30 September 2016

Notes to Online Appendices

This document (available at http://www.chepa.org/research-papers/valuing-health-outcomes/online-appendices) contains Online Appendices for Feeny, David, Murray Krahn, Lisa A. Prosser, and Joshua A. Salomon, "Valuing Health Outcomes," Chapter 7 in Neumann, Peter J., Gillian D. Sanders, Louise B. Russell, Joanna E. Siegel, and Theodore G. Ganiats, eds., *Cost-Effectiveness in Health and Medicine*, Second Edition, New York: Oxford University Press, 2016, pp. 167-199.

The notations in brackets that follow each Online Appendix title identify the section and approximate page number of the main text in Chapter 7 to which each online appendix corresponds. Several appendices correspond to more than one section in the main text. Acknowledgments and related exposition are found in Neumann et al. (2016). The Online Appendices are meant to be read in conjunction with the main text of Chapter 7.

Online Appendix 7.1. Brief Description of Major Generic Preference-Based Multi-attribute Measures Used in Cost-Effectiveness Analysis [Section 7.1.3, p. 169]

The following is meant to provide a brief description and evaluation of the usefulness of four major generic preference-based measures that have been widely used in cost-effectiveness analyses: EuroQol 5D (EQ-5D), the Health Utilities Index (HUI), the Quality of Well-Being Scale (QWB) and the Short-Form 6D (SF-6D). More extensive descriptions of these measures and their measurement properties can be found in a number of sources including Brazier et al. (2007); Drummond et al. (2005); Feeny (2005a); Feeny (2005b); Hawthorne and Richardson (2001); and Rowen and Brazier (2011). For each measure a basic description of the measure and its use will be presented, evidence on its measurement properties and scoring function will be reviewed, and a summary of its advantages and disadvantages will be provided. Finally, a brief description of the most recent version of the disability weighting system used in the 2010 Global Burden of Disease Study to estimate disability-adjusted life years (DALYs) will be provided.

The performance of the four major generic preference-based measures tends to vary by clinical context. In general, there is substantial evidence for the construct validity of each of these measures. Further, in general, there is evidence of the responsiveness of these four measures in a wide variety of applications. Typically these four measures are of the same order of magnitude of responsiveness as other generic measures of health status and generally less responsive than disease- and condition-specific measures (Guyatt et al. 1999; Wiebe et al. 2003).

EQ-5D. The EQ-5D was developed collaboratively in multiple languages by investigators from a number of European countries. The original version of the EQ-5D included five attributes (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) with three levels for each (no problem, moderate problem, severe problem). The EQ-5D has been widely used in clinical studies and population health surveys, including the US Medical Expenditure Panel Survey (https://meps.ahrq.gov/mepsweb/) 2000, 2001, and 2002 (Sullivan et al. 2005), and two other nationally representative surveys in the United States (Fryback et al. 2007; Luo et al. 2005). The original three-level version is subject to substantial ceiling effects and attenuated responsiveness (Brazier et al. 2004; Insinga and Fryback 2003; Longworth and Rowen 2013; Turner et al. 2013).

The new five-level version was designed to improve EQ-5D’s performance (Pickard et al. 2007). Indeed, evidence is emerging that indicates that the 5L system is less subject to ceiling and floor effects than the 3L version (see, e.g., Buchholz et al. 2015; Janssen et al. 2013; Jia et al.

2014; Lee et al. 2013; Pickard et al. 2007). Using an interim scoring system for the 5L system (van Hout et al. 2012), Jia et al. (2014) and Lee et al. (2013) provide some initial evidence that the 5L system enhances responsiveness. Efforts are underway to create scoring functions for the 5L system (Oppe et al. 2014). As the 5L scoring functions emerge, and as experience using the 5L system accumulates, evidence on its measurement properties will become available.

A disadvantage of the EQ-5D is its poor coverage of problems with vision and hearing (Tosh et al. 2012) and many mental health problems (Brazier 2010; Richardson et al. 2015). As noted in Online Appendix 7.4, there is a substantial body of evidence that indicates that there are important interactions in preferences among attributes in the EQ-5D system and that therefore the widely used linear additive scoring functions for EQ-5D may be less than ideal. In addition, there is evidence that the 10-year time horizon used to elicit the preference scores used for the estimation of many EQ-5D scoring functions may introduce distortions (Heintz et al. 2013; van Nooten et al. 2009; van Nooten et al. 2014). An advantage of the EQ-5D is the availability of a scoring function based on time-tradeoff scores elicited in the United States (Shaw et al. 2005).

Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3). HUI3 is the most widely used of the HUI suite of instruments. HUI3 includes eight attributes (vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain/discomfort) with five or six levels per attribute (Feeny et al. 2002; Furlong et al. 2001; Horsman et al. 2003; Torrance et al. 1996). HUI3 has been widely used in clinical studies and population health surveys, including four in the United States (Fryback et al. 2007; Luo et al. 2005; National Institute on Aging n.d; Sanmartin et al. 2004).

In population health surveys HUI3 is subject to modest ceiling effects relative to the QWB and SF-6D but considerably less than the EQ-5D-3L. An advantage of HUI3 is its inclusion of vision (cataracts, macular degeneration), hearing (cochlear implants), speech (stroke, stuttering), cognition (stroke, dementia, Alzheimer's disease), and dexterity (arthritis, stroke). A characteristic of HUI3 is that its health-status classification system does not include social interaction, although it does include attributes for which impairments are often associated with limitations in social interaction (vision, hearing, speech, cognition, and emotion). HUI3 seems to be free of floor effects. The HUI3 multiplicative scoring function is based on visual analog scores and standard gamble scores elicited from a representative community sample in Hamilton, Ontario, Canada. HUI2 was originally developed to assess health-related quality of life in children with cancer but has also been widely used in studies of adults. Because of its focus on child health, it continues to be used in a number of pediatric settings.

Quality of Well-Being Scale (QWB). The QWB was the first widely used multi-attribute measure (Kaplan and Bush 1982; Patrick et al. 1973). Subsequently the QWB was revised (Kaplan and Anderson 1996). The revised system includes three attributes (mobility, physical activity, social activity) with three levels each and a 27-item symptom/problem complex that covers a wide range of symptoms and problems, often providing rich detail on health status. The QWB has been widely used in clinical studies and population health surveys, including two in the United States (Fryback et al. 1997; Fryback et al. 2007). The QWB seems to be free of both floor and ceiling effects. The linear additive QWB scoring function is based on visual analog scores.

The Short-Form 6D (SF-6D). SF-6D includes a health-status classification system and multi-attribute scoring function based on the widely used Short-Form 36 (Brazier et al. 2002) and Short-Form 12 questionnaires (Brazier and Roberts 2004). Role Physical and Role Emotion were combined to form Role Limitation in SF-6D; the other five attributes are Physical Function, Social Functioning, Pain, Mental Health, and Vitality, with four to six levels per attribute for the system based on SF-36 and three to five for the system based on SF-12. Selim et al. (2011) provide a scoring function for the RAND 12-item questionnaire used by the US Veterans Administration. The SF measures have been widely used in clinical studies and population health surveys in the United States, including Fryback et al. (2007). The linear additive scoring function was estimated using standard gamble scores from a community sample in the United Kingdom. There are ongoing efforts to create a scoring system based on US preferences (Craig et al. 2013). An advantage of SF-6D is that it seems to be free of ceiling effects. Another advantage of SF-6D is its inclusion of vitality. In a head-to-head longitudinal study of patients with chronic epilepsy that included EQ-5D, HUI2, HUI3, and SF-6D, Langfitt and colleagues (2006) concluded that SF-6D is preferred because it includes more of the attributes affected by epilepsy than do the other measures. Interestingly, Fiest and colleagues (2014), in a study of surgical interventions for epilepsy, found that HUI3 was more responsive than SF-36 but less responsive than the disease-specific measure, the Quality of Life in Epilepsy (QOLIE). A disadvantage of SF-6D is widespread reports of floor effects (Brazier et al. 2004; Brazier et al. 2010; Feeny et al. 2003; Fisk et al. 2005; Harrison et al. 2009; Hatoum et al. 2004; Mortimer and Segal 2008; O'Brien et al. 2003; Turner et al. 2013).

Disability Weights for 2010 Global Burden of Disease Study. The disability weights are used to quantify the non-fatal health losses in the estimation of disability-adjusted life-years (DALYs). The elicitation of preferences to estimate the disability weights was based on a paired comparison approach (discrete choice) (Salomon et al. 2012). Data were collected in face-to-face interviews in Bangladesh, Indonesia, Peru, and Tanzania; in telephone interviews in the United States; and in an open-access web-based survey. Data were then analyzed using probit regression based on a random utility model. As experience with these new disability weights accumulates, information on the measurement properties of the system will emerge.

Online Appendix 7.2: Practical and Ethical Limitations of the QALY Approach [Section 7.2, p. 172]

Several practical and ethical limitations to the QALY approach should be noted. First, QALYs may not accurately reflect the burden of short-lived but intense experiences. For instance, extreme pain experienced briefly during a dental procedure might be a tiny fraction of the overall time period and thus, regardless of how low the utility score was, would have little influence on the overall estimate of QALYs. Such experiences might nonetheless be regarded by patients as being very important and may influence decision making. The path-state approach discussed in Section 7.1.4 of the main text is one approach for handling this problem. Second, a number of important objections to QALYs have been discussed in the literature, including the issue of whether or not QALYs discriminate against the disabled and that conventional QALYs can favor interventions that provided marginal gains to many people at the expense of interventions that would provide substantial gains to a smaller number of people. Some have argued that treating those with highly impaired health should be valued more highly than treating those with good baseline health. Many are troubled by the inter-personal comparisons of utility implicit in calculating QALYs. For instance, the German health technology assessment agency, the Institute for Quality and Efficiency in Health Care, rejects the use of QALYs to make comparisons across therapeutic areas due to concerns “regarding solidarity, equity, and fairness” (Institute for Quality and Efficiency in Health Care 2009, p. 3). These issues are explored in Chapter 12 of the main text [“Ethical and Distributive Considerations”].

Online Appendix 7.3: Are QALYs Utilities? [Section 7.2, p. 172]

If QALYs were utilities, then the maximization of QALYs gained subject to a budget constraint would be consistent with the maximization of expected utility and using welfare theory from economics (see Chapter 2 of the main text ["Theoretical Foundations of Cost-Effectiveness Analysis in Health and Medicine"]). Under rather restrictive assumptions the QALY can be interpreted as a utility (Drummond et al. 2005; Torrance and Feeny 1989). These assumptions include: that the quantity and quality of life are mutually utility independent (i.e., the preference for one of these attributes is independent of the level for the other attribute); that there is a constant proportional tradeoff (i.e., the proportion of remaining life span that one would trade for a specified quality improvement is independent of the remaining duration of life); and that the utility function for additional life years is linear with time. There is substantial empirical evidence on people's preferences that does not support these assumptions (see, e.g., Attema and Brouwer 2012; Beresniak et al. 2015; Loomes and McKenzie 1989; Spencer 2003; and Treadwell 1998). Many analysts do not regard QALYs as utilities and instead view them as an "index" number that is monotonically related to utility (Garber et al. 1996). These analysts regard QALYs as a useful measure of health and gains in health. The maximization of QALYs gained subject to a budget constraint can been regarded as an appropriate objective for health policy.

