Cited as:

Kaplan, R.M. (1995). Utility assessment for estimating quality-adjusted life years. In F. Sloan (ed.) Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies. Boston: Cambridge University Press, pp. 31-60.

CHAPTER 3

Utility assessment for estimating quality-adjusted life years

Robert M. Kaplan

Programs in health care have varying objectives. The objective of prenatal care might be a reduction in infant mortality. Rheumatologists strive to make their patients more functional, whereas primary care providers often focus on shortening the cycle of acute illness. All of these providers are attempting to improve the health of their patients. However, they each measure health in a different way. Comparing the productivity of a rheumatologist with that of a neonatologist may be like comparing apples to oranges.

The diversity of outcomes to health care has led many analysts to focus on the simplest common ground, typically, mortality or life expectancy. Those who are alive are statistically coded as 1, and those who are dead are statistically coded as 0. Mortality allows the comparison between different diseases. For example, we can state the life expectancy of those who will eventually die of heart disease and compare it to the life expectancy of those who eventually die of cancer. The difficulty is that everyone who remains alive is given the same score. A person confined to bed with an irreversible coma is alive and is counted the same as someone who is actively playing volleyball at a picnic. Utility assessment, on the other hand, allows the quantification of levels of wellness on the continuum anchored by death and optimum function.

This chapter reviews the concept of utility in relation to the evaluation of cost–effectiveness of pharmaceutical products. The concept of quality-adjusted life years and the related concept of utility are first reviewed. Then, methods of utility assessment are considered. Differences in economic and psychological approaches to utility assessment are reviewed and evaluated, as well as practical issues relevant to whose preferences should be used in the model. Finally, applications of cost–effectiveness models in resource allocation are reviewed.

Conceptual framework

To evaluate health-related quality of life, one must consider all of the different ways that illness and its treatment affect outcomes. Health concerns can be reduced to two categories: life duration and quality of life. Individuals are concerned about illness, disability, and effects of treatment because they can affect life expectancy and quality of life. Assessment of a pharmaceutical treatment should consider a few basic questions:

1. Does the illness or its treatment make life last a shorter duration of time?
2. Does the condition or its treatment make life less desirable and, if so, how much less desirable?
3. What are the duration effects: how much life is lost or how long is the period of undesirable health effects?

This chapter focuses on the second issue. Determining how illness or treatment affects desirability of life is a matter of preference or utility. Such evaluations require that health states be compared to one another.

Within the last few years interest has been growing in using quality of life data to help evaluate the benefits of health care programs. In cost–effectiveness analysis, the benefits of medical care, behavioral interventions, or preventive programs can be expressed in terms of well years. Others have chosen to describe outcomes in quality-adjusted life years (QALYs; Weinstein and Stason 1976) or health years of life (Russell 1986). The term "QALY" has become most popular and is therefore used here. QALYs integrate mortality and morbidity to express health status in terms of equivalents of well years of life. If a woman dies of lupus at age fifty and one would have expected her to live to age seventy-five, the disease was associated with twenty-five lost life years. If 100 women died at age fifty (and also had a life expectancy of seventy-five years), 2,500 (100 × 25 years) life years would be lost. Yet, death is not the only outcome of concern in lupus. The disease leaves many adults somewhat disabled over long periods of time. Although still alive, the quality of their lives has diminished. QALYs take into consideration the quality-of-life consequences of these illnesses. For example, a disease that reduces quality of life by one-half will take away 0.5 QALYs over the course of one year. If it affects two people, it will take away 1 QALY (2 × 0.5) over a one-year period. A medical treatment that improves quality of life by 0.2 for each of five individuals will result in the equivalent of 1 QALY if the benefit is maintained over a one-year period. This system has the advantage of considering both benefits and side effects of programs in terms of the common QALY units. Although QALYs are typically assessed for patients, they can be measured for others, including caregivers who are placed at risk because they experience stressful life events.

The concept of relative importance

Dimensions of quality of life

Nearly all health-related quality-of-life measures have multiple dimensions, such as pain and lack of mobility. The exact dimensions vary from measure to measure. There is considerable debate in the field about which dimensions should be included (Wiklund et al. 1992). For example, the most commonly included dimensions are physical functioning, role functioning, and mental health. The Medical Outcomes Study (MOS) includes eight health concepts (Stewart and Ware 1993). Although many questionnaires include different dimensions, they still may be tapping the same constructs. For example, a measure without a mental health component does not necessarily neglect mental health. Mental health symptoms may be included and the impact of mental health, cognitive functioning, or mental retardation may be represented in questions about role functioning. Some measures have multiple dimensions for mental health symptoms, whereas others include fewer items and ask about problems in general. Although a common strategy is to report outcomes along multiple dimensions, it is not clear that multiple dimensions are more capable of detecting clinical differences. This remains an empirical question for systematic analysis.

Relative importance of dimensions

Most treatments have side effects as well as benefits. Generally, the frequencies of various side effects are tabulated. Thus, a medication to control high blood pressure might be associated with low probabilities of dizziness, tiredness, impotence, and shortness of breath. The major challenge is in determining what it means when someone experiences a side effect. Should the patient who feels sleepy discontinue the medication? How do we determine whether or not observable side effects are important? Should a patient with insulin dependent diabetes mellitus (IDDM) discontinue therapy because he or she develops skin problems at the injection sites? Clearly, local irritation is a side effect of treatment. But without treatment the patient would die. Often the issue is not whether treatment causes side effects, but how we should place these side effects within the perspective of total health. Ultimately, we must decide whether treatment produces a net benefit or a net deficit in health status.

Many measures of health-related quality of life simply tabulate frequencies for different symptoms or represent health status using profiles of outcomes. A representation of three hypothetical treatment profiles is shown in Figure 3.1. It is common in the presentation of these profiles to connect the points, even though increments on the category axis (x-axis) are not meaningful. T-scores

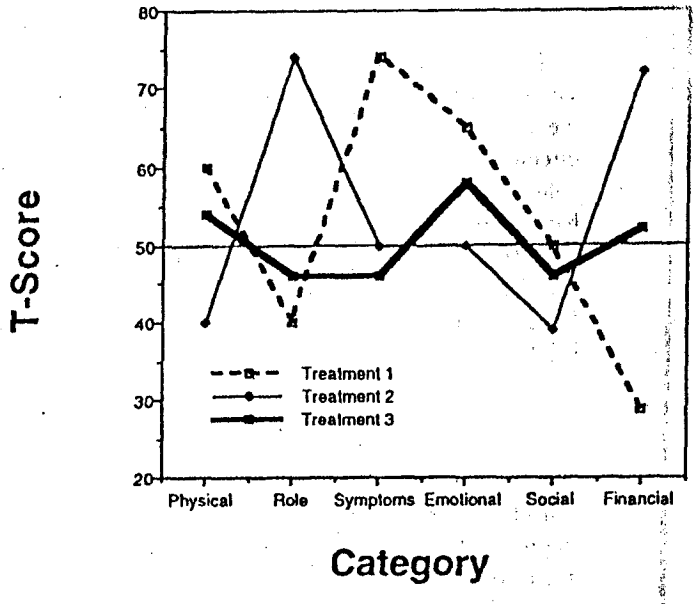

Figure 3.1. Comparison of profiles for three hypothetical treatments. (*Source*: Kaplan and Coons 1992, 31.)

(*y*-axis) are standardized scores with a mean of 50 and a standard deviation of 10. Treatment 1 may produce benefits for physical functioning but decrements for role functioning. Treatment 2 may produce decrements for physical functioning but increments for role functioning. This information may be valuable for diagnostic purposes. However, ultimately, clinicians make some general interpretations of the profile by applying a weighting system. They might decide that they are more concerned about physical, rather than role, functioning or vice versa. Judging the relative importance of various dimensions is common and typically is done implicitly, arbitrarily, and in an idiosyncratic way. Physicians may ignore a particular test result or a particular symptom because another one is more important to them. The process by which relative importance is evaluated can be studied explicitly and be part of the overall model.

