Pergamon

Soc. Sci. Med. Vol. 39, No. 1, pp. 7–21, 1994

0277-9536/94 $7.00 + 0.00

COST UTILITY ANALYSIS: WHAT SHOULD BE MEASURED?

J. RICHARDSON

National Centre for Health Program Evaluation, Monash University, Yarra House, Fairfield Hospital, Yarra Bend Road, Fairfield, Victoria 3078, Australia

Abstract—The paper re-examines the issue of the appropriate unit for measuring output in cost utility analysis and the technique that will measure it. There are two main themes. The first is that utility, as it is often conceived and quantified, is not an appropriate basis for measurement. Consequently, a question arises concerning the selection of an appropriate unit of measurement. The second theme is that there is a need to establish criteria for the evaluation of measurement units. Four criteria are proposed which follow from commonly accepted social objectives and from the requirements of a measurement unit. It is concluded that, as judged by these criteria, the measurement units produced by the time trade-off and person trade-off (equivalence) techniques are more satisfactory than the units produced by the rating scale, magnitude estimation or the standard gamble.

Key words—cost utility analysis, quality adjusted life years, standard gamble

1. INTRODUCTION

Much of the appeal of cost utility analysis (CUA) must be attributed to the fact that its unit of output is the quality adjusted life year—the QALY. As most would agree that the quality of life—QoL—should be a consideration in decisions about the allocation of resources, its inclusion in economic analysis appears to represent a significant methodological advance. However, while the name 'QALY' was a successful device for promoting CUA it obscures the fact that the term 'quality' has a variety of meanings* and that the measurement units produced by different scaling techniques† may be measuring different concepts of quality or a concept that is an inappropriate basis for the allocation of resources. The most common approach to this problem has been to define 'quality' as 'utility'. Several authors have noted [2, 3] that the QALY is not a pure measure of utility but a number of life years weighted by an index of utility. This differs from a pure measure of utility since the index assigns the same numerical value for each individual's healthy year, irrespective of variation in the utility of full health between individuals. The unit of output is therefore the 'utility adjusted life year', which may nor may not be as intuitively appealing as the 'quality adjusted life year'. A more important issue than nomenclature is whether or not the various techniques employed in CUA succeed in measuring any of the concepts of utility described in the literature and, more fundamentally, whether these concepts of utility are an appropriate basis for decisions about the allocation of resources. It is surprising that the latter issue has not received some attention. Elsewhere in the economics literature there has been a vigorous debate about the nature and meaning of 'utility' and this would suggest that a consideration of its use in CUA is desirable [4–8].

In selecting a scaling technique that is an appropriate basis for resource allocation, a common approach, particularly amongst economists, is to equate 'utility' with the utility described by the von Neumann and Morgenstern axioms ('N–M utility') [9]. This leads to the conclusion that the standard gamble (SG), is the appropriate gold standard for measurement as it is the technique that purports to measure this concept of utility. A weaker version of this argument recognises that the assumptions underlying the measurement of N–M utility are not strongly supported by empirical evidence but still concludes that the SG comes closest to the measurement of utility; imperfect theoretical support is claimed to be better than *ad hoc* measurement and the

*Mosteller notes that "we know now that several different measures fly under the banner of quality of life—perhaps three to seven different kinds of measures.... The needed advance is to put them in some agreed pattern, to clump them, to name the clumps, and then to tell what each is especially good for and not to fight with one another about what is really a quality of life measure" [1, p. 282].

†Throughout the paper the term 'scaling' is used to describe the technique by which a health state description is converted into a numerical value or score. The five examples discussed are the standard gamble, time trade off, rating scale, magnitude estimation, and person trade-off (equivalence technique). They are described in the Appendix. The term 'measurement unit' is used to describe the outcome of the application of the technique. For example, QALYs refer to the product of life years and an index of 'utility' produced by a scaling technique. Healthy year equivalents are produced with the standard gamble in combination with a multi-year health state scenario.

standard gamble, like a medical procedure, is carried out under risk.

In Section 2 of the paper it is first argued that different authors are often referring to different concepts when they refer to 'utility'. Basing a unit of output upon 'utility' is therefore ambiguous unless the particular concept is specified. Second, while it is well known that the axioms from which N–M utility is derived are empirically flawed, it is argued that the *theoretical* basis for N–M utility is also defective. Consequently, the SG should not be regarded as a gold standard for measurement but as one of several scaling techniques to be evaluated on an equal basis with other techniques.

In view of this discussion it is argued in Section 3 that the analysis of alternative measurement units and scaling techniques should not commence with the assumption that QALYs should be based upon 'utility', somehow defined. Rather, it is argued that possible measurement units should be evaluated using explicit criteria. Four are suggested that are necessary to ensure that a measurement unit satisfies the purported objective of a QALY, namely, the combination of the quantity and quality of life into a single unit appropriate for use in CUA. Finally, these criteria are applied to the measurement units (QALYs) produced by five of the most commonly used scaling techniques. The conclusion reached is that there are strong grounds for accepting the time trade-off or the person trade-off (equivalence technique) as the preferred method for scaling health states when the objective is to determine the allocation of scarce resources.

2. CONCEPTS OF UTILITY

Positive analysis

The question of what should be measured in the economic evaluation of health programs may appear to have a self-evident answer, namely 'utility'. However, the use and meaning of this concept have varied over time, and it is generally used differently by psychologists and economists. There is not even agreement on its meaning amongst economists. For example Allais notes that at the 1982 Oslo Conference on Utility and Risk "by far the most heatedly debated issues were those on the concept of utility" [4, p. 8]. The debate, recorded in the proceedings, revealed very significant differences in the way in which it is conceptualised by economists [8].

The use of the term 'utility' has been the subject of several reviews [10, 7]. These reveal that historically there have been at least four distinct concepts. Initially, the term referred to a psychological concept of welfare or well-being. The concept was derived directly from Bentham's [11] belief that nature has placed us under two masters, pleasure and pain and that, while the intensity of these sensations might vary, the nature of the resulting 'utility' is essentially the same irrespective of its source. This view was adopted with little modification by economists. Viner [12], for example, claims that "the utility theory of value is primarily an attempt to explain price determination in psychological terms" while Strotz [7, p. 84] viewed utility as "a psychological entity measurable in its own right" (in [7, p. 130]). This remained the predominant view of European economists in the second half of the nineteenth century including Jevens, Menger, Walrus and Marshall [7]. Outside the economics profession 'utility' is often used in this original sense of a psychological and measurable quantity. For example, in a recent review of the measurement of health state preferences Froberg and Kane [13] commence with the statement that "in this paper, preferences or utilities refer to levels of subjective satisfaction, distress, or desirability that people associate with a particular health state. Other synonyms for this level of subjective satisfaction are quality of life, weight, or rating of the health state" [13, p. 346]. As with other psychological concepts these attributes cannot be directly observed but only inferred. The concept itself is a construct and the functional relationship between the construct and external evidence must be embodied in psycho-physical theory [14].

The second concept of utility emerged from the work of Edgeworth, Fisher, Pareto and Slutsky, and more recently Hicks and Allen [15]. These writers argued that for the purposes of consumption theory it is unnecessary to define utility in psychological terms. Rather, a satisfactory theory could be established on the basis of an ordinal ranking of preferences and the term 'utility' was used to indicate the order of preferences or indifference between options. While removing the psychological connotations, this also reduced the value of the concept outside the framework of positive consumption theory. Thus, Graaff asserts that "to say that [a person's] welfare would be higher in A than in B is to say *no more than* he would choose A rather than B if he were allowed to make the choice" [16, p. 5, emphasis added]. Referring to the same instrumentalist concept, Philips recognises its more restrictive application when he argues that "the utility function is a formal concept, useful to the economist, not to the consumer . . . the economist wants to create a tool useful for correct description of observed behaviour . . . In the limit one might say that the utility function exists because we postulated it" [17, p. 26]. As utility is defined in terms of revealed preferences, this use of the concept does not require a behavioural theory to establish a functional relationship between a psychological construct and observed behaviour. Consequently, any construct devised to measure the psychological concept could result in a utility ranking that differed from one based upon revealed preferences.

Third, and related to each of the first two concepts, utility has been treated as an index of the strength of choice, and as having cardinal properties. Thus Allais argues that "some, including myself even believe that

it [cardinal utility] can be defined independently of any random choice by reference to the intensity of preferences... Others deny it any existence or any operational value. Still others hesitate to state a categorical judgement; this is apparently the case of K. Arrow who answered: 'I am not sure—maybe it exists' to one of my questions" [4, p. 28]. The existence of a cardinal property indicates a significant departure from the previous concept. Utility is no longer simply an instrument to assist with the description and analysis of behaviour. Rather, the third concept which is derived from preference utilitarianism is related to a potentially measurable characteristic of an individual: the *intensity* of preferences. In his initial work on the subject in 1735 Bernoulli, along with other nineteenth century economists, similarly conceived of utility as having cardinal properties and as being measurable. However, while the third concept is related to a person's psychological characteristics it differs from the first, broader concept of psychological well-being or quality of life as it relates only to the intensity of *preferences*. Individuals could, potentially, have a preference for an activity that would reduce their quality of life as independently conceived and measured.