Online Appendix 7.4: Disability-Adjusted Live Years [Section 7.2, p. 172]

DALYs share several of the key features of QALYs, in that they are summary measures of population health that combine information on mortality with information on non-fatal outcomes; that the unit of account is time, which is what provides commensurability between mortality and non-fatal outcomes; and that the latter are incorporated in the measures using weights that are measured on a continuum between ideal health and states equivalent to dead. The primary application of DALYs is in quantifying the burden of disease, and for this reason DALYs are quantified in terms of health losses. Although primarily developed to measure disease burden, DALYs are also commonly used as the metric for measuring health benefits from interventions (in terms of reductions in burden expressed as DALYs averted), particularly in analyses in low- and middle-income countries. Both the World Health Organization, and more recently recommendations for the Gates Foundation Reference Case, have specified DALYs as the metric in cost-effectiveness analyses.

Two key value choices in DALYs that have attracted considerable attention and provided contrasts to QALYs are the weighting of health outcomes differentially depending on the age at which they are experienced, and the "disability weights" attached to different non-fatal outcomes. In the case of the former, until recently the standard formulation of DALYs included an "age weighting" function that resulted in years lived at young adult ages being more heavily weighted than years lived either in childhood or at the oldest ages, but in the most recent iteration of the Global Burden of Disease Study, there has been a move to uniform age weights (Murray et al. 2012). With regard to the disability weights assigned to every non-fatal outcome, the approach used in the Global Burden of Disease Study has evolved. In the most recent revision, weights are based on a large empirical data collection effort that has included household surveys in nine countries and a large open-access Internet survey, with a total respondent sample of more than 60,000 people worldwide (Salomon et al. 2012; see also Online Appendix 7.1). The primary basis for estimating weights in this study comes from survey responses to paired comparison questions in which respondents consider two hypothetical persons described briefly in terms of levels of functioning on key dimensions of health and any other salient symptoms of a particular condition, and then indicate which of the two people they would regard as being healthier. The analytic techniques for translating these ordinal responses into cardinal weights are described in Online Appendix 7.9.

A number of earlier criticisms of DALYs remain relevant (Gold et al. 2002). First, in the QALY approach the focus is on the value attached to the health state, not the value attached to the burden associated with a disease. Second, to date DALYs do not handle comorbidities. Yet in the context of healthcare interventions, many of those treated suffer from more than one chronic condition (Tinetti et al. 2012). Further, DALYs, in general, do not capture the side effects of treatments, an important omission in many clinical contexts. Finally, the estimation of the disability weights is based on the person tradeoff approach, to which some analysts have objected. A potentially useful application of DALYs is in the evaluation of interventions involving disease prevention, providing estimates of the DALYs avoided.

Online Appendix 7.5: How to Estimate Multi-Attribute Utility Functions (MAUFs) to Provide Utility Scoring Systems [Section 7.4, p. 173]

As noted in the chapter, one major approach for obtaining utility scores based on community values is to employ a generic preference-based multi-attribute instrument. Patients complete a questionnaire based on the health-status description system of an instrument. The health states are then valued using the scoring system based on a multi-attribute utility function (MAUF) estimated for that instrument. The resulting scores then reflect the self-reported health status of patients valued using the preferences of members of the general population.

Recall that in the multi-attribute approach health status is comprised of a number of attributes. For instance, for the Health Utilities Index Mark 3 (HUI3) there are eight attributes (vision, hearing, speech, ambulation, dexterity, cognition, emotion, and pain and discomfort) with five or six levels per attribute ranging from highly impaired (“so unhappy that life is not worthwhile”) to normal or unimpaired (“happy and interested in life”). The health status of an individual at a point in time is then described as an n-element (8-element for HUI3) vector with one level for each of the attributes.

This appendix provides a summary of the methods used to estimate MAUFs for these multi-attribute measures. Three major functional forms have been used to estimate MAUFs: the linear additive; the multiplicative; and the multi-linear (Keeney 1988; Keeney and Raiffa 1993). The linear additive is the simplest and most easily estimated functional form. But it makes the strongest assumptions about the structure of preferences among the attributes, namely that there are no preference interactions among the attributes. (We will discuss relevant empirical evidence on this issue below.) The multiplicative functional form includes an omnibus preference interaction term; the attributes are all preference complements or all preference substitutes. The multi-linear functional form allows for pairs of attributes to be preference complements, other pairs to be preference substitutes, and other pairs for which there are no preference interactions. These much less restrictive assumptions come at the cost of considerably more challenges for empirical estimation.

Multi-attribute utility theory (MAUT) provides guidance on the choice of functional form. Within that framework there have been two major estimation approaches: the decomposed approach and the statistical inference approach. The first step in the decomposed approach is to estimate single-attribute utility functions for each attribute by eliciting preferences for each level within each attribute. Then respondents are asked to evaluate multi-attribute corner states, states in which the attribute in question is at its worst level, while all the other attributes are at their best level; the resulting score provides an indication of the weight attached to that attribute.

Simultaneous equations are solved to provide an estimate of the magnitude and sign of the omnibus preference interaction term.

In contrast, the statistical inference approach relies on more familiar linear regression models in which directly elicited utility scores for various health states are the dependent variable and the levels within the attributes in that system serve as independent variables. In practice, frequently additional *ad hoc* terms have also been included. Some investigators have used experimental designs to select that health states for which valuations will be obtained (Brazier et al. 2002); in many cases the experimental design was chosen to be able to identify all of the parameters of a linear additive function.

Recently a number of investigators have used Bayesian approaches for the estimation of MAUFs. For example, Kharroubi and colleagues (2007) estimated a linear additive scoring function for the Short-Form 6D using the Bayesian approach; see also Kharroubi et al. (2013). Similarly, Kharroubi and McCabe (2008) have estimated a MAUF for the HUI2 system using a Bayesian approach. The authors argue that the non-parametric Bayesian approach often results in lower prediction error. In practice the Bayesian non-parametric approach has demonstrated some advantage, but the reliance on the linear additive framework has attenuated some of its potential advantages.

To date the most frequently used approach for the estimation of MAUFs has been the statistical inference approach based on the linear additive functional form. The decomposed approach based on the multiplicative model has also been frequently used.

Empirical Evidence on Functional Form. It is important to examine the empirical evidence on how well the widely used functional forms, the linear additive and the multiplicative, perform. A more detailed exposition appears in Online Appendix 7.6. Briefly, there is substantial evidence that there are important interactions in preferences across attributes. This evidence calls into question the validity of linear additive functions that assume that there are no interactions in preferences among the attributes.

Transforming Visual Analog Scores into Standard Gamble or Time-Tradeoff Scores. In a number of scoring function projects, standard gamble (or time-tradeoff) scores were collected for only a subset of the health states being evaluated, while visual analog scale scores were obtained for all of the health states. Examples of this approach include HUI1 (Torrance et al. 1982), HUI2 (Torrance et al. 1996), and HUI3 (Feeny et al. 2002). Torrance and colleagues (2001) report results from a number of studies that used the power function to transform visual analog scores into standard gamble scores. In the HUI3 project, the power function more accurately predicted directly measured standard gamble scores than did a spline function. In contrast, McCabe and colleagues (2004) report that a cubic function out-performed the power function. The criterion used here is agreement between directly measured scores and transformed scores. The method used to transform scores typically will have an important impact of the performance of the MAUF.

Assessment of the Performance of MAUFs. A wide variety of techniques have been used to assess the internal and external validity (out-of-sample predictive validity) of estimates of MAUFs. Various measures of goodness-of-fit, depending on the estimation method used, have been used to evaluate the results of regression analyses.

With respect to internal and external validity, investigators have examined within-sample (internal validity) and out-of-sample (external validity) accuracy in predicting directly measured utility scores. Mean absolute differences between predicted and observed scores have been calculated (see, e.g., Dolan 1997 and Shaw et al. 2005); similarly, mean error and root mean square error have been calculated. In the HUI3 scoring function project respondents were randomly allocated to the modeling survey that provided the information for estimating the HUI3 MAUF or the direct survey. In the direct survey, standard gamble scores were obtained for 73 HUI3 health states; none of these scores was used to estimate the MAUF. Out-of-sample prediction was then assessed by calculating agreement for the 73 health states for scores predicted by the MAUF and directly measured scores; the intra-class correlation coefficient was 0.88 (Feeny et al. 2002), indicating a very good level of agreement between predicted and directly measured scores (Altman 1991).

A related approach to assessing the performance of MAUFs is to compare scores for a respondent's current health derived from them to directly obtained time-tradeoff or standard gamble scores for the respondent's current health obtained at the same time. In one prospective study, patients waiting for or undergoing elective total hip arthroplasty were asked to complete the questionnaire for HUI2 and HUI3 and at the same time to provide a standard gamble score for their current health (Feeny et al. 2003). Mean standard gamble, HUI2, and HUI3 scores across all assessments were 0.65, 0.66, and 0.55, respectively. At the group level there was substantial agreement between directly measured standard gamble and HUI2 scores, while HUI3 scores were systematically lower. It is also worth noting that agreement at the individual level was, in general, poor.

Similar results were observed in a long-term follow-up of teenaged survivors of extremely low birth weight and a control group of full-term births (Feeny et al. 2004). Mean standard gamble, HUI2, and HUI3 scores were 0.91, 0.92, and 0.84, respectively. Again at the group level, standard gamble and HUI2 scores were very similar, while HUI3 scores were systematically lower. At the individual level, agreement between standard gamble and HUI scores was fair.

The Generalizability of MAUFs. General population samples as well as other more focused samples have been used to estimate MAUFs for widely used generic preference-based measures. How generalizable are these functions? Do scoring functions based on community preferences for the same instrument differ among countries? The evidence is mixed. For the Quality of Well Being Scale (QWB), the original estimates from the early 1970s were based on community preferences in San Diego, California. Estimates derived from preference elicitation from patients with arthritis in the northeastern United States (Balaban et al. 1986) and estimates derived from a community sample in Trinidad and Tobago (Hector et al. 2010) are very similar to the original San Diego results; see also Kaplan (1994). Similarly, estimates for HUI2 based on a community sample are very similar to estimates based on a sample of parents of children with a life-threatening cancer (Wang et al. 2002). The UK (McCabe et al. 2004) and Canadian (Torrance et al. 1996) HUI2 scoring functions are similar. Estimated scoring functions for HUI3 based on results in Canada (Feeny et al. 2002), the Netherlands (Raat et al. 2004), France (Le Galès et al. 2002), and Spain (Ruiz et al. 2003) are very similar. Salomon and colleagues, in a study that elicited preference scores from more than five countries, note that “we have reported compelling evidence that contradicts the prevailing hypothesis that assessments of disability must vary widely across samples with diverse cultural, educational, environmental, or demographic circumstances” (Salomon et al. 2012, p. 2139).

