When QALYs Don't Work: Measuring and Using General Time Equivalents

F. Reed Johnson
Senior Fellow and Principal Economist
RTI Health Solutions

PO Box 12194
Research Triangle Park, NC 27709
(919) 541-5958
frjohnson@rti.org

RTI(h)(s)™

[Measures That Count]

When QALYs Don't Work: Measuring and Using General Time Equivalents

F. Reed Johnson
Senior Fellow and Principal Economist
RTI Health Solutions

PO Box 12194
Research Triangle Park, NC 27709
(919) 541-5958
frjohnson@rti.org

Rev. 9/20/04

When QALYs Don't Work

Quality adjusted life years (QALYs) are a common metric used to measure health impacts of diseases and their treatments. Utility weights are values between zero and one that indicate the relative importance or seriousness of health outcomes. They integrate mortality, morbidity, and preferences into a single index number representing the equivalent years of perfect health. For example, a health state with a utility weight of 0.5 indicates that two years in that health state is equivalent to one year of perfect health.

While this idea has the advantage of simplicity, using QALYs as a measure of effectiveness in cost-utility analysis can be problematic because such simplicity requires unrealistic assumptions about patient preferences. Because of the assumptions required and methods available for constructing utility weights, QALYs often don't work in many situations that are relevant for health-care decision making. For example:

- When utility weights are applied to acute conditions such as a few hours of pain, calculated QALY gains often are tiny, although avoiding the acute condition may be clearly important to the patients affected.
- When utility weights are applied to new treatments that involve such improvements as dosage, mild side-effects, or convenience, QALY gains also often are tiny, although studies indicate patients may strongly prefer the new treatment relative to existing alternatives.

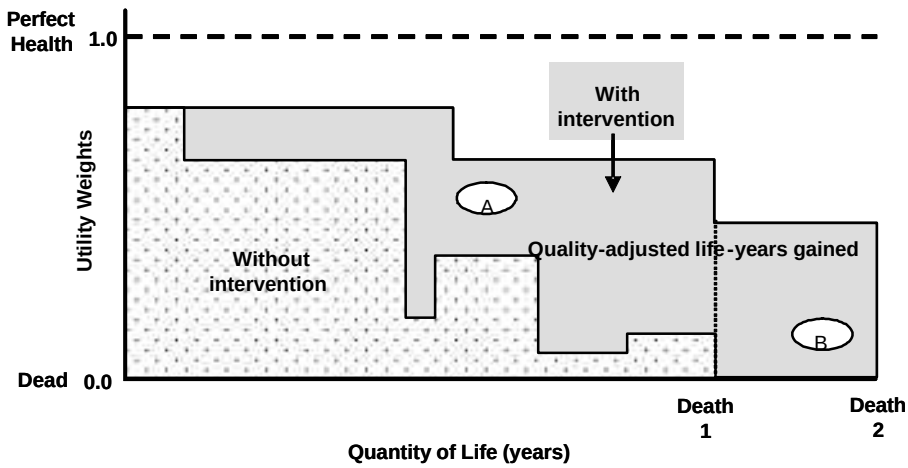
QALYs do not work well in circumstances where QALY assumptions misrepresent patients' actual preferences and perceptions. QALY assumptions are much stronger than those generally imposed in utility studies other than health applications. This paper discusses the implications of these assumptions when the underlying utility function actually has a more realistic form, proposes an alternative approach based on principles and methods used generally in other areas of applied economics, and illustrates this approach using empirical estimates for acute respiratory and cardio-vascular symptoms.

QALY Assumptions

While QALYs provide a simple, useful method for incorporating quality-of-life differences in economic analyses of medical interventions, reducing complex relationships among dynamic health states to a single index requires several well-known assumptions

(Pliskin, et al., 1980; Drummond, et al., 1997; Bootman, Townsend, and McGhan, 1996).[1] Together, these assumptions allow us to calculate increases in quality-adjusted life-years gained as shown in Figure 1. The intervention changes the lifetime health profile so that the patient gains area A quality improvement over the pre-intervention life expectancy of Death 1 years plus area B for the longevity extension to year Death 2. Given a set of fixed utility weights corresponding to all possible health conditions, calculating the areas A and B is a matter of multiplying constants times durations of expected clinical outcomes.