If one accepts that preference, or utility, assessment is central to valuing a service relative to its cost, several conceptual issues must be considered (Froberg and Kane 1989a, 1989b, 1989c, 1989d). For example, a variety of approaches to the measurement of preference can yield different results (see the Froberg and Kane studies for a review). However, these differences are to be expected: the various approaches to preference assessment are based on dif-

The concept of utility

The concept of QALYs has been in the literature for nearly twenty-
Perhaps the first application was suggested by Fanshel and Bush (197
later Torrance (1976) introduced a conceptually similar model. Since then, a variety of applications have appeared.

Despite the differences in approach, some important assumptions are similar. All approaches set one completely healthy year of life at 1. Years of life at less than optimal health are scored as less than 1. The basic assumption is that two years scored as 0.5 add up to the equivalent of one year of complete wellness. Similarly, four years scored as 0.25 are equivalent to one completely well year of life. A treatment that boosts a patient's health from 0.5 to 0.75 produces the equivalent of 0.25 QALYs. If applied to four individuals, and the duration of the treatment effect is one year, the effect of the treatment would be equivalent to one completely well year of life. The disagreement is not over the QALY concept but rather over how the weights for cases between 0 and 1 are obtained.

Health utility assessment has its roots in the classic work of von Neumann and Morgenstern (1944). Their mathematical decision theory characterized how a rational individual should make decisions when faced with uncertain outcomes. Von Neumann and Morgenstern outlined axioms of choice that have become basic foundations of decision analysis in business, government, and health care. This work was expanded upon by Raiffa (1968) and several others (see reviews by Bell and Farquhar 1986; Howard 1988). Torrance and Feeney (1989), who reviewed the history of utility theory and its applications to health outcome assessment, argued that the use of the term "utility theory" by von Neumann and Morgenstern was unfortunate. Their reference to utility differs from the more common uses by economists that emphasize consumer satisfaction with commodities that are received with certainty. Nineteenth century philosophers and economists assumed the existence of cardinal (or interval level) utilities for these functions. A characteristic of cardinal utilities is that they can be averaged across individuals and ultimately used in aggregates as the basis of utilitarian social policy.

By the turn of the century, Pareto challenged the value of cardinal utilities and demonstrated that ordinal utilities could represent consumer choice (Bator 1957). Arrow (1951) further argued that there are inconsistencies in individual preferences under certainty and that meaningful cardinal preferences cannot be measured and may not even exist. As a result, many economists have come to doubt the value of preference ratings (Nord 1991).

Perhaps the most important statement against the aggregation of individual preferences was Arrow's impossibility theorem (Arrow 1951). In this classic work, Arrow considered the expected group decision based on the individual preferences of the group members. After laying out a set of very reasonable assumptions about how an aggregate decision should not contradict the apparent preferences of group members, Arrow demonstrated how aggregate decisions can violate the apparent will of the individual decision makers.

Arrow's impossibility theorem may not be applicable to the aggregation of utilities in the assessment of QALYs for several reasons. First, utility expressions for QALYs are expressions of probabilistic outcomes, not goods received with certainty. Von Neumann and Morgenstern emphasized decisions under uncertainty, an approach theoretically distinct from Arrow's. The traditional criticisms of economists are directed toward decisions to obtain certain, rather than uncertain, outcomes (Torrance and Feeney 1989). Second, Arrow assumed that the metric underlying utility was not meaningful and not standardized across individuals. Substantial psychometric evidence now suggests that preferences can be measured using scales with meaningful interval or ratio properties. When cardinal (interval) utilities are used instead of rankings, many of the potential problems in the impossibility theorem are avoided (Keeney 1976).

Different approaches to the calculation of QALYs are based on very different underlying assumptions. One approach considers the duration of time someone is in a particular health state as conceptually independent from the utility for the state (Weinstein and Stason 1976; Kaplan and Anderson 1990). Another approach merges duration of stay and utility (Torrance and Feeney 1989). This distinction is central to understanding the difference in approaches and affects the evidence required to validate the utility assessment procedure.

In the approach advocated by Kaplan and Anderson (1990) and Weinstein and Stason (1976), utilities for health states are obtained at a single point in time. For example, suppose that the state of confinement to a wheelchair is assigned a weight of 0.5. The patients in this state are observed over the course of time to empirically determine their transitions to other states of wellness. If they remain in the state for one year, then they would lose the equivalent of 0.5 well years of life. The key to this approach is that the preference concerns only a single point in time and does not acknowledge duration and that the transition is determined through observation or expert judgment. The alternative approach emphasized by Torrance and Feeney (1989) and others (e.g., Nord 1992) obtains preference for both health state and duration. These approaches also consider the more complex problems of uncertainty. Thus, they are consistent with the von Neumann and Morgenstern notion of decision under uncertainty, in which probabilities and trade-offs are considered explicitly by the judge.

Methods for assessing utility

Different techniques have been used to assess these utilities for health states. These techniques will be summarized briefly, and then comparisons between the techniques will be considered. Some analysts do not measure utilities directly. Instead, they evaluate health outcome by simply assigning a reasonable utility (Weinstein and Stason 1983). However, most current approaches have respondents assign weights to different health states on a scale ranging from 0 (for dead) to 1 (for wellness). The most common techniques include category rating scales, magnitude estimations, the standard gamble, the time trade-off, and the equivalence person trade-off, each of which will be described briefly.

Rating scales

Rating scales provide simple techniques for assigning numerical values to objects. There are several methods for obtaining rating scale information. One approach, the category scale, is a simple partition method in which subjects are requested to assign a number to each case selected from a set of numbered categories representing equal intervals. This method, exemplified by the familiar ten-point rating scale, is efficient, easy to use, and applicable in a large number of settings. Typically, the subject reads the description of a case and rates it on a ten-point scale ranging from 0 for dead to 10 for asymptotic optimum function. End points of the scale are typically well defined; instructions, as the sample in Box 3.1 indicates, are straightforward. Another common rating method, the visual analogue method, shows subjects a line, typically 100 centimeters in length, with the end points well defined. The subject's task is to mark the line to indicate where their preference rests in relation to the two poles.

Appropriate applications of rating scale reflect contemporary developments in the cognitive sciences. Judgment/decision theory has been dominated by the belief that human decisions follow principles of optimality and rationality. Considerable research, however, has challenged the normative models that have attempted to demonstrate rational choice. Cognitive theories such as information integration theory (Anderson 1990) provide better explanations of the cognitive process of judgment. Information integration theory includes two constructs: integration and valuation. Integration describes the cognitive algebra of mentally combining multiple pieces of information during the judgment process. Valuation refers to the weight applied to a particular piece of information. Estimation of these weights requires a theory of measurement. Normative studies of decision making often use arbitrary weights, whereas the cognitive theory requires estimates of subjective value parameters. Although expected

Table 3.1. *Selected results from comparative valuation studies*

Study	*N*	Kind of subjects	Selected results					State
			SG	RS	ME	PTO	TTO	
Torrance 1976	43	Students	.75	.61			.76	Not indicated
			.73	.58			.70	
			.60	.44			.63	
			.44	.26			.38	
Bombardier et al. 1982	52	Health care personnel, patients, family	.85	.65			.78	Needs walking stick
			.81	.47			.58	Needs walking frame
			.64	.29			.41	Needs supervision when walking
			.55	.18			.28	Needs one assistant for walking
			.38	.08			.11	Needs two assistants
Llewellyn-Thomas et al. 1984	64	Patients	.92	.74				Tired; sleepless
			.84	.68				Unable to work; some pain
			.75	.53				Limited walking; unable to work; tired
			.66	.47				In house; unable to work; vomiting
			.30	.30				In bed in hospital; needs help for self-care; trouble remembering
Read et al. 1984	60	Doctors	.90	.72			.83	Moderate angina
			.71	.35			.53	Severe angina
Richardson 1991	46	Health care personnel	.86	.75			.80	Breast cancer: Removed breast; unconcerned
			.44	.48			.41	Removed breast; stiff arm; tired; anxious; difficulties with sex
			.19	.24			16	Cancer spread; constant pain; tired; expecting not to live long
Patrick et al. 1973	30	Students		.78	.85	.71		Skin defect
				.60	.66	.58		Pain in abdomen; limited in social activities
				.50	.54	.42		Visual impairment; limited in traveling and social activities
				.37	.46	.36		Needs wheelchair; unable to work
				.28	.36	.32		In hospital; limited walking; back pain; needs help for self-care; loss of consciousness