Fourth, 'von Neumann–Morgenstern (N–M) utility' is obtained from a standard gamble in which individuals indicate the probability that leaves them indifferent between the state to be evaluated and a gamble in which the outcomes are two reference states. ('Full health' and 'death' are often used in the evaluation of health states.) By assuming the von Neumann–Morgenstern axioms to be correct it is possible to derive a utility value from the probabilities in the gamble.* It is not clear whether the axioms were initially part of an attempt to quantify the intensity of preferences under conditions of risk (and the indications are that they were not—see below), or if the N–M utility index was, as Baumol [18] asserts, no more than an operational device for predicting a person's choice between lottery tickets.† But more recently N–M utility appears to have been explained only in terms of the *process* of its quantification and not in terms of an underlying concept. Lane for example, argues that "utilities are easy to interpret because they are defined directly in terms of a specific trade-off between the consequences and the problematic choice between two reference consequences" [19, p. 591]. The fact that it is not easy to interpret N–M utility is illustrated by the fact that Lane's interpretation is defective. The trade-off referred to, and quantified by a probability, only indicates N–M utility if the N–M axioms are assumed to be true.

Joan Robinson [20] argues that the first two of these concepts of utility have been confused. Baumol [18] points out that the last two concepts have been wrongly equated. It is likely that the great appeal of N–M utility in the context of CUA is derived from such a conflation of concepts. The intuitive attraction to utility as a measure of QoL is probably derived from the first, psychological, interpretation. Confidence in its empirical relevance is probably derived from the large literature that has usefully employed utility functions as part of the framework for positive analysis. This confidence and intuitive appeal probably explains the attraction of N–M utility in the context of CUA and elsewhere.

Utility or value

The fact that the concept of utility has mutated does not disqualify it from being a sensible maximand in one form or other. It only indicates the need to ensure that the concept is clearly understood and correctly measured. The process for deriving N–M utility via the standard gamble is clear but the concept is not. A possible interpretation is that it measures cardinal utility in the third sense discussed above, but under conditions of risk; that is, it provides a cardinal index of the strength of choice under risk. A distinction is often drawn between 'value', which is the result of decision-making in a risk-free environment, and 'utility', which is revealed under conditions of risk. The claim is often made that the N–M axioms describe decision making under risk and that this is the relevant context for cost utility analysis, as the outcomes of medical interventions are always uncertain. To clarify the claim, consider the following equation:

$$A = p.\ J(Y_1) + (1-p).\ J(Y_2) \qquad (1)$$

In equation (1), p and $(1-p)$ are the probabilities of two outcomes Y_1 and Y_2, which are assessed as being worth $J(Y_1)$ and $J(Y_2)$ respectively. The final quantitative assessment of this prospect according to the rule incorporated in the equation is A, the expected value. The procedure embodies an assessment under risk. If J is a concave function then $A < J(pY_1 + (1-p)Y_2)$—with a diminishing marginal valuation of Y, the outcome under risk will be less favourable than a riskless, actuarially equivalent, prospect. However, the specific claim of those who wish to distinguish 'utility' from 'value' is that, in addition to the effect that is captured by the concavity of the function J, the magnitude $J(Y)$ must *also* be assessed under conditions of risk. It would not be sufficient to measure Y_1 and Y_2 under conditions of certainty and to incorporate risk solely through the use of the formula. For example, in the case of an evaluation of a medical intervention for which there was a 10% chance of death and a 90% chance of life

*N–M utility is the sub-set of what Allais refers to as neo-Bernoullian utility in which the probabilities are objective.

†Fishburn also notes that von Neumann and Morgenstern distinguish the concept of utility from its numerical value: When von Neumann and Morgenstern talk about what we refer to today as their linear utility function they almost always talk about *numerical utility* or numerical valuation (values) of utility [17, p. 137].

in a health state, S, it would not be correct to evaluate S under certainty (using, for example, the time trade-off technique) and then to weight its value by 0.9 to obtain quality adjusted life years. Rather, it is claimed that S should be evaluated under risk, using the standard gamble.

The reason for the assessment of Y under risk is that there is a 'specific utility of gambling' or a 'specific utility of risk' arising from risk *per se* as distinct from the utility obtained in any riskless state or combination of riskless states. In Pope's terminology [21, 22], there is a "pre outcome period" before the result of the gamble or risk is known or experienced. During this period there may be a variety of emotional factors operating that are relevant to utility and directly attributable to the uncertainty of the outcome. These may include like or dislike of excitement and danger, the anticipation of regret or elation, tension, curiosity, wonder, hope, fear or worry. The relevant issue here is whether or not the N–M axioms allow for this specific utility of risk. In the CUA literature and elsewhere it has been explicitly argued that the axioms take full account of an individual's risk behaviour [23, 24]. If they do not, then serious doubt arises about the use of the N–M standard gamble as the gold standard of measurement in cost utility analysis.

Von Neumann and Morgenstern did not believe that their axioms accounted for the specific utility of risk. This is explicit in the introduction to their 1944 analysis of game theory:

> The conceptual and practical difficulties of the notion of utility and particularly of the attempts to describe it as a number, are well known and their treatment is not among the primary objectives of this work . . . Let it be said at once that the standpoint of the present book on this very important and very interesting question will be mainly opportunistic. We wish to concentrate on one problem—which is not that of the measurement of utilities and of preferences—and we shall therefore attempt to simplify all other characteristics as far as reasonably possible [9, p. 28].

Because of the ensuing misunderstanding of their intentions Morgenstern was forced to reiterate the point. In a posthumously published article he wrote:

> I want to make it absolutely clear that I believe—as von Neumann did—that there may be a pleasure of gambling, of taking chances, a love of assuming risks, etc. But what we did say and what I do feel I have to repeat even today after so many efforts have been made by so many learned men, is that the matter is still very elusive. I know of no axiomatic system worth its name that specifically incorporates a specific pleasure or utility of gambling together with a general theory of utility . . . I am not saying that it is impossible to achieve it in a scientifically rigorous manner. I am only saying (as we did in 1944) that this is a very deep matter [25, p. 181].

The von Neumann–Morgenstern view of their own theory has not been universally accepted. Harsanyi for example argues:

> Fundamentally, the answer is that the decision maker's 'gambling temperament' has already been allowed for in defining his von Neumann–Morgenstern (vN–M) utility function. Therefore, if the utilities of the various possible outcomes are measured in (vN–M utility) units, then the expected utility of a lottery ticket will already fully reflect the decision maker's positive or negative (or neutral) attitude towards risk [26, p. 155].

Similarly, in the health economics literature Gafni argues that "the concept of risk attitude . . . the concept of time preference . . . [and] the concept of quantity effect . . . are all accounted for in individuals" answers to lottery questions in health . . . [24, p. 4].

Harsanyi and others have subsequently changed their view. Pope describes the sequence of events as follows:

> Prior to the late 1940s all contributors to mainstream decision literature recognised that the expected utility procedure omits risk taking considerations arising *directly* from not knowing the outcome. Consequently, there were many efforts to generalise the procedures so as to remedy the defect. But, a mistaken view that the expected utility procedure includes all risk taking considerations took hold. This view even came to be known as the classical interpretation of the expected utility procedure. It appeared in numerous publications up into the early 1980s, and led to confusing changes in terminology. By the mid 1980s, proofs of the erroneous nature of this view had gained wide currency, and the mistaken interpretation of the expected utility procedure is now comparatively rarely encountered [22, pp. 197–198].*

There are good reasons for a return to the pre-1940 position. In equation (1) earlier the value of the prospect, A, is the result of the independent evaluation of two outcomes and then their combination. That is, there are three distinct steps, viz: (1) the evaluation of Y_1; (2) the evaluation of Y_2; and (3) the combination of $J(Y_1)$ and $J(Y_2)$ to obtain their expected value. If steps 1 and 2 are carried out using the standard gamble and if the procedure allows for a specific utility of risk, then $J(Y_1)$ and $J(Y_2)$ will differ from an evaluation of Y_1 and Y_2 carried out under certainty. If the evaluation under certainty produced values $V(Y_1)$ and $V(Y_2)$ respectively, then

$$J(Y_1) = V(Y_1) + g_1$$

$$J(Y_2) = V(Y_2) + g_2$$

where g_1 and g_2 are interpreted as the differences attributable to a specific utility of risk. Under certainty the evaluation of Y_1 in step 1 has to be independent of Y_2 or any other possible value of Y (since by definition Y_2 is not a possible outcome). Consequently, g_1 cannot be a function of Y_2 and,

