CENTRE FOR HEALTH PROGRAM EVALUATION

WORKING PAPER 5

Cost Utility Analysis: What Should be Measured; Utility, Value or Healthy Year Equivalents?

Jeff Richardson

Paper presented to the 1990 Second World Congress on Health Economics University of Zurich, Switzerland
September 10 - 14, 1990
ISSN 1038-9547
ISBN 1 875677 12 7

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of Public Health and Community Medicine at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE receives core funding from the National Health and Medical Research Council and Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

ABSTRACT

This paper re-examines the question of the appropriate unit of measurement in cost utility analysis and the technique that will measure it. There are two main themes. The first is an examination of the use of `utility' as an object of measurement. It is argued that, as it is often conceived and measured, `utility' is not an appropriate unit. The second major theme of the paper is the selection and evaluation of a unit of measurement if the usual application of utility theory is rejected.

Because of its appeal to utility theory, the standard gamble has been widely accepted as the gold standard for measurement. More specifically, the standard gamble is believed to measure utility directly while other techniques simply measure `value' or are pragmatic devices for replicating the results of the somewhat complex standard gamble. In the first part of the paper it is argued that the standard gamble cannot be supported in this way because the theory from which it is derived is unsatisfactory. Three issues are considered. The first is the historical ambiguity in the concept of `utility' itself. The second is the validity of the distinction between utility and value which is used to demonstrate the superiority of the standard gamble. Third, there is a consideration of both the evidence and the normative argument used to defend utility maximisation as envisaged by the proponents of the standard gamble.

In the second part of the paper the view that there can be no unambiguous unit of output is discussed and rejected. It is suggested that this view may have arisen as some of the measurement techniques in CUA have been adapted from psychometrics where the object of the measurement is often imprecisely defined. Four criteria are suggested which appear to satisfy social objectives and the practical requirements of CUA. These are that there is a social consensus; that an increase in the number of the units corresponds with a socially desired outcome; that the units have a meaningful interval property; that they are readily understood and that it is possible to map the various health outcomes from an intervention into the unit. These criteria are applied to five of the major measurement techniques currently in use including the standard gamble. It is tentatively suggested that Healthy Year Equivalents are probably the best available units of output but that these should be derived directly from the time trade off or equivalent technique.

In the final section of the paper there is a consideration of some of the recent criticisms of cost utility analysis and particularly those associated with equity and distributional objectives. It is argued that critics have implicitly demanded more from a unit of output than is reasonable. When its inevitable limitations are recognised cost utility analysis in its present form fulfils an important role in the analysis of resource allocation.

* I would like to acknowledge the valuable comments by Dr Robin Pope on the sections dealing with risk theory. The paper has also benefited greatly from comments and discussions with Jane Hall and Colin Burrows. Any errors in the paper are part of the authors's own contribution!

Cost Utility Analysis: What Should be Measured; Utility, Value or Healthy Year Equivalents?

1 Introduction

Much of the appeal of Cost Utility Analysis (CUA) must be attributed to the fact that it uses as its unit of output the Quality Adjusted Life Year — the `QALY'. As the quality of life is indisputably relevant to the allocation of resources, few would argue that the adjustment of life years for quality represents a methodological advance. However, while the choice of the title was a successful marketing device for CUA it obscures the fact that quality has no precise meaning[1] and that different techniques used in the literature may be measuring a different concept of quality or may not be measuring quality in any meaningful sense at all. One approach to this issue has been simply to define quality as utility. The real unit of output is then the utility adjusted life year, which may nor may not be as intuitively appealing as the quality adjusted life year. A second approach is to determine whether utility as measured is correlated with other measures believed to reflect quality (Churchill, 1984; Evans, 1985, 1987). This approach, of course, begs the difficult issue of what should be measured. It is argued below that this questions has not been satisfactorily addressed and that in some cases the results of applying a measurement technique are deeply ambiguous.

There have been two broad approaches to the quantification of health. First, particular health scenarios or vignettes have been presented to subjects and their assessments of these health states have been measured. Secondly, more generally applicable multi-attribute scales have

[1] Mosteller (1989) 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', p. 282.

been constructed for a variety of purposes. Several have been derived specifically to construct an index of utility whose function is to help determine the appropriate allocation of resources.[2] As with the single scenario technique these scales require a value to be placed upon the different combinations of attributes which is appropriate for the purpose of cost utility analysis, namely resource allocation. The present paper re-examines the question of the appropriate unit of measurement for such an evaluation and the appropriate technique for measuring it. There are two main themes. The first is an examination of the use of `utility' as an object of measurement. It is argued that, as it is often conceived and measured, utility is not an appropriate unit and that the standard gamble is not an acceptable gold standard for measurement. The second major theme of the paper is the selection and evaluation of a unit of measurement if the usual appeal to utility theory is rejected.

Because it is based upon the expected utility hypothesis the standard gamble has been widely accepted as the gold standard for measurement. More specifically, the standard gamble is believed to measure utility directly, while other techniques simply measure value, or are pragmatic devices for replicating the results of the somewhat complex standard gamble. Validity in these cases is measured by comparison with the standard gamble (see for example Torrance, 1986). In Section 2 there is a review of the different concepts of utility and, in particular, the concept upon which the standard gamble is based. It is argued that the present theoretical foundations of the technique are untenable and, in particular, that the distinction between utility and value reflects an incorrect understanding of the measurement at risk. These conclusions imply that there is no basis for the view that the standard gamble should be the gold standard for measurement or that utility, as envisaged by proponents of the standard gamble, should be the unit of measurement.

The presumption that we should maximise utility and the subsequent debate over what this is or should be is a good example of the essentialist method of definition and the confused and unproductive consequences of this method described and criticised by Popper (1963). The alternative analytical approach is to determine objectives and subsequently, where useful, attach definitions to clarify new or unique concepts. Possibly as a result of the vagaries of some of the concepts measured in psychometrics, there appears to be a view that in health state measurement there can be no precise or clear unit of output. This view is discussed and largely rejected. In Section 3, criteria are suggested for selecting the unit of output in cost utility analysis and in Section 4 these are applied to five of the major measurement techniques currently in use. The arguments for the expected utility procedure's standard gamble are considered and, in particular, the two stage application of the procedure suggested by Mehrez and Gafni to derive healthy year equivalents - HYE's - is criticised. It is argued that HYE's are probably the best available units of output but that these should be derived directly from the Time Trade Off or Equivalence Technique. Finally, in Section 5 some of the unresolved questions associated with measurement are considered and, in particular, the issues of equity and distribution and the claim that decision making must, necessarily, be political.

---

[2] For a discussion of the health related MAU scales see Drummond et al (1986), Rosser and Kind (1978). For a discussion of health scales more generally, see McDowell and Newell (1987).