Study	*N*	Kind of subjects	Selected results SG	RS	ME	PTO	TTO	State
Kaplan et al. 1979	54	Psychology students		.93	.44			Polluted air
				.67	.13			Limited walking; pain in arms and/or legs
				.49	.06			Needs wheelchair; needs help for self care; large burn
				.25	.02			Small child; in bed; loss of consciousness
Sintonen 1981	60	Colleagues		.61	.72			Difficulties in moving outdoors
				.45	.51			Needs help outdoors
				.25	.34			Needs help indoors also
				.09	.15			Bedridden
				.04	.04			Unconscious
Buxton et al. 1987	121	Health care personnel, university staff			.997		.72	Breast cancer: Removed part of breast; occasionally concerned
					.994		.70	Removed breast; occasionally concerned
					.987		.68	Removed breast; occasionally concerned, also about appearance
					.917		.27	Removed part of breast; stiffness of arm; engulfed by fear; unable to meet people
				.910		.38		Removed whole breast; otherwise as previous case
Nord 1991, 1992[a]	22	General public		.71		.985		Moderate pain; depressed
				.65		.98		Unable to work; moderate pain
				.30		.97		Unable to work; limited leisure activity; moderate pain; depressed
				.20		.90		Problems with walking; unable to work; limited leisure activity; strong pain; depressed

Note: *N* = number; SG = standard gamble methods; RS = rating scale methods; ME = magnitude estimation methods; PTO = person trade-off methods; TTO = time trade-off methods.

Magnitude estimation values are obtained by applying the Rosser/Kind index (Rosser and Kind 1978).

[a]The person trade-off values are transformed from raw scores published in Nord 1991. This study did not include the state "dead." The transformations to a 1–0 scale are based on a subsequent separate valuation of "dead," still using person trade-off (Nord 1992).

Source: Nord 1992.

uncertainty. By contrast, several methods more explicitly consider decision under uncertainty. The standard gamble offers a choice between two alternatives: living in health state A with certainty or taking a gamble on treatment, for which the outcome is uncertain (Fig. 3.2). The respondent is told that treatment will lead to perfect health with a probability of p or immediate death with a probability of $1 - p$ (choice B). The health state described in A is intermediate between wellness and death. The probability (p) is varied until the subject is indifferent between choices A and B.

An attractive feature of the standard gamble is that it is based on the axioms of utility theory. The choice between a certain outcome and a gamble conforms to the exercises originally proposed by von Neumann and Morgenstern. Although the interval properties of the data obtained using the gamble have been assumed, they have not been empirically demonstrated (Froberg and Kane 1989b). A variety of other problems with the gamble have also become apparent. For example, it has often been stated that the standard gamble has face validity because it approximates choices made by patients (Mulley 1989). However, treatment of most chronic diseases does not approximate the gamble. There is no product that will make a patient with arthritis better; nor is there one that is likely to result in immediate death. In other words, the decision-making experience of the patient is not likely to include an option that has a realistic gamble. Further, the cognitive demands of the task are high.

Time trade-off

The concept of probability is difficult for most respondents and requires the use of visual aids or props to assist in the interview. Thus, an alternative to the

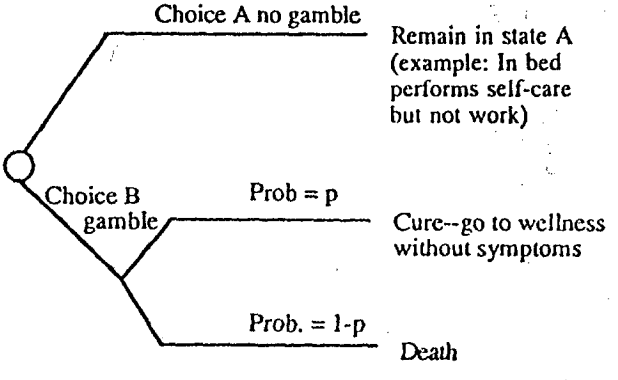

Figure 3.2. Illustration of the standard gamble. (Adapted from Torrance and Feeny 1989.)

standard gamble, also consistent with the von Neumann and Morgenstern axioms of choice, uses a trade-off in time. Here, the subject is offered a choice of living for a defined amount of time in perfect health or a variable amount of time in an alternative state that is less desirable. Presumably, all subjects would choose a year of wellness versus a year with some health problem. However, by reducing the time of wellness and leaving the time in the sub-optimal health state fixed (such as one year), an indifference point can be determined. For example, a subject may rate being in a wheelchair for two years as equivalent to perfect wellness for one year. The time trade-off is theoretically appealing because it asks subjects to explicitly state their preferences in terms of life year equivalents.

Person trade-off

Finally, a person trade-off technique allows comparisons of the numbers of people helped in different states. For example, respondents might be asked to evaluate the equivalencies between the number of persons helped by different programs. They might be asked how many persons in state B must be helped to provide a benefit equivalent to helping one person in state A. From a policy perspective, the person trade-off directly seeks information similar to that required as the basis for policy decision.

Comparisons of the methods

Several articles, reviewed by Nord (1992), have compared utilities for health states as captured by different methods. In general, standard gamble and time trade-off methods give higher values than rating scales in most, but not all, studies (Table 3.1). In about half of the studies reported, time trade-off yields lower utilities than standard gamble. In one of the earlier studies, Patrick, Bush, and Chen (1973) found that person trade-off methods gave the same results as rating scales. However, these findings were not confirmed in more recent studies (Nord 1991). Magnitude estimation has produced highly variable results across studies (Nord 1992). Such variability of results across studies is hardly surprising. The methods differ substantially in the questions posed to respondents.

Psychological versus economic models

Psychometric models divide the decision process into component parts. Health states are observed and categorized. Utilities are observed and categorized. Preferences are obtained as weights for these health states and the ratings apply to a particular point in time and are analogous to consumer preferences under

Table 3.1. *Selected results from comparative valuation studies*

Study	*N*	Kind of subjects	Selected results SG	RS	ME	PTO	TTO	State
Torrance 1976	43	Students	.75	.61			.76	Not indicated
			.73	.58			.70	
			.60	.44			.63	
			.44	.26			.38	
Bombardier et al. 1982	52	Health care personnel, patients, family	.85	.65			.78	Needs walking stick
			.81	.47			.58	Needs walking frame
			.64	.29			.41	Needs supervision when walking
			.55	.18			.28	Needs one assistant for walking
			.38	.08			.11	Needs two assistants
Llewellyn-Thomas et al. 1984	64	Patients	.92	.74				Tired; sleepless
			.84	.68				Unable to work; some pain
			.75	.53				Limited walking; unable to work; tired
			.66	.47				In house; unable to work; vomiting
			.30	.30				In bed in hospital; needs help for self-care; trouble remembering
Read et al. 1984	60	Doctors	.90	.72			.83	Moderate angina
			.71	.35			.53	Severe angina
Richardson 1991	46	Health care personnel	.86	.75			.80	Breast cancer: Removed breast; unconcerned
			.44	.48			.41	Removed breast; stiff arm; tired; anxious; difficulties with sex
			.19	.24			16	Cancer spread; constant pain; tired; expecting not to live long
Patrick et al. 1973	30	Students		.78	.85	.71		Skin defect
				.60	.66	.58		Pain in abdomen; limited in social activities
				.50	.54	.42		Visual impairment; limited in traveling and social activities
				.37	.46	.36		Needs wheelchair; unable to work
				.28	.36	.32		In hospital; limited walking; back pain; needs help for self-care; loss of consciousness