*In support of the historical interpretation given here, Harsanyi [27], was quoted earlier as believing that N–M utility incorporates the specific utility of risk (which he terms 'process utilities'), Harsanyi argues that "even though risk taking behaviour in the real world in many cases will involve both types of utilities, it is clear from von Neumann and Morgenstern's own words . . . that their theory is meant to abstract from all process utilities (which they call the 'specific utility of gambling') and is meant to apply only to situations where these process utilities are unimportant" [27, p. 307].

conversely, g_2 cannot be a function of Y_1. This is inconsistent with the notion of a specific utility of risk which arises from Y_1 *not* being the only possible outcome for Y. Consider the uncertain prospect which includes *both* Y_1 and Y_2 as possible outcomes. The evaluation of the prospect will include a specific utility of risk which arises from the knowledge that *both* Y_1 and Y_2 are possible outcomes: there could be no regret or elation, hope or fear associated with outcome Y_1 if there had been no alternative outcome. Further, the intensity of the regret or elation would, in general, be different if the prospect consisted of a third outcome Y_3 and not Y_2. This may be seen by allowing Y_2 to differ from Y_1 by an arbitrarily small amount. As this amount approaches zero the basis for elation or regret is eliminated. This dependence of the specific utility of risk on the difference between Y_1 and Y_2 cannot be taken into account in either g_1 or g_2 nor their combination. Likewise, the emotions giving rise to a specific utility of risk such as hope and fear arise because there are different possible outcomes and this excludes the possibility that these reactions can be taken into account in $J(Y_1)$ and $J(Y_2)$.

The combination rule embodied in equation (1) cannot combine g_1 and g_2 in a way that allows for an additional specific utility of risk above and beyond what is embodied in the evaluation of Y_1 and Y_2 and the assumption that it can do so results in a contradictory outcome. This occurs if $Y_1 = Y_2 = Y_0$. It follows that in equation (1), $A = J(Y_0)$. Risk is eliminated and the function J represents decision making under certainty. Hence $g_1 = g_2 = g = 0$. Since g_1 and g_2 are independent of Y_2 and Y_1 respectively, this must be a general result. In all other cases where $Y_1 \neq Y_2$, equation (1) is claimed to represent decision making under risk. However, the one rule cannot simultaneously represent behaviour under risk and under certainty. This is equivalent to the conclusion that the combination rule does not permit the existence of a specific utility of risk that describes risk behaviour generally.

The conclusion drawn by Allais [4] from a more generalised proof is that N–M utility measured under uncertainty is the same as cardinal utility measured under certainty (the third concept in Section 2). The chief idea behind his proof was also put forward by Pope in 1983 [21]. More recently, Bouyssou and Vansnick [28] have demonstrated that the N–M utility function is not only identical to cardinal utility under certainty ('classical utility') but that every N–M function and every classical utility function must be a linear transform of every other (i.e. they differ by a scaling factor only). The authors note that this amounts to negating any specific element due to the introduction of risk in a choice situation and to reducing the concept of risk aversion to no more than the idea of decreasing marginal utility [28, pp. 109–110].

The conclusion is significant. It implies that the standard gamble should not be given special status. At best it may measure cardinal utility under certainty. At worst it introduces an additional, random element (g_1 and g_2 in the example) whose relationship to the specific utility of the risk associated with a medical procedure is unknown.

It is perhaps unsurprising that it has been hard to model the specific (positive or negative) utility of risk—or the (dis)pleasure of gambling, the (dis)pleasure of taking a risk, the direct dependence of utility on risk, the utility of the mere act of taking a chance or the specific utility of gambling as it has been variously called). Empirical evidence suggests that the emotions that contribute to the specific (dis)utility of risk are varied and complex and that their importance is dependent upon the context of the risk. For example, in her review of the subject, Pope reports that people are more prepared to take risks when the choice is voluntary, when avoiding the bad outcome depends partly on their own skills and control of the situation, when the bad outcome affects a less vulnerable subgroup of the household or nation, and when the general social atmosphere applauds risk taking [22, p. 15]. Further, the process of decision making does not always conform even approximately to N–M behaviour with some random variation attributable to an add-on utility of risk. Rather, heuristics are often adopted that are specific to particular contexts.

The more general reviews of the expected utility (EU) literature and the N–M axioms reveal that they are empirically flawed to such an extent that they cannot be assumed in any given context unless independently shown to be valid. Schoemaker, for example, concludes that:

> EU theory fails on at least three counts. First, people do not structure problems as holistically and comprehensively as EU theory suggests. Second, they do not process information, especially probabilities, according to the EU rule. Finally, EU theory as an 'as if' model, poorly predicts choice behaviour in laboratory situations. Hence it is doubtful that the EU theory should, or could serve as a general descriptive model [29, p. 552].

Froberg and Kane similarly note that at the individual level, expected utility maximisation is more the exception than the rule, at least for the types of decision tasks examined [13, p. 464].

One response to this evidence has been an attempt to reformulate the axioms in a way that avoids criticism. To date, this has not been achieved satisfactorily.* Fishburn concludes his review of these

*For example Machina [30] has attempted to avoid the use of the independence axiom, which has been a target for particular empirical criticism. The theory does not provide a convincing basis for a return to the expected utility hypothesis. Allais [5] and Pope [31] have argued that Machina's argument has been widely misunderstood, that it contains mathematical errors and that with any truly testable interpretation of his well defined and testable maxim, this collapses to the conventional expected utility hypothesis.

developments with the belief that "the next two decades or so will see numerous refinements, applications and experimental analysis of the types of representations described here" [6, p. 280]. The conclusion is driven by the inadequacies of the present axioms and their inability to systematically represent the utility of gambling. Another response has been to argue that the axioms proposed to date are imperfect but that they are the best available. The argument is unpersuasive. There is no methodological imperative for the adoption of an axomatic approach to the analysis of behaviour and very compelling reasons for its rejection in an applied context if a satisfactory, empirically robust set of axioms cannot be found. False assumptions result in highly unreliable conclusions.

The normative argument

The alternative interpretation of N–M utility offered by Torrance and Feeny [32] amongst others, is that its purpose is normative—that it indicates what individuals *should* do even if the outcome does not correspond with their own choice. They argue that:

> The von Neumann–Morgenstern utility theory is a normative model for individual decision making under uncertainty [32, p. 562]. . . . The theory and the methods of measurement were developed as a normative (prescriptive) model for individual decision-making under uncertainty. The model is general; it applies to decision-making in all fields, including health [32, p. 560].

The historical interpretation of von Neumann–Morgenstern is open to serious doubt.* Despite this, the interpretation of the axioms as having normative significance has had considerable appeal. Marschak [34] argued that the axioms defined rational behaviour and that their repeated application would ensure that "the probability that the achieved utility differs from the maximum achievable utility by an arbitrarily small number approaches unity [34, p. 139]. Ramsey [35] went further and argued that violation of the axioms to take account of the specific utility of risk signified inconsistent behaviour that would result in the decision-maker's failure to survive in a competitive environment.

The usefulness of this normative interpretation is questionable. The axioms result in the expected value of utility being accepted as the maximand for decisions under risk. But as both J. M. Keynes and Allais have noted, the outcome of a risky prospect is not its expectation (in [22]). If an outcome is sufficiently unpleasant it is not irrational to adopt a rule that avoids the outcome or, perhaps, to adopt a rule that maximises the value of the worst possible outcome. There is nothing specifically rational about adopting the probabilities of outcomes as weights in a one-off decision. Marschak's appeal to repeated outcomes is irrelevant in such cases and somewhat odd in view of the fact that Bernoulli initially introduced the expected utility hypothesis specifically to explain one-off choices, where expected values were not a possible outcome.

A more important defect in the normative argument is that, as noted earlier, the N–M axioms do not take account of the specific utility of risk and of the associated emotions. If these are relevant to welfare and can be taken into account in a decision rule, it is rational to do so. Thus, for example, Baumol argues "it is not the purpose of the Neumann Morgenstern Utility Index to set up any sort of measure of introspective pleasure intensity", and Arrow notes that "the utilities assigned are not in any sense to be interpreted as some intrinsic amount of good in the outcome" (in [7, p. 134]). It is not unreasonable for a rational individual to seek to maximise the good in the outcome, or the introspective pleasure intensity. Harsanyi finally accepts this when he argues that "it is only in such situations [where the specific utility of gambling is unimportant] that the N–M axioms represent acceptable rationality requirements. In particular, whenever process utilities [the specific utility of gambling] are important, the compound lottery axiom and their independence axioms lose their plausibility" [27, p. 307].

Finally, and perhaps most fundamentally, if N–M utility can be defined and understood only in terms of its axioms then the normative justification for their use involves a logical tautology. The usual argument is that behaviour consistent with the axioms should be adopted because this is the behaviour that maximises utility. But maximising utility means nothing more than adopting behaviour that is consistent with the axioms.

The conclusion of this section is that N–M utility cannot be accepted as the basis for the measurement of output in CUA because of its status in established economic theory. In both its positive and normative forms the theoretical arguments for using N–M utility as a gold standard are subject to serious defects.