2 Concepts of Utility

Positive Analysis

For most economists the question of what should be measured and maximised has a self-evident answer, namely, `utility'. This generally results in the adoption of the expected utility procedure's andard gamble as the gold standard for measurement. However, there has been considerable confusion about what this is. Allais (1984), for example, notes that at the 1983 Oslo Conference on Risk and Utility `by far the most heatedly debated issues were those on the concept of utility' (p.8). The debate at the conference focused upon the properties of utility and what it does or does not include. That is, the conference revealed that there is no consensus over the use or even the concept of utility. The issue is highly pertinent in the context of Cost Utility Analysis. An economist advocating the maximisation of utility might legitimately be asked by a non economist what precisely this means.

Historically there have been at least four concepts. Initially, utility referred to welfare or well-being. Viner (1925), for example, claims that `the utility theory of value is primarily an attempt to explain price determination in psychological terms'. Secondly, utility was re-defined in functional terms to provide a framework for the positive analysis of consumer choice. In this framework utility functions determine an ordinal ranking of preferences. Thus, Graff (1967) asserts that `to say that his 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' (p.5, emphasis added). Graff would probably concede that a similar concept applied to women's preferences. Referring to the same concept, Phlips (1974) 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' (p. 26).

Thirdly, 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 (1984) 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' (p. 28).

Finally, `Neo-Bernoullian Utility' is obtained under conditions of risk by the linear summation of (objective or subjective) probabilities times quantities that are defined by a set of axioms. Von Neumann—Morgenstern (N-M) utility is the best known example and derived by assuming the applicability of the Von Neumann-Morgenstern axioms of `rational behaviour'. 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), 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 (1957), 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' (p. 591). It is indicative of the fact that utility in the (N—M) sense is not easy to interpret that Lane's explanation

is defective. The trade off referred to and which is quantified by a probability only indicates (N—M) utility if the (N—M) axioms are assumed to apply.

Joan Robinson (1965) noted that the first two of these concepts of utility have been confused and conflated. It is likely that the great appeal of utility in the context of Cost Utility Analysis is derived from such a confusion. The intuitive attraction to utility as an outcome measure is probably derived from the first, psychological, interpretation. Confidence in its empirical relevance is probably derived from the vast literature which has usefully employed utility functions as a framework for positive analysis. This confidence and intuitive appeal is probably responsible for the relatively uncritical and unfounded acceptance of (N—M) utility in the context of CUA.

Should we Maximise (N-M) Utility?

The fact that the concept of utility has mutated and has sometimes been confused does not disqualify it from being a sensible maximand in one form or other. It simply indicates the need to ensure that in a particular context the concept is clearly understood and correctly measured. The process for deriving (N—M) utility via the standard gamble is clear at least to researchers, if not always to subjects but the concept is not. One 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. The two issues which arise from this interpretation are, firstly, whether the axioms can, in fact, describe decision making under risk and, secondly, whether the empirical evidence suggests that they describe it accurately.

Utility vs Value

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 made that the (N-M) axioms describe choice under risk and this is the relevant context for cost utility analysis, as the outcomes of medical interventions are always uncertain.
To clarify the claim, consider the equation below:

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

In this, 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 procedure embodies an assessment under conditions of 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. Thus, it is argued that if there was to be an evaluation of a medical intervention for which there was a 10% chance of death and a 90% chance of life in a particular health state, S, it would not be correct to evaluate S under conditions of 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, S should be evaluated under conditions of risk using, for example, the standard gamble.

The reason for the assessment of Y under conditions of 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 (1983, 1989a) terminology, 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 financial and emotional factors operating which are relevant to utility — the emotional factors may include boredom, dislike or like of excitement and danger, the anticipation of regret or elation, tension, curiosity, wonder, hope, fear or worry. These emotions cannot be experienced unless there is uncertainty. The relevant issue here is whether or not the (N—M) axioms allow for this specific utility of risk. 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 measured this element of risk. This is explicit in the introduction to their 1944 analysis of games 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.*
>
> *Von Neumann and Morgenstern, 1947, 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 vigorous manner. I am only saying (as we did in 1944) that this is a very deep matter.*
>
> *Morgenstern, 1979, p. 181.*

The Von Neumann—Morgenstern view of their own theory has not been universally accepted. Harsanyi (1977, p.155) for example argues that:-

> *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.*
>
> *Harsanyi, cited in Watkins, 1977, p.155.*

Harsanyi and others have subsequently changed their view. Pope (1989) describes the sequence of historical 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.*[3]
>
> *Pope, 1989 p.11-12.*

There are good reasons for a return to the pre-1940 position. The assumption that the (N—M) axioms can include the specific utility of gambling or risk leads to an apparently contradictory conclusion. In equation (1) above, suppose that this factor is included in $J(Y_1)$ and $J(Y_2)$. It therefore includes the evaluation of all the elements relevant to decision making under risk and it could be equated with `utility'. Now suppose that $Y_1 = Y_2 = Y_0$. It follows that $A = J(Y_0)$. Risk is eliminated and the function J represents decision making with certainty. That is, the same function simultaneously represents decision making under risk and under certainty, and this excludes the possible existence of a specific utility of gambling.

The conclusion drawn by Allais (1984) from a generalised version of this argument is that the utility measured by the (N-M) axioms is the same as the cardinal utility measured under certainty (the third concept discussed above). The chief idea behind his proof was also put forward by Pope in 1983. More recently, Bouyssou and Vansnick (1988) 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 the classical idea of decreasing marginal utility used in economics' (p.109-110).

Empirical Evidence

It is not surprising that it has been hard to model the specific utility of risk (or the "pleasure of gambling", the "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 the not surprising conclusion that as the emotions that contribute to the specific utility of risk are varied and complex their importance is dependent upon the context of the risk. For example, in her review of the subject, Pope (1989) reports that "people are more ready to take extra risks when this is voluntary, when avoiding the bad outcome depends partly on

---

[3] In support of the historical interpretation given here, Harsanyi [1983], quoted earlier as believing that vN—M utility incorporates the specific utility of risk, 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.' Harsanyi, 1983, p.307 original emphases.

their own skills and degree of control of the situation, when the bad outcome affects a less vulnerable sub-group of the household or nation, and when the general social atmosphere applauds risk taking" (p. 15). Further, the process of decision making does not always conform approximately to (N—M) behaviour with some random variation attributable to an add-on "utility of risk". Rather, specific heuristics appear to be adopted in particular contexts.

