

Bravata, D. M., Nelson, L. M., Garber, A. M., et al. Invariance and inconsistency in utility ratings. Med Decis Making 2005; 25(2): 158-67.

Brazier, J. (2005). Current state of the art in preference-based measures of health and avenues for further research. The University of Sheffield School of Health and Related Research, Health Economics and Decision Science Discussion Paper Series. Sheffield, University of Sheffield: 1-21.

Brazier, J. and Deverill, M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. Health Econ 1999; 8(1): 41-51.

Brazier, J., Deverill, M. and Green, C. A review of the use of health status measures in economic evaluation. J Health Serv Res Policy 1999; 4(3): 174-84.

Brazier, J., Deverill, M., Green, C., et al. A review of the use of health status measures in economic evaluation. Health Technol Assess 1999; 3(9): i-iv, 1-164.

Brazier, J., Roberts, J. and Deverill, M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002; 21(2): 271-92.

Brazier, J., Roberts, J., Tsuchiya, A., et al. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004; 13(9): 873-84.

Brazier, J. E. and Roberts, J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004; 42(9): 851-9.

Brazier, J. E., Walters, S. J., Nicholl, J. P., et al. Using the SF-36 and Euroqol on an elderly population. Qual Life Res 1996; 5(2): 195-204.

Chapman, R. H., Stone, P. W., Sandberg, E. A., et al. A comprehensive league table of cost-utility ratios and a sub-table of "panel-worthy" studies. Med Decis Making 2000; 20(4): 451-67.

Conner-Spady, B. P. and Suarez-Almazor, M. E. M. D. P. Variation in the Estimation of Quality-adjusted Life-years by Different Preference-based Instruments. Medical Care July 2003; 41(7): 791-801.

Coons, S. J., Rao, S., Keininger, D. L., et al. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000; 17(1): 13-35.

Delquie, P. (2006). "Assess." from http://faculty.insead.edu/delquie/ASSESS.htm.

Dolan, P., Gudex, C., Kind, P., et al. Valuing health states: a comparison of methods. J Health Econ 1996; 15(2): 209-31.

Dorman, P., Slattery, J., Farrell, B., et al. Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke. United Kingdom Collaborators in the International Stroke Trial. Stroke 1998; 29(1): 63-8.

Drummond, M. Introducing economic and quality of life measurements into clinical studies. Ann Med 2001; 33(5): 344-9.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., et al. (2005). Methods for the Economic Evaluation of Health Care Programmes. Oxford, Oxford University Press.

EuroQoLGroup (2005). EQ-5D A Generic, Single Index Measure (Promotional Bulletin).

Fairbank, J. C., Couper, J., Davies, J. B., et al. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66(8): 271-3.

Feeny, D. The Health Utilities Index: A tool for assessing health benefits. PRO Newsletter 2005; 34(Spring): 2-6.

Feeny, D., Furlong, W., Torrance, G. W., et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002; 40(2): 113-28.

Feeny, D., Torrance, G. and Furlong, W. (1996). Health Utilities Index. Quality of Life and Pharmacoeconomics in Clinical

Trials. B. Spilker. Philadelphia, Lippincott-Raven: 239-52.

Fisk, J. D., Brown, M. G., Sketris, I. S., et al. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. J Neurol Neurosurg Psychiatry 2005; 76(1): 58-63.

Fryback, D. G., Lawrence, W. F., Martin, P. A., et al. Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study. Med Decis Making 1997; 17(1): 1-9.

Gandek, B., Ware, J. E., Aaronson, N. K., et al. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998; 51(11): 1171-8.

Gold, M. R., Siegel, J. E., Russell, L. B., et al. (1996). Cost-Effectiveness in Health and Medicine. New York, Oxford University Press.

Gold, M. R., Stevenson, D. and Fryback, D. G. HALYS and QALYS and DALYS, Oh My: similarities and differences in summary measures of population Health. Annu Rev Public Health 2002; 23: 115-34.