Study	*N*	Kind of subjects	Selected results SG	RS	ME	PTO	TTO	State
Kaplan et al. 1979	54	Psychology students		.93	.44			Polluted air
				.67	.13			Limited walking; pain in arms and/or legs
				.49	.06			Needs wheelchair; needs help for self care; large burn
				.25	.02			Small child; in bed; loss of consciousness
Sintonen 1981	60	Colleagues		.61	.72			Difficulties in moving outdoors
				.45	.51			Needs help outdoors
				.25	.34			Needs help indoors also
				.09	.15			Bedridden
				.04	.04			Unconscious
Buxton et al. 1987	121	Health care personnel, university staff			.997		.72	Breast cancer: Removed part of breast; occasionally concerned
					.994		.70	Removed breast; occasionally concerned
					.987		.68	Removed breast; occasionally concerned, also about appearance
					.917		.27	Removed part of breast; stiffness of arm; engulfed by fear; unable to meet people
				.910		.38		Removed whole breast; otherwise as previous case
Nord 1991, 1992[a]	22	General public		.71		.985		Moderate pain; depressed
				.65		.98		Unable to work; moderate pain
				.30		.97		Unable to work; limited leisure activity; moderate pain; depressed
				.20		.90		Problems with walking; unable to work; limited leisure activity; strong pain; depressed

Note: *N* = number; SG = standard gamble methods; RS = rating scale methods; ME = magnitude estimation methods; PTO = person trade-off methods; TTO = time trade-off methods.

Magnitude estimation values are obtained by applying the Rosser/Kind index (Rosser and Kind 1978).

[a]The person trade-off values are transformed from raw scores published in Nord 1991. This study did not include the state "dead." The transformations to a 1–0 scale are based on a subsequent separate valuation of "dead," still using person trade-off (Nord 1992).

Source: Nord 1992.

certainty. Probabilities are a separate dimension and are determined empirically. These models combine the empirically determined probabilities and the preferences. Psychologists and economists differ in their views about the most appropriate model. Economists have challenged the psychometric approaches (Richardson 1991; Nord 1992), emphasizing that data obtained using rating scales cannot be aggregated. They acknowledge that rating scales may provide ordinal data but contend that they do not provide interval level information necessary for aggregation. These judgments under certainty are subject to all of the difficulties outlined by Arrow (1951).

Psychologists have also challenged the use of rating scales. For example, Stevens (1966) questioned the assumption that subjective impressions can be discriminated equally at each level of a scale. He claimed that the rating scale method is biased because subjects will attempt to use categories equally often, thus spreading their responses when the cases are actually close together and compressing them when the true values are actually far apart. These biases would suggest that numbers obtained on rating scales cannot have meaning.

Armed with these arguments, economists have proposed standard gamble or time trade-off methods as validity criteria for rating scales. The basic assumption is that methods that conform to the von Neumann and Morgenstern axioms assess true utility. If rating scales produce results inconsistent with these utilities, they must be representing preferences incorrectly. As compelling as these arguments are, they disregard a substantial literature analyzing the process of human judgment.

Cognitive limitations

Evidence for the standard gamble and time trade-off techniques

Since the standard gamble technique meets the axiomatic requirements of the von Neumann and Morgenstern theory of decision under uncertainty, some experts believe that the gamble should serve as a gold standard for evaluating other methods. However, there are several concerns about the standard gamble and related techniques. One of the most important has been raised by Tversky, Slovic, and Kahneman (1990). In a series of laboratory experiments, these investigators demonstrated that subjects tend to reverse their previously revealed preferences. For example, in one experiment, subjects were presented with two lotteries. The lotteries had two outcomes: a cash prize or no win at all. In one lottery, the cash prize involved a high probability of winning a small amount of money, while the other lottery offered a cash prize with a low probability of winning a large amount of money. The participants were then asked to state the minimum price they would be willing to accept to sell each bet.

In the next phase of the experiment, the subjects were presented with pairs of bets. In each case, they were offered bets with a high probability of a low payoff versus a low probability of a high payoff. In some cases, the comparison was with the bet against a sure thing. In these cases, one of the options paid a specified sum of money with a probability of 1. This established the pricing. If subjects behave rationally, the alternatives should produce the same estimated value of the bets. However, they did not, and significant reversals occurred, such as a person choosing the high-probability/low-payoff bet over the low-probability/high-payoff bet but assigning the high-probability/low-payoff bet a lower selling price. In fact, 46 percent of subjects showed some reversal. The explanation for these results is that the subjects used inappropriate psychological representations and simplifying heuristics. How a question is framed can have a significant impact upon choice because it can evoke these inappropriate cognitive strategies. In general, humans are poor processors of probabilistic information. When confronted with complex decisions, they use simplifying rules that often misdirect decisions (Kahneman and Tversky 1984).

Several studies have documented unexpected preferences using standard gamble or time trade-off methodologies. For example, MacKeigan (1990) found that patients preferred immediate death to being in a state of mild to moderate dysfunction for three months. Apparently, some subjects misunderstand the nature of the trade-off or felt that any impaired quality of life is not worth enduring. McNeil, Weichselbaum, and Pauker (1981) obtained similar results. They found that if survival was less than five years, subjects were unwilling to trade any years of life to avoid losing their normal speech. These results suggest that either patients have unexpected preferences or that they have difficulty using the trade-off methodologies. Cognitive psychologists have suggested explanations for these problems. Some methods, such as the standard gamble, require only simple trade-offs. They may not require complex processing of probability information. However, other information processing biases may distort judgment. For instance, humans employ an anchoring and adjustment heuristic in decision making. Typically, information is processed in a serial fashion. Subjects begin with one piece of information and adjust their judgment as more information is added. However, experiments have suggested that the adjustments are often inadequate or biased (Kahneman and Tversky 1984). Use of the gamble and trade-off methods could evoke bias due to the anchoring and adjustment heuristic.

Other explanations for the inconsistent results in studies using trade-off methods have been proposed. Some studies have been poorly designed or conducted. For example, there have been problems in studies that request a choice between a mild disability and a very large disability. Often patients will not make this trade. However, a careful application of the methodology would identify a smaller trade-off that the patient would take. Some of the problems

may be avoided with careful application of the methodology (Torrance and Feeny 1989).

Evidence for rating scales

Several lines of evidence argue against the use of rating scales. As noted above, rating scales are theoretically inconsistent with the utility under uncertainty provisions of the von Neumann and Morgenstern theory. From principles of microeconomic theory, rating scales should not produce data that can be aggregated. When compared against the theoretically more appealing standard gamble and time trade-off methods, rating scales produce different results. In addition, the use of rating scales has been challenged by psychophysicists who also argue that these methods produce, at best, ordinal level data (Stevens 1966).

Recent psychological research challenges these criticisms of rating scales (Anderson 1990). Although rating methods are subject to serious biases, most of these biases can be controlled. For example, it has been argued that subjects have a tendency to use each category in a ten-point rating scale equally often. Thus, for stimuli that are close together, subjects will use all categories from 0 through 10 on a ten-point rating scale. Similarly, for cases that represent broader variability in true wellness, subjects will also use the entire range. As a result, it has been argued that any numbers obtained from rating scales are meaningless (Parducci 1968). However, systematic studies of health case descriptions do not confirm this property. Kaplan and Ernst (1983), for example, were unable to document these context effects for health case descriptions. The real issue is whether or not rating scales can produce meaningful data. Most studies evaluating utilities have sought to demonstrate convergent validity (Revicki and Kaplan 1993). Convergent validity is achieved when different methods produce the same results. Many investigators have emphasized the standard gamble because they feel that it is theoretically more sound (Nord 1992).