The more general reviews of the literature on the Von Neumann-Morgenstern 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 (see for example, Schoemaker, 1982). One response to this evidence has been an attempt to reformulate the axioms in a way which avoids criticism. To date, this has not been achieved satisfactorily[4]. Another response has been to argue that while the axioms are imperfect they are the best available. However, this presupposes that modelling and measurement must be based upon fundamental behavioural axioms. While this may be an attractive objective - and it is explicitly rejected by many - there is no methodological imperative for the adoption of this approach to analysis and very compelling reasons for its rejection in an applied context if a satisfactory, empirically robust, set of axioms cannot be found.

The Normative Argument

The alternative interpretation of the N—M Standard Gamble and its underlying assumptions offered by Torrance and Feeny (1989), is that von Neumann—Morgenstern utility is normative — it indicates what individuals should do even if the outcome does not correspond with their own choice. Thus, they argue that "the theory and measurement methods were developed by von Neumann—Morgenstern (1944) as a normative (prescriptive) model for individual decision-making under uncertainty (p. 2)... The model was a normative one. That is, they prescribed how a rational individual *ought* to make decisions when faced with uncertain outcomes" (p. 4). The historical interpretation of von Neumann—Morgenstern is open to serious doubt.[5] Despite this, the interpretation of the axioms as having normative significance has had considerable appeal. Marschak (1950, p.139) argued that axioms defined rational behaviour and that their repeated application would ensure that "the probability that the achieved utility differs from the maximum

---

[4] For example Machina 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 (1988) and Pope (1989b) 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" maximand, this collapses to the conventional expected utility hypothesis.

[5] Savage, for example, writes `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], Savage, L.J., *The Foundations of Statistics*, 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 safely say that it has not been accomplished in the extensive literature about the topic', von Neumann and Morgenstern, 1947, p. 9.

achievable utility by an arbitrarily small number approaches unity". Ramsey (1950) went even 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 Keynes and Allais have noted[6], the outcome of a risky prospect is not its expectation. If an outcome is sufficiently unpleasant it is not irrational to adopt a rule that avoids the outcome or, perhaps, to adopt a rule which maximises the value of the worst possible outcome. There is nothing particularly 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 to explain one-off choices.

A variant of the Marschak argument is that choices consistent with the axioms will result in the best possible outcome for a large group of individuals. However, any one individual may rationally opt for an alternative choice and it is not clear under these conditions in what sense it would be rational for groups to adopt a different decision rule from the individual.

A more important criticism of the normative argument is that, as noted above, the (N—M) axioms do not take account of the specific utility of risk and all of the associated emotions. If these are relevant to welfare and can be taken into account in a decision rule it is scarcely rational to fail to do so. Thus, for example, Harsanyi (1983) argues that "it is only in such situations (where the specific utility of gambling is unimportant) that the vN—M axioms represent acceptable rationality requirements. In particular, whenever process utilities (the specific utility of gambling) are important, their compound lottery axiom and their independence axioms lose their plausibility" (p. 307).

Finally, and perhaps most fundamentally, if (N—M) utility can only be defined and understood 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 which maximises utility. But maximising utility means nothing more than adopting behaviour which is consistent with the axioms.

---

[6] Quoted in Pope 1989.

3. CRITERIA FOR EVALUATING THE UNIT OF OUTPUT

In his summation of the 1983 Oslo Conference on Utility and Risk Theory, Maurice Allais (1984) observes that `most of the conflicts noted seem to derive from the use of the same words, probability, random variables, chance, utility, rationality, etc., to designate entirely different concepts, (p. 6). Allais borrows the following quotation from Claude Bernard to illustrate his concern over the potential hazards arising from the misuse of definition.

> 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.
>
> Claude Bernard, quoted in Allais, 1984, p.5.

Both authors are echoing the Popperian concern over the inverted role of definition or what Popper (1963) describes as the "essentialist method of definition". In this, some word, X "is presumed to define some inherent essence or nature of a thing' (p. 20), which is presumed to be fundamental to the understanding of an issue. Debate and confusion arise over the question of "what x really is" or "what are its properties?". In the present context, X is utility. The starting-point conviction that utility should be maximised and the subsequent debate over its correct definition, its meaning and properties, exemplifies the confusion arising from the essentialist method of definition.[7]

The alternative analytical approach 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 maximised?" 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 involve one of the concepts of utility discussed here then the term can be used unambiguously with its meaning defined by the context of the rules. 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.

The issue of the clarity of objectives is important not simply in the context of utility but more generally. A number of the techniques currently used in CUA and their theoretical rational have evolved from a secondary disciplinary base, namely psychophysics and psychometrics. In these, the subject of the analysis has necessarily been less precise than in the physical sciences. In the former discipline the measurement of individuals responses to physical stimuli such as sound or

---

[7] The ambiguous meaning of "utility" has been recognised by some health economists. For example, Labelle, Feeny and Torrance (1989) state that "the precise definition of utility has long eluded economists and decision scientists" (p. 9). This recognition has not had an effect upon the belief that maximising this undefined utility is the appropriate starting point for the analysis of resource allocation.

light is at least anchored to a physical phenomena where there is a precise unit of measurement. Subsequently, the techniques used in psychophysics were adapted to psychometrics where measurement was extended to the quantification of such nebulous concepts as the *seriousness* of crime, the *preference* for an occupation, and the *attractiveness* of a product. The absence of a physical counterpart to the concept being measured results in a problem for the determination of validity and invites the potentially unrewarding task of examining the meaning of the works used to describe the concept. That is, it suggests the need to answer questions of the form "what is X?" (the concept being measured) "what are its properties and are the properties measured by the particular scale truly properties of x?' This process with its implicit assumption that there is a correct concept of X - `an inherent essence or true nature" - is quite distinct, from an examination of the particular properties determined or required by the relevant theory or task to be solved and then deciding, in the light of this, whether they should be measured.

The measurement of health states clearly has a strong family resemblance to some of the issues examined in the psychometric literature. The general concepts of ill health and the quality of life are no more precisely defined and their meaning may be subject to as much individual variability as the other concepts noted above. The implicitly conclusion of some researchers appears to be that because there is variation in individual opinion about what constitutes good or bad health then the units of measurement in CUA can be no more precise than in these other cases of psychometric measurement where a degree of conceptual imprecision - vagueness - is inevitable.[8]

This approach is unduly pessimistic. It is by no means obvious that a unit of measurement cannot or has not been found which eliminates, at least to a significant degree, the vagueness of some psychometric measures. Because the theoretical interest of economists has been distracted by the arguments for and against (NM) utility this issue has not been explored in the economics literature.[9]

The criteria for evaluating a unit of measurement should follow from the social objectives and from the practical requirements of such a measure. Four criteria are suggested here. They do not purport to be exhaustive and some of their limitations will be discussed later. It is suggested, however, that they should be regarded as necessary if not sufficient conditions to be met. First, as the measure is to be used for economic evaluation, a fundamental requirement is that more units are considered to be better than less and projects are to be preferred when, all else equal, there is

---

[8] For example, Kind (1988) 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 (p. 11,12).