Figure 1. Increase in QALYs following an intervention

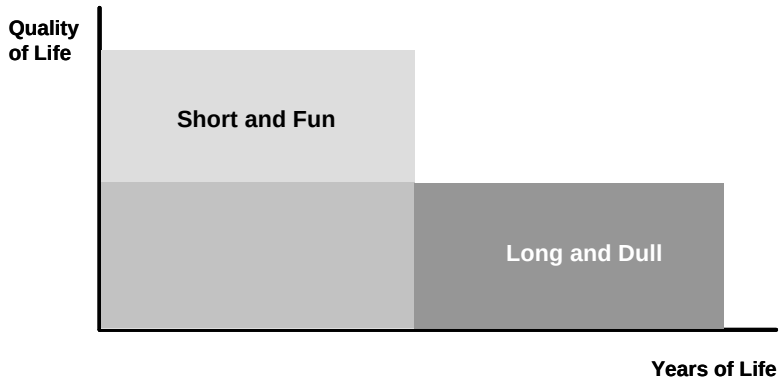


Unfortunately, the linearity that makes Figure 1 possible is a very strict requirement which is inconsistent with economic theory and most observed behavior (Johannesson, 1996). The mismatch between QALY assumptions and how patients actually experience ill health may lead to incorrect conclusions. Figure 2 illustrates one possible consequence of QALY assumptions. Although the areas under the two life profiles are the same, most people would not regard a short, but happy life as having exactly the same utility as a long, but dull life. This illustration suggests that experienced utility may be rather more complicated than that described by a set of constant weights. Thus experienced utility is likely to include preferences for health-care process or context factors as well as be nonlinear in time. Such

[1] These assumptions include mutual utility independence, constant and proportional tradeoffs, and risk neutrality. These assumptions result in linear utility functions with constant slopes that provide the utility weights for

factors often are important determinants of patient satisfaction related to alternative health-care interventions. Some examples include privacy of health-history information, dosage frequency for attention-deficit medications, diagnostic information such as genetic testing that may not affect therapeutic decisions, and differential convenience of devices such as insulin syringes and pens.

Figure 2. Two lives with equal QALYs



Finally, preferences for various health outcomes and health care process factors are likely to vary across patients with different health histories, ages, genders, education, income, and caregiving settings. If such factors are correlated with a set of clinical outcomes, then conventional QALY calculations may further distort experienced utility.

Bias in QALY Estimates

Figure 3 illustrates how nonlinearity and inclusion of nonclinical utility factors may bias QALY calculations. Starting from a condition of normal health, suppose a patient experiences a spell of ill health of duration ?Time. The corresponding reduction in QALYs is shown by U_Q, which has constant slope (Utility Weight – 1)/ ?Time. However, following assumptions normally employed in economics, "disutility" curve U_1 describes utility as nonlinear.[2] ?QALYs understate the actual utility loss by (U_{QALY} – U_{nonlin}). If utility also depends on nonclinical factors, as shown by curve U_2, then actual utility falls further to U_{other}. Finally, if the duration of

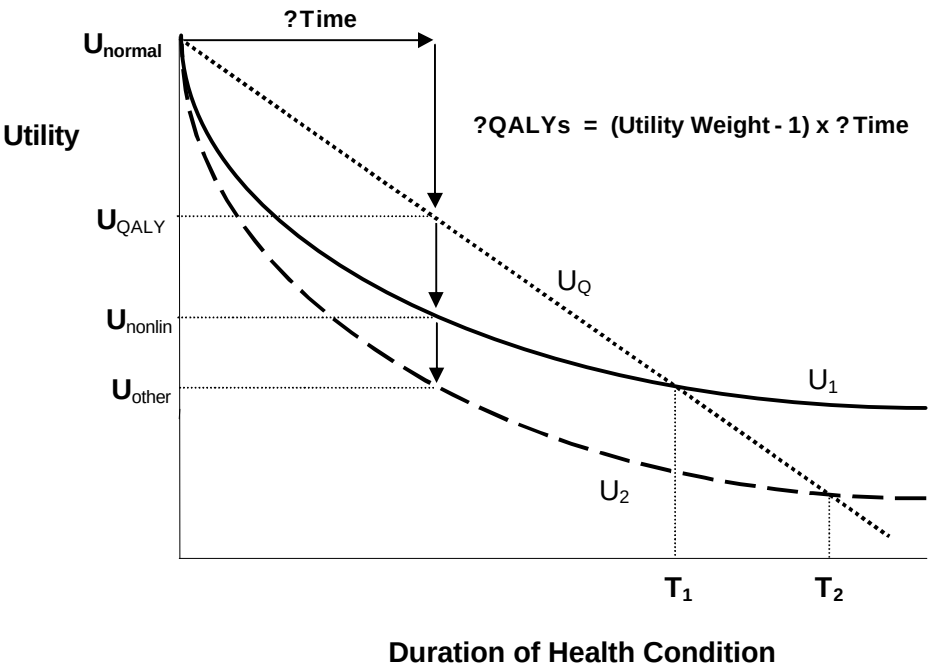
calculating QALYs.