Only recently have empirical tests evaluating the various approaches been conducted. An empirical test of scale property has been introduced within the last few years (Anderson 1990). The model takes into consideration the psychological process used in evaluating cases. Typically, utility assessment involves a global judgment of a case that is usually made up of multiple attributes. Common attributes of health status are shown in Table 3.2.

When the attributes of the case are systematically varied, parameters of the judgment process can be estimated. Substantial evidence suggests that human judges most often integrate multiattribute information using an averaging rule (Anderson 1990). The averaging rule yields an additive model of human judgment. This averaging process has been validated in specific experimental tests (see Anderson 1990 for a three-volume review of the evidence). Once the averaging process has been established, an analysis of variance model can be

Table 3.2. *Examples of quality-of-life health attributes and levels*

Attribute	Level
Physical function	No limitations
	Mild or moderate limitations
	Severe limitations
Social function	No limitations
	Mild or moderate limitations
	Severe limitations
Emotion well-being	No limitations
	Mild or moderate limitations
	Severe limitations
Pain	No pain
	Mild or moderate pain
	Severe pain
Cognitive ability	No limitations
	Mild or moderate limitations
	Severe limitations

Source: OTA 1992a.

used to evaluate the scale properties. Typically, this is done by systematically varying components of case descriptions as rows and columns in an experimental design. Global judgments are obtained for each cell within the resulting matrix. The analysis of variance model allows parameter estimation for scale values and weights.

According to the functional measurement model, the absence of a significant interaction effect in the analysis of variance establishes the interval property, assuming that the subjects are combining information using an averaging rule. The difference between utilities for two items that differ only by one attribute should be equal to the difference between two other items that differ only by that one attribute. Figure 3.3 shows several applications of the functional measurement test for health case descriptions. These data confirm a large number of other studies that have also shown the interval property for rating scales (Anderson 1990). However, studies have failed to confirm the interval property for magnitude estimation (Kaplan, Bush, and Berry 1979) or for trade-off methodologies (Zhu and Anderson 1991). The axioms underlying the functional measurement model have been published (Luce 1981).

It is of interest that the rating scale debate raged for nearly a century among psychophysicists. It was widely believed that rating scale methods could not

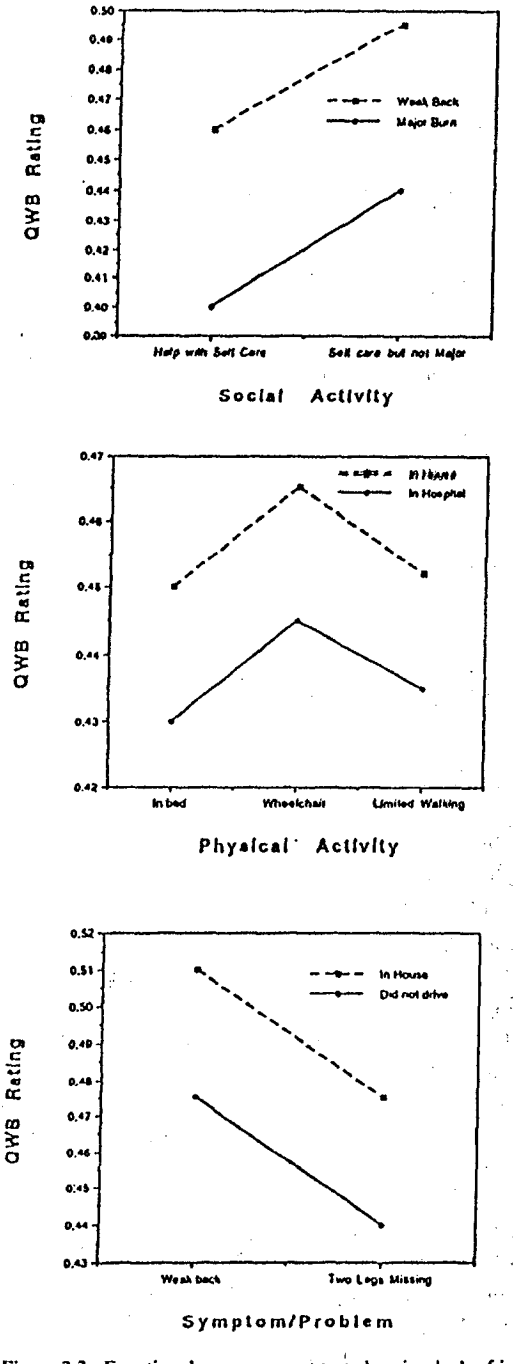

Figure 3.3. Functional measurement test showing lack of interaction among items that differ by the same levels on each two attributes (QWB = Quality of Well-Being Scale).

produce meaningful data, whereas methods requiring dichotomous choice or ratings of subjective intervals were regarded as true psychological metrics. Recent evidence, however, has challenged these beliefs and has confirmed that rating scales can produce data on an interval-level scale, and many psychophysicists have come to accept these methods (Anderson 1990).

In summary, there is substantial debate about which technique should be used to acquire utility information. Results obtained from different methods do not correspond although they typically have a high degree of similarity in the ranks they assign to outcomes. However, the differences in preferences yielded by different methods can result in different allocation of resources if the preferences are not obtained on a linear or interval response scale. For example, suppose that the difference between the effect of a drug and a placebo is 0.05 units of well-being as assessed by rating scales and 0.02 as measured by magnitude estimation. The benefit would have to last twenty years to produce 1 QALY if rating-scale utilities were used and fifty years if magnitude estimation utilities were used. Aggregation of benefits necessarily requires an underlying linear response scale in which equal differences at different points along the response scale are equally meaningful. For example, the difference between 0.2 and 0.3 (0.1 QALY if the duration is one year) must have the same meaning as the difference between 0.7 and 0.8. A treatment that boosts patients from 0.2 to 0.3 must be considered of equal benefit to a treatment that brings patients from 0.7 to 0.8. Confirmation of this scale property has been presented for rating scales but not for the other methods.

Another difference between methods is the inclusion of information about uncertainty in the judgment process. Time trade-off, standard gambles, and person trade-off all theoretically include some judgment about duration of stay in a health state. Magnitude estimation and rating scales typically separate utility at a point in time from probability. Considerably more theoretical and empirical work will be necessary to evaluate these differences of approach.

Whose preferences should be used in the model?

Choices between alternatives in health care necessarily involve preference judgments. For example, deciding what services to include in a basic benefits package is an exercise in value, choice, or preference. Preference is expressed at many levels in the health care decision process. For example, an older man may decide to cope with the symptoms of urinary retention in order to avoid the ordeal and risk of prostate surgery. A physician may order expensive tests to ensure against the very low probability that a rare condition will be missed. Or an administrator may decide to allocate resources to prevention for large numbers of people instead of devoting the same resources to organ transplants for a smaller number.

In cost–utility analysis, preferences are used to express the relative impor-

tance of various health outcomes. There is a subjective or qualitative component to health outcome. Whether one prefers a headache or an upset stomach caused by its remedy is a value judgment. Not all symptoms are of equal importance. Most patients would prefer a mild itch to vomiting. Models of how well treatments work and models that compare or rank treatments implicitly include these judgments. Models require a precise numerical expression of this preference. Cost–utility analysis explicitly includes a preference component to represent these trade-offs.

The model advocated by our group incorporates preferences from random samples of the general population (Kaplan 1993a). The rationale is that although administrators ultimately choose between alternative programs, preferences should represent the will of the general public, not administrators.

Some critics of cost–utility analysis begin with the assumption that preferences differ. For example, in most areas of preference assessment, it is easy to identify differences between different groups or different individuals. It might be argued that judgments about net health benefits for white Anglo men should not be applied to Hispanic men, who may give different weight to some symptoms. We all have different preferences for movies, clothing, and political candidates. It is often assumed that differences must extend to health states and that the entire analysis will be highly dependent upon the particular group that provided the preference data. Allocation of resources to Medicaid recipients, for example, should not depend on preferences from both Medicaid recipients and nonrecipients (Daniels 1991). Other analysts have suggested that preference weights from the general population should not be applied to any particular patient group. Rather, patient preferences from every individual group must be obtained.