Goldstein, M. K. (2002). Quality of life assessment software. AMIA Symposium.

Goldstein, M. K., Clarke, A. E., Michelson, D., et al. Developing and testing a multimedia presentation of a health-state description. Med Decis Making 1994; 14(4): 336-44.

Goldstein, M. K., Michelson, D., Clarke, A. E., et al. A multimedia preference-assessment tool for functional outcomes. Proc Annu Symp Comput Appl Med Care 1993: 844-8.

Gonzales, E., Eckman, M. and Pauker, S. "Gambler": a computer workstation for patient utility assessment. Med Decis Making 1992; 12: 350.

Green, C., Brazier, J. and Deverill, M. Valuing health-related quality of life. A review of health state valuation techniques. Pharmacoeconomics 2000; 17(2): 151-65.

Grootendorst, P., Feeny, D. and Furlong, W. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. Med Care 2000; 38(3): 290-9.

Guillemin, F. The value of utility: assumptions underlying preferences and quality adjusted life years. J Rheumatol 1999; 26(9): 1861-3.

Hays, R. D., Anderson, R. T. and Revicki, D. (1998). Assessing reliability and validity of measurement in clinical trials. Quality of Life Assessment in Clinical Trials: Methods and Practice. M. J. Staquet, Hays, R. D. and Fayers, P. M. New York, Oxford University Press: 169-182.

Haywood, K. L., Garratt, A. M., Dziedzic, K., et al. Generic measures of health-related quality of life in ankylosing spondylitis: reliability, validity and responsiveness. Rheumatology (Oxford) 2002; 41(12): 1380-7.

Hinz, A., Klaiberg, A., Brahler, E., et al. [The Quality of Life Questionnaire EQ-5D: modelling and norm values for the general population]. Psychother Psychosom Med Psychol 2006; 56(2): 42-8.

Horsman, J., Furlong, W., Feeny, D., et al. The Health Utilities Index (HUI(R)): concepts, measurement properties and applications. Health Qual Life Outcomes 2003; 1(1): 54.

HUInc. (2004, 02/17/2004). "Health Utilities Inc: An update." Retrieved 07/28/2006, from http://www.healthutilities.com/.

Hurst, N. P., Jobanputra, P., Hunter, M., et al. Validity of Euroqol--a generic health status instrument--in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. Br J Rheumatol 1994; 33(7): 655-62.

Jansen, S. J., Stiggelbout, A. M., Wakker, P. P., et al. Unstable preferences: a shift in valuation or an effect of the elicitation procedure? Med Decis Making 2000; 20(1): 62-71.

Johnson, J. A. and Coons, S. J. Comparison of the EQ-5D and SF-12 in an adult US sample. Qual Life Res 1998; 7(2): 155-66.

Johnson, J. A. and Pickard, A. S. Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. Med Care 2000; 38(1): 115-21.

Jones, C. A., Feeny, D. and Eng, K. Test-retest reliability of health utilities index scores: evidence from hip fracture. Int J Technol Assess Health Care 2005; 21(3): 393-8.

Kaarlola, A., Pettila, V. and Kekki, P. Performance of two measures of general health-related quality of life, the EQ-5D and the RAND-36 among critically ill patients. Intensive Care Med 2004; 30(12): 2245-52.

Kaplan, R. The Minimally Clinically Important Difference in Generic Utility-Based Measures. COPD: J Chron Obs Pul Dis 2005; 2(1): 91-97.

Kaplan, R., Sieber, W. J. and Ganiats, T. G. The Quality of Well-Being Scale: comparison of the interviewer-administered version with a self-administered questionnaire. Psychol Health 1997; 12: 783-791.

Kaplan, R. M., Anderson, J. P., Patterson, T. L., et al. Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Center. Psychosom Med 1995; 57(2): 138-47.

Kaplan, R. M., Bush, J. W. and Berry, C. C. Health status: types of validity and the index of well-being. Health Serv Res 1976; 11(4): 478-507.