[2] Economists generally assume preferences are subject to the law of diminishing marginal utility. That means that additional units of consumption yield less additional utility than previous units. In other words, the first cup of coffee of the day provides more satisfaction than the tenth cup of coffee the same day.

the illness is long enough, U_Q crosses U_1 and U_2. When U_Q is below nonlinear utility, QALYs overstate the utility loss of the health condition.

Note, however, that linear QALYs measure utility losses correctly at durations T_1 and T_2. This result suggests that bias in QALY measures is in part simply the familiar problem of extrapolating outside the range of the data used in estimation.

Figure 3. Bias in QALY Estimates

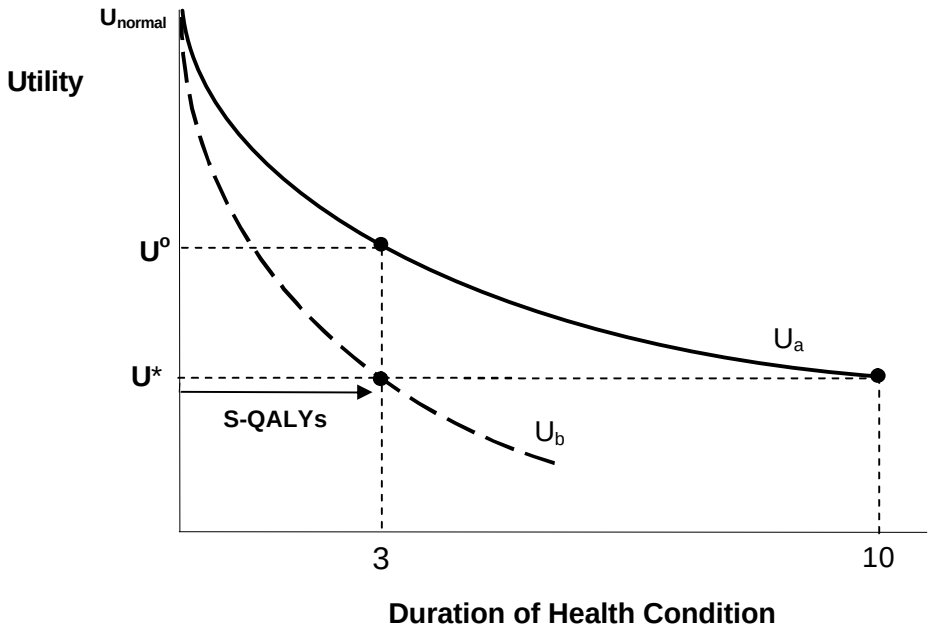


A Better QALY

We can use the basic idea of utility equivalents to construct a much better time-equivalent measure. We call this measure "super QALYs" (S-QALYs). Figure 4 illustrates how to construct this general time-equivalent measure. U_a is the utility of some relative mild condition such as a mild case of the common cold and U_b is the utility of a more severe condition such as acute bronchitis. A patient with given personal characteristics experiences a loss of utility from U_{normal} to U^o if the duration of the mild condition is 3 days. Now suppose

Having the serious condition for 3 days reduces utility to U*. To decrease utility to the same level as that experienced with the serious condition, the time spent in the mild condition would have to be 10 days. Thus 3 days in the serious condition has the same utility, or S-QALYs, as 10 days in the mild reference condition. That indicates that bronchitis has an average "weight" of 0.3 per day, relative to mild respiratory symptoms. However, this weight is not constant, but depends on both the reference health condition and the reference duration.