The difference between instrumental and terminal preferences (Rokeach 1973) is important to understanding this debate. The difference between instrumental and terminal preference is analogous to the difference between means and ends. Instrumental preferences describe the means by which various assets are attained. For instance, socialists and capitalists hold different instrumental values with regard to the methods for achieving an optimally functioning society. Different individuals may have different preferences for how they would like to achieve happiness, and evidence suggests that social and demographic groups vary considerably on instrumental values.

Terminal values are the ends, or general states of being, that individuals seek to achieve. The Rokeach (1973) classic study of values demonstrated that there is very little variability among social groups for terminal preferences. There is less reason to believe that different social or ethnic groups will have different preferences for health outcomes. All groups agree that it is better to live free of pain than to experience pain. Freedom from disability is universally preferred over disability states. It is often suggested that individuals with particular

disabilities have adapted to them. However, when asked, those with disabilities would prefer not to have them. If disability states were preferred to non-disability states, there would be no motivation to develop interventions to help those with problems causing disabilities.

Although critics commonly assume substantial variability in preferences, the evidence for differential preference is weak at best. An early study demonstrated some statistically significant, but very small, differences in preferences among social and ethnic groups (Kaplan, Bush, and Berry 1979). Other studies have found little evidence for preference difference between patients and the general population. For example, Balaban and colleagues (1986) compared preference weights obtained from arthritis patients with those obtained from the general population in San Diego. They found remarkable correspondence for ratings of cases involving arthritis patients (Fig. 3.4). The mean value for each of thirty scenarios rated by arthritis patients almost perfectly predicted the mean values for the same scenario provided by the general population in San Diego. A similar study of cancer patients by Nerenz and colleagues (1990) found that preference weights from Wisconsin cancer patients were very similar to those obtained from the San Diego general population.

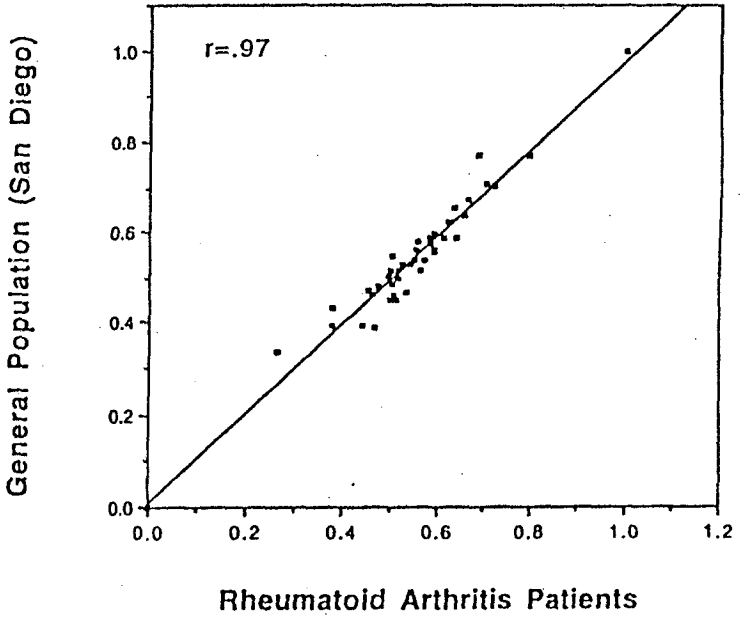

Figure 3.4. Comparison of rheumatoid arthritis patients versus general population. (*Source*: Balaban, Fagi, Goldfarb, and Nettler 1986.)

Also, preferences appear not to vary by location. Patrick and his colleagues (1985) found essentially no differences between preference for another health status measure among study subjects in the United Kingdom and in Seattle. Kaplan (1991) compared residents of the Navaho Nation living in rural Arizona with the general population in San Diego and found few differences. Differences between San Diego citizens evaluated in the 1970s and Oregon citizens evaluated in the 1990s were small even though the weights obtained by the Oregon Health Services Commission were based on a different scaling methodology and different wording of case descriptions (Kaplan, DeBon, and Anderson 1991).

A scaling methodology similar to that used by the Oregon Health Sevices Commission was used by the EuroQol Group in a series of European communities. The data from those studies suggest that differences in preference among the European communities are small and nonsignificant. In one analysis, ratings from European sites were similar to those obtained from respondents in San Diego and estimated approximate San Diego preferences for these cases (Fig. 3.5).

Clearly, overall preferences for health states appear to be quite similar.

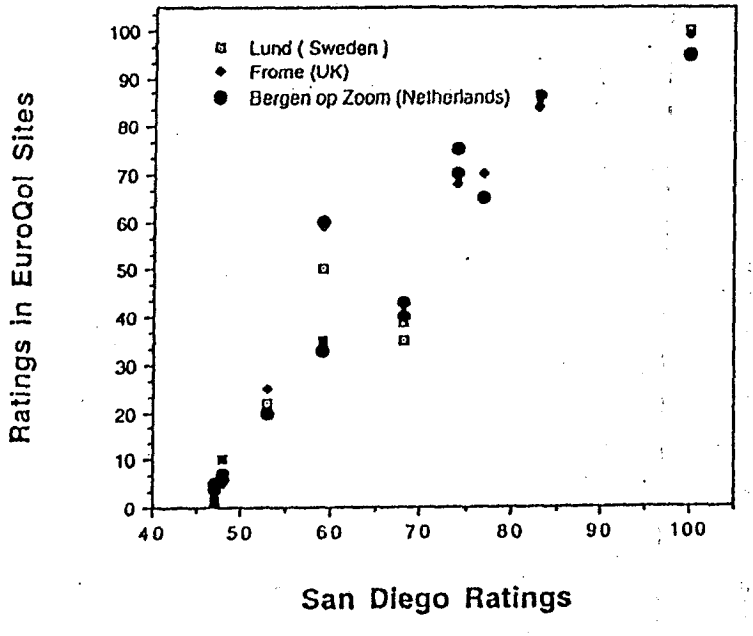

Figure 3.5. Comparison of San Diego case ratings and ratings from sites in EuroQol utilities studies. (*Source*: Balaban, Fagi, Goldfarb, and Nettler 1986.)

There may be considerable variability in preferences for certain particular health states (Mulley 1989), but averaged across individuals, with some exceptions (Kaplan, Bush, and Berry 1978; Kaplan 1993b), the mean preference for different cases in different groups is remarkably similar. Further analysis is required to determine whether these small differences affect the conclusions of various analyses.

Fixed versus variable preference models

Most approaches to utility assessment use the mean preference for a particular case to represent all individuals so characterized. For example, suppose that the average utility for being in a wheelchair, being limited in major activities, and having missing limbs is 0.50. The models would assign the same number to all individuals who occupy that state. In individual decision models, however, decisions might be different if the patient's own utilities were used (O'Connor and Nease 1993).

As appealing as individual decision analysis can be, such analysis is time consuming. Also, Clancy, Cebul, and Williams (1988) demonstrated that the use of individual preferences rarely leads to different treatment courses than would be obtained from the use of aggregate preferences.

Application and criticism: the Oregon experiment

In 1987 a young boy in Oregon developed acute leukemia and his physicians decided that he needed a bone marrow transplant. In addition to his serious illness, the boy became the victim of a new change in the Oregon Medicaid program. With the state unable to afford many basic health services, there was some concern about whether the underfunded public program should be paying for very expensive organ transplantation procedures. A grassroots citizens group, known as Oregon Health Decisions, had created strong support for new approaches to resource allocation. The state legislature determined that thirty-four transplants to Medicaid patients during 1987–9 used the same financial resources as prenatal care and delivery for 1,500 pregnant women. The legislature recognized that they could use their limited resources to provide a small benefit to the large number of pregnant women instead of providing a larger benefit to a small number of people needing organ transplantation. The case attracted substantial media attention and forced the Oregon legislature to grapple with some very serious questions. During the debate, the family of the young leukemia sufferer attempted to raise money for the transplant, but the boy died before he could get the medical procedure.