In contrast, there often appear to be important differences in scoring functions for the EQ-5D estimated in different countries (Knies et al. 2009; Xie et al. 2014; Oremus et al. 2014; Norman et al. 2009). For instance, while there is some agreement in time-tradeoff scores for mildly impaired states when comparing results from UK and US studies, there is considerable divergence in scores for more severely affected states (Johnson et al. 2005). Furthermore, the UK (Dolan 1997) and US (Shaw et al. 2005) scoring functions for EQ-5D differ in a number of important ways. For instance, the score for the all-worst EQ-5D state using the UK scoring function is -0.59, while the score for the same state based on the US scoring function is -0.11. The UK and Japanese functions also differ importantly (Tsuchiya et al. 2002). Similarly, Badia and colleagues (2001) report important differences in the UK and Spanish scoring functions.

The conventional wisdom is that, all other things being equal, it would be desirable to employ a scoring function based on community preferences from the country in which the results of the cost-effectiveness analysis will be used. In practice, for many instruments, experience to date indicates that results would probably not differ importantly if a scoring function from another country were used (HUI2, HUI3, QWB), while for other instruments results probably would differ importantly (EQ-5D-3L). Further, the use of country-specific scoring functions (sometime of uneven quality) reduces the scope for transnational comparisons of study results. The consistent use (at least in a sensitivity analysis) of the scoring system for an instrument based on a high-quality study would enhance the comparability among studies conducted in different countries.

Online Appendix 7.6: Empirical Evidence on Functional Form [Section 7.4, p. 174]

The linear additive functional form has been employed to estimate a number of multi-attribute utility functions (MAUFs) for the EQ-5D system (Badia et al. 2001; Dolan 1997; Greiner et al. 2003; Shaw et al. 2005). In all of these cases the investigators found that adding *ad hoc* terms substantially improved fit. For example, for the estimation of a scoring function for EQ-5D based on community preferences in the United Kingdom, Dolan (1997) added an *ad hoc* term for any level other than Level 1 and an *ad hoc* term for any attribute at Level 3. Similarly, Brazier and colleagues (2002) and Brazier and Roberts (2004) for the SF-6D add a term when any attribute was at its most severe level.

That investigators found that adding *ad hoc* additional terms enhanced the fit is indirect evidence that suggests that the linear additive functional form may be inappropriate. This inference is consistent with the results reported by Busschbach and colleagues (1999), who used the multiplicative functional form to estimate a MAUF for the EQ-5D-3L system. Their results rejected the linear additive form and found that the attributes were preference complements. Brazier and colleagues (2011) and Yang and colleagues (2014) also present evidence that is inconsistent with the linear additive model.

As noted earlier, the linear additive form is a special case of the multiplicative. Results from the estimation of the HUI1 (Torrance et al. 1982), HUI2 (McCabe et al. 2004; Torrance et al. 1996; Wang et al. 2002), HUI3 (Feeny et al. 2002; Le Galès et al. 2002; Raat et al. 2004; Ruiz et al. 2003), Assessment of Quality of Life (AQoL) (Hawthorne and Richardson 2001; Hawthorne et al. 2001), and more recent AQoL-8D (Richardson et al. 2015) all reject the linear additive functional form in favor of the multiplicative and report preference complementarity among the attributes included in their systems (see also Montejo et al. 2011; Salomon et al. 2003). Similar results rejecting the linear additive in favor of the multiplicative are reported for a number of condition-specific MAUFs (Beusterien et al. 2005; Lo et al. 2006; Revicki et al. 1998a; Revicki et al. 1998b). These studies reporting on results for four generic preference-based multi-attribute measures and four condition-specific preference-based multi-attribute measures call into question the validity of the reliance on the linear additive functional form. In the context of estimating utility functions based on data from discrete-choice experiments, van der Pol and colleagues (2014) present evidence that non-linear specifications fit the data better than the linear additive functional form.

Online Appendix 7.7: Technical Details of Direct Utility Measurement Approaches [Section 7.5, p. 174]

Importance of Visual Aids in Direct Elicitation Tasks. There has been an increased understanding of the importance of visual aids in assisting respondents in the valuation task. (Figures depicting the standard gamble, time-tradeoff, and visual analogue scale and a number of sample visual aids can be found in Drummond et al. 2005; Feeny 2005a; Feeny 2005b; and Furlong et al. 1990.) The usefulness of visual aids in improving understanding of small probabilities was demonstrated in the early 2000s through research on contingent valuation (Corso et al. 2000). The importance of visual aids is also supported by more recent research by Zikmund-Fisher and colleagues (2014) that demonstrates the need for visual aids to overcome many respondents' low numeracy skills. Early standard gamble surveys were typically administered face-to-face using chance boards to assist respondents in conceptualizing the choices. Now such surveys are often conducted via computer either face-to-face using a tablet or over the Internet; both of these approaches can support the inclusion of visual aids to assist interpretation of the presented choices.

Violation of Constant Proportional Tradeoffs. For the time-tradeoff technique, research conducted since the original Panel has demonstrated an effect of time preference on time-tradeoff scores (Attema and Brouwer 2010; Bleichrodt and Johannesson 1997; van der Pol and Roux 2005). In other words, the amount of time traded has not been shown to be proportional when different lengths of time are used as the denominator tradeoff amount (e.g., 10 years, 20 years, or 40 years), which results in different utility scores. Further, it has been noted that the time-tradeoff question confounds preferences for the health states themselves with time preference; this is because the years of life that are “sacrificed” in the time-tradeoff come at the end of the life span and, therefore, may be valued less (because they are farther into the future). A method of correcting for time preference in the analysis of time-tradeoff data has been suggested (Johannesson et al. 1994), but has seldom been used.

Reliability and Validity of Direct Measurement Approaches. Relatively little additional research has examined the measurement properties (reliability, validity) of direct elicitation methods. Earlier work demonstrates that test-retest reliability for these methods ranges from 0.63 to 0.8 (Froberg and Kane 1989; Nease et al. 1995). Recent studies report similar levels of reliability. Recent evidence has also suggested that the standard gamble and time-tradeoff methods have ceiling effects for mild health conditions.

Chronic and Temporary Health States. Chained approaches use a modification of either the standard gamble or time-tradeoff in which the lower or upper bound is an intermediate state instead of perfect health or dead. The resulting value is then transformed onto the 0 to 1 utility scale by asking an additional “chaining” question, in which the respondent values the intermediate health state on the “perfect health” to “being dead” scale. The waiting tradeoff was designed to assess states associated with diagnostic testing; in it the utility score is derived from the respondent’s waiting time for an ideal test compared with not waiting for the actual test. The sleep tradeoff, in which the “dead” state is replaced by a “dreamless sleep,” and modified time-tradeoff, in which the conventional time-tradeoff is combined with an open-ended response, were developed to improve respondents’ understanding of the task. Further research is needed to measure the performance of these techniques and to guide the selection of methods when traditional elicitation methods are difficult to apply.

States Worse than Dead (WTD). In light of challenges around states WTD, additional alternatives for analyzing time-tradeoff data continue to emerge, such as strategies focusing on median rather than mean valuations (Lamers 2007; Li and Fu 2009; Shaw et al. 2010) or regression-based approaches that lead to alternative estimators under the rubric of "episodic random utility" (Craig and Busschbach 2009), but consensus on an ideal approach remains elusive. Under these circumstances, other research has focused on revised time-tradeoff question formats such as the "lead-time approach" (e.g., Buckingham and Devlin 2006; Devlin et al. 2011; Robinson and Spencer 2006).

Special Populations: Eliciting Preference Scores from Children. Direct elicitation approaches also pose additional methodological challenges for children's health. Briefly, school-age children and adolescents older than 12 years may be able to respond to standard gamble and time-tradeoff questions (Juniper et al. 1997). However, younger children require proxy respondents, and the natural proxy respondent, a parent, may not be able to serve as an unbiased respondent. There are other challenges, such as how or whether it is possible to include aspects of child health that do not represent easily defined domains, such as the opportunity for normal growth and development, or a change in the preference attached to a domain as a child ages. Moreover, existing health state utility instruments may be used for valuing pre-teen or adolescent health until new instruments, such as the EQ-5D-Y or CHU 9D, are available for use. An additional consideration is that for community-perspective ratings, the community would ideally also include children; however, it will not be feasible to include children of younger ages. For children, the appropriate valuation approach should be determined by the age of the child, whether existing instruments adequately cover the anticipated domains of the specific state of health or change in health being considered for valuation, and the feasibility of collecting primary data.

Psychophysical Approaches. This section provides additional details on the approach and psychometric properties for utility measurement methods derived from the psychophysical tradition: the *paired-comparison* approach, and *rating scale* methods. Alternative methods, such as magnitude scaling, have not been frequently used and are not discussed here.

Intra-rater reliability of rating scale techniques has ranged from 0.70 to 0.94 (Froberg and Kane 1989). Correlation of test-retest reliability at 1 week using a rating scale approach has been reported as 0.77 (O'Connor et al. 1987); at 1 year, another study reported a correlation of 0.49, comparing unfavorably with test-retest reliability of the time-tradeoff technique (Torrance 1976). Schunemann and colleagues (2007) report test-retest reliability of 0.86 for the Feeling Thermometer.

Rating scale methods are highly familiar to most people from a variety of everyday experiences in which they are asked to provide information on an array of experiences (e.g., sporting events, movies, levels of pain) using this technique. It has been suggested that the cognitive burden for respondents is lower than with other techniques. However, empirical work has shown that people have difficulty directly assigning a number to feelings about health states (Patrick et al. 1994). In addition, some investigators have found that individuals are unable to provide an explanation of the relationship of their responses on a rating scale to the concepts of welfare or utility that would be the foundation of decisions about resource allocation (Richardson 1994).

Issues include end-of-scale aversion and context effects (Streiner and Norman 1995; Torrance et al. 2001). Rating scale methods typically yield health utility scores that are "closer to the middle of the range" and that can be lower or higher than scores elicited using standard gamble or time-tradeoff methods. Category scaling is considered to be limited by its use of a fixed number of categories in the scaling task; people are held to be capable of making much more accurate judgments of the relative magnitude of stimuli than category scales permit (McDowell and Newell 1996).