3. CRITERIA FOR EVALUATING THE UNIT OF OUTPUT

There has been little discussion in the literature about alternative gold standards or criteria for selecting a QoL scaling technique that is appropriate for economic evaluation. In the psychometrics tradition, emphasis is almost exclusively upon the process of validation with the implicit or explicit assumption

*Savage for example, writes that "one idea now held by me that I think that Von Neumann and Morgenstern do not explicitly support and that so far as I know they might not wish to have attributed to them is the normative interpretation of the theory [of expected utility]" [33, p. 97]. This is consistent with von Neumann–Morgenstern's own introductory comments on the requirements for rationality: "It may safely be stated that there exists, at present, no satisfactory treatment of the question of rational behaviour. There may, for example, exist several ways by which to reach the optimal position but . . . [an analysis of this] is an exceedingly difficult task, and we may safely say that it has not been accomplished in the extensive literature about the topic" [9, p. 9].

that criterion validity cannot be achieved. It is not surprising that within this tradition there appears to have been little, if any, consideration of the characteristics that would make a measurement unit appropriate for the specific requirements of an economic analysis. Economists have been largely distracted by the debate over N–M utility. Even where this has not occurred analysis invariably assumes that the unit of output in CUA should be 'utility' or be based upon 'utility' [3]. It is suggested below that an alternative analytical approach should be adopted which avoids the ambiguities associated with the definition of utility.

These ambiguities are not confined to CUA. In his summation of the 1983 Oslo Conference on Utility and Risk Theory, Allais observes that 'most of the conflicts seemed to have arisen from the use of the same words to designate entirely different concepts—words such as 'probability', 'random variables', 'chance', 'utility', 'rationality' [4, p. 6]. Allais cites the following passage from Claude Bernard to express his concern.

> In creating a word to define a phenomenon, the idea it expresses is generally specified at that time together with its exact meaning. However, with the passage of time and the progress of science, the meaning of the word changes for some but keeps its initial significance for others. As a result there is often such a discordance that persons employing the same word mean very different ideas... if we focus on words rather than phenomena we stray quickly from reality [4, p. 5].

Both authors are echoing the Popperian argument about the inverted role of definition or what Popper describes as 'the essentialist method of definition'. In this, some word, X "is presumed to define some inherent essence or nature of a thing" [36, p. 20], which is presumed to be fundamental to the understanding of an issue. Debate focuses upon the question of 'what X really is or should be' or 'what are the properties of X?'. As noted by Bernard, ambiguity arises because the term is eventually used to describe different concepts but, because of the common terminology, there is a conviction that it describes some more universal concept whose properties need exploration and clarification. In the present context, 'X' is 'utility'. Because of the useful application of a concept of utility elsewhere in economic theory there has been an assumption that some general concept of utility must be the appropriate basis for output in the context of CUA. The resulting ambiguity has been recognised by some health economist [3]. For example, Labelle *et al.* argued that "the precise definition of utility has long eluded economists and decision scientist" [37, p. 29] and Culyer notes that "a... puzzle that pervades some welfarist theory is the meaning to be attached to the word 'utility'" [38, p. 293].

The alternative approach to definition is to determine, first, which concepts are relevant for a proposed solution, method or hypothesis, and subsequently to use definitions to abbreviate the description of the concept. In the present case the fundamental question should not be 'What is utility, what are its properties and how can it be measured?' but 'What objectives does the society seek to achieve through its health programs and how may they be measured?' If these objectives or the analysis arising from them employs one of the concepts of utility discussed earlier then the term can be used unambiguously with its meaning defined by the context of the analysis. In other words, the definition of utility or any other unit of output should follow from, and not be the initial subject of, the analysis.

Determining social objectives is a complex task. In cost benefit analysis it is simplified by distinguishing between productive efficiency and distributional objectives. The quantity produced is treated as an input into a social welfare function which takes into account distributional and other ethical considerations. However, the separation of these objectives does not imply that the unit for measuring production in CBA, the dollar, is value free. Rather, it implies widespread support for the judgment that, all else equal, what individuals are prepared to pay for a commodity is often an acceptable basis for measuring value. The acceptance of the dollar as a unit of measurement is also related to the fact that it is easily understood in terms of its purchasing power over other goods and serices.

The separation of the issues of distribution and productive efficiency has particular appeal in the health sector where questions of equity and distribution are highly contentious. The approach has been discussed by Wagstaff [3] who argues that the separation may overcome many of the legitimate criticisms of the QALY approach. While Wagstaff explicitly recognises that utility adjusted life years (QALYs) are not a measure of utility in the 'welfarist/utilitarian' tradition they are, nevertheless, treated as the appropriate unit for measuring the quantity of health. The appropriate technique for measuring utility is not addressed. The remainder of the present paper is concerned with the prior issue of the evaluation of measurement units and the scaling techniques that produce them.

As the discipline of psychometrics is concerned with the measurement of imprecise psychological concepts it is not surprising that its scaling techniques were first used to measure preferences for 'health' and 'quality of life'. These techniques—the rating scale and magnitude estimation—remain the basis of a significant proportion of CUA. It is probably for this reason that some researchers have concluded that the units of measurement in CUA can be no more precise than in other cases of psychometric measurement. A degree of conceptual imprecision—vagueness—is inevitably introduced by the treatment of 'utility' in its original sense, i.e. as a psychological construct linked in some unobserved way to behaviour. For example, Kind argues that "the efficiency with which any

scaling procedure is able to capture and represent personal preferences for health states is largely unknown, since no standard values have been, or are likely to be promulgated" [39, pp. 11, 12]. Similarly Froberg and Kane argue that "in scaling preferences we are concerned with an abstract variable or 'construct' rather than an observable one. To define this abstract variable and determine what a particular scaling method actually measures requires methods of construct validation" [13, pp. 466–467].

The conclusions are unduly pessimistic. The fact that the concepts of 'QoL' and 'preferences' may be abstract does not imply that an individual's expression of 'preferences' and the scale for measuring those preferences must be correspondingly abstract. Analogously, the concept of 'consumer preferences' or 'tastes' in consumer theory is abstract and its direct measurement would encounter significant difficulties. However, the external expression of these tastes as 'revealed preferences' through dollar expenditures normally provides a satisfactory basis for resource allocation. In the CUA literature it has yet to be demonstrated that a unit of measurement appropriate for the *allocation of resources* cannot or has not been found that eliminates, at least to a significant degree, the vagueness of the psychometric measures.

As noted, this issue has received little attention in the economics literature* and the appropriate criteria for the evaluation of the unit of output have been almost ignored. An important exception is the recent suggestion by Nord [42] that as the purpose of measurement is to quantify preferences for different outcomes then a criterion test of validity may be the test of 'reflective equilibrium', i.e. determining whether values revealed by various instruments correspond with the values that are directly elicited from trade-off questions. This amounts to the suggestion that such trade-off questions should replace the standard gamble as the gold standard. Criteria are proposed below which could be used to evaluate Nord's suggestion.

The criteria for selecting a unit of output should follow from the social objectives and from the practical requirements of such a measurement unit. Four are suggested here. They are pragmatic in the sense that they are not derived from an established theory—at present there is no such acceptable theory. The criteria, requirements, do not purport to be exhaustive. It is argued, however, that they should be regarded as necessary, if not sufficient, conditions that should be met.

As the purpose of measurement is to obtain a basis for the allocation of resources, the first criterion is that, *all else equal*, more units are considered to be better than fewer. As economic evaluation must reflect social values, this implies that there should be a broad consensus that the units correspond with a socially desired outcome. In cost effectiveness analysis the use of 'lives saved' or 'life years gained' are examples where there would be a broad social consensus that, *all else equal*, more units would be preferred to fewer. Similarly, in CBA 'dollars' meet the criterion as more dollars represent an increased capacity to obtain satisfaction from goods and services elsewhere.

Secondly, the unit of measurement should have, as far as possible, a clear and unambiguous meaning. The end point of an economic evaluation should be information that is persuasive: it should help to convince decision makers that a program is, or is not, desirable. This is less likely to occur if the measure does not appear to have intrinsic plausibility or if the measure is incomprehensible to all but a small group of evaluation experts. More importantly, it is unlikely that projects will be ranked entirely in terms of their costs and benefits as defined by the chosen units. Rather, distributional and political objectives are likely to intervene and when trade-offs are made between conflicting objectives it is necessary to understand clearly what the trade-off entails.

The need for easy comprehension is noted by Mosteller when he reports his personal experience with 'talented lay people' responsible for allocating resources between alternative medical technologies. He notes that "they wanted to know what different technologies will produce ... what the benefits and losses would be, but they do not like to have these complicated problems summed up in single numbers. In using quality adjusted life years or any other cost benefit analysis summaries, they felt something was being concealed from them, and they did not understand how the work was being done" [1, p. 285].