[9] Torrance (1976) 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" (p. 129). Torrance, however, subsequently accepts that validity may be determined by correspondence with the outcome of the standard gamble (Torrance, 1986).

a lower cost per unit. In other words, there should be a broad consensus that the measure corresponds with a socially desired outcome.

Secondly, and arising in part from the need for social consensus, the unit of measurement should have, as far as possible, a clear and unambiguous meaning. The end point of an economic evaluation of a health program 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. Decision makers (who are often untrained in economics) need to understand, assess and appreciate what is obtained in exchange for expenditures. 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 (1989) 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" (p. 285).

The third criteria is that the unit of measurement has a *meaningful* interval property to permit the summation of benefits. The term "meaningful" is used here to emphasise that the property should be recognisably related to the magnitude of the benefit and not be an artefact of the scale.[10] Taking an extreme example, individuals could be asked to rate from 0 to 10 the "circularity" of particular 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 figures 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.

The final requirement suggested here is that it should be possible to map the health outcomes of an intervention into the unit of measurement and that the unit should be sensitive to variation in the relevant dimensions of the outcome. Fulfilling this practical requirement is dependent upon the existence of suitable techniques for carrying out the mapping. It is at this point that the more commonly discussed measurement criteria become relevant. The technique employed must have

---

[10] Individuals may be asked to rate virtually anything about which they have an opinion on a numerical scale that has an interval property. While it may be argued that the intensity of individuals' preferences has an interval property in some sense, it does not follow that the magnitude being assessed also has this property. Unless the object of measurement is indeed the individual preferences, the apparent interval property may be misleading.

the usual properties of validity and reliability. But these properties are secondary to the determination of what it is that should be measured.

These criteria are very largely fulfilled by the use of life years as a unit of output in cost effectiveness analysis. The unit is easily understood. All else equal, there is a social consensus that more life years are to be preferred to less and there is a clear interval property — all else equal the difference between 1 and 3 life years would be accepted as the same as the difference between 5 and 7 life years. Life years do not, however, fulfil the fourth criterion. The most relevant dimension of outcome is often the quality of life, but life years per se do not vary with this.

Similarly, "dollars" fulfil several of the criteria above. They are easily understood: More are to be preferred to less and they have a meaningful interval property. Using the usual measurement technique - the willingness to pay revealed in various contexts including interviews - they are sensitive to quality. However, two defects appear to be fatal. First, it is the intrinsic to the measurement technique that income and wealth will affect results in a way that is not generally acceptable. Secondly, willingness to pay studies in practice have given highly inconsistent results.

4. PRESENT MEASUREMENT TECHNIQUES[11]

QALYS would satisfy the criteria if they were what they purported to be, namely, homogeneous years obtained by a clear and meaningful index of quality. But the QALYs measured by the current gold standard do not purport to measure quality. Rather, they measure utility adjusted life years and the interpretation of these is, at best, very unclear. Despite this, the existing techniques have considerable appeal. They are related to individual preferences and it is widely accepted that these should be the basis for economic decision making. The units obtained by applying the techniques for utility measurement may, therefore, be assessed, not in terms of their correspondence to utility theory, but in terms of the four criteria discussed above. There are two relevant questions. First, abstracting from practical issues such as comprehension and wording of a health state description, what will a technique measure? More particularly, do the units measured by the techniques meet the first three criteria? Secondly, do practical problems alter the answer to (1)? The second question is empirical and it is the first that is 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).[12] A possible reason for this is that the latter two techniques involve choice, whereas the RS and ME do not. A second and related reason is that the units on these scales could have a different meaning than the units on the TTO and SG. It must, therefore, be asked what these units in fact mean or measure.

The RS gives a distance along a calibrated linear scale which a subject believes indicates, in some sense, the value or worth of a health state relative to the reference points on the scale. This outcome is in centimetres or a fraction of a linear distance. Despite their antiquity in the psychometric literature, neither of these are meaningful units as they must be translated. But it is not clear what they should be translated into. The question that the rating scale leaves unanswered concerns the functional relationship between the units of the scale and welfare, utility or the acceptable trade-off with healthy years or whatever it is that the scale reflects. At best, it is difficult to attribute a meaningful interval property to the scale (in the sense discussed earlier). At worst, the meaning of the scale itself is ambiguous.

Similarly, with ME subjects may be asked "how many times worse is one state than another [reference] state?". 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 "feeling" response, the meaning of which can vary from person to person. As different individuals are likely to use different huristics to answer the question, it is again difficult to place a clear meaning on the final scale and its units.

---

[11] For a description of the techniques discussed here see Torrance, 1986.

[12] For review and further empirical results see Richardson et al (1990).

This ambiguity is 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.[13] Some have claimed that, in principle, the two scales should give the same result (Brooks, 1988). This would only be true if, upon introspection, individuals could observe a clearly calibrated scale of preferences which could be read in an analogous way to the reading of an external scale. If there is no such internal, calibrated scale then there will be no necessary correspondence between results. Rather, they may reflect the framing effect of the question, the spreading or compression effects of the response, in addition to the true preferences. Even the individual's capacity to relate their preferences to such a scale has been questioned.[14]

It has been observed, empirically, on repeated occasions that the two scales produce different results (Kaplin et al, 1979, McDowell and Newell, 1987, Kine, 1989). The unavoidable conclusion is that one or other of the scales does not have the interval property that is required.[15] A further possibility is that neither has the property in the form discussed earlier and that the apparent interval property of the techniques is an illusion generated by the literal meaning of the language which respondents have been required to use, but a meaning that does not correspond with any real attribute that would be observed if we were able, somehow, to see below the veneer of the language. As both techniques produce units with no independent meaning there appears to be little possibility of evaluating these scales.

The two techniques do have the superficially attractive property that they abstract from time the questions asked and the scales used need not make any reference to the duration of a health state. However, the property is more likely to detract from the validity of the techniques. The health states to be evaluated do have a time dimension. Suppression of this information cannot enhance a health state description except under special circumstances viz., when the ratio of the value of time in a health state to the value of the same time in full health does not vary with time. Under other circumstances the numbers provided by the latter two scales must change with the duration of the health state implicitly assumed by respondents. This may vary from individual to individual and differ from the time period that is relevant for a particular assessment.

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 and, as discussed earlier, they are not. The outcome from the technique is, literally, a probability, p, which makes an individual indifferent between a certain outcome and a probabilistic choice. However, even if it is accepted that the (N—M) axioms are not *generally* true, the index p may be defended

---

[13] For a review of the debate see Kaplin and Ernst (1983).