The problems with financing Medicaid in Oregon are similar to those faced by essentially all other American states. The costs of health care are expanding

much more rapidly than are the budgets for Medicaid. One alternative is to change eligibility criteria and remove some individuals from the Medicaid rolls. Oregon also recognized that American health care was not a two-tiered system but rather a three-tiered system. The three-tiered system included people who had regular insurance and could pay for their care; people enrolled in Medicaid, and a growing third tier of people who had no health insurance at all. In 1991, it was acknowledged that this third tier represented about one-fifth the population of the state. In Oregon, that accounts for about 450,000 citizens. And the number of uninsured is steadily increasing. Collectively, Oregon citizens spent approximately $6 billion on health care in 1989, three times what they spent in state income taxes (Kitzhaber 1993).

Stimulated by the community support from Oregon Health Decisions, Oregon concluded that they (and most other states) were rationing health care. Oregon passed three pieces of legislation to attack this issue. This chapter focuses most specifically on Senate Bill 27. This bill mandated that health services be prioritized in order to eliminate services that did not provide benefit.

A Health Services Commission was created to develop the prioritized list. This commission obtained several sources of information. First, they held public hearings to learn about preferences for medical care in the Oregon communities. These meetings helped clarify how citizens viewed medical services. Various approaches to care were rated and discussed. On the basis of forty-seven town meetings that were attended by more than a thousand people, thirteen community values emerged. These values included prevention, cost–effectiveness, quality of life, ability to function, and length of life. The major lesson from the community meetings was that citizens wanted primary care services. Further, the people consistently argued that the state should forgo expensive heroic treatments for individuals or small groups in order to offer basic services for everyone. To pay for preventive services, it was necessary to reduce spending elsewhere.

A major portion of the commissioners' activity was to evaluate services using the Quality of Well-Being (QWB) Scale from the General Health Policy Model (Kaplan and Anderson 1990). The commissioners could not possibly have conducted clinical trials for each of the many condition–treatment pairs (see Table 3.3). So the commission formed a medical committee that had expertise in essentially all specialty areas and had the participation of nearly all of the major provider groups in the state. Working together, the committee estimated the expected benefit from 709 condition–treatment pairs. The QWB Scale also requires preference weights. These weights are not medical expert judgments but should be obtained from community peers. Oregon citizens were particularly concerned about using weights from California to assign priorities in their state. Thus, 1,000 Oregon citizens participated in a telephone survey conducted by Oregon State University. This exercise became a central issue in the evaluation of the proposed program.

Table 3.3 *Examples of condition–treatment pairs*

Condition	Treatment
Rectal prolapse	Partial colectomy
Osteoporosis	Medical therapy
Ophthalmic injury	Closure
Obesity	Nutritional and lifestyle counseling

In 1990, the commission released its first prioritized list. Unfortunately, many of the rankings seemed counterintuitive, and the approach drew serious criticism in the popular press. As a result, the system was reorganized according to three basic categories of care: essential, very important, and valuable to certain individuals. Within these major groupings were seventeen subcategories. The commission decided to place greatest emphasis on problems that were acute and treatable yet potentially fatal if untreated. In these cases treatment prevents death and there is full recovery. Examples include appendectomy for appendicitis and nonsurgical treatment for whooping cough. Other categories classified as essential were maternity care, treatment for conditions that prevents death but does not allow full recovery, and preventive care for children. Nine categories were classified as essential. Listed as very important were treatments for nonfatal conditions that would return the individual to a previous state of health. Included in this category were acute nonfatal one-time treatments that might improve quality of life: hip replacements, cornea transplants, and so on. At the bottom of the list were treatments for fatal or nonfatal conditions that did not improve quality of life or extend life, including progressive treatments for the end stages of diseases such as cancer and AIDS or care for conditions in which the treatments were known to be ineffective. In the revised approach, the commission decided to ignore cost information and to allow their own subjective judgments to influence the rankings on the list. Conditions selected from the top, middle, and the bottom of the list are summarized in Table 3.4.

To implement the proposal, Oregon needed a waiver from the U.S. Department of Health and Human Services (DHHS). However, in August 1992, the DHHS rejected Oregon's application for a waiver on the grounds that the Oregon proposal violated the Americans with Disabilities Act of 1990 which became law in July 1992. The DHHS's position was that the Oregon preference survey on quality of life quantified stereotyped assumptions about persons with disability. According to the statement scholars have found that people without disability systematically undervalue the quality of life of those with disabilities. A paper by Hadorn (1991) and an analysis by the U.S. Office of Technology

Table 3.4. *Examples of condition–treatment pairs from top, middle and bottom of list*

Top 10

1. Medical treatment for bacterial pneumonia
2. Medical treatment of tuberculosis
3. Medical or surgical treatment for peritonitis
4. Removal of foreign body from pharynx, larynx, trachea bronchus, or esophagus
5. Appendectomy
6. Repair of ruptured intestine
7. Repair of hernia with obstruction and/or gangrene
8. Medical therapy for croup syndrome
9. Medical therapy for acute orbital cellulitis
10. Surgery for ectopic pregnancy

Middle 10

350. Repair of open wounds
351. Drainage and medical therapy for abscessed cysts of Bartholin's gland
352. Medical therapy for polynodal cyst with abscess
353. Medical therapy for acute thyroiditis
354. Medical therapy for acute otitis media
355. Pressure equalization tubes or tonsillectomy and adenoidectomy for chronic otitis media
356. Surgical treatment for cholesteatoma
357. Medical therapy for sinusitis
358. Medical therapy for acute conjunctivitis
359. Medical therapy for spina bifida without hydrocephalus

Bottom 10

700. Mastopexy for gynecomastia
701. Medical and surgical therapy for cyst of the kidney
702. Medical therapy for end stage HIV disease (comfort care excluded – it is high on list)
703. Surgery for chronic pancreatitis
704. Medical therapy for superficial wounds without infection
705. Medical therapy for constitutional aplastic anemia
706. Surgical treatment for prolapsed urethral mucosa
707. Paracentesis of aqueous humor for central retinal artery occlusion
708. Life support for extremely low birth weight (<500 g) and under 23 week gestation
709. Life support for anencephaly

Assessment (OTA 1992a) were cited to support this statement. However, the great bulk of the evidence summarized earlier in this chapter was ignored. Using Oregon data, utility differences across groups are small. For example, those who have ever been in a wheelchair versus those never in a wheelchair (Fig. 3.6), men and women (Fig. 3.7), and those insured and uninsured for health care (Fig. 3.8) have very similar utilities for thirty-one cases rated.

The DHHS's decision failed to acknowledge that resource allocation designs necessarily require human judgment. Ultimately, decisions are made by patients, physicians, administrators, or their surrogates. Oregon clearly recognized this and attempted to separate aspects of human judgment. For example, when decisions required medical knowledge, they depended upon a medical committee. When the decisions required in-depth understanding of human values, they depended on discussions held in open forums in Oregon towns. When the judgments involved an assessment of quality of life for those with either symptoms or disabilities, they depended on the preference of Oregon citizens. This exercise was unusual because all of these judgments were made publicly using methods that could be replicated by others.