Online Appendix 7.8: Collecting Ordinal Information [Section 7.5.2, p. 174]

There are a variety of different types of ordinal information corresponding to different modes of data collection, including:

- Discrete choice data are elicited by asking respondents to choose between two or more alternatives (health states) typically described by their levels along several dimensions. In the context of the health valuations, these choices are stated choices by which respondents indicate their selection from amongst a set of alternatives. The framing of the choice may be in terms of which state the respondent would choose to live in for some defined amount of time, or in terms of a judgment as to which state is associated with the best health level overall. Discrete choices may take the form of paired comparison, in which respondents indicate the preferred option between the two alternatives, or they may be presented as choices of the most preferred alternative amongst a choice set including more than two alternatives.
- Respondents may be asked to provide a complete rank ordering of a set of health states from the best to the worst (or vice versa). This information may be elicited either through an open-ended sorting task, or through a more structured interview protocol.

- Less commonly used in health valuations for cost-effectiveness analysis, ordered categorical response scales constitute an alternative data collection mode, although this type of information differs from rankings or paired comparisons in that no direct comparison is made between health states. Instead, respondents are asked to rate each health state individually in terms of how good or bad they regard the level of health associated with that state overall, using a defined set of response categories. For example, respondents may be asked to characterize the overall health associated with a particular state as excellent, very good, good, fair, or poor.

The starting point for imputing interval-scale (cardinal) scores from ordinal information is an assumption that choices over sets of items are related to latent cardinal values that are distributed around the mean levels for each item. Under this framework, a person may choose an item with a lower mean value than another item due to individual variability or random error. The frequency of these reversals is related to the proximity of the mean values for different items on the latent scale. Mean values that are far apart, in other words, will produce greater agreement in preferences than mean values that are close together.

A similar logic applies to a complete ordering of states if we regard this ordering as resulting from a series of discrete choices. For example, the ordering of three items, A, B, and C, may be regarded as a sequence of discrete choices, either through paired comparisons (A over B, A over C, and B over C) or choices within subsets (A from the set {A, B, C}, then B from the set {B, C}). The key assumption that allows this translation is called Luce’s choice axiom, or independence from irrelevant alternatives, which we will discuss further in Online Appendix 7.9.

Online Appendix 7.9: Alternative Approaches for Estimating MAUFs: Based on Ordinal Data [Section 7.5.2, p. 175]

As described above, there has been a recent increase in interest in estimating health valuations from ordinal data, such as those collected from ranking or discrete-choice exercises. As with data collected using more conventional elicitation methods such as the standard gamble and the time trade-off, ordinal data collection techniques may also be used in an overall estimation framework for MAUFs. In such a case, discrete choice and ranking exercises could take as stimuli health states described using generic descriptive systems such as EQ-5D, HUI, QWB, or SF-6D, but could also take other stimuli, e.g., condition-specific health-state descriptions. If stimuli are described by a standardized system with multiple dimensions, then modeling choices as a function of these dimensions allows estimation of scoring functions for the descriptive system, as described in Online Appendix 7.5. Modeling approaches to estimating MAUFs based on ordinal data are typically based on the random utility model, attributed to Luce (1959) and McFadden (1974). This model has two components: (1) a statistical model that describes the probability of ranking a particular health state higher than another, given the (unobserved) cardinal utility associated with each health state; and (2) a valuation function that relates the mean utility for a given health state to a set of explanatory variables. The latter is analogous in models of interval-scale data to the MAUFs described above. Readers are referred to several studies in the literature which describe the modeling approach in detail (e.g., McCabe et al. 2006; Salomon et al. 2003). A prominent example of the use of this approach in developing a MAUF is the recent work to develop a scoring system for the five-level version of the EQ-5D (Oppe et al. 2014).

Online Appendix 7.10: Alternative Approaches for Estimating Utility Scores [Section 7.8, p. 181]

There are situations in which the analyst cannot rely on multi-attribute measures as the source for utility scores. Many cost-effectiveness studies do not include the collection of primary data for the valuation of health outcomes. Further, in some cases the analyst has evidence from previous studies that generic preference-based measures do a poor job of distinguishing differences in health-related quality of life by severity (or other factors) and/or were not responsive in that clinical context.

A variety of alternative approaches have been developed to generate utility scores in these situations. There are two broad categories of approaches: (1) mapping or cross-walks and (2) the development of condition-specific preference-based measures (discussed above).

The Mapping Approach. The mapping approach involves imputing scores for a generic preference-based measure on the basis of information from a non-preference based measure. Mappings have been developed both for generic and condition-specific measures. In order to create the mapping algorithm based on the statistical association between the two measures, the analyst uses one or more data sets that include both measures. For instance Fryback and colleagues (1997), using data from the Beaver Dam Health Outcomes Study, created a regression-based equation to impute Quality of Well Being (QWB) scores from Short-Form 36 (SF-36) data. Similarly, Nichol and colleagues (2001) imputed HUI2 scores from SF-36 data. Revicki and colleagues (2009) developed a regression-based model to predict EQ-5D scores from patient-reported outcomes measurement information system (PROMIS) global items. A number of systematic reviews and examinations of the methodologies of mapping studies have been published (Ades et al. 2013; Brazier and Tsuchyia 2010; Brazier et al. 2010; Chan et al.

2014; Lin et al. 2013; Longworth and Rowen 2013; Mortimer and Segal 2008; Payakachat et al. 2014; Petrou et al. 2015).

A variation is the "bolt on" approach, in which a relevant attribute is added to an existing generic preference-based measure to accommodate its importance in that clinical context (Yang et al. 2015). This requires the development of a new augmented scoring function, typically based on a linear additive MAUF. An example is the addition of sleep to the EQ-5D system (Yang et al. 2014). Interestingly, the inclusion of sleep had little effect on the MAUF. Brazier and colleagues (2011) suggest that the "bolt on" approach works only if the underlying scoring function is linear additive. They report on an effort to add a pain and discomfort attribute to a condition-specific asthma measure and note that the effect of adding the new attribute to the overall scores was "not simply additive in its impact on health state values" (p. 250). In a more recent study, Yang et al. (2015, p. 58) conclude that "therefore an additive model to incorporate the bolt-on is likely to be inadequate."

A number of generalizations emerge from these reviews of mapping approaches. First, it is important to have strong empirical evidence that generic preference-based measures perform poorly in this context. An expert evaluation of content validity is not sufficient evidence to justify developing a mapping algorithm. Second, a key is the extent of overlap in attributes between the non-preference-based measure and the target preference-based measure. If there is little overlap between the two measures, the accuracy of predicted preference-based utility scores is compromised. In particular for mappings based on condition-specific measures, the fact that those measures typically do not reflect the side effects of treatment or burdens of comorbidities, can be a serious drawback.

Brazier and colleagues (2010, p. 220) summarize the role of mapping algorithms as follows: “the use of mapping functions is always a second-best solution to using a preference-based generic measure in the first place (or arguably using a preference-weighted condition-specific measure), but is often necessary for pragmatic reasons.”

Online Appendix 7.11: Alternative Approach Based on the Item-Response Theory Approach to Assessing Health Status and the Patient Reported Outcomes Measurement Information System (PROMIS) Project [Section 7.8, p. 181]

Recent applications of the item-response theory (IRT) approach to the assessment of health status, especially when linked to computer adaptive testing, provide the potential for highly efficient and precise assessments of health status that involve little burden for respondents. IRT has been widely used in educational testing for decades and has recently been applied to the assessment of health status. IRT describes in probabilistic terms “the relationship between a person’s response to a survey question and his or her level of the ‘latent variable’ being measured by the scale” (Reeve and Fayers 2005, p. 55). The latent variable, in the present context, is the domain or attribute of health status being assessed.

With computer adaptive testing, a respondent who, for instance, indicates that she is able to walk a block on flat ground without any difficulty need not be asked if she can walk from her bedroom to her front door. Instead, the respondent would be asked about her ability to participate in more strenuous activities in order to assess her level of physical function.

IRT models are unidimensional; that is, “the set of items measure a single continuous latent construct” (Reeves and Fayers 2005, p. 62). Begun in 2004, PROMIS (http://www.nihpromis.org/?AspxAutoDetectCookieSupport=1#1) has been applying IRT methods to develop item banks and scales for a broad array of domains or attributes, including anger, anxiety, depression, applied cognition-abilities, applied cognition-general concerns, pain-behavior, pain-interference, physical function, satisfaction with participation in social roles, fatigue, and sleep disturbance (Cella et al. 2010). As evidence on the test-retest reliability, cross-sectional construct validity, and responsiveness of the PROMIS scales accumulates, it should become clearer whether or not the PROMIS system provides an effective and efficient way to collect data on health status. Further, techniques for multi-dimensional IRT and other advanced methods are being developed and are beginning to be applied in health assessment (Reise and Revicki 2015).

Of course, to use such data in cost-effectiveness analysis, the analyst would need to link a preference-based scoring system to PROMIS data. There are a number of ongoing projects to develop preference-based scoring systems for various PROMIS scales (Craig et al. 2014a; Craig et al. 2014b). Revicki and colleagues (2009) have developed a regression-based model to predict EQ-5D-3L scores from PROMIS global items. The use of PROMIS scales linked to preference-based scoring systems will likely provide another useful approach for assessing and valuing outcomes for cost-effectiveness analyses. Krabbe (2013) describes an elegant model to combine the IRT and multi-attribute utility paradigms.

Online Appendix 7.12: Use of Patient-Derived Data Outside of Cost-Effectiveness Analysis [Section 7.9.2, p. 183]

The Health-Related Quality-of-Life Measurement Tradition. Health measurement outside of the field of cost-effectiveness analysis has focused almost exclusively on patient-reported outcomes. Within the field of “health-related quality of life” or “patient-reported outcomes” measurement, it is taken as given that patients are the experts in reporting how disease and treatments have affected their health (Feeny et al. 1990; Guyatt et al. 1993; Streiner and Norman

1995). To the extent to which the “source of valuation” question is discussed within this tradition, the question that arises is whether any source other than patients can be used reliably. For interventions involving children or the elderly, proxy respondents are sometimes used, and there is a significant literature evaluating the concordance between proxy and patient respondents (Epstein et al. 1989; Naglie 2007; Neumann et al. 2000; Rothman et al. 1991). Thus, it is important to note that the major research tradition in outcomes research starts from a position almost diametrically opposed to the mainstream view in cost-effectiveness analysis, and places patients’ reports regarding their own health at the center of the health measurement and valuation enterprise. Of course, that patients are often better informed about the health states associated with their condition and its treatment does not imply that their values are different or more valid.

The Growing Role of Patient Preferences in Clinical and Social Decision Making. Bensing (2000) has argued that patient-centered medicine is one of the two paradigms that has dominated modern medicine. The trend toward patient-centered medicine is characterized by an increased emphasis on patient experience rather than the patient’s disease, and an increased role for patients in decision making. The establishment of the Patient Centered Outcomes Research Institute (PCORI) in the United States and the increasing number of journals (e.g., *The Patient*), academic societies (International Shared Decision Making Conference), and publications devoted to content and methods associated with patient choice, patient preference measurement, and the growing role of patient preferences within the clinical practice guideline movement all testify to the increasing role that patients’ values and preferences are playing within the field of clinical decision making.