The third criterion is that the unit of measurement has a *meaningful* interval property to permit the summation of benefits. Since CUA is concerned with *changes* in health states, a ratio property is not necessary. However, *differences* in the numerical value of the units must be directly and proportionally related to differences in the benefits measured by the units. The term 'meaningful' is used here to emphasise that the property should be recognisably related to the magnitude of the benefit received and not be an artefact of the scale. In the previous example both 'life years' and 'lives saved' have a meaningful property. The difference between two lives and one life can be readily understood. *All else equal*, the former would represent twice the benefit of the latter and, in the absence of budget constraints or other relevant

*Torrance briefly considers this issue and argues that "proper weights should be non arbitrary, community based, scientifically measured values reflecting the relative desirability of [strength of preferences for] the various states of health. This requires the availability of a measurement instrument (or instruments) of proven reliability and validity which can be used on the general public to quantify the preferences for the relevant states of health. No such instrument has been reported in the literature to date" [40, p. 129]. Torrance, however, subsequently accepts that validity may be determined by correspondence with the outcome of the standard gamble [41].

considerations, obtaining two lives would be expected to justify twice the expenditure. Similarly, and again setting aside the issue of budget constraints, the willingness to pay a maximum of $\$y$ for one type of QoL improvement and $\$2y$ for another improvement may be understood in terms of the value of other goods and services that an individual is prepared to forego to obtain the QoL improvements and the latter case may be meaningfully interpreted as indicating twice the benefit obtained by the former, as judged by the individual.

By contrast with these examples, and taking an extreme case of an artefactual property, individuals could be asked to rate from 0 to 10 the 'circularity' of geometric figures. Respondents may oblige and attach numbers to shapes. These might be indicative of an ordinal ranking as judged by some unknown criterion. However, it would have little meaning to argue that a shift from 0 to 2.0 on the scale indicated the same increase in 'circularity' as a shift from 7.0 to 9.0. Rather, the numerical values would be a product of the scale and some unknown criterion of the respondent. In this case there is little semblance of meaning to the interval property unless the individual's criterion of circularity is known. Froberg and Kane [13] note that a limitation of many approaches to measurement is that the assumption of an interval scale is based upon definition. Both Mulkay *et al.* [43] and Nord [44] similarly express concern that the numerical properties of utility weights may be artificial products of the survey instrument. Elsewhere in the literature the issue has received limited attention.

The criterion discussed above might be termed the 'weak interval requirement' or criterion. It is derived primarily from the need to ensure that as QoL improves, quantitative comparisons can be made that are not artefactual. However, a much more demanding interval criterion must be met if QALYs are to fulfil the purpose usually assigned to them in CUA. In this, no distinction is drawn between QALYs gained because of life extension or life improvement. In the conventional QALY calculation which separates life years from the index of quality, a 10% increase in QALYs may be achieved by a 10% increase in either the QoL (utility) index or the life years. This implies that the units for measuring QoL should not simply meet the weak property but that an increase in the QoL index should be equivalent in some meaningful way to a similar percentage increase in life years. The most important (and controversial) contribution of QALYs is that they trade-off QoL and life itself. The strong interval criterion is, in effect, that the basis for this trade-off should be visible and comprehensible.

In the earlier example, the willingness to pay for health improvement results, under certain circumstances, in the unit of payment, the dollar, having a weak interval property. If the additional (probably false) assumptions were made that individuals were well informed and that risky choices were unaltered by linear transformation then the dollar value of life could validly be inferred from the willingness to pay for a reduced risk of death (contingent valuation). Under these circumstances dollars would meet the strong interval criterion for a measurement unit. An increased willingness to pay for health improvement and for life *per se* would have an equivalent meaning in terms of the strength of an individual's preferences.

The fourth criterion is that there should be a simple and practical scaling technique for mapping health states into the unit of measurement and that the technique should be sensitive to variation in health states, reliable and valid (i.e. that it should measure the units that it purports to measure). These are the measurement properties most commonly discussed in the literature [13, 41, 45]. The requirements are a prerequisite to, but separable from the practical application of the second criterion. Unless a satisfactory scaling technique exists the units of measurement may be theoretically attractive but unpersuasive in practice. For example, it may be conceptually appealing to ask individuals to convert into a utility index a whole-of-life scenario in which health and social states change with age and disease progression. However, in the absence of a reliable and valid technique for converting such scenarios into a numerical score the approach can not be operationalised. This fourth criterion is again a recognition of the fact that economic evaluation must eventually be an applied discipline.

4. EXISTING SCALING TECHNIQUES

This discussion of criteria may seem to imply that a number of measurement units have been proposed in the literature, each with clearly distinguished properties that could be evaluated with the proposed criteria. In fact, practitioners of CUA have given little explicit attention to the properties of the measurement units used. With the exception of the healthy year equivalent (discussed later) these units have consisted of weighted life years, where the weights have purported to represent indices of 'QoL' or 'utility'. The various scaling techniques used to produce these weights have been assumed appropriate for the task of producing homogeneous 'quality adjusted life years'. As evidenced by the comparison of QALY results in 'league tables' it has also been commonly assumed that the different techniques produce the same or similar adjustment weights. Some have explicitly argued that the different techniques produce comparable results or at least results that can be reconciled by simple transformation [13, 32, 41]. Empirical comparisons do not support this belief (for reviews see [42, 46]). The differences suggest that the techniques involve qualitatively different cognitive tasks [14, 42] and that the adjusted life years produced by each technique may represent distinct measurement unit.

The five most commonly used scaling techniques

are described in the appendix. It is possible to discriminate between the units produced by each of these techniques—the adjusted life years—not in terms of their correspondence with utility theory—but by the extent to which they satisfy the criteria discussed above. Such an evaluation has two parts. First, abstracting from practical issues such as sensitivity and reliability, what will a technique measure? (Criteria 1–3). Second, do practical problems alter or constrain the choice of technique? (Criterion 4). This final question is empirical and is not considered here.

Rating scale and magnitude estimation

Both the rating scale (RS) and magnitude estimation (ME) have produced results in the cost utility literature that are empirically different from the time trade-off (TTO) and standard gamble (SG). A proximate reason for this is that the latter two techniques involve choice, whereas the RS and ME do not [42]. It is known that individual responses are sensitive to context and the framing of questions. It is therefore possible that differences in scale values are entirely attributable to these effects. However, the second possibility is that the techniques do not measure the same underlying quantity or that one or both of the units produced by the techniques do not meet criteria 2 or 3. That is, the techniques may not lead to a unit with a clear meaning and interval property.

Taken literally, the scores on a RS give the *distances* along a calibrated linear scale which represent to subjects, in some sense, the value or worth of a health state relative to the reference points on the scale. In psychometric theory these scores are related to a psychological construct by an unknown functional relationship [14]. The question that the rating scale leaves unanswered concerns the functional relationship between the units of the scale on the one hand, and on the other hand the welfare, utility or attribute that is believed to be appropriate as the basis for resource allocation. Empirical inquiry into this issue indicates that individuals are not themselves able to provide an explanation of their own responses in these terms [44, 47, 50]. This suggests that the RS does not meet the weak interval criterion, that there should be a meaningful interval property.* As the RS is normally administered in a way that makes no reference to a relationship between scale values, life and length of life, it is even less likely that the results from the RS could meet the strong interval criterion.

With ME, subjects are asked a question of the form 'How many times worse is one state than another [reference] state?'. The response is unconstrained. As there is no universally accepted scale for health states, the meaning of the question is deeply obscure. Presumably subjects must translate the question into an equivalent TTO or RS question to produce an answer. Alternatively, subjects may give a purely subjective response which does not purport to have an independent meaning. As different individuals are likely to use different heuristics to answer the question, it is again difficult to place a clear meaning on the final scale and the resulting QALY units (Criterion 2). The phrase 'how many times worse' may result in values which are appropriate for a trade-off between quantity and quality of life (Criterion 3). To date there is little evidence or reason to encourage this belief.

These ambiguities are not confined to the context of health. In the psychometrics literature there have been significant differences about the meaning and appropriateness of each of the scales, the existence and the meaning of an interval property.* Some have claimed that the two scales should give the same result [51]. Others have agreed with the suggestion first made by Marks [52] that the two scales measure fundamentally different psycho-physical process. In virtually all empirical studies in the general psychometric literature the results of category rating scales are not linearly related to magnitude estimates [14]. This result has also been observed in the context of health measurement [46, 53, 54]. The unavoidable conclusion is that one or other of the scales does not have the interval property that is required by the third criterion.‡ A further possibility is that neither has the property required and that the apparent interval property of each of the techniques is an artefact imposed by the scale that respondents have been required to use. In the conclusion to a recent survey of psycho-physical scaling, Gescheider notes that "perhaps the most perplexing and certainly one of the oldest problems in psycho-physics is the observed non-linear relations between scales obtained by different methods. Whether these non-linearities are due to cognitive—judgment factors or to sensory—perceptual factors is yet to be determined. Two scaling procedures that apply to the same perceptual dimension cannot both be valid" [14, p. 194]. The comment underscores the difficulty in placing a meaning on the units obtained from these scales and the difficulty in judging their suitability as a basis for allocating resources.