[14] See Kind and Rosser (1988)

[15] 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, if either, scale has the required interval property. Further, there does not appear to be a single power function which will transform health outcome results consistently (see Richardson et al 1990).

as being an acceptable unit of outcome in the present context. First, it reflects choice and, 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. Thirdly, it can be argued that, by contrast with the TTO and ET, which also measure choice, the SG necessarily reflects an attitude towards risk and that the medical interventions to be assessed must also involve risk. These arguments amount to an assertion that in the present context the specific utility of risk is either quantitatively insignificant, that it results in the addition or subtraction of a constant amount from all probabilities, 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 which can be easily comprehended by those untrained in statistics, the results of the standard gamble can be translated into healthy year equivalents (Mehrez and Gafni, 1986b).

While there is some force in these arguments there are serious defects in the standard gamble as a pragmatic measurement device. It is questionable whether the interval property is a true indication of the intensity of an individual's choice in a meaningful sense or whether it is again a property of a scale which is imposed upon ordinal values. Empirical evidence suggests that people have difficulty understanding probabilities, especially extreme values. Further, it is open to serious doubt whether the risk entailed by the standard gamble improves or detracts from the integrity of the information obtained. 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. As noted earlier, it is claimed that the risk context of the (N—N) standard gamble is part of the technique for measuring the "utility" (as opposed to the "value") of the health state. It is not intended as a means of measuring the importance of the probability of transition into that state.

The less rigorous defence of the standard gamble is that it improves the decision context, as the individual is aware of the existence of risk in some general sense. 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. This context is utterly different 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 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. The value of 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 this term is used in its general sense. 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.

Another way of describing this problem is as follows. The equation described by the SG is:

$$p.\,U_1 + (1\text{-}p).\,U_0 + U_g = U_s + U_u$$

where U_1, U_0 and U_s represent the utilities of full health, death, and health state S respectively. U_g is the specific utility of gambling in the context of the SG and U_u is the disutility of uncertainty associated with a health state S. The pragmatic argument for the SG is that U_g and U_s may cancel out. This would only be a fortuitous result. Further, even this representation of the SG assumes that the components of the gamble — U_0, U_1, U_g — may be combined with a simple linear function.

It is true that, as they are presently used, the other techniques discussed here abstract from risk altogether. However, abstraction from risk per se is not a defect. The defect is in the abstraction from any risk or uncertainty associated with the intervention being evaluated. The introduction of irrelevant considerations cannot improve assessment.[16] In principle, the measurement of risk could be achieved by the inclusion of a statement of the risks and uncertainties in the health state description. There would, of course, be a practical constraint resulting from 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.

Mehrez and Gafni (1989b) use a two stage procedure to derive healthy year equivalents (HYE's) from the SG. By converting from a probability to a HYE the unit becomes more comprehensible as meets the second criterion above. In the first stage of the procedure the utility of a fixed number of years, n, in a state of ill health 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 which are equivalent to the n years if ill health. That is, the two stages take the form:

Stage 1: p.(full health for n years) = n sick years ..... p varies
+(1-p) death

Stage 2: p.(full health for n years) = h healthy years ..... h varies,
+(1-p) death

However, the logical format of the two stages is:

[16] If this were not so, the TTO 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!

$$a = b$$

$$a = c$$

from which $b = c$

In other words, unless the initial equations for the SG 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:

$$a = b - r_1$$

$$a = c - r_2$$

$$b = c + (r_1 - r_2)$$

where r_1 and r_2 are the specific (dis) utilities of gambling in the two stages. It is possible that $r_1 = r_2$, in which case there is, once again, no purpose in the two-stage procedure. If $r_1 \neq r_2$ the question raised above 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 in the argument. 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 in a realistic setting, that is, r_1 captures the specific utility of gambling, then the net effect $(r_1 - r_2)$ cannot also do so unless $r_2 = 0$. There is no reason to believe that this is so.

In sum, the SG is a technique which assesses a riskless health state by means of a gamble. The technique cannot produce a probability which 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 (dislike) for the specific risk embodied in the technique, and it will reflect the unknown decision rule applied by individuals in this particular context. There does not appear to be, at present, a satisfactory technique for translating the outcome of the standard gamble into an easily understood unit as required by the second criterion discussed earlier.

Time Trade off Equivalence Technique

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 all the 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. In 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 TTO and ET is that these two relationships are collapsed into one and the health state to be evaluated is contrasted directly with a reference state. Thus, the TTO directly measures the number of healthy years that are equally valued — considered equivalent to — a given period in a health state. As there is no additional scale used in the calculation, one important source of potential error and ambiguity is removed. Consequently, the following scenario is possible with each of the techniques discussed earlier but not with the TTO or ET: an individual might indicate with the use of one of these earlier instruments that, as measured, the utility of n years in health state S_1 was superior to the utility of m years in a health state S_2. The same individual could, in practice, indicate directly that their actual preference was for the latter and not the former option.

The units produced by the TTO and by the ET have both an interval and a ratio property. That is, there is a clear meaning to the statement that (all else equal) six healthy year equivalents are double three healthy year equivalents, namely, that the duration of the flow of benefits of living six years is twice the duration of three years and there is an objective standard for measuring duration. Similarly, there is a clear, comprehensible, meaning to the outcome of the ET. By definition, X people in one health state are equivalent to Y people in another - possibly full health. If the number of years in these states is specified then the meaning is very similar to the meaning of the TTO. The unit of measurement may be interpreted as the life year and when the comparison state is full health the unit is the healthy year equivalent.

The principal conceptual difference between the healthy year equivalent implied by the ET and the healthy year equivalent derived by the TTO arises if the TTO is derived from an interview in which an individual is asked to imagine that (s)he actually experiences the health state being evaluated. The personal values revealed are, in principle, closer to the outcome of consumer sovereignty than the arms length evaluation of others welfare revealed by the ET. Results of the two techniques could vary because of a systematic difference between choice criteria applied to personal decisions and those applied to social decisions. The selection of techniques would depend, in part, upon whose values were considered to be relevant. However, in both cases the healthy year equivalent derived remains a comprehensible unit. The difference arises from the question "who should determine the equivalents"

Interpretation of the outcome from the time trade off is somewhat confounded by the fact that, while the unit upon which the measurement is based is clear, the benefits envisaged for the assessment are at different points in time. They cannot, therefore, be simply added to healthy years calculated for a different period of time. They must be first discounted. That is, if n years in a health state, S, are equivalent to m years of full health, the relevant benefit is the present value of the healthy year equivalents, namely d_i, where d_i is the discount factor for a period i years in the future. (This adjustment to the results of the TTO has not, generally, been carried out in the studies reported in the literature. See Richardson, et al, 1990). The situation is exactly analogous to the conceptualisation of a flow of (inflation adjusted) dollars over time. The concept of a dollar is clear and is determined by the quantity of goods and services it can purchase at that point in time. However, dollars obtained at different points in time cannot be compared directly. The need to discount dollars before comparison is made does not, however, affect the clarity of the concept of a dollar.