With respect to decision making for health systems, the role of patients is also growing. Many health technology assessment (HTA) agencies, for example, have expanded the role of patient and consumer participation. First, HTA agencies are increasingly gathering direct information about patient perspectives. This can take the form of soliciting input from patients or patient groups regarding their experiences (see, e.g., National Institute for Health and Care Excellence [NICE] 2013). Or it may take the form of formal reviews of published qualitative and quantitative studies of patients' experiences (Brooker et al. 2013). Second, HTA agencies are increasingly attempting to bring patients into the entire process of technology evaluation, starting with scoping the problem, gathering evidence, assessing value, appraising the evidence, developing recommendations, and ensuring dissemination of findings (Facey et al. 2010; Hailey and Nordwall 2006). In the Hailey and Nordwall (2006) survey of 37 HTA agencies conducted in 2005, the authors found that 57% already involved consumers/patients in some way and 83% intended to involve consumes in the future. Involvement included topic formulation, preparation of assessments, preparation of summaries of HTA results, and dissemination.

References to Online Appendices

Ades, A. E., Buobing Lu, and Jason J. Madan, “Which health-related quality of life outcome when planning randomized trials: disease-specific or generic, or both? A common factor approach.” *Value in Health*, Vol. 16, No. 1, January-February, 2013, pp. 185-194.

Altman, DG, *Practical Statistics for Medical Research*. London: Chapman & Hall, 1991.

Attema, Arthur E., and Werner F. Brouwer, “The value of correcting values: influence and importance of correcting TTO scores for time preference.” *Value in Health*, Vol. 13, No. 8, December, 2010, pp. 879-884.

Attema, Arthur E., and Werner F. Brouwer, “A test of independence in discounting from quality of life.” *Journal of Health Economics*, Vol. 31, No. 1, January, 2012, pp. 22-34.

Badia, Xavier, Montserrat Roset, Michael Herdman, and Paul Kind, “A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states.” *Medical Decision Making*, Vol. 21, No. 1, January-February, 2001, pp. 7-16.

Balaban, Donald J., Philip C. Sagi and Neil I. Goldfarb, “Weights for scoring the quality of well-being instrument among rheumatoid arthritics: a comparison to general population weights.” *Medical Care*, Vol. 24, No. 11, Nov. 1986, pp. 973-980.

Bensing, Jozien, “Bridging the Gap. The separate worlds of evidence-based medicine and patient-centered care.” *Patient Education and Counseling*, Vol. 39, No. 1, January, 2000, pp. 17-25.

Beresniak, Ariel, Antonieta Medina-Lara, Jean Paul Auray, Alain De Wever, Jean-Claude Praet, Rossana Tarricone, Aleksandra Torbica, Danielle Dupont, Michel Lamure, and Gerard Duru, “Validation of the underlying assumptions of the quality-adjusted life-years outcome: results from the ECHOUTCOME European project.” *Pharmacoeconomics*, Vol. 33, No. 1, January, 2015, pp. 61-69; published online September 18, 2014.

Beusterien, Kathleen, Nigel Leigh, Calayane Jackson, Robert Miller, Kevin Mayo, and Dennis Revicki, “Integrating preferences into health status assessment for amyotrophic lateral sclerosis: the ALS Utility Index.” *Amyotrophic Lateral Sclerosis*, Vol. 6, No. 3, September, 2005, pp. 169-176.

Bleichrodt, Han, and Magnus Johannesson, “An experimental test of a theoretical foundation for rating-scale valuations.” *Medical Decision Making*, Vol. 17, No. 2, April-June, 1997, pp. 208-216.

Brazier, John, Jennifer Roberts, and Mark Deverill, “The estimation of a preference-based measure of health status from the SF-36.” *Journal of Health Economics*, Vol. 21, No. 2, March, 2002, pp. 271-292.

Brazier, John E., and Jennifer Roberts, “The estimation of a preference-based measure of health from the SF-12.” *Medical Care*, Vol. 42, No. 9, September, 2004, pp. 851-859.

Brazier, John, Jennifer Roberts, Aki Tsuchiya, and Jan Busschbach, “A comparison of the EQ-5D and SF-6D across seven patient groups.” *Health Economics*, Vol. 13, No. 9, September, 2004, pp. 873-884.

Brazier, John, Julie Ratcliffe, Joshua A. Salomon, and Aki Tsuchiya, *Measuring and Valuing Health Benefits for Economic Evaluation*. Oxford: Oxford University Press, 2007.

Brazier, John, and Aki Tsuchyia, “Preference-based condition-specific measures of health: what happens to cross programme comparability?” *Health Economics*, Vol. 19, No. 2, February, 2010, pp. 125-129.

Brazier, John E., Yaling Yang, Aki Tsuchiya, and Donna Louise Rowen, “A review of mapping (or cross walking) non-preference based measures of health to generic preference-based measures.” *European Journal of Health Economics*, Vol. 11, No. 2, April, 2010, pp. 212-225; published online July 8, 2009.

Brazier, John, “Is the EQ-5D fit for purpose in mental health.” *British Journal of Psychiatry,* Vol. 197, No. 5, November, 2010, pp. 348-349.

Brazier, John, Donna Rowen, Aki Tsuchiya, Yaling Yang, and Tracy A. Young, “The impact of adding an extra dimension to a preference-based measure.” *Social Science & Medicine,* Vol. 73, No. 2, July, 2011, pp. 245-253.

Brooker, Ann-Sylvia, Steven Carcone, William Witteman, and Murray Krahn, “Quantitative patient preference evidence for health technology assessment: a case study.” *International Journal of Technology Assessment in Health Care,* Vol. 29, No3, July, 2013, pp. 290-300.

Buchholz, Ines, Kirsten Thielker, You-Shan Feng, Peter Kupatz, and Thomas Kohlmann, “Measuring change in health over time using the EQ-5D 3L and 5L: a head-to-head comparison of measurement properties and sensitivity to change in a German inpatient rehabilitation sample.” *Quality of Life Research,* Vol. 24, No. 4, April, 2015, pp. 829-835.

Buckingham, Ken, and Nancy Devlin, “A theoretical framework for TTO valuations of health.” *Health Economics,* Vol. 15, No. 10, October, 2006, pp. 1149-1154.

Busschbach, Jan J. V., Joseph McDonnell, Marie-Louise Essink-Bot, and Ben A. Van Hout, “Estimating parametric relationships between health description and health valuations with an application of the EuroQol EQ-5D.” *Journal of Health Economics*, Vol. 18, No. 5, October, 1999, pp. 551-571.

Cella, David, William Riley, Arthur Stone, Nan Rothrock, Bryce Reeve, Susan Yount, Dagmar Amtmann, Rita Bode, Daniel Buysse, Seung Choi, Karon Cook, Robert DeVellis, Darren DeWalt, James F. Fries, Richard Gershon, Elizabeth A. Hahn, Ji-Shei Lai, Paul Pilkonis, Dennis

Revicki, Matthais Rose, Kevin Weinfurt, Ron Hays, on behalf of the PROMIS Cooperative Group, “The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008.” *Journal of Clinical Epidemiology*, Vol. 63, No. 11, November, 2010, pp. 1179-1194.

Chan, Kelvin K. W., Andrew R. Willan, Michael Gupta, and Eleanor Pullenayegum, “Underestimation of uncertainties in health utilities derived from mapping algorithms involving health-related quality-of-life measures and potential remedies.” *Medical Decision Making*, Vol. 34, No. 7, October, 2014, pp. 863-872.

Corso, Phaedra S., James K., Hammitt, and John D. Graham, *Valuing Mortality-Risk Reduction: Using Visual Aids to Improve the Validity of Contingent Valuation*. Centers for Disease Control and Harvard School of Public Health, 2000.

Craig, Benjamin M., and Jan J. V. Busschbach, The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation.” *Population Health Metrics*, Vol. 7, No. 3, January 13, 2009; http://www.pophealthmetrics.com/content/7/1/3.

Craig, Benjamin M., A. Simon Pickard, Elly Stolk, and John E. Brazier, “US valuation of the SF-6D.” *Medical Decision Making*, Vol. 33, No. 6, August, 2013, pp. 793-803; published online April 29, 2013.

Craig, Benjamin M., Bryce B. Reeve, David Cella, Ron D. Hays, Alan S. Pickard, and Dennis A. Revicki, “Demographic differences in health preferences in the United States.” *Medical Care*, Vol. 52, No. 4, April, 2014a, pp. 307-313; published online December 26, 2013.

Craig, Benjamin M., Bryce B. Reeve, Paul M. Brown, David Cella, Ron D. Hays, Joseph Lipscomb, A. Simon Pickard, and Dennis Revicki, “US valuation of health outcomes measured using the PROMIS-29.” *Value in Health*, Vol. 17, No. 8, December, 2014b, pp. 846-853.

Devlin, Nancy J., Aki Tsuchiya, Ken Buckingham, and Carl Tilling, “A uniform time trade off method for states better and worse than dead: feasibility study of the ‘lead time’ approach.” *Health Economics*, Vol. 20, No. 3, March, 2011, pp. 348-361; published online April 29, 2010.

Dolan, Paul, “Modeling valuations for EuroQol health states.” *Medical Care*, Vol. 35, No. 11, November, 1997, pp. 1095-1108.

Drummond, Michael F., Mark J. Sculpher, George W. Torrance, Bernie O’Brien, and Greg L. Stoddart, *Methods for the Economic Evaluation of Health Care Programmes.* Third Edition. Oxford: Oxford University Press, 2005.

Epstein, Arnold M., Judith A. Hall, Janet Tognetti, Linda H. Son, and Loring Conant, Jr., “Using proxies to evaluate quality of life: can they provide valid information about patients’ health status and satisfaction with medical care?” *Medical Care*, Vol. 27, Supplement 3, March, 1989, pp. S91-S98.

Facey, Karen, Antoine Boivin, Javier Gracia, Helle Ploug Hansen, Alessandra Lo Scalzo, Jean Mossman, and Ann Single, “Patients’ perspectives in health technology assessment: a route to robust evidence and fair deliberation.” *International Journal of Technology Assessment in Health Care*, Vol. 26, No.3, July, 2010, pp. 334-340.

Feeny, David, Roberta Labelle, and George Torrance, “Integrating Economic Evaluations and Quality-of-Life Assessments,” in Bert Spilker ed., *Quality of Life Assessments in Clinical Trials*. New York, Raven Press, 1990, pp. 71-83.

Feeny, David, William Furlong, George W. Torrance, Charles H. Goldsmith, Zenglong Zhu, Sonja DePauw, Margaret Denton, and Michael Boyle, “Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 system.” *Medical Care*, Vol. 40, No. 2, February, 2002, pp. 113-128.

Feeny, David, Chris Blanchard, Jeffrey L. Mahon, Robert Bourne, Cecil Rorabeck, Larry Stitt, and Susan Webster-Bogaert, “Comparing community-preference based and direct standard gamble utility scores: evidence from elective total hip arthroplasty.” *International Journal of Technology Assessment in Health Care*, Vol. 19, No. 2, Spring, 2003, pp. 362-372.