Figure 4. Super QALYs



Empirical Methods for Measuring Time Equivalents

Measuring Conventional Utility Weights

QALYs are intended to measure people's preferences for health over different time periods. Thus it is necessary to survey these preferences systematically. Three different methods have been used to construct the utility weights used to calculate QALYs for both general population and patient samples: rating scale, standard gamble, and time tradeoff. Most rating or "thermometer" scales ask subjects to rate a specified health state on a scale ranging from death to full health. The results then are normalized to range between 0 and 1 to obtain the resulting utility weight. Although the rating scale is easy to administer, it involves no tradeoffs among health states or between health states and time. Thus it has no utility-theoretic basis (Louviere, 1988). In addition, empirical experiments have shown only a loose, and highly non-linear, relationship between rating scale utility estimates and utility weights derived using time trade-off techniques (O'Leary et al. 1995; Torrance et al., 2001).

Standard gamble (SG), which is derived from von Neumann and Morgenstern's (1944) theory of utility under uncertainty, is a method for eliciting theoretically consistent utility weights. Respondents are asked to choose between a pair of health states *A* and *B*. *A* represents a chronic condition of reduced health. *B* is a gamble between full health (probability p) or immediate death (probability 1-p). The probability is varied until the respondent is indifferent between health state *A* and health state *B*. This indifference probability represents the utility weight for health-state *A*, subject to the assumptions discussed previously. However, many subjects are not able to reliably manipulate probability relations and the SG task itself does not correspond to realistic or familiar evaluation problems. In addition, many subjects refuse to accept tradeoffs between mortality risks and acute, self-limiting conditions.

The third commonly used method of measuring theoretically consistent utility weights is time tradeoff (TTO). Subjects are asked to evaluate living the remainder of their life in an impaired health state versus living a shorter number of years in full health. The number of years in full health is varied until the respondent is indifferent between the two choices. The utility weight is the ratio between healthy years and impaired years at the point of indifference.

TTO appears to be an easier task than SG for subjects, but poses similarly unrealistic choices and requires evaluating longevity tradeoffs even when reduced life expectancy is not a relevant health outcome.

Measuring Actual Utility

Methods for collecting the tradeoff information necessary to specify the illness disutility curves in Figure 4 are readily available in the form of well-established conjoint or choice-format stated-preference survey techniques. Analysts use conjoint analysis (CA) to quantify preferences for a variety of market and nonmarket goods and services. These include medical interventions, pharmaceutical treatments, and environmental health risks (Bryan et al. 1988; Johnson et al. 2000; Johnson et al. 1998; Johnson and Desvousges 1997; Ryan and Hughes 1997; Viscusi, Magat, and Huber 1991; Wittink and Cattin 1989).

CA recognizes that commodities have value because of their characteristics or attributes. People have preferences for each attribute and are willing to accept tradeoffs among them. CA encourages subjects to explore their preferences for various attribute combinations through a series of judgments. CA uses the observed pattern of answers to the series of tradeoff tasks to assess the implicit weights subjects assign to various outcome attributes. CA methods are sufficiently flexible to measure the nonlinearities and incorporate both clinical and nonclinical factors that influence perceived health utility.

Using a choice-format conjoint survey, Johnson et al., (2000) estimated elicited nonlinear health utility functions for several acute respiratory and cardio-vascular symptoms with durations between 1 and 10 days. Figure 5 is an example of the choice-format tradeoff task used in this study. Including both cost and duration makes it possible to derive both nonlinear time-equivalent and money-equivalent measures of health preferences.