Kaplan, R. M., Bush, J. W. and Berry, C. C. Health status index: category rating versus magnitude estimation for measuring levels of well-being. Med Care 1979; 17(5): 501-25.

Kaplan, R. M., Ganiats, T. G., Sieber, W. J., et al. The Quality of Well-Being Scale: critical similarities and differences with SF-36. Int J Qual Health Care 1998; 10(6): 463-5.

Kerrigan, C. L., Collins, E. D., Kneeland, T. S., et al. Measuring health state preferences in women with breast hypertrophy. Plast Reconstr Surg 2000; 106(2): 280-8.

Kind, P. (1996). The EuroQol instrument: an index of health-related quality of life. Quality of Life and Pharmacoeconomics in Clinical Trials. B. Spilker: 191-201.

Kind, P., Dolan, P., Gudex, C., et al. Variations in population health status: results from a United Kingdom national questionnaire survey. Bmj 1998; 316(7133): 736-41.

Konig, H. H., Ulshofer, A., Gregor, M., et al. Validation of the EuroQol questionnaire in patients with inflammatory bowel disease. Eur J Gastroenterol Hepatol 2002; 14(11): 1205-15.

Kupferberg, D. H., Kaplan, R. M., Slymen, D. J., et al. Minimal clinically important difference for the UCSD Shortness of Breath Questionnaire. J Cardiopulm Rehabil 2005; 25(6): 370-7.

Lenert, L. and Kaplan, R. M. Validity and interpretation of preference-based measures of health-related quality of life. Med Care 2000; 38(9 Suppl): II138-50.

Lenert, L. A., Cher, D. J., Goldstein, M. K., et al. The effect of search procedures on utility elicitations. Med Decis Making 1998; 18(1): 76-83.

Lenert, L. A., Feddersen, M., Sturley, A., et al. Adverse effects of medications and trade-offs between length of life and quality of life in human immunodeficiency virus infection. Am J Med 2002; 113(3): 229-32.

Lenert, L. A., Sturley, A. and Watson, M. E. iMPACT3: Internet-based development and administration of utility elicitation protocols. Med Decis Making 2002; 22(6): 464-74.

Luo, N., Chew, L. H., Fong, K. Y., et al. A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease. J Rheumatol 2003; 30(10): 2268-74.

Marra, C. A., Rashidi, A. A., Guh, D., et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005; 14(5): 1333-44.

Marra, C. A., Woolcott, J. C., Kopec, J. A., et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005; 60(7): 1571-82.

McDonough, C. M., Grove, M. R., Tosteson, T. D., et al. Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among spine patient outcomes research trial (SPORT) participants. Qual Life Res 2005; 14(5): 1321-32.

McDowell, I. and Newell, C. (1996). Measuring Health: A Guide to Rating Scales and Questionnaires. New York, Oxford University Press.

McHorney, C. A., Ware, J. E., Jr. and Raczek, A. E. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993; 31(3): 247-63.

MedicalOutcomesTrust. (2006). "I Want SF." Retrieved 8/2/2006, from http://www.sf-36.org/wantsf.aspx?id=1.

Naglie, G., Tomlinson, G., Tansey, C., et al. Utility-based Quality of Life measures in Alzheimer's disease. Qual Life Res 2006; 15(4): 631-43.

Nichol, M. B., Sengupta, N. and Globe, D. R. Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36. Med Decis Making 2001; 21(2): 105-12.

Nordin, M., Alexandre, N. M. and Campello, M. Measures for low back pain: a proposal for clinical use. Rev Lat Am Enfermagem 2003; 11(2): 152-5.

Oga, T., Nishimura, K., Tsukino, M., et al. A comparison of the responsiveness of different generic health status measures in patients with asthma. Qual Life Res 2003; 12(5): 555-63.