Feeny, David, William Furlong, Saroj Saigal, and Jian Sun, “Comparing directly measured standard gamble scores to HUI2 and HUI3 utility scores: group and individual-level comparisons.” *Social Science & Medicine*, Vol. 58, No. 4, February, 2004, pp. 799-809.

Feeny, David, “The Roles for Preference-Based Measures in Support of Cancer Research and Policy,” Chapter 4 in Joseph Lipscomb, Carolyn Cook Gotay, and Claire Snyder, eds., *Outcomes Assessment in Cancer: Measures, Methods, and Applications*, New York, Cambridge University Press, 2005a, pp. 69-92.

Feeny, David, “Preference-Based Measures: Utility and Quality-Adjusted Life Years,” Chapter 6.2 in Peter Fayers and Ron Hays, eds., *Assessing Quality of Life in Clinical Trials*, Second Edition, Oxford: Oxford University Press, 2005b, pp. 405-429.

Fiest, Kirsten M., Tolulope T. Sajobi, and Samuel Wiebe, "Epilepsy surgery and meaningful improvements in quality of life: results from a randomized controlled trial." *Epilepsia*, Vol. 55, No. 6, June, 2014, pp. 886-892, published online April 15, 2014.

Fisk, J. D., M. G. Brown, I. S. Sketris, L. M. Metz, T. J. Murray, and K. J. Stadnyk, "A comparison of health utility measures for the evaluation of multiple sclerosis treatments." *Journal of Neurology, Neurosurgery, and Psychiatry*, Vol. 76, No. 1, January, 2005, pp. 58-63.

Froberg, Debra G., and Robert L. Kane, "Methodology for measuring health state preferences - II: scaling methods." *Journal of Clinical Epidemiology*, Vol. 42, No. 5, 1989, pp. 459-471.

Fryback, Dennis G., William F. Lawrence, Patricia A. Martin, Ronald Klein, and Barbara Klein, "Predicting quality of well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study." *Medical Decision Making*, Vol. 17, No. 1, January-March, 1997, pp. 1-9.

Fryback, Dennis G., Nancy Cross Dunham, Mari Palta, Janel Hanmer, Jennifer Buechner, Dasha Cherepanov, Shani Herrington, Ron D. Hays, Robert M. Kaplan, Ted Ganiats, David Feeny, and Paul Kind, "U.S. norms for six generic health-related quality of life indexes from the National Health Measurement Study." *Medical Care*, Vol. 45, No. 12, December, 2007, pp. 1162- 1170.

Furlong, William, David Feeny, George W. Torrance, Ronald Barr, and John Horsman, "Guide to Design and Development of Health-State Utility Instrumentation." McMaster University Centre for Health Economics and Policy Analysis Working Paper No 90-9, June 1990.

Furlong William J., David H. Feeny, George W. Torrance, and Ronald D. Barr, "The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies." *Annals of Medicine*, Vol. 33, No. 5, July, 2001, pp. 375-384.

Garber, A.M., M. C. Weinstein, G. W. Torrance, and M. S. Kamlet, “Theoretical Foundations of Cost-Effectiveness Analysis,” Chapter 2 in Marthe R. Gold, Joanna E. Siegel, Louise B. Russell, and Milton C. Weinstein, eds., *Cost-Effectiveness in Health and Medicine*. First Edition. New York: Oxford University Press, 1996, pp. 25-53.

Gold, Marthe R., David Stevenson, and Dennis G. Fryback, “HALYs and QALYs and DALYs, Oh My: similarities and differences in summary measures of population health.” *Annual Review of Public Health*, Vol. 23, 2002, pp. 115-134; published online October 15, 2001.

Greiner, Wolfgang, Tom Weiknen, Martin Nieuwenhuizen, Siem Oppe, Xavia Badia, Jan Busschbach, Martin Buxton, Paul Dolan, Paul Kind, Paul Krabbe, Arto Ohinmaa, David Parkin, Montserat Roset, Harri Sintonen, Aki Tsuchiya, and Frank de Charro, “A single european currency for EQ-5D health states: results from a six-country study.” *European Journal of Health Economics*, Vol. 4, No. 3, September, 2003, pp. 222-231.

Guyatt, Gordon, H., David H. Feeny, and Donald L. Patrick, “Measuring health-related quality of life.” *Annals of Internal Medicine*, Vol. 118, No. 8, April 15, 1993, 622-629.

Guyatt, Gordon H., Derek R. King, David H. Feeny, David Stubbing, and Roger S. Goldstein, “Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation.” *Journal of Clinical Epidemiology*, Vol. 52, No. 3, March, 1999, pp. 187-192.

Hailey, David, and Margareta Nordwall, “Survey on the involvement of consumers in health technology assessment programs.” *International Journal of Technology Assessment in Health Care*, Vol. 22, No. 4, Fall, 2006, pp. 497-499.

Harrison, M.J., L. M. Davies, N. J. Bansback, M. J. McCoy, S. M. M. Verstappen, K. Watson, and D. P. M. Symmons, “The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis.” *Quality of Life Research*, Vol. 18, No. 9, November, 2009, pp. 1195-205.

Hatoum, Hind T., John E. Brazier, and Kasem S. Akhras, “Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting.” *Value and Health*, Vol. 7, No. 5, September, 2004, pp. 602-609.

Hawthorne, Graeme, and Jeff Richardson, “Measuring the value of program outcomes: a review of multiattribute utility measures.” *Expert Reviews in Pharmacoeconomics Research*, Vol. 1, No. 2, 2001, pp. 215-228.

Hawthorne, Graeme, Jeff Richardson, and Neil Atherton Day, “A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments.” *Annals of Medicine*, Vol. 33, No. 5, July, 2001, pp. 358-370.

Hector, Richard D., Sr, John P. Anderson, Rosemarie C. P. Paul, Robert E. Weiss, Ron D. Hays, and Robert M. Kaplan, “Health state preferences are equivalent in the United States and Trinidad and Tobago.” *Quality of Life Research*, Vol. 19, No. 5, June, 2010, pp. 729-738.

Heintz, Emelie, Marieke Krol, and Lars-Ake Levin, “The impact of patients’ subjective life expectancy on time tradeoff valuations.” *Medical Decision Making*, Vol. 33, No. 2, February, 2013, pp. 261-270.

Horsman, John, William Furlong, David Feeny, and George Torrance, "The Health Utilities Index (HUI®): concepts, measurement properties and applications." *Health and Quality of Life Outcomes* [electronic journal], Vol. 1, October, 2003, p 54, http://www.hqlo.com/content/1/1/54.

Insinga, Ralph P., and Dennis G. Fryback, "Understanding differences between self-ratings and population ratings for health in the EuroQOL." *Quality of Life Research*, Vol. 12, No. 6, September, 2003, pp. 611-619.

Institute for Quality and Efficiency in Health Care, *General Methods for the Assessment of the Relation of Benefits to Costs*, Version 1.0, Cologne, Germany, October 19, 2009.

Janssen, M. F., A. Simon Pickard, Dominik Golicki, Claire Gudex, Maciej Niewada, Lucianna Scalone, Paul Swinburn, and Jan Busschbach, "Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study." *Quality of Life Research*, Vol. 22, No. 7, September, 2013, pp. 1717-1727.

Jia, Y. X., F. Q. Cui, L. Li, D. L. Zhang, G. M. Zhang, F. Z. Wang, X. H. Gong, H. Zheng, Z. H. Wu, N. Miao, X. J. Sun, L. Zhang, J. J. Lv, and F. Yang, "Comparison between the EQ-5D-5L and the EQ-5D-3L in patients with hepatitis B." *Quality of Life Research*, Vol. 23, No. 8, October, 2014, pp. 2355-2363.

Johannesson, Magnus, Joseph S. Pliskin, and Milton C. Weinstein, "A Note on QALYs, time tradeoff, and discounting." *Medical Decision Making*, Vol. 14, No. 2, April-June, 1994, pp. 188-193.

Johnson, Jeffrey A., Nan Luo, James W. Shaw, Paul Kind, and Stephen Joel Coons, "Valuations of EQ-5D health states: are the United States and United Kingdom different?" *Medical Care*, Vol. 43, No. 3, March, 2005, pp. 221-228.

Juniper, Elizabeth F., Gordon H. Guyatt, David H. Feeny, Lauren E. Griffith, and Penelope J. Ferrie, "Minimum skills required by children to complete health-related quality of life instruments for asthma: comparison of measurement properties." *European Respiratory Journal*, Vol. 10, No. 10, October, 1997, pp. 2285-2294.

Kaplan, Robert M., "Value judgement in the Oregon Medicaid experiment." *Medical Care*, Vol. 32, No. 10, 1994, pp. 975-988.

Kaplan Robert M., and James W. Bush, "Health related quality of life measurement for evaluation research and policy analysis." *Health Psychology*, Vol. 1, 1982, pp. 61-80.

Kaplan, Robert M., and John P. Anderson, "The General Health Policy Model: An Integrated Approach," Chapter 32 in Bert Spilker, ed., *Quality of Life and Pharmacoeconomics in Clinical Trials*. Second Edition. Philadelphia: Lippincott-Raven Press, 1996, pp. 309-322.

Keeney, Ralph L., "Building models of values." *European Journal of Operational Research*, Vol. 37, No. 2, November, 1988, pp. 149-157.

Keeney, Ralph L., and Howard Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Cambridge University Press, 1993. First Edition, John Wiley & Sons, 1976.

Kharroubi, Samer A., John E. Brazier, Jennifer Roberts, and Anthony O'Hagan, "Modelling SF-6D health state preference data using a nonparametric Bayesian method." *Journal of Health Economics*, Vol. 26, No. 3, May, 2007, pp. 597-612.

Kharroubi, Samer A., and Christopher McCabe, "Modeling HUI 2 health state preference data using a nonparametric Bayesian method." *Medical Decision Making*, Vol. 28, No. 6, November-December, 2008, pp. 875-887.

Kharroubi, Samer A., John E. Brazier, and Sarah McGhee, "Modeling SF-6D Hong Kong standard gamble health state preference data using a nonparametric Bayesian approach." *Value in Health*, Vol. 16, No. 6, September-October, 2013, pp. 1032-1045.

Knies, S., S. M. Evers, M. J. Candel, J. L. Severens, and A. J. Ament, "Utilities of the EQ-5D: transferable or not?" *Pharmacoeconomics*, Vol. 27, No. 9, 2009, pp. 767-779.

Krabbe, Paul F. M., "A generalized measurement model to quantify health: the multi-attribute preference response model." *PLOS One*, Vol. 8, No. 11, November 21, 2013, e794.94.

Lamers, Leida M., "The transformation of utilities for health states worse than death." *Medical Care*, Vol. 45, No. 3, March, 2007, pp. 238-244.