*The method of successive intervals is one approach that is claimed to produce an interval property [48]. The technique essentially adjusts interval widths in order to impose a normal distribution upon data. The interval property does not appear to have any meaning that is independent of the construction. Consequently it is questionable whether this adjustment to scale values improves their suitability as a basis for *resource allocation*.

†For a review of the debate see Ref. [13] and for empirical evidence see Ref. [49].

‡It has generally been found that results from the RS and ME are related by a power function. This does not help resolve the issue of which scale, if either, has the required interval property. Further, there does not appear to be a single power function that will transform health-outcome results consistently [46].

The N–M standard gamble

The N–M standard gamble (SG) purports to measure von Neumann–Morgenstern utility. More precisely, the SG would measure N–M utility if the N–M axioms were acceptable but, as discussed in Section 2, they are not supported by the empirical evidence. The outcome from the technique is, strictly, a probability, p, which makes an individual indifferent between a certain outcome and a probabilistic choice. However, even if the N–M axioms are not *generally* true, SG based QALYs may be acceptable measurement units as judged by the criteria proposed here. First, they reflect choice. All else equal, higher values of p should be preferred. Secondly, probabilities have an interval property. There is an objective sense in which, for example, an increase in probability from 0.2 to 0.4 is quantitatively equivalent to an increase from 0.7 to 0.9.* As the probability is obtained by a direct comparison between quality of life and expected length of life, the units obtained might also meet the strong interval criteria. Thirdly, it can be argued that, by contrast with the TTO and PTO which also measure choice, the SG necessarily reflects an attitude towards risk and that any medical interventions to be assessed must also involve risk. This amounts to an assertion that in *the present context* the specific utility of risk is either quantitatively insignificant, or that it is a positive advantage by increasing the realism of the choice context. Finally, while it may be conceded that probabilities are not a unit that can easily be comprehended by those untrained in statistics, the results of the standard gamble can be translated into more readily understood 'Healthy Year Equivalents' (HYEs) [55].

Despite these properties there are serious defects in the standard gamble as a measurement instrument. The strong interval criterion would only be met if the N–M axioms were correct.† It is also questionable whether the weak interval property is met and differences between probabilities are a true measure of the differences in the intensity of an individual's choice or whether the interval property is a property of the scale that is imposed on ordinal values. Empirical evidence suggests that people have difficulty understanding the objective meaning of probabilities, especially extreme values [29]. Finally, it is open to doubt whether the risk situation created by the standard gamble improves or detracts from the realism of the choice context and, consequently, whether it adds to, or subtracts from the meaningfulness of the interval property.

This last issue is particularly important as it is the risk situation created by the SG that has made the technique attractive to many analysts. However, while it is true that many medical interventions involve risk, this usually takes the form of a particular (known or unknown) probability of a transition to a particular health state. By contrast, the risk introduced by the (N–M) standard gamble is part of the technique for measuring the 'utility' (as opposed to the 'value') of the health state itself. It is *not* intended as a means of measuring the importance of the probability of transition into or out of that state. The counter argument is that the SG at least introduces an element of 'risk' if only in some general sense and that measurement 'under risk' is desirable. However, the risk modelled by the usual SG is the result of a singularly unrealistic situation in which the individual faces instant death as a possible outcome from one of the two choices. The context is totally dissimilar from a health scenario involving the possibility (with unknown probability) of, for example, some non-life-threatening chronic disease. The empirical evidence on risk behaviour referred to in Section 2 reveals such a diverse, context-specific range of behaviour that these two situations must be regarded as being quite distinct. Further, the value of p in the SG depends primarily upon the unpleasantness of the health state, S, which is described under conditions of certainty. In reality, S may occur in conjunction with very significant uncertainty or with negligible uncertainty. Yet the same SG is believed to capture the essence of both risk contexts. Clearly p cannot reflect real-world uncertainty when information about the nature and magnitude of this is not given to subjects. While particular examples can be found where the risk of death during an operation may correspond fairly closely to the risk embodied in the SG, in many—and probably most—instances the only similarity between the risk in the SG and the real-world health state is that the lack of certainty in both cases can be loosely described as 'risk'. When the usual distinction is made between 'risk' and 'uncertainty' even this similarity may end, as individuals often experience the latter and not the former.

As they are presently used, the other techniques employed in CUA abstract from risk. However the abstraction *per se* is not a defect. The defect is in the abstraction from the risk or uncertainty *associated with* the intervention being evaluated. The introduction of irrelevant risk considerations cannot improve

*There has been extensive debate about the meaning of the term 'probability' [4]. The standard gamble is often implemented with the assistance of a 'probability wheel' which presents probabilities as being similar to the chance of winning or losing on a roulette table [41]. When envisaged this way, an increase in the probability from 0.2 to 0.4 and from 0.7 to 0.9, both raise the likelihood of a successful outcome by the same amount. This objective interpretation of probabilities—which is explicit in the N–M axioms—should not be linked to the issue of diminishing marginal utility. Health states may or may not have to be improved disproportionately (in some other sense) in order to increase (N–M) utility by the same incremental amount as (N–M) utility rises. The interval property refers to the *probability* used for measuring preferences and not to the health state *per se*.

†In order to extrapolate from a choice involving a probabilistic outcome to a choice in which non probabilistic life—QoL comparisons are made, preferences under risk must be invariant under linear transformation.

assessment.* In principle, the ideal measurement could be achieved by the inclusion of a statement about relevant risks and uncertainties in the health state description. There would, of course, be a practical constraint because of the complexity of the resulting description and the subject's capacity to comprehend risk and uncertainty in the context of a limited interview. The trade-off between this and other aspects of the description sacrificed in order to include the risk statement would depend upon the relative importance of these factors as assessed at the pre interview stage.

A variation of the SG proposed by Mehrez and Gafni [56] is to use a two stage procedure to derive healthy year equivalents (HYEs) from the SG. By converting from a probability to a HYE the measurement unit appears to become more acceptable as judged by the second criterion.† In the first stage the utility of a fixed number of years in a state of ill health, n, is evaluated with the SG to obtain an index of utility from the value of the probability, p. In the second stage, p is held constant at the level found in stage 1 and a gamble conducted to determine the number of healthy years that are equivalent to the n years of ill health. That is, the two stages take the following form:

Stage 1:

p. (full health for n years) $=n$ sick years ...
$+(1-p)$ death — p varies, n fixed

Stage 2:

p. (full health for n years) $=h$ healthy years ...
$+(1-p)$ death — h varies, p fixed

The logical format of the two stages is:

$$a = b$$
$$a = c$$
$$\text{from which} \quad b = c$$

Unless the Stage 1 equation for the SG (i.e. $a = b$) omits some magnitude, there is no purpose in using the SG and a direct trade-off between sick years and healthy years would be appropriate. The true situation implied by the need for the two-stage procedure is that:

$$a = b - g_1$$
$$a = c - g_2$$
$$\text{from which} \quad b = c + (g_1 - g_2)$$

where g_1 and g_2 are the specific (dis) utilities of gambling in the two stages. It is possible that $g_1 = g_2$, in which case there is, once again, no purpose in the two-stage procedure. If $g_1 \neq g_2$ the previous question must be answered, namely, whether the specific (dis) utilities of gambling in this context capture the behaviour of individuals in dissimilar risk contexts. There is, however, an additional anomaly. Stage 2 is simply a device for translating utility into healthy year equivalents. It does not measure some element of behaviour, as in stage 1. If stage 1 is believed to capture the essence of risk behaviour, that is, that g_1 captures the specific utility of gambling, then the net effect $(g_1 - g_2)$ cannot also do so unless $g_2 = 0$. There is no reason to believe that this is so.

In sum, the SG is a technique that assesses a riskless health state by means of a gamble. The technique cannot produce a probability that reflects the single value (utility) of the health state while simultaneously reflecting all of the individuals' diverse responses to all possible levels of risk (and uncertainty): large and small risks; present and future risks; risks with respect to life, morbidity and financial status; and risk in a multitude of different contexts. The SG will, however, reflect the preference (like or dislike) for the specific risk embodied in the technique, and it will reflect the unknown decision rule applied by individuals in this particular context. The inclusion of this irrelevant and random element implies that SG based QALYs will not meet criterion 3. There does not appear to be, at present, a satisfactory technique for translating the outcome of the SG into an easily understood unit as required by criterion 2.

Time trade-off and person trade-off

The end point of cost utility analysis, at least in its present form, is the derivation of homogeneous life years as a unit of output. A common feature of the three techniques discussed above is that they carry out this task in a way that requires two separate functional relationships. The first is the relationship between the health state and the scale; the second is the relationship between the scale and the homogeneous year of life envisaged by CUA. The common criticism of each of these techniques has been that the nature of the second relationship is not known. In the case of the RS there is an unknown relationship between the linear distances recorded on the scale and the true preference for life years. With the SG the existence of a specific utility of risk confounds the relationship between the probability revealed in the gamble and the preferences for (risky or riskless) life years.