Figure 5. Example of Choice-Format Tradeoff Task

Category	Initial Condition	Alternative A	Alternative B
Duration of episode	5 days	1 day	Your level of health as it is today. You do not experience this episode.
Symptom	Shortness of breath and swelling in ankles and feet	Shortness of breath and swelling in ankles and feet	
Daily Activities	• Are in hospital • Need help caring for yourself	• CANNOT leave your house, go to work, go to school, do housework, or participate in social or recreational activities • Have SOME physical limitations • Need help caring for yourself	
Additional costs to your household	$0	$300	$500

1 Prefer Initial Condition	2 Prefer Alternative A	3 Prefer Alternative B

Please circle a number from 1 to 3 that best reflects your choice if faced with these options.

Figure 6 plots disutility curves for three conditions, with maximum duration equal to 10 days. The logarithmic specification for duration, chosen based on specification tests, gives strongly diminishing increments in utility over durations for these acute symptoms, in contrast to QALY assumptions.

Figure 6. Empirical Utility Functions for Three Symptom/Activity Limitation Conditions

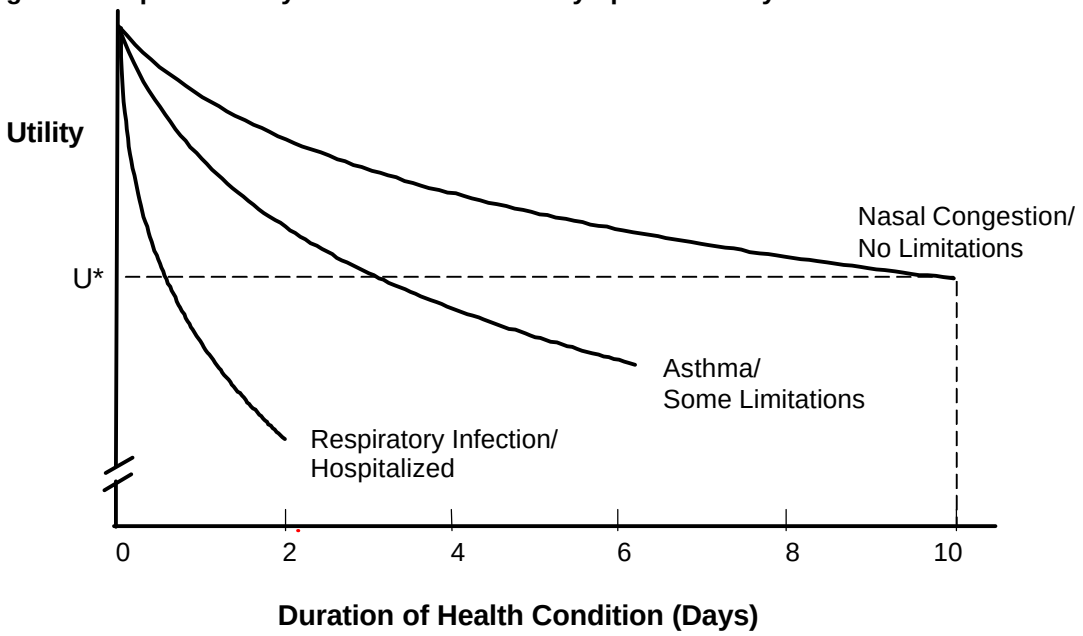


Table 1 compares some illustrative estimates for a reference duration of 5 days, using nasal congestion with no activity restriction as the reference condition. Conventional linear QALYs (actually quality-adjusted life *days*) shown are derived from the Quality of Well Being (QWB) score (Kaplan and Anderson, 1996, et al.) multiplied by 5 days.[3] S-QALYs use nasal congestion with no activity limitations as the reference condition, while conventional QALYs use perfect health. Nevertheless, the S-QALY reference condition is very close to a QWB weight of 1.

Conventional QALYs range between 2.35 days for 5 hospitalized days to 3.85 for 5 days of the mildest outcome. In contrast S-QALYs range between 0.17 days for the most severe outcome to 4.5 for the mildest outcome. As predicted by theory, conventional QALYs appear to understate milder conditions and overstate more serious conditions. They also are less sensitive to variations in outcomes. The worst condition in the table is about 61% (2.35/3.85) of the best condition measured in QALYs, but is only about 4% (0.17/4.5) of the

[3] The annualized QWB values are equal to the 5-day values multiplied by 5/365.

best condition measured in S-QALYs. Willingness to pay (WTP) to avoid the best condition is about 6% of WTP to avoid the worst condition.