Langfitt, J. T., B. G. Vickrey, M. P. McDermott, S. Messing, A. T. Berg, S. S. Spencer, M. R. Sperling, C. W. Bazil, and S. Shinnar, "Validity and responsiveness of generic preference-based HRQOL instruments in chronic epilepsy." *Quality of Life Research*, Vol. 15, No. 5, June, 2006, pp. 899-914.

Le Galès, Catherine, Catherine Buron, Nathalie Costet, Sophia Rosman, and Pr. Gérard Slama, "Development of a preference-weighted health status classification system in France: the Health Utilities Index." *Health Care Management Science*, Vol. 5, Nol 1, 2002, pp. 41-51.

Lee, Chun Fan, Nan Luo, Raymond Ng, Nan Soon Wong, Yoon Sim Yap, Soo Kien Lo, Whay Kuang Chia, Alethea Yee, Lalit Krisnha, Celest Wong, Cynthia Goh, and Yin Bun Cheung, "Comparison of the measurement properties between a short and generic instrument, the 5-level EuroQol Group's 5-dimension (EQ-5D-5) questionnaire, and a longer and disease-specific instrument, the Functional Assessment of Cancer Therapy —Breat (FACT-B), in Asian breast cancer patients." *Quality of Life Research*, Vol. 22, No. 7, September, 2013, pp. 1745-1751.

Li, Liang, and Alex Z. Fu, "Some methodological issues with the analysis of preference-based EQ-5D index scores." *Health Services and Outcomes Research Methodology*, Vol. 9, No. 3, September, 2009, pp. 162-176.

Lin, Fang-Ju, Louise Longworth, and A. Simon Pickard, "Evaluation of content on EQ-5D as compared to disease-specific utility measures." *Quality of Life Research*, Vol. 22, No. 4, May, 2013, pp. 853-874; published online June 23, 2012.

Lo, Phoebe S. Y., Michael C. F. Tong, Dennis A. Revicki, Ching Chyi Lee, John K. S. Woo, Henry C. K. Lam, and C. Andrew van Hasselt, "Rhinitis Symptom Utility Index (RSUI) in Chinese subjects: a multiattribute patient-preference approach." *Quality of Life Research*, Vol. 13, No. 5, June, 2006, pp. 877-887.

Longworth, Louise, and Donna Rowen, “Mapping to obtain EQ-5D utility values in NICE Health Technology Assessments.” *Value in Health*, Vol. 16, No. 1, January-February, 2013, pp. 202-210.

Loomes, Graham, and Lynda McKenzie, “The use of QALYs in health care decision making.” *Social Science & Medicine*, Vol. 28, No. 4, 1989, pp. 299-308.

Luce, R. Duncan, *Individual Choice Behavior*. New York, John Wiley & Sons, 1959.

Luo, Nan, Jeffrey A. Johnson, James W. Shaw, David Feeny, and Stephen Joel Coons, “Self-reported health status of the general adult us population as assessed by the EQ-5D and Health Utilities Index.” *Medical Care*, Vol. 43, No. 11, November, 2005, pp. 1078-1086.

McCabe, Christopher, Katherine Stevens, and John Brazier, “Utility Values for the Health Utility Index Mark 2: An Empirical Assessment of Alternative Mapping Functions,” Sheffield Health Economics Group Discussion Paper 04/1, 2004.

McCabe, Christopher, John Brazier, Peter Gilks, Aki Tsuchiya, Jennifer Roberts, Anthony O’Hagan, and Katherine Stevens, “Using rank data to estimate health state utility models.” *Journal of Health Economics*, Vol. 25, No. 3, May, 2006, pp. 418-431.

McDowell, Ian, and Claire Newell, *Measuring Health: A Guide to Rating Scales and Questionnaires*. Second Edition. New York: Oxford University Press, 1996.

McFadden, Daniel, “Conditional Logit Analysis of Qualitative Choice Behavior,” in Paul Zarembka, ed., *Frontiers in Econometrics*. New York, Academic Press, 1974, pp. 105-142.

Montejo, Angel Luis, Javier Correas-Lauffer, Jorge Maurino, Guillermo Villa, Pablo Rebollo, Teresa Diez, and Louis Cordero, "Estimation of a multiattribute utility function for the Spanish version of the TooL questionnaire." *Value in Health*, Vol. 14, No. 4, June, 2011, p 564-570.

Mortimer, Duncan, and Leonie Segal, "Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights." *Medical Decision Making*, Vol. 28, No. 1, January-February, 2008, pp. 66-89.

Murray, Christopher J. L., Majid Ezzati, Abraham D. Flaxman, Stephen Lim, Rafael Lozano, Catherine Michaud, Mohsen Naghavi, Joshua A. Salomon, Kenji Shibuya, Theo Vox, Daniel Wikler, and Alan D. Lopez, "GBD 2010: design, definitions, and metrics." *Lancet*, Vol. 380, No. 9859, December 15, 2012, pp. 2063-2066.

Naglie, Gary, "Quality of life in dementia." *Canadian Journal of Neurological Science*, Vol 34, Supplement 1, March, 2007, pp. S57-S61.

National Institute for Health and Care Excellence (NICE), "Contributing to Clinical Guidelines – A Guide for Patients and Carers, Factsheet 4: Support for Patients and Carers Involved in Developing a Guideline," March 2013, https://www.nice.org.uk/media/default/About/NICE-Communities/Public-involvement/Developing-NICE-guidance/Factsheet-4-contribute-to-developing-clinical-guidelines.pdf.

National Institute on Aging, "Health and Retirement Study: A Longitudinal Study of Health, Retirement, and Aging" [web page], n.d., http://hrsonline.isr.umich.edu/index.php?p=avail.

Nease, Robert F., Terry Kneeland, Gerald T. O'Connor, Walton Sumner, Carolyn Lumpkins, Linda Shaw, David Pryor, and Harold C. Sox, "Variation in patient utilities for outcomes of the management of chronic stable angina. Implications for clinical practice guidelines." *Journal of the American Medical Association*, Vol. 273, No. 15, April 19, 1995, p 1185-1190; "Correction." *Journal of the American Medical Association*, Vol. 274, No. 8, August 23, 1995, p 612.

Neumann, Peter J., Sally S. Araki, and Elaine M. Gutterman, "The use of proxy respondents in studies of older adults: lessons, challenges, and opportunities." *Journal of the American Geriatric Society*, Vol. 48, No. 12, December, 2000, pp. 1646-1654.

Neumann, Peter J., Gillian D. Sanders, Louise B. Russell, Joanna E. Siegel, and Theodore G. Ganiats, eds., *Cost-Effectiveness in Health and Medicine*, Second Edition, New York: Oxford University Press, 2016.

Nichol, Michael B., Nishan Sengupta, and Denise R. Globe, "Evaluating quality-adjusted life years: estimation of the Health Utility Index (HUI) from the SF-36." *Medical Decision Making*, Vol. 21, No. 2, March-April, 2001, pp. 105-112.

Norman, Richard, Paula Cronin, Rosalie Viney, Madeleine King, Deborah Street, and Julie Ratcliffe, "International comparisons in valuing EQ-5D health states: a review and analysis." *Value in Health*, Vol. 12, No. 8, November-December, 2009, pp. 1194-1200.

O'Brien, Bernie J., Marian Spath, Gordon Blackhouse, J. L. Severens, Paul Dorian, and John Brazier, "A view from the bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index." *Health Economics*, Vol. 12, No. 11, November, 2003, pp. 975-981.

O'Connor, Annette M., N. F. Boyd, P. Warde, L. Stolbach, and J. E. Till, "Eliciting preferences for alternative drug therapies in oncology: Influence of treatment outcome description, elicitation technique and treatment experience on preferences." *Journal of Chronic Conditions,* Vol. 40, No. 8, 1987, pp. 811-818.

Oppe, Mark, Nancy J. Devlin, Ben van Hout, Paul F. M. Krabbe, and Frank de Charro, "A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol." *Value in Health*, Vol. 17, No. 4, June, 2014, pp. 445-453.

Oremus, Mark, Jean-Eric Tarride, Natasha Clayton, Canadian Willingness-to-Pay Study Group, and Parminder Raina, "Health utility scores in Alzheimer's disease: differences based on calculation with American and Canadian preference weights." *Value in Health*, Vol. 17, No. 1, January-February, 2014, pp. 77-83.

Patrick, Donald L., James Bush, and Milton Chen, "Methods for measuring levels of well-being for a health status index." *Health Services Research*, Vol. 8, No. 3, Fall, 1973, pp. 228-245.

Patrick, Donald L., Helene E. Starks, Kevin C. Cain, Richard F. Uhlmann, and Robert A. Pearlman, "Measuring preferences for health states worse than death." *Medical Decision Making*, Vol. 14, No. 1, January-March, 1994, pp. 9-18.

Payakachat, Nalin, J. Mick Tilford, Karen Kuhlthau, N. Job van Exel, Erica Kovacs, Jayne Bellando, Jeffrey M. Pyne, and Werner B. F. Brouwer, "Predicting health utilities for children with autism spectrum disorders." *Autism Research*, Vol. 7, No. 6, December, 2014, pp. 649-663, published online September 25, 2014.

Petrou, Stavros, Oliver Rivero-Arias, Helen Dakin, Louise Longworth, Mark Oppe, Robert Froud, and Alastair Gray, "Preferred reporting items for studies mapping onto preference-based outcome measures: The MAPS statement." *Health and Quality of Life Outcomes,* Vol. 13, No.

106, published online 01 August 2015; http://www.hqlo.com/content/pdf/s12955-015-0305-6.pdf.

Pickard, A. Simon, Maria C. De Leon, Thomas Kohlmann, David Cella, and Sarah Rosenbloom, "Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients." *Medical Care*, Vol. 45, No. 3, March, 2007, pp. 259-263.

Raat, Hein, Gouke Bonsel, Reinoud Gemke, Erik Verrips, Paul Krabbe, and Marie-Louise Essink-Bot, "Feasibility and reliability of a mailed questionnaire to obtain visual analogue scale valuations for health states defined by the Health Utilities Index Mark 3." *Medical Care*, Vol. 42, No. 1, January, 2004, pp. 13-18.

Reeve, Bryce, and Peter Fayers, "Applying Item Response Theory Modelling for Evaluating Questionnaire Item and Scale Properties," in Fayers, Peter, and Ron Hays, eds., *Assessing Quality of Life in Clinical Trials*, Second Edition. Oxford: Oxford University Press, 2005, pp. 55-73.

Reise, Steven Paul, and Dennis A. Revicki, eds., *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York: Routledge/Taylor & Francis Group, 2015.

Revicki, Dennis A., Nancy K. Leidy, F. Brennan-Deimer, C. Thompson, and A. Togias, "Development and preliminary validation of a multi-attribute Rhinitis Symptom Utility Index." *Quality of Life Research*, Vol. 7, No. 8, December, 1998a, pp. 693-702.