The great appeal of both the time trade-off (TTO) and the person trade-off (PTO) is that these two relationships are collapsed into one and the health state to be evaluated is contrasted directly with a reference state which acts as a measurement unit. Thus, the TTO directly measures the number of healthy years that are equally valued as—considered equivalent to—a given period in a health state. As there is no additional scale used in the calculation, one important source of ambiguity is removed. Consequently, the following scenario is possible with each of

*If this were not so, the other techniques could be improved by including in the health-state description the fact that there would be an equal but unknown chance of either a $10,000 bonus or fine, irrespective of other considerations!

†The distinctive feature of the HYE is not the use of the SG nor the two stage procedure discussed here, but the derivation of units from multi-year scenarios.

the three techniques discussed earlier, but it is not possible with the TTO or PTO: an individual might imply through the use of one of these instruments that the utility of n years in health state S_1 was superior to the utility of n years in a health state S_2 but the same individual could indicate directly that his or her *preference* was for the latter and not the former option. In Nord's terminology [42], the earlier techniques could fail the test of reflective equilibrium.

The units produced by the TTO and by the PTO have an interval property that satisfies both the weak and the strong interval criteria. There is a clear meaning to the statement that (all else equal) six healthy years are double three healthy years, i.e. that the duration of the flow of benefits of living six years is twice the duration of living three years and that there is an objective standard for measuring duration vis, calender years.* Similarly, there is a clear, comprehensible, meaning to the outcome of the PTO. Returning X people to full health from one health state is considered equivalent to Y people experiencing full health rather than another (reference) state—possibly death. If the number of years in these states is specified, and the reference state is death the meaning is very similar to the meaning of the TTO. The unit of measurement may be interpreted as the healthy or normal year of life. With both techniques the measurement units are obtained by requiring subjects to trade-off directly the quality and quantity of life. This satisfies the strong interval criterion.

Despite their similarities the two techniques are conceptually different. TTO values are derived from an interview in which an individual is asked to imagine that (s)he personally experiences the health state being evaluated. With the PTO, individuals are asked to make an impersonal judgement about changes that could affect others. The personal values revealed by the TTO are, in principle, closer to the outcome of consumer sovereignty than the arm's-length evaluation judgements revealed by the PTO. Consequently, results from the two techniques could vary because of a systematic difference between choice criteria applied to personal decisions and those applied to social decisions. In principle, the choice between the two techniques would therefore depend upon a judgement concerning the importance of incorporating libertarian or paternalistic values as the basis for measuring the quantity of output. The choice would also be affected by practical measurement issues (Criterion 4). These have not been addressed in this paper.

*The existence of an interval property defined in terms of duration of the benefits does not imply that the sum total of the psychological or other benefits obtained during this period have a similar interval property. The intensity of benefits may vary from person to person [3] and the relationship between psychological and physical time is obscure [57]. This implies that neither the TTO nor PTO embody purely utilitarian values. However, this is true of QALYs generally [3] and the criterion proposed here is the existence of a meaningful—intelligible—interval property and not simple utilitarism.

5. CONCLUSION

Cost utility analysis must take into account both the quantity of output produced and its distribution. The latter issue requires an explicit consideration of the social welfare function which has not been undertaken here. The former issue implies the existence of a satisfactory unit of measurement. With some dissension, it has been generally accepted by economists that QALYs represent such a unit. By implication it has also been accepted that the various scaling techniques used to produce QALY values are all measuring the same quantity, namely 'utility'.

This latter assumption is hard to sustain in the face of the empirical evidence now available. Rather, it appears that different techniques are measuring different 'quantities' that are the result of distinct psycho-social processes. This implies that a choice between the techniques cannot be avoided. This raises the general question posed in the title of the present paper namely, 'what should be measured to achieve the objectives of CUA?'.

The first conclusion reached in the paper is that the answer to this question should not simply be 'utility'. This would not identify which, if any, of the four concepts of 'utility' discussed here was the appropriate choice. In the literature, the term has been used inconsistently. For some it has retained its original utilitarian meaning of subjective satisfaction or well-being. For others, the term is now understood in terms of preference utilitarianism which makes no reference to risk. Yet another group accepts the von Neumann–Morgenstern approach in which 'utility' may only be measured 'under risk' and is defined by a particular set of behavioural axioms. It has been argued here that the special status given to this last approach because of its basis in economic theory cannot be justified. In addition to conflicting empirical evidence, the *theoretical* basis is flawed.

The second conclusion of the paper is that because of this unsatisfactory situation the requirements of a measurement unit for CUA should be explicitly considered. The analysis should not commence with the presumption that 'utility' is the natural basis for measurement and that all that is required is a clarification of the concept. This approach elevates differences in the use of words to the status of a 'conceptual issue'. The alternative approach is to evaluate possible measurement units in terms of explicit criteria. These should include the usual criteria of sensitivity, validity and reliability but, in addition, include criteria relevant to the specific requirements of CUA. As these include the combination of quantity and quality of life into a single unit, it has been argued here that the unit must not simply have a meaningful interval property but one where the nature of the trade-off between quantity and quality of life can be readily

comprehended: that is, when the value of the unit increases because of the change in the quality of life, and when it increases by an equal amount because of the quantity of life the reason for the equality can be understood.

The application of this and the other suggested criteria leads to the final major conclusion that the TTO and PTO techniques fulfil the requirements of CUA to a greater extent than the RS, ME or SG. The conclusion is tentative as the issues of sensitivity, validity and reliability have not been considered. Neither the RS nor ME techniques produce clearly meaningful units. Consequently, the relationship between increased QALY values arising from increased quantity and quality of life is obscure when the QoL is measured by these techniques. By contrast, the SG is based upon an explicit trade-off between expected quantity and quality of life. However, the interval property of the SG based QALY depends upon the subject's capacity to act consistently as objective probabilities vary, that is, to act in accordance with the empirically invalid N–M axioms of choice. Additionally, the existence of idiosyncratic, context dependent risk behaviour introduces an extraneous factor into the measurement task. As a consequence, the units produced by the SG are contaminated and both the meaning of the units and their interval property are called into question.

Neither the TTO nor the PTO are obviously subject to these defects. Unlike the other three techniques there is no intermediate scale imposed between the health state and the final measurement unit. Direct comparisons are made which require subjects to consider explicitly the trade-off between quantity and quality of life in such a way that the strong interval criterion is fulfilled. As with all of the measurement units proposed to date, there remains an unanswered question about the relationship between stated and revealed preferences, that is, the issue of validity is not satisfactorily resolved. Additionally, the reliability and sensitivity of the TTO and PTO have received little attention in the literature. Despite these qualifications, the units produced by these techniques fulfil the prerequisites identified here as essential for an output measure for cost utility analysis.