While the TTO makes the time dimension very obvious, other techniques must also discount future values either explicitly or implicitly. For example, the equivalances between x people in a health state and y in full health may be obtained without any reference to time or with an explicit statement that the health states will last for only one year. However, as morbid health states often last longer than a year, future health must be evaluated either explicitly by describing and evaluating the full multi-period health scenario or by making the rather tenuous assumption that the value of a morbid health state is independent of time and then discounting future years where the undiscounted value of a future year has been determined from a timeless health-state scenario[17].

A further, potentially serious, criticism of these techniques is that there may be no agreement on the meaning of full health and therefore no common understanding of what is entailed by a healthy year. "Healthy years", taken literally, imply years with no ill health and these are unusual. As people age their physical well-being deteriorates to a greater or lesser extent. In the extreme, elderly or chronically ill persons may be unable to recall full health and interpret the concept meaningfully. The problem is not unique to the TTO and ET. As the other techniques used to produce QALYs use full health and death as anchor points, the problem is general. One possible solution is to replace the "healthy year" with the "normal year" or to contrast an ill state scenario with a defined, standard scenario describing normal health as done by Hall et al (1990). The a priori arguments for this approach are, however, inconclusive. The health state chosen as the unit of measurement, or as an anchor point, should be easily understood by the person interviewed. It should also be understood by those interpreting the results. "Normal" health is more realistic but it is not necessarily the more easily visualised by subjects. People may be unaware of their expected future health or may (correctly in some cases) consider themselves to be exceptional, not normal. In particular, the young will generally have only experienced full health (i.e. no serious sickness) and may find it difficult to visualise the sorts of aches, pains, illnesses and chronic conditions that may subsequently characterise normal health. Further, a "normal" year may have different meanings to different subjects unless described in the interview. Unless all studies then adopt a common description, cross study comparisons may be invalidated. The unreality of the healthy year is not a defect if it is well understood and it was regarded as a unit of measurement, not as an option.

In sum, the choice of the healthy or the normal year as a unit of measurement is an empirical issue. If the former is easily understood it could remain as the reference state. If evidence suggests unacceptable variation in the understanding of the concept then a reference state would have to be defined during the interview.

---

[17] For a discussion of the relationship between utility and time, see Loomes and McKenzie (1989), Mehrez and Gafni (1989), Richardson et al (1990).

5. ETHICS AND DISTRIBUTIONAL ISSUES

The conclusion to be drawn from the previous sections is that there are neither strong theoretical nor pragmatic grounds for the adoption of the standard gamble as a gold standard of measurement. Of the alternatives available, the units which satisfy the suggested criteria most fully are those produced by the time trade off and equivalence techniques. As both of these are based upon equivalences and do not involve the measurement of quality per se, the nomenclature of Mehrez and Gafni - the Health Year Equivalent (HYE) - is more descriptive of the unit (although Mehrez and Gafni (1989a) had a different purpose for introducing the term). It has the additional advantage of avoiding the ambiguity which surrounds the term which was noted in the introduction.

Some recent contributions to the literature may appear to cast doubt upon the relevance of the issues discussed here and suggest that there are a number of prior questions that must be answered before CUA could be regarded as an acceptable basis for resource allocation.[18] These issues could be grouped, somewhat loosely, under the two headings: "the scope and measurability of benefits" and "distribution and other ethical issues".

Loomes and McKenzie (1989) note that the individual who is the direct subject of an intervention is not necessarily the only, or even the chief, beneficiary. This may be the person's spouse, parent or friend. They cite the example of a handicapped child or elderly person who may be worse off through institutionalisation but whose "erstwhile principle carers" may gain significantly. Similarly, in their analysis of neonatal intensive care units Boyle et al (1983) do not include the benefits (or costs) to parents of the survival of their (possibly chronically ill) infant.

A second set of issues involve the inclusion or exclusion of such intangible benefits as information and the reduced level of anxiety which might follow from a true negative result in a screening program. For example, it is argued by Mooney and Olsen (1989) that a recognition of such process variables as patient autonomy, information and how the information is provided to the individual are all relevant to well being.

These issues do not detract from the relevance of CUA. In any economic evaluation a decision must be made about the scope of the benefits to be included. Some will be excluded on the purely pragmatic grounds that there are finite research funds and prior examination of the issues indicate that the benefits to be excluded appear to be quantitatively negligible. The omission of quantitatively significant benefits is a criticism of the research team and not the technique. The more serious issue is whether or not some benefits are capable of conversion into QALYs or HYE's. In principle, there is no reason why measurement should be confined to the person receiving treatment nor that intangibles should not be included in health state descriptions. The ability of the various cost utility techniques to measure such benefits is an empirical issue which has not been explored. It is quite possible that the techniques will not be particularly useful with some categories of benefits. But is would be inappropriate to conclude that this invalidated the techniques generally.

---

[18] See in particular Loomes and McKenzie (1989) and Carr-Hill (1989)

Issues concerning the scope of benefits to be included in an *analysis* become ethical and not simply technical when consideration is given to the social constituency to be included in the examination of benefits. The question here is not whether a particular benefit to an accepted member of the social group should be included but who or what should constitute the membership of the social group. The issue becomes acute when considering procedures which affect birth. It was implicitly assumed by Boyle et al (1983) that unborn but conceived persons *should* be included in a social constituency. By contrast, future children were excluded who, with a measurable degree of statistical certainty, would have been conceived to replace the low birth weight children who would have died without intensive care. The inclusion of both groups would almost certainly result in negative benefits from neonatal intensive care as measured by QALYs. The inclusion of the QALYs produced by neonatal intensive care as a benefit contrasts with the outcome of prenatal diagnosis and genetic counselling where the intention of the intervention is to prevent the birth of particular babies, and presumably, allow their replacement with healthier infants. Kuhse and Singer (1988) note that "there is surely something very odd about the fact that until birth, the benefits of a health program can be seen in terms of preventing the existence of a human being, whereas if an identical child is actually born, the benefits of a medical program are seen in terms of the prolongation of the infant's life. To the best of our knowledge, no justification for such a distinction can be found in any textbook of economics' (p. 110). If QALYs or HYE's were the only unit of output and their maximisation was the only ethical principle prenatal diagnosis and genetic counselling would have to be proscribed. Preference would always be given to procedures favouring young women entering the age of childbearing as their health would maximise future life years. It is clear from these examples that CUA does not provide a complete answer to all ethical issues. Like other areas of economic evaluation it has limitations. Once again, however, their recognition does not require the abandoning of CUA generally.