Table 1. QALYs, S-QALYs and WTP for 5 Days of the Indicated Health Outcome

Outcomes		QALYs* (Days)	S-QALYs§ (Days)	WTP§ for improve-ment to best outcome (CAN$)	WTP per QALY (CAN$)	WTP per S-QALY (CAN$)
Symptom	**Activity Restriction**					
Stuffy/runny nose and sore throat	You have some physical limitations (trouble bending, stooping, or doing vigorous activities)	3.85	4.5	30	26	60
Stuffy/runny nose and sore throat	You cannot leave your house, go to work, go to school, do housework, participate in social or recreational activities, and you have some physical limitations	2.93	2.4	189	91	73
Coughing or wheezing with fever, chills or aching all over	You cannot leave your house, go to work, go to school, do housework, participate in social or recreational activities, and you have some physical limitations	2.49	1.0	368	147	92
Coughing or wheezing with fever, chills or aching all over	You are in hospital and need help caring for yourself (feeding, bathing, dressing, toilet)	2.35	0.17	542	205	112

* QWB score x 5 days
§ Calculated from estimates in Johnson et al. (2000).

Advantages and Disadvantages of S-QALYs

The strong assumptions underlying conventional QALYs may give decision makers the impression that health improvements can be measured with a metric that looks like a ruler. In fact, QALYs function more like shoe sizes than rulers. Two different shoes that are labeled with the same size often do not fit the same foot because feet are complex shapes. Utility functions typically are nonlinear and include multiple factors that describe individual preferences over alternative outcomes and processes. Nevertheless, conventional QALY assumptions are attractive because they make empirical estimation easy. Unfortunately, both the theoretical and empirical examples indicate that distortions from QALY assumptions can be significant.

S-QALYs have significant advantages over conventional time-equivalent measures.

- Tradeoffs among illness severity and duration depend realistically on patients' personal characteristics and their actual experienced utility losses over time.
- S-QALYs can incorporate factors influencing patient satisfaction other than clinical outcomes.
- S-QALYs can accurately measure utility losses for both acute and chronic conditions.
- S-QALYs do not require subjects to trade off mortality risks when mortality is not a realistic disease endpoint.

Nevertheless, these improvements in validity of time-equivalent measures come at the cost of greater complexity.

- Relative to applying a previously estimated QALY utility weight, estimating S-QALYs is far more expensive, complicated, and time consuming.
- It is no longer possible to use a single utility-weight constant to aggregate time equivalents over time and across patients.
- The reference condition is no longer perfect health, but some mild ill-health condition. While such a condition may be a more realistic comparison, the measure loses some of the intuitive appeal of conventional QALYs.

The additional effort to obtain more accurate measures of patients' actual preferences may be justified when QALYs simply don't work. As we have seen, QALYs don't work when their strong, simplifying assumptions neglect important features of the treatment and context that are nevertheless important to patients. QALYs also work poorly when utility weights are applied to durations quite different that those used to estimate the weights. Of course, extrapolating outside the range of the data is regarded as inappropriate for other kinds of quantitative analysis, but is commonly practiced in health utility analysis.

Fortunately, the methods used to estimate full, nonlinear health utility functions are similar to familiar conjoint methods used in health economics, market research, transportation economics, and environmental economics. The resulting estimates allow us to adjust time equivalent measures for all features and context factors that are important to patients, as well as the personal characteristics of the patients themselves. As Gafni and Birch (1995) once commented:

"When asking the public to assist in determining health priorities, we should use techniques that allow people to reveal their true preferences. If not, why bother asking them at all? "

References and Additional Reading

Bala, M., Wood, L., Zarkin, G., Norton, E., Gafni, A., O'Brien B., 1999. Are Health States "Timeless"? The Case of the Standard Gamble Method. Journal Clinical Epidemiology 52, 1047--1053.

Bala, M., Wood, L., Zarkin, G., Norton, E., Gafni, A., O'Brien, B., 1998. Valuing Outcomes in Health Care: A Comparison of Willingness to Pay and Quality-Adjusted Life-Years. Journal Clinical Epidemiology 51, 667--676.