REFERENCES

1. Mosteller F. Final panel: comments on the Conference on Advances in Health Status Assessment. *Med. Care* **27,** Suppl. S2H2–S286, 1989.
2. Mooney G. and Olsen J. A. QALYs: where next? *Providing Health Care: The Economics of Alternative Systems of Finance and Delivery* (Edited by McGuire A. and Fenn Pond Mayhew K.). Oxford University Press, Oxford, University Press, Oxford, 1991.
3. Wagstaff A. QALYs and the equity–efficiency trade-off. *J. Hlth Econ.* **10,** 21–41, 1991.
4. Allais M. The foundations of the theory of utility and risk. Some central points of the discussion at the Oslo Conference. Cited in *Progress in Utility and Risk Theory* (Edited by Hagen O. and Wenstop F.). D. Reidel, Dordrecht, 1984.
5. Allais M. A new neo-bernoullian theory: the machina theory. A critical analysis. *Risk Decision and Rationality* (Edited by Munier B.). D. Reidel, Dordrecht, 1988.
6. Fishburn P. C. Expected utility; an anniversary and a new era. *J. Risk Uncertainty* **1,** 267–283, 1988.
7. Fishburn P. C. Retrospective on the utility theory of von Neumann and Morgenstern. *J. Risk Uncertainty* **2,** 127–158, 1989.
8. Hagen O. and Wenstop F. (Eds) *Progress in Utility and Risk Theory*. D. Reidel, Dordrecht, 1984.
9. von Neumann J. and Morgenstern O. *Theory of Games and Economic Behaviour*. Princeton University Press, Princeton, 1944.
10. Stigler G. J. The development of a utility theory 1, 2. *J. Polit. Econ.* **58,** 307–327, 373–396, 1950.
11. Bentham J. *Introduction to the Principles of Morals and Legislation*, 1789.
12. Viner J. Utility concept in value theory and its critics. In *Utility Theory: A Book of Readings* (Edited by Page J.) Wiley, New York, 1925.
13. Froberg D. G. and Kane R. L. Methodology for measuring health state preferences. *J. clin. Epidemiol.* **42,** 345–354, 459–471, 585–592, 675–685, 1989.
14. Gescheider G. A. Psycho-physical scaling. *Am. Rev. Psychol.* **39,** 169–200, 1988.
15. Hicks J. R. and Allen R. G. D. A reconsideration of the theory of value. *Economica*, pp. 52–75 Feb., pp. 196–216 May, 1934.
16. Graaff J. *Theoretical Welfare Economics*. Cambridge University Press, Cambridge, 1967.
17. Philips L. *Applied Consumption Theory*. North Holland, Amsterdam, 1974.
18. Baumol W. J. *Economic Theory and Operations Analysis*. Prentice Hall, Englewood Cliffs, NJ, 1972.
19. Lane D. A. Utility, decision, and quality of life. *J. chron. Dis.* **40,** 585–591, 1987.
20. Robinson J. *The Accumulation of Capital*, 2nd edn. Macmillan, London, 1965.
21. Pope R. The pre outcome period and the utility of gambling. In *Foundations of Utility and Risk Theory with Applications* (Edited by Stigum B. and Wenstop F.), pp. 137–177. D. Reidal, Dordrecht, 1983.
22. Pope R. E. Additional perspectives on modelling health insurance decisions. In *Economics and Health* (Edited by Selby Smith C.). Public Sector Management Institute, Monash University, 1988.
23. Gafni A. and Torrance G. W. Risk attitude and time preference in health. *Management Sci.* **30,** 440–451, 1984.
24. Gafni A. The standard gamble method; what is being measured and how it is interpreted. CHEPA *Working Paper Series*, 91–94. MacMaster University, Hamilton, Ontario, Canada, 1991.
25. Morgenstern O. Some reflections on utility. *Expected Utility Hypothesis and the Allais Paradox*. (Edited by Allais M. and Hagen O.). D. Reidel, Dordrecht, 1979.
26. Watkins J. W. N. Towards a unified decision theory: a non-Bayesian approach. In *Fundamental Problems in the Special Sciences* (Edited by Butts and Hintikka), pp. 347–379. Reidel, Dordrecht, 1977.
27. Harsanyi J. Use of subjective probabilities in games theory. In *Foundations of Utility and Risk Theory with Applications* (Edited by Stigum B. and Wenstop F.). Reidel, Dordrecht, 1983.
28. Bouyssou D. and Vansnick J. C. A note on the relationships between utility and value functions. In *Risk, Decision and Rationality* (Edited by Munier B. R.). Reidel, Dordrecht, 1988.
29. Shoemaker P. The expected utility model: its variants, purposes, evidence and limitations. *J. econ. Lit.* **20,** 1529–563, 1982.
30. Machina M. Expected utility analysis without the independence axiom. *Econometrica* **50,** 277–323, 1982.
31. Pope R. E. Machina's decision model: an empty box? Mimeograph, Department of Economics, University of New South Wales, Campbell ACT, Australia, 1989.

32. Torrance G. W. and Feeny D. Utilities and quality adjusted life years. *Int. J. Technol. Assess. Hlth Care* **5**, 559–575, 1989.
33. Savage L. J. *The Foundations of Statistics*, 2nd revised edn. Dover, New York, 1954.
34. Marschak J. Rational behaviour, uncertain prospects, and measurable utility. *Econometrica* **18**, 11–141, 1950.
35. Ramsey F. Truth and probability. In *The Foundations of Mathematics and other Logical Essays* (Edited by Braithwaite R.). Humanities Press, New York, 1950.
36. Popper K. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge and Kegan Paul, New York, 1963.
37. Labelle R., Feeny D. and Torrance G. Conceptual foundations of health status and quality of life utility measures mimeograph, Department of Clinical Epidemiology and Biostatistics/Centre for Health Economics and Policy Analysis, McMaster University, 1989.
38. Culyer A. J. The normative economics of health care, finance and provision. *Providing Health Care: The Economics of Alternative Systems of Finance and Delivery*. (Edited by McGuire A., Fenn P. and Mayhew K.). Oxford University Press, Oxford, 1991.
39. Kind P. The development of health indices. *Measuring Health: A Practical Approach* (Edited by Teeling Smith G.). J. Wiley, New York, 1988.
40. Torrance G. W. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio Econ. Planning Sci.* **10**, 129–136, 1976.
41. Torrance G. W. Measurement of health-state utilities for economic appraisal: a review. *J. Hlth Econ.* **5**, 1–30, 1986.
42. Nord E. Methods for quality adjustment of life years. *Soc. Sci. Med.* **34**, 5, 1992.
43. Mulkay M., Ashmore M. and Pinch T. Measuring the quality of life. *Sociology* **21**, 541–564, 1987.
44. Nord E. A comment on the meaning of numerical valuations of health states. *Soc. Sci. Med.* **30**, 943–944, 1990.
45. McDowell I. and Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford University Press, New York, 1987.
46. Richardson J., Hall J. and Salkeld G. Cost utility analysis: the compatibility of measurement techniques and the measurement of utility through time. In *Economics and Health: Proceedings of the Eleventh Australian Conference of Health Economists* (Edited by Selby Smith C.). Public Sector Management Institute, Monash University, 1990.
47. Morris J. and Durand M. A. *Category Rating Methods: Numerical and Verbal Scales—Results from a Pilot Study*. Mimeograph, Centre for Health Economics, University of York, Heslington, 1989.
48. Blischke W. R., Bush J. W. and Kaplan R. M. Successive intervals analysis of preferences measures in a health status index. *Hlth Services Res. Summer*, pp. 181–198, 1975.
49. Kaplan R. M. and Ernst J. A. Do category rating scales produce biased preference weights for a health index? Med. Care **21** 193–207, 1983.
50. Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int. J. Hlth Plan. Management* **6**, 234–242, 1991.
51. Brooks R. G. *Scaling in Health Status Measurement: An Outline Guide and Commentary*. Institute of Health Economics Report, The Swedish Institute for Health Economics, Lund, 1988.
52. Marks L. E. On scales of sensation: prolegomena to any future psychophysics that will come forth as science. *Percept. Psychophys.* **16**, 358–376, 1974.
53. Kaplan R. M., Bush J. W. and Berry C. C. Category rating versus magnitude estimation for measuring levels of well being. *Med. Care* **27**, 501–521, 1979.
54. Kind P. The design and construction of quality of life measures. Discussion Paper 43. Centre for Health Economics, Health Economics Consortium, University of York, United Kingdom, 1989.
55. Mehrez A. and Gafni A. Quality-adjusted life years, utility theory, and healthy-year equivalents. *Med. Decision Making* **9**, 142–149, 1989.
56. Mehrez A. and Gafni A. Healthy year equivalents: how to measure them using the standard gamble. *Med. Decision Making* **11**, 140–146, 1991.
57. Burrows C. and Brown K. Time perception: some implications for the development of scale values in measuring health status and quality of life. In *Economics and Health: Proceedings of the Thirteenth Australian Conference of Health Economists* (Edited by Selby Smith C.). Public Sector Management Institute, Monash University, Melbourne, 1991.
58. Loomes G. and McKenzie L. The use of QALYs in health care decision making. *Soc. Sci. Med.* **28**, 299–308, 1989.
59. Gafni A. and Birch S. Equity considerations in utility based measures of health outcomes in economic appraisals: an adjustment algorithm. *J. Hlth Econ.* **10**, 329–342, 1991.

APPENDIX

*Utility Measurement Techniques**

Rating scale (RS)

A typical rating scale consists of a line with clearly defined end points. The most preferred health state is placed at one end of the line and the least preferred at the other. The remaining health states are placed between these two, in order of their preference, so that the intervals between the placements correspond to the differences in preference as perceived by the subject.

Magnitude estimation (ME)

The subjects are asked to provide the ratio of undesirability of pairs of health states. For example, is state A two or three times worse than the other state B? If state B is judged to be x times worse than state A, the undesirability (disutility) of state B is x times that of state A. A series of questions allows all states to be located on the undesirability scale.

Standard gamble (SG)

The subject is offered two alternatives. Alternative 1 is a treatment with two possible outcomes: either the patient is returned to normal health and lives for an additional t years (probability p), or the patient dies immediately (probability $1-p$). Alternative 2 has the certain outcome of chronic state i for t years. Probability p is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is p.

Time trade-off (TTO)

Two alternatives are offered. Alternative 1 is state i for time t followed by death; alternative 2 is full (or normal) health for time x. Time x is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is given by $h_i = x/t$.

Person trade-off (PTO): (equivalence technique)

The subject is asked a question of the following kind: 'If there are x people in adverse health situation A and y people in adverse health situation B, and if you can only help [cure] one group, which group would you choose?' One of the numbers x or y can be varied until the subject finds the two groups equivalent in terms of needing or deserving help. The undesirability (disutility) of situation B is x/y times as great as that of situation A.

*Descriptions are summarised from more detailed descriptions in Ref. [41]. See also Ref. [51].