The principle ethical question to be resolved in CUA is the way in which individual valuations should be aggregated and, therefore, the way in which interpersonal comparisons are to be made. As noted by Torrance (1986) the basic assumption made is that "the difference in utility between being dead and being healthy is set equal across people. In this way the method is egalitarian ... each individual's health is countered equally" (p. 17). In Mooney's (1989) phrase this implies a form of "quasi-utilitarianism" in which the maximand is not total utility but a weighted average of individual's utilities where the weights are designed to treat individuals equally irrespective of the absolute intensity of their preferences.

Such quasi-utilitarianism necessarily conflicts with other ethical bases. Libertarians would reject an aggregation rule which constrained an individual's right to reveal their own preference, that is, through their willingness to pay. More generally, those subscribing to a deontological view of ethics would argue that resource allocation should be determined, not only by consequence but by "ethical rules" and "human rights". Thus, for example, in his critique of QALYs, Harris (1987) argues that the only priority in health cae should be the preservation of life and that all have an equal right to life no matter what its length or quality. It his critique of this paper Williams (1987) notes the incompatibility of the ethical bases and simply argues that "at the end of the day we simply have to stand up and be counted as to which set of principles we wish to have underpin the way the health care system works" (p. 123).

Proponents of CUA must commence with the view that there will be widespread acceptance of the ethical principle that, all else equal, more healthy year equivalents should be preferred to less.

However, this principle leaves unresolved the important question of how often all else will, in fact, be equal or at least quantitatively insignificant. The authors cited earlier point to a number of distributional considerations which might, potentially, be quantitatively important. These include:

- the possibility that, for example, five HYEs for an individual may not be considered equivalent to one HYE for five individuals.
- the (related) possibility that a more equal distribution of benefits might increase social welfare because of a perceived increase in the access to health care.
- the possibility that a different social valuation may be placed upon a life year at a different stage of a person's life. The survey by Wright (1986) may be cited to support the view that life years are considered to be more important by a cross section of the population when the individual concerned is raising children than when the individual is commencing school, at the peak of their earning capacity or retired.

Distributional considerations such as these may be included in the analysis of resource allocation in one of two ways. First, there may be an attempt to justify the valuation attached to the life year so that it represents a social and not an individual value. Secondly, the distributional question may be separated from the issue of the *production* of healthy year equivalents as defined by the TTO or the preferred measurement instrument.[19] The recent critiques of CUA implicitly assume that the first is the only acceptable approach. The failure of the practitioners of CUA to embody all factors relevant to social welfare in a single utility weight is only a failure if this is the objective. Similarly, practitioners of general cost benefit analysis have failed to produce such an index. Despite early attempts to achieve weighting schemes to convert dollars into utility it is now the standard practice in CBA to separate issues of production and distribution.

There are also compelling theoretical and practical reasons for adopting this separation in CUA. As evidenced in the present paper there has been a distinct lack of conceptual clarity even with the comparatively straightforward question of the unit of production. The issues associated with distribution and equity are even less tractable and an attempt to fuse them with an index of individual utility is likely, in its application, to cause terminal confusion. There is also no consensus on the process by which issues of equity and distribution should be resolved. There is an implication in the literature that such issues should be determined by seeking the opinions of a representative cross-section of the community. Majority voting is not, however, universally accepted as a basis for ethical decision making in all contexts. It is also likely that social ranking would be subject to Arrow's voting paradox.

The potential obstacles facing the construction and use of a set of social utility weights are probably insurmountable. First, there would be an issue of selecting an appropriate panel of judges. There is already debate in the literature about whose opinions should be represented. Secondly, even with an agreed process of achieved social consensus, the construction of a reliable and valid set of weights would be difficult. Loomes and McKenzie (1989), for example, note that "any attempt to derive a societal value ... would be confounded by the problem of

---

[19] It may be argued that the HYE is not a pure unit of production as it has, as noted earlier, a value basis. The same however is true in cost benefit analysis where market based dollars are used as a measure of output.

deciding whether this value ought to be an aggregate of the values which individuals place on their *own* years of life, or whether a societal value should comprise the values which people place on everyone else's life years.

Thirdly, the result of such an exercise would almost certainly violate the third criterion suggested earlier. The meaning of the composite index of equity/output would be difficult to interpret. Higher values would indicate a preferred state, but the interval property required in the comparison of costs and benefits would almost certainly become uninterpretable. It is doubtful if such an index would be acceptable to Mosteller's "talented lay people". Finally, as different countries have different social values, there would be no possibility of cross national comparison. Even within countries it is likely that equity and distributive objectives would be interpreted differently by different research teams and their results would cease to be comparable. This would seriously devalue the cumulative research effort as the objective of cost utility-cost effectiveness analysis is the comparison and ranking of different projects.

6. CONCLUSIONS

There remains the unanswered question of whether or not we should maximise either utility or value. It has been argued here that the distinction generally made is, at best, unhelpful. The utility implied by the von Neumann and Morgenstern axioms must be the same as cardinal utility under certainty. Consequently, the derivation of HYE's with the time trade off or equivalence techniques may be analysed in terms of utility and, all else equal, the maximisation of HYE's for a given budget may also be viewed as utility maximisation. However, from this perspective, the maximisation of utility is a device which is useful for the analysis of independently determined social objectives. The analysis is not driven by the preconception that the social objective is utility maximisation, somehow defined.

It is clear that whatever its nomenclature the unit used for the evaluation of output will not resolve all of the issues associated with the allocation of resources in the health sector. The conclusion to be drawn from the previous section is that it is, at best, premature to contemplate the construction of a single set of weights which purport to measure social utility; at worst it is an ill-advised objective. Measures discussed in this paper may not resolve all technical, ethical and distributive issues but they do embody a sufficiently important component of the decision process - namely individual preferences - that they convey important information to decision makers and in a way that is easy to interpret. On occasions, other factors will not be considered important and resource decisions could be based entirely upon the ranking of the cost utility ratio. On other occasions, distributional or ethical considerations may intervene and a political component in the decision process will be required. But as with all economic analyses, the answers given by CUA should narrow the breadth of the disagreement and limit the scope for arbitrary decision making.

REFERENCES

Allais, M., 1984, "The Foundations of the Theory of Utility and Risk. Some Central Points of the Discussion at the Oslo Conference" in Hagen, O., and Wenstop, F. (eds) *Progress in Utility and Risk Theory*, D. Reidel Publishing Company.