Poissant, L., Mayo, N. E., Wood-Dauphinee, S., et al. The development and preliminary validation of a Preference-Based Stroke Index (PBSI). Health Qual Life Outcomes 2003; 1(1): 43.

Ramsey, S., Willke, R., Briggs, A., et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. Value Health 2005; 8(5): 521-33.

Riazi, A., Cano, S. J., Cooper, J. M., et al. Coordinating outcomes measurement in ataxia research: Do some widely used generic rating scales tick the boxes? Mov Disord 2006.

Ross, P. L., Littenberg, B., Fearn, P., et al. Paper standard gamble: a paper-based measure of standard gamble utility for current health. Int J Technol Assess Health Care 2003; 19(1): 135-47.

Salkeld, G., Cameron, I. D., Cumming, R. G., et al. Quality of life related to fear of falling and hip fracture in older women: a time trade off study. Bmj 2000; 320(7231): 341-6.

Schweikert, B., Hahmann, H. and Leidl, R. Validation of the EuroQol questionnaire in cardiac rehabilitation. Heart 2006; 92(1): 62-7.

Scientific Advisory Committee of the Medical Outcomes Trust Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002; 11(3): 193-205.

Shaw, J. W., Johnson, J. A. and Coons, S. J. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care 2005; 43(3): 203-20.

Sieber, W. J., Groessl, W. J., David, K. M., et al. (2004). Quality of Well-Being Self-Administered (QWB-SA) Scale. San Diego: 1-38.

Siegel, J. E. S., Weinstein, M. C. P., Russell, L. B. P., et al. Recommendations for Reporting Cost-effectiveness Analyses. JAMA 1996; 23(30): 1339-1341.

Sims, T. L., Garber, A. M., Miller, D. E., et al. Multimedia quality of life assessment: advances with FLAIR. AMIA Annu Symp Proc 2005: 694-8.

Slevin, M. L., Stubbs, L., Plant, H. J., et al. Attitudes to chemotherapy: comparing views of patients with cancer with those of doctors, nurses, and general public. Bmj 1990; 300(6737): 1458-60.

Sonnenberg, F. A. and Beck, J. R. Markov models in medical decision making: a practical guide. Med Decis Making 1993; 13(4): 322-38.

Spitzer, W. O., Dobson, A. J., Hall, J., et al. Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. J Chronic Dis 1981; 34(12): 585-97.

Stalmeier, P. F., Goldstein, M. K., Holmes, A. M., et al. What should be reported in a methods section on utility assessment? Med Decis Making 2001; 21(3): 200-7.

Stavem, K. Reliability, validity and responsiveness of two multiattribute utility measures in patients with chronic obstructive pulmonary disease. Qual Life Res 1999; 8(1-2): 45-54.

Stavem, K., Bjornaes, H. and Lossius, M. I. Properties of the 15D and EQ-5D utility measures in a community sample of people with epilepsy. Epilepsy Res 2001; 44(2-3): 179-89.

Stavem, K., Froland, S. S. and Hellum, K. B. Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS. Qual Life Res 2005; 14(4): 971-80.

Sullivan, P. W., Lawrence, W. F. and Ghushchyan, V. A national catalog of preference-based scores for chronic conditions in the United States. Med Care 2005; 43(7): 736-49.

Sumner, W., Nease, R. and Littenberg, B. U-titer: a utility assessment tool. Proc Annu Symp Comput Appl Med Care 1991: 701-5.

Sung, L., Greenberg, M. L., Doyle, J. J., et al. Construct validation of the Health Utilities Index and the Child Health Questionnaire in children undergoing cancer chemotherapy. Br J Cancer 2003; 88(8): 1185-90.

Thoma, A., Sprague, S., Veltri, K., et al. Methodology and measurement properties of health-related quality of life instruments: a prospective study of patients undergoing breast reduction surgery. Health Qual Life Outcomes 2005; 3: 44.

Torrance, G. W. Measurement of health state utilities for economic appraisal. J Health Econ 1986; 5(1): 1-30.