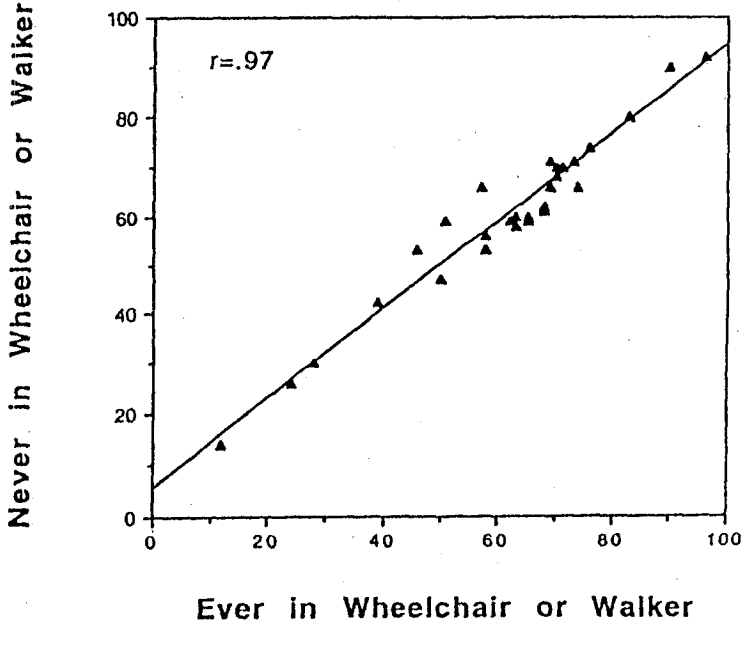

Figure 3.6. Ratings of thirty-one cases by those who have ever been confined to a wheelchair or walker and those who have not. (Data from Oregon Health Services Commission, Oregon State University.)

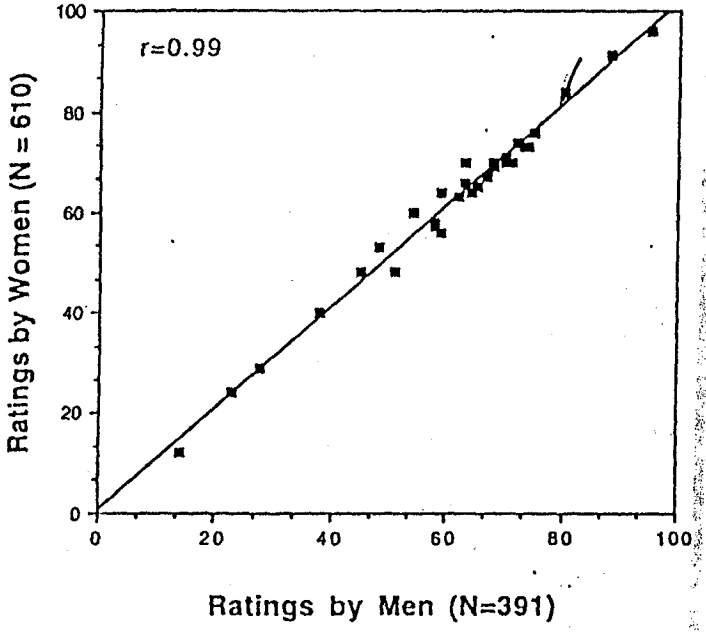

Figure 3.7. Ratings of thirty-one cases by men and women in Oregon.

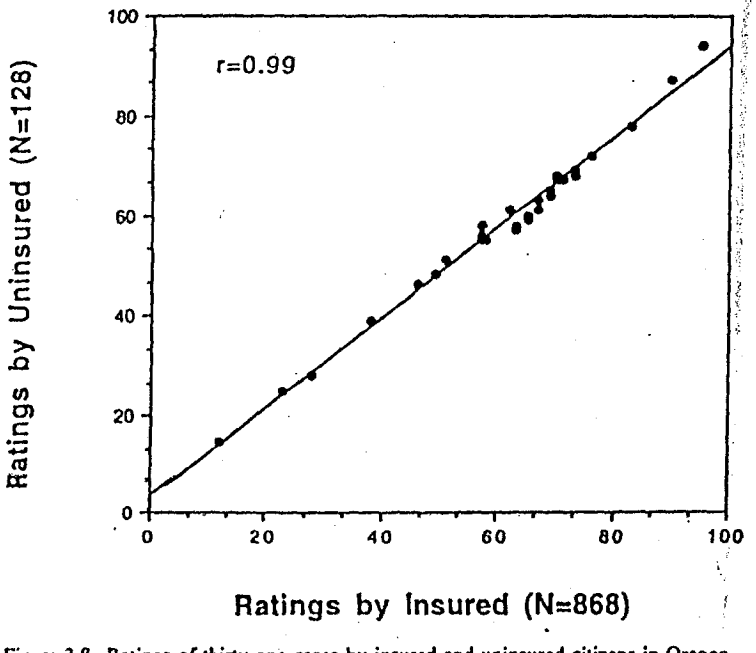

Figure 3.8. Ratings of thirty-one cases by insured and uninsured citizens in Oregon.

The analysis underlying the rejection of the Oregon application was not only misinformed, it was incorrect. It assumed that there would be discrimination against persons with disabilities because treatment could not improve their chronic problems. However, this analysis made a very serious conceptual error. Effectiveness of treatment is based on estimated course of the illness with and without treatment. A treatment that sustains life, even without improvements in quality of life, produces very substantial benefits. For example, suppose a person is in an accident that leaves him or her in a state rated 0.5 with treatment or in a state rated 0.0 (death) without treatment. According to the Oregon model, the treatment will produce 0.50 QALYs (calculated as 0.50 – 0.00) for each year the person remains in that state. That is a powerful treatment effect in comparison to most alternatives. The crucial element is that the treatment works. The system does attempt to exclude treatments that neither extend life nor make patients better. In other words, the targets for elimination are only treatments that use resources and make no difference.

The DHHS also misrepresented the meaning of quality-of-life scores. They assumed that having a low quality-of-life score was discriminatory because people with disabilities and those without disabilities would not be rated the same. However, the assumption contradicts the notion that people with disabilities need medical services. People who are at optimum health (1 on the QWB Scale) need fewer services than those who occupy lower levels. Quantifying these differences allows us to set priorities for future resource allocation. If, for the sake of argument, we decide to score people with disabilities 1, it would follow that we should not provide services for these individuals, because they have already achieved the optimum level of wellness. Scores lower than 1 suggest that resources should be used to improve these conditions.

Instead of debating these issues, Oregon chose to resubmit their application with the utility portion of the model excluded. Their revised waiver application considered probability of death and probability of moving from a symptomatic to an asymptomatic state. By giving up the utility component of the model, Oregon ignores the fact that health states are valued.

Summary

Cost–utility studies depend on measures of utility. In addition to the issue of whose preferences are obtained, we must also consider how preferences are measured. Economists and psychologists differ on their desired approach to preference measurement. Economists favor approaches based on expected utility theory. The axioms of choice (von Neumann and Morgenstern 1944) depend upon certain assumptions about gambling or trade-off. Thus, economists only acknowledge utility assessment methods that formally consider economic trades (Torrance 1986). The advantage of these methods is that they clearly are

linked to economic theory. However, there are also some important disadvantages. For example, Kahneman and Tversky (1984) have shown empirically that many of the assumptions that underlie economic measurements of choice are open to challenge. Human information processors do poorly at integrating complex probability information when making decisions that involve risk. Further, economic analysis assumes that choices accurately correspond to the way rational humans assemble information.

A substantial literature from experimental psychology questions these assumptions. In particular, Anderson (1990) has presented evidence suggesting that methods commonly used for economic analysis do not represent the underlying true preference continuum. Newer research by economists employs integrated cognitive models (Viscusi 1989), and contemporary research by psychologists consider economic models of choice. However, significantly more exchange between economists and psychologists is needed to resolve the theoretical and practical difficulties of utility assessment.

In summary, a review of the literature on utility assessment suggests that preferences can be explicitly considered in a cost–utility analysis. A variety of studies have evaluated the generalizability (Kaplan, Bush, and Berry 1976), the validity, and the reliability of the preference measures (Kaplan, Bush and Berry 1976, 1979; Froberg and Kane 1989c). Methodological studies have tested some of the specific concerns about rating scale methods (Kaplan 1982; Kaplan and Ernst 1983). Preference differences across groups appear to be small and are not sufficiently large to justify their use in influencing policy decisions. This review of the evidence indicates that rating scales provide an appropriate method for utility assessment.