Revicki, Dennis A., Nancy K. Leidy, F. Brennan-Deimer, S. Sorensen, and A. Togias, "Integrating patient preferences into health outcomes assessment: the multi-attribute Asthma Symptom Utility Index." *Chest*, Vol. 114, No. 4, 1998b, pp. 998-1007.

Revicki, Dennis A., Ariane K. Kawata, Neesha Harnam, Wen-Hung Chen, Ron D. Hays, and David Cella, "Predicting EuroQol (EQ-5D) scores from the patient-reported outcomes measurement information system (PROMIS) global items and domain item banks in a United States sample." *Quality of Life Research*, Vol. 18, No. 6, August, 2009, pp. 783-791.

Richardson, J., "Cost utility analysis: what should be measured?" *Social Science & Medicine*, Vol. 39, No. 1, 1994, pp. 1-24.

Richardson, Jeff, Kompal Sinha, Angelo Iezzi, Munier A. Khan, "Modeling utility weights for the Assessment of Quality of Life (AQoL)-8D." *Quality of Life Research*, Vol. 23, No. 8, October, 2014, pp. 2395-2404.

Richardson, Jeff, Munir A. Khan, Angelo Iezzi, and Aimee Maxwell, "Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI3, 15D, QWB, and AQoL-8D multiattribute utility instruments." *Medical Decision Making*, Vol. 35, No. 3, April, 2015, pp. 276-291, published online August 26, 2014.

Robinson, Angela, and Anne Spencer, "Exploring challenges to TTO utilities: valuing states worse than dead." *Health Economics*, Vol. 15, No. 4, April, 2006, pp. 393-402.

Rothman, Margaret L., Susan C. Hedrick, Kris A. Bulcroft, David H. Hickam, and Laurence Z. Rubenstein, "The validity of proxy-generated scores as measures of patient health status." *Medical Care*, Vol. 29, No. 2, February, 1991, pp. 115-124.

Rowen, Donna, and John Brazier, “Chapter 33. Health Utility Measurement,” in Sherry Glied and Peter C. Smith, eds., *The Oxford Handbook of Health Economics*, Oxford: Oxford University Press, 2011, pp. 788-813.

Ruiz, Miguel, Javier Rejas, Javier Soto, Antonio Pardo, and Irene Rebollo, “Adaptación y validación del Health Utilities Index Mark 3 al castellano y baremos de corrección en la población española.” *Medicina clinica (Barcelona)*, Vol. 120, No. 3, 2003, pp. 89-96.

Salomon, Joshua A., Christopher J. L. Murray, T. Bedirhan Ustun, and Somnath Chatterji, “Health State Valuations in Summary Measures of Population Health,” in Christopher J. L. Murray and David B. Evans, eds., *Health Systems Performance Assessment: Debates, Methods and Empiricism*. Geneva: World Health Organization, 2003, pp. 409-433.

Salomon, Joshua A., Theo Vos, Daniel R. Hogan, Michael Gagnon, Mohsen Naghavi, Ali Mokdad, et al., “Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010.” *Lancet*, Vol. 380, No. 9859, December 15, 2012, pp. 2129-2143.

Sanmartin, Claudia, Edward Ng, Debra Blackwell, Jane Gentleman, Michael Martinez, and Catherine Simile, “Joint Canada/United States Survey of Health, 2002-03,” Statistics Canada Catologue 82M0022-XIE, Ottawa, Minister of Industry, 2004.

Schunemann, Holger J., Geoff Norman, Milo A. Puhan, Elisabeth Stahl, Lauren Griffith, Diane Heels-Ansdell, Victor M. Montori, Ingela Wiklund, Roger Goldstein, M. Jeffrey Mador, and Gordon H. Guyatt, “Application of generalizability theory confirmed lower reliability of the standard gamble than the feeling thermometer." *Journal of Clinical Epidemiology*, Vol. 60, No. 12, December, 2007, pp. 1256-1262.

Selim, A.J., W. Rogers, S. X. Qian, J. Brazier, and L. E. Kazis, "A preference-based measure of health: the VR-6D derived from the veterans RAND 12-Item Health Survey." *Quality of Life Research*, Vol. 20, No. 8, October, 2011, pp. 1337-1347; published online February 19, 2011.

Shaw, James, W., Jeffrey A. Johnson, and Stephen Joel Coons, "US valuation of the EQ-5D health states: development and testing of the D1 valuation model." *Medical Care*, Vol. 43, No. 3, March, 2005, pp. 203-220.

Shaw, James, W., A. Simon Pickard, Shengsheng Yu, Shijie Chen Vincent G. Iannacchione, Jeffrey A. Johnson, and Stephen Joel Coons, "A median model for predicting United States population-based EQ-5D health state preferences." *Value in Health*, Vol. 13, No. 2, March-April, 2010, pp. 278-288.

Spencer, Anne, "A test of the QALY model when health varies over time." *Social Science & Medicine*, Vol. 57, No. 9, November, 2003, pp. 1697-1706.

Streiner, David L., and Geoffrey R. Norman, *Health Measurement Scales. A Practical Guide to their Development and Use*. Second Edition. Oxford: Oxford University Press, 1995.

Sullivan, Patrick W., William F. Lawrence, and Vahram Ghushchyan, "A national catolog of preference-based scores for chronic conditions in the United States." *Medical Care*, Vol. 43, No. 7, July, 2005, pp. 736-749.

Tinetti, Mary E., Terri R. Fried, and Cynthia M. Boyd, "Designing health care for the most common chronic conditions—multimorbidity." *Journal of the American Medical Association*, Vol. 307, No. 23, June 20, 2012, pp. 2493-2494.

Torrance, George W., "Social preferences for health states: an empirical evaluation of three measurement techniques." *Socio-Economic Planning Sciences*, Vol. 10, No. 3, 1976, pp. 129-136.

Torrance, George W., Michael H. Boyle, Sargent P. Horwood, "Application of multi-attribute utility theory to measure social preferences for health states." *Operations Research*, Vol. 30, No. 6, November-December, 1982, pp. 1042-1069.

Torrance, George W., and David Feeny, "Utilities and quality-adjusted life years." *International Journal of Technology Assessment in Health Care*, Vol. 5, No. 4, 1989, pp. 559-575.

Torrance, George W., David H. Feeny, William J. Furlong, Ronald D. Barr, Yueming Zhang, and Qinan Wang, "multi-attribute preference functions for a comprehensive health status classification system: Health Utilities Index Mark 2." *Medical Care*, Vol. 34, No. 7, July 1996, pp. 702-722.

Torrance, George W., David Feeny, William Furlong, "Visual analogue scales: do they have a role in the measurement of preferences for health states?" *Medical Decision Making*, Vol. 21, No. 4, July-August, 2001, pp. 329-334.

Tosh, Jonathan, John Brazier, Philippa Evans, and Louise Longworth, "A review of generic preference-based measures of health-related quality of life in visual disorders." *Value in Health,* Vol. 15, No. 1, January-February, 2012, pp. 118-127.

Treadwell, Jonathan R., “Test of preferential independence in the QALY model.” *Medical Decision Making*, Vol. 18, No. 4, October-December, 1998, pp. 418-428.

Tsuchiya, Aki, Shunya Ikeda, Naoki Ikegami, Shuzo Nishimura, Ikuro Sakai, Takashi Fukuda, Chisato Hamashima, Akinori Hisashige, and Makoto Tamura, “Estimating an EQ-5D population value set: the case of Japan.” *Health Economics*, Vol. 11, No. 4, June, 2002, pp. 341-353.

Turner, Nicholas, John Campbell, Tim J. Peters, Nicola Wiles, and Sandra Hollinghurst, “A comparison of four different approaches to measuring health utility in depressed patient.” *Health and Quality of Life Outcomes*, Vol. 11, No. 81, May 9, 2013; http://www.hqlo.com/content/11/1/81.

van der Pol, Marjon, and Larissa Roux, “Time preference bias in time trade-off.” *European Journal of Health Economics*, Vol. 6, No. 2, June, 2005, pp. 107-111.

van der Pol, Marjon, Gillian Currie, Seija Kromm, and Mandy Ryan, “Specification of utility function in discrete choice experiments.” *Value in Health*, Vol. 17, No.2, March/April, 2014, pp. 297-301.

van Hout, Ben, M. F. Janssen, You-Shan Feng, Thomas Kohlmann, Jan Busschbach, Dominik Golicki, Andrew Lloyd, Luciana Scalone, Paul Kind, and A. Simon Pickard, “Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets.” *Value in Health*, Vol. 15, No. 5, July-August, 2012, pp. 708-715.

van Nooten, F. E., X. Koolman, and Werner B. F. Brouwer, “The influence of subjective life expectancy on health state valuations using a 10 year TTO.” *Health Economics*, Vol. 18, No. 5, May, 2009, pp. 549-558.

van Nooten, F. E., X. Koolman, J. J. V. Busschbach, and Werner B. F. Brouwer, “Thirty down, only ten to go? Awareness and influence of a 10-year time frame in TTO.” *Quality of Life Research*, Vol. 23, No. 2, March, 2014, pp. 377-384; published online August 14, 2013.

Wang, Qinan, William Furlong, David Feeny, George Torrance, Ronald Barr, “How robust is the Health Utilities Index Mark 2 utility function?” *Medical Decision Making*, Vol. 22, No. 4, July-August, 2002, pp. 350-358.

Wiebe, Samuel, Gordon Guyatt, Bruce Weaver, Suzan Matijevic, and Casey Sidwell, “Comparative responsiveness of generic and specific quality-of-life instruments.” *Journal of Clinical Epidemiology*, Vol. 56, No. 1, January, 2003, pp. 52-60.

Xie, Feng, Kathryn Gaebel, Kuhan Perampaladas, Brett Doble, Eleanor Pullenayegum, “Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist.” *Medical Decision Making*, Vol. 34, No. 1, January, 2014, pp. 8-20; published online March 22, 2013.

Yang, Yaling, John Brazier, and Aki Tsuchiya, “Effect of adding a sleep dimension to the EQ-5D descriptive system: a ‘bolt-on’ experiment.” *Medical Decision Making*, Vol. 34, No. 1, January, 2014, pp. 42-53; published online March 22, 2013.

Yang, Yaling, Donna Rowen, John Brazier, Aki Tsuchiya, Tracey Young, and Louise Longworth, “An exploratory study to test the impact on three ‘bolt-on’ items to the EQ-5D.” *Value in Health*, Vol. 18, No. 1, January, 2015, pp. 52-60.

Zikmund-Fisher, Brian J., Holly O. Witteman, Mark Dickson, Andrea Fuhrel-Forbis, Valerie C. Kahn, Nicole L. Exe, Melissa Valerio, Lisa G. Holtzman, Laura D. Scherer, and Angela

Fagerlin, “Blocks, ovals, or people? Icon type affects risk perception and recall of pictographs.” *Medical Decision Making*, Vol. 34, No. 4, May, 2014, pp. 443-453.