Torrance, G. W. Utility approach to measuring health-related quality of life. J Chronic Dis 1987; 40(6): 593-603.

Torrance, G. W. Preferences for health outcomes and cost-utility analysis. Am J Manag Care 1997; 3 Suppl: S8-20.

Torrance, G. W., Furlong, W., Feeny, D., et al. Multi-attribute preference functions. Health Utilities Index. Pharmacoeconomics 1995; 7(6): 503-20.

van Stel, H. F. and Buskens, E. Comparison of the SF-6D and the EQ-5D in patients with coronary heart disease. Health Qual Life Outcomes 2006; 4: 20.

Vitale, M. G., Levy, D. E., Johnson, M. G., et al. Assessment of quality of life in adolescent patients with orthopaedic problems: are adult measures appropriate? J Pediatr Orthop 2001; 21(5): 622-8.

Walters, S. J. and Brazier, J. E. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003; 1(1): 4.

Walters, S. J. and Brazier, J. E. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005; 14(6): 1523-32.

Wang, H., Kindig, D. A. and Mullahy, J. Variation in Chinese population health related quality of life: results from a EuroQol study in Beijing, China. Qual Life Res 2005; 14(1): 119-32.

Ware, J., Jr., Kosinski, M. and Keller, S. D. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996; 34(3): 220-33.

Ware, J. E., Jr. and Sherbourne, C. D. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30(6): 473-83.

Weinstein, M. C., Siegel, J. E., Gold, M. R., et al. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. Jama 1996; 276(15): 1253-8.

Wiebe, S., Eliasziw, M. and Matijevic, S. Changes in quality of life in epilepsy: how large must they be to be real? Epilepsia 2001; 42(1): 113-8.

Wu, A. W., Jacobson, K. L., Frick, K. D., et al. Validity and responsiveness of the euroqol as a measure of health-related quality of life in people enrolled in an AIDS clinical trial. Qual Life Res 2002; 11(3): 273-82.

Appendices

Appendix 1

"Sample" questions from selected dimensions of described multi-attribute health status classification systems. Please note: These tables do not include all questions in each domain. The full survey instruments must be accessed from the developer or other public access as indicated in the text.

Table A1: Health Utilities Index (Mark3) (HUI)

Dimensions (8): Question Framework	Levels (5 or 6 in each dimension)
Vision: Which one of the following best describes your ability, during the past week to see well enough to read ordinary newsprint?	1. Able to see well-enough without glasses……. 2. Able to see well-enough with glasses…..
Ambulation: Which one of the following best describes your ability, during the past week, to walk?.....	1. Able to walk around the neighborhood without difficulty…….. 2. Able to walk around the neighborhood with difficulty……...
Pain: Which one of the following best describes the pain and discomfort you have experienced during the past week?	1. Free of pain and discomfort 2. Mild to moderate pain that prevents no activities

Table A2: EuroQol EQ-5D

Dimensions (5): Question Framework	Levels (3 for each dimension)
Mobility: Which statements best describe your own state of health today?	1. No problems walking 2. Some problems walking about
Pain/discomfort: Which statements best describe your own state of health today?	1. No pain or discomfort 2. Moderate pain or discomfort
Anxiety/depression Which statements best describe your own state of health today?	1. Not anxious or depressed 2. Moderately anxious or depressed

Table A3: Quality of well-being scale (QWB)

Dimensions (4)	Levels (2-3 for each dimension)
Mobility	1. Over the last 3 days what days did you drive a motor vehicle (please fill in all days that apply) 2. Over the last three days which days did you use public transportation such as……(please fill in all days that apply)
Physical Activity	1. Over the last 3 days did you have trouble climbing stairs…(please fill in all days that apply) 2. Over the last 3 days did you avoid walking, have trouble walking…(please fill in all days that apply)
Usual activity	1. Over the last 3 days because of any physical or emotional health reasons, on which days did you avoid, need help with, or were limited in doing……(please fill in all days that apply) 2. Over the last 3 days because of any physical or emotional health reasons, on which days did you avoid or feel limited in doing some of your usual activities (please fill in all days that apply)