Bala, M., Zarkin, G., 1999. Are QALYs an appropriate measure for valuing morbidity in acute diseases? Health Economic Letters 3, 10--14.

Bleichrodt, H., 1995. QALYs and HYEs: Under what conditions are they equivalent? Journal of Health Economics 14, 17--37.

Bleichrodt, H., Herrero, C., Pinto, J., 2002. A proposal to solve the comparability problem in cost-utility analysis. Journal of Health Economics 21, 397--403.

Bleichrodt, H., Johannesson, M., 1997. An Experimental Test of the Theoretical Foundation for Rating-Scale Valuations. Medical Decision Making 17 2, 208--216.

Bleichrodt, H., Quiggin, J., 1999. Life-cycle preferences over consumption and health: when is cost-effectiveness analysis equivalent to cost–benefit analysis? Journal of Health Economics 18 6, 681--708.

Broome, J., 1993. Qalys. Journal of Public Economics 50, 149--167.

Bryan, Stirling, Martin Buxton, Robert Sheldon, and Alison Grant. 1998. "Magnetic Resonance Imaging for the Investigation of Knee Injuries: An Investigation of Preferences." Health Economics 7:595-603.

Buckingham, K., 1993. A note on HYE. Journal of Health Economics 12, 311--323.

Culyer, A., A, Wagstaff., 1993. QALYs versus HYEs. Journal of Health Economics 12, 311--323.

Dey, Aloke. 1985. Orthogonal Fractional Factorial Designs. New York: Halstead Press.

Dolan, P., Edlin, R., 2002. Is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis? Journal of Health Economics 21, 827--843.

Dolan, P., Olsen, J., 2000.The equity-efficiency trade-off across different streams of health. Working Paper, University of Sheffield.

Drummond, M., O'Brien, B., Stoddard, G., Torrance, G., 1997. Methods for the Economic Evaluation of Health Care Programmes, Second Edition. Oxford University Press, United Kingdom.

Feeny, D., Torrance, G., 1989. Incorporating utility based quality of life assessment measures in clinical trials. Medical Care 27 3, S190--S204.

Franic, D., Pathak, D., Gafni, A., 2003. Are health states 'timeless'? A case study of an acute condition: post-chemotherapy nausea and vomiting. Journal of Evaluation in Clinical Practice 9 1, 69--82.

Gafni, A., Birch, S., 1995. Preferences for outcomes in economic evaluation: An economic approach to addressing economic problems. Social Science and Medicine 40, 767-776.

Gafni, A., Birch, S., Mehrez, A., 1993. Economics, health and health economics: HYEs versus QALYs. Journal of Health Economics 12, 325--339.

Gold, M., Siegel, J., Russell, L., Weinstein, M., 1996. Cost-Effectiveness in Health and Medicine: The Report of the Panel on Cost-Effectiveness in Health and Medicine. Oxford University Press, New York.

Hammitt, J., QALYs vs WTP. 2002. Risk Analysis 22, 985—1001.

Hirth, R., Mchernew, E., Miller, A., Fendrick, W., Weissert. 2000. Willingness to Pay for a QALY: In Search of a Standard. Medical Decision Making 20, 332--342.

Jansen, S., Stigglebout, A., Wakker, P., Vliet Vlieland, T., Leer, J., Nooy, M., Kievit, J., 1998. Patients utilities for cancer treatments: a study of the chained procedure for standard gamble and time trade off. Medical Decision Making 18, 391--399.

Johannesson, M., 1995. Qalys:. A Comment. Journal of Public Economics, 56, 327--328.

Johannesson, M., 1996. Theory and Methods of Economic Evaluation of Health Care. Kluwer Academic Publishers, The Netherlands.

Johannesson, M., Pliskin, J.S., Weinstein, M.C., 1993. Are healthy years equivalents an improvement over quality adjusted life years? Medical Decision Making 13, 281--286.

Johnson, F. R., Fries, E., and Banzhaf, H. S. 1997. "Valuing Morbidity: An Integration of the Willingness-to-Pay and Health-Status Index Literatures." Journal of Health Economics 16:641-665.