Table A4: Short form 6D (SF-6D)

Dimensions (6): Question framework	Levels (4-6 for each dimension)
Physical functioning Select one statement in each group to show which best describes your health	1. Your health does not limit you in vigorous activities 2. Your health limits you a little in vigorous activities
Pain	1. You have no pain 2. You have pain but it does not interfere……
Social functioning	1. Your health limits your social activities none of the time 2. Your health limits your social activities a little of the time

Appendix 2

Table A5: List of available software to measure utilities with direct methods

Software	Reference	Link
U-Titer/U-Titer II	(Sumner, Nease et al. 1991)	http://ilya.wustl.edu/~utiter/UtiterDemo/.
U-Maker	(Sonnenberg and Beck 1993)	
ProSPEC	(Bayoumi 2003)	http://individual.utoronto.ca/bayoumi/prospec/
iMPACT2, iMPACT3, iMPACT4	(Lenert, Feddersen et al. 2002) (Lenert, Sturley et al. 2002)	http://preferences.ucsd.edu http://sourceforge.net/projects/impact4
Gambler	(Gonzales, Eckman et al. 1992)	
FLAIR1, FLAIR2	(Goldstein, Michelson et al. 1993; Goldstein, Clarke et al. 1994) (Goldstein 2002; Sims, Garber et al. 2005). (Sims, Garber et al. 2005)	http://healthpolicy.stanford.edu/research/functional_life_and_independence_research_flair_project/
Assess	(Delquie 2006)	http://faculty.insead.edu/delquie/ASSESS.htm

Table A6: Resources for indirect measurement systems**: For a general introduction to various measurement systems see the Medical Outcomes Trust's website: http://www.outcomes-trust.org/instruments.htm**

System	Reference	Link
Health Utilities Index (HUI) Mark 2 and 3	(HUInc 2004)	www.healthutilities.com.
EuroQol (EQ-5D)	(EuroQoLGroup 2005)	www.euroqol.org. http://www.pdsurg.bham.ac.uk/PDFs/EuroQoL.pdf
Quality of Well-Being Scale (QWB)	(Kaplan, Sieber et al. 1997; Kaplan, Ganiats et al. 1998)	http://medicine.ucsd.edu/fpm/hoap/qwb.htm
SF-6D	(Brazier, Roberts et al. 2002)	http://www.sf-36.org/tools/sf36.shtml http://www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d/obtain.html

Appendix 3

Table A7: Additional resources

Resource	Reference	Link
Tables of preference or utility weights by disease category		http://www.tufts-nemc.org/cearegistry/data/default.asp
Database of cost-effectiveness ratios		http://www.tufts-nemc.org/cearegistry/data/default.asp
U.K. National Health Service Health Technology Assessment Program		http://www.hta.nhsweb.nhs.uk/
U.K. National Health Service National Institute for Health and Clinical Excellence (NICE)		http://www.nice.org.uk/
International Society for Pharmacoeconomics and Outcomes Research (ISPOR)	(Ramsey, Willke et al. 2005). (Drummond, Sculpher et al. 2005).	http://www.ispor.org/
Guidelines for reporting results of preference assessment	(Siegel, Weinstein et al. 1996; Stalmeier, Goldstein et al. 2001).	
Comparisons of preference-based measures of health by Brazier	(Brazier, Deverill et al. 1999; Brazier 2005)	http://www.hta.nhsweb.nhs.uk/ProjectData/3_publication_listings_ALL.asp (vol.3 number 9) http://www.shef.ac.uk/content/1/c6/01/87/47/0505FT.pdf
USPSTF	(Gold, Siegel et al. 1996)	
Interactive textbook from NIH on utility assessment		http://symptomresearch.nih.gov/chapter_24/sec10/cmgs10pg1.htm