Johnson, F. R., Banzhaf, M. and Desvousges, W. 2000. "Willingness to Pay for Improved Respiratory and Cardiovascular Health: A Multiple-Format Stated-Preference Approach." Health Economics 9:295–317.

Johnson, F. R., Desvousges, W. H., Ruby, M., Stieb, D., and De Civita, P. 1998. "Eliciting Stated Preferences: An Application to Willingness to Pay for Longevity." Medical Decision Making 18 Suppl:S57-S67.

Johnson, F. R. and Desvousges, W. H. 1997. "Estimating Stated Preferences With Rated-Pair Data: Environmental, Health, and Employment Effects of Energy Programs." Journal of Environmental Economics and Management 34:79–99.

Kaplan, R.M., Anderson, J.P., 1996. The General Health Policy Model: An Integrated Approach in: Spilker, B. (Eds.), Quality of Life and Pharmacoeconomics in Clinical Trials. Lippincott Williams & Wilkins, Philadelphia.

Klose, T., 2003. A utility-theoretic model for QALYs and willingness to pay. Health Economics 12 1, 17--31.

Lenert, L.A., Cher, D.J., Goldstein, M.K., Bergen, M.R., Garber A.M., 1998 The effect of search procedures on utility elicitations. Medical Decision Making 18, 76--83.

Louviere, J.J. 1988. Analyzing Decision Making: Metric Conjoint Analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–067. Sage Publications, California.

Mehrez, A., Gafni, A., 1989. Quality Adjusted Life Years, Utility Theory, and Healthy-Years Equivalents. Medical Decision Making 9,142--149.

Mehrez, A., Gafni, A., 1991. The healthy years equivalents: How to measure them using the standard gamble approach. Medical Decision Making11, 140--146.

Mehrez, A., Gafni, A., 1993. Healthy-years equivalents versus quality-adjusted life years: In pursuit of progress. Medical Decision Making 13, 287--292.

Nord, E. 1992. Methods for Quality Adjustment of Life Years. Social Science and Medicine 34, 559--569.

O'Leary, J.F., Fairclough, D.L., Jankowski, M.K., Weeks, J.C., 1995. Comparison of Time-tradeoff Utilities and Rating Scale Values of Cancer Patients and Their Relatives: Evidence for a Possible Plateau Relationship. Medical Decision Making 15,132--137.

Pliskin J., Shepard D., Weinstein M., 1980. Utility functions for life years and health status. Oper Res 28, 206–224.

Prieto, L., Sacristan, J.A., 2003. Problems and solutions in calculating quality-adjusted life years (QALYs), Health and Quality of Life Outcomes 1, 80.

Richardson, J., 1994. Cost Utility Analysis: What Should be Measured? Social Science and Medicine 39, 7--21.

Ryan, M., Farrar, S., 2000. Using conjoint analysis to elicit preferences for health care. British Medical Journal 320 7248,1530 .

Ryan, Mandy, and Jenny Hughes. 1997. "Using Conjoint Analysis to Assess Women's Preferences for Miscarriage Management." Health Economics 6:261-273.

Ryan, Mandy, and Jenny Hughes. 1997. "Using Conjoint Analysis to Value Alternative Models of Antenatal Care." Soc. Sci. Med. 44(3):371-380.

Torrance, G.W., 1986. Measurement of Health State Utilities for Economic Appraisal: A Review. Journal of Health Economics 5, 1--30.

Torrance, G.W., Feeny, D.K., Furlong, W.J., Barr, R.D., Zhang, Y., Wang Q. 1996. Multi-attribute Utility Function for a Comprehensive Health Status Classification System: Health Utilities Index Mark 2. Medical Care 34, 702--722.

Viscusi, W. Kip, Wesley A. Magat, and Joel Huber. 1991. "Pricing Environmental Health Risks: Survey Assessments of Risk-Risk and Risk-Dollar Trade-offs for Chronic Bronchitis." Journal of Environmental Economics and Management 21:32–51.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, New Jersey.

Weinstein, M.C., Stason, W.B., 1976. Hypertension: A Policy Perspective. Harvard University Press, Massachusetts.

Wittink, Dick R., and Philippe Cattin. 1989. "Commercial Use of Conjoint Analysis: An Update." Journal of Marketing 53:91–96.