Allais, M., 1988, "A New Neo-Bernoullian Theory: The Machina Theory. A Critical Analysis", in Munier, B., *Risk Decision and Rationality*, D. Reidel Publishing Company Dordrecht.

Bouyssou D., and Vansnick J.C., 1988, "A Note on the Relationships Between Utility and Value Functions", in Munier, B.R., *Risk, Decision and Rationality*, D. Reidel Publishing Company.

Boyle, M.J., Torrance, G.W., Sinclair, J.C. and Horwood, S., 1983, "Economic Evaluation of Neonatal Intensive Care of Very Low Birth Weight Infants", *New England Journal of Medicine*, 308: 1330-1337.

Brooks, R.G., 1988 "Scaling in Health Status Measurement: An Outline, Guide and Commentary", Institute of Health Economics Report, 1988: 4, The Swedish Institute for Health Economics.

Carr-Hill, R.A., 1989, "Assumptions of the QALY Procedures", *Social Science and Medicine*, 29, 3, 469-477.

Churchill, et al, 1984, "Cost Effectiveness Analysis Comparing Continuous Ambulatory Peritoneal Dialysis to Hospital Haemodialysis", *Medical Decision Making*, 4, 20-23.

Drummond, M.F., Stoddart, G.L. and Torrance, G.W., 1986, "Methods for the Economic Evaluation of Health Care Programs', *Oxford Medical Publications*.

Evans, R.W., et al, 1985 "The Quality of Life of Kidney and Heart Transplant Recipients", *Transplant Proc.* 17, 1579-82.

Evans, R.W., et al, 1987, "The Quality of Life of Kidney and Heart Transplant Recipients", *Transplant Proc.* 17, 1579-82.

Graaff, J., 1967, *Theoretical Welfare Economics*, Cambridge University Press, Cambridge.

Hall, J., Gerard, K., Salkeld, G. and Richardson, J., 1990, "A Cost Utility Analysis of Mammographic Screening for Breast Cancer in Australia", Paper presented to the 2nd World Congress on Health Economics, University of Zurich, Switzerland, September 10-14, 1990.

Harris, J., 1987, "QALYfying the Value of Life", *Journal of Medical Ethics*, 13, 117-123.

Harsanyi, J., 1977," *Rational Behaviour and Bargaining Equilibrium in Games and Social Institutions*", Cambridge University Press.

Harsanyi, J., 1983, "Use of Subjective Probabilities in Games Theory", in Stigum, B. and Wenstop, F., (ed), *Foundations of Utility and Risk Theory with Applications*, Reidel Press.

Kaplin, R.M., Bush, J.W. and Berry, C.C., 1979, "Category Rating versus Magnitude Estimation for Measuring Levels of Well Being", *Medical Care*, May 1979, 27, 5, 501-521.

Kaplin, R.M., and Ernst, J.A., 1983, "Do Category Rating Scales produce biased preference weights for a health index?", *Medical Care*, 21, 193-207.

Kind, P., 1988, "The Development of Health Indices", in Teeling Smith, G., *Measuring Health: A Practical Approach*, John Wiley and Sons.

Kind, P. and Rosser, R., 1988, "The Quantification of Health", *European Journal of Social Psychology*, 18, 63-77.

Kind P., 1989, "The Design and Construction of Quality of Life Measures", Discussion Paper 43, Centre for Health Economics, Health Economics Consortium, University of York.

Kuhse, H. and Singer, P., 1988, "Age and the Allocation of Medical Resources", *The Journal of Medicine and Philosophy*, 13, 101-116.

Labelle, R., Feeny, D. and Torrance, G., 1989, "Conceptual Foundations of Health Status and Quality of Life Utility Measures", NIMEO, Department of Clinical Epidemiology and Biostatistics/Centre for Health Economics and Policy Analysis, McMaster University.

Lane, D.A., 1987, "Utility, Decision, and Quality of Life", *Journal of Chronic Diseases*, 40, 6, 585—591.

Loomes, G. and McKenzie, L., 1989, "The Use of QALYs in health care decision making", *Social Science and Medicine,* 28, 4, 299—308.

Machina, M., 1982, "Expected Utility" Analysis Without the Independence Axiom", *Econometrica*, 50, 277—323.

Marschak, J., 1950, "Rational Behaviour, Uncertain Prospects, and Measurable Utility", *Econometrica*, 18, 11—141.

McDowell, I. and Newell, C., 1987, "Measuring Health: A guide to Rating Scales and Questionnaires", Oxford University Press, New York, 1987.

Mehrez, A. and Gafni, A., 1989a, "Quality-adjusted life years, utility theory, and health-years equivalents", *Medical Decision Making*, 9, 2, 142—149.

Mehrez, A. and Gafni, A., 1989b, "Healthy Year Equivalents: How to Measure Them Using The Standard Gamble", forthcoming Working Paper CHEPA, Department of Epidemiology and Biostatistics, McMaster University.

Mooney, G. and Olsen, J.A., 1989, "QALYs: Where Next", Mimeo, *Institute of Social Medicine*, University of Cophenhagen.

Morgenstern, O., 1979, "Some Reflections on Utility", in Allais, M. and Hagen, O., *Expected Utility Hypothesis and the Allais Paradox*, D. Reidel Publishing Company.

Mosteller, F., 1989, "Finale Panel: Comments on the Conference on Advances in Health Status Assessment', *Medical Care,* 27, 3, Supplement, S2H2—S286.

Philips, L., 1974, *Applied Consumption Theory*, North Holland Publishing Co., Amsterdam.

Pope, R., 1983, "The Pre Outcome Period and the Utility of Gambling", in Stigum B. and Wenstop, F., (eds), *Foundations of Utility and Risk Theory with Applications*, D. Reidal Dorecht, 137—177.

Pope, R.E., 1989a, "Additional Perspectives on Modelling Health Insurance Decisions", in Selby Smith C., 1989, *Economics and Health*, Public Sector Management Institute, Monash University.

Pope, R.E., 1989b, "Machina's Decision Model: An Empty Box?", Memeo, Department of Economics, University of New South Wales, Campbell ACT, Australia.

Popper, K., 1963, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Routledge and Kegan Paul.

Ramsey, F., 1950, "Truth and Probability", in Braithwaite, R. (ed), The Foundations of Mathematics *and other Logical Essays*, Humanities Press, New York.

Richardson, J., Hall, J., Salkeld, G., 1990, "`Cost Utility Analysis: The Compatibility of Measurement Techniques and the Measurement of Utility Through Time", *in Selby Smith, C., 1990*, *Economics and Health: Proceedings of the Eleventh Australian Conference of Health Economists*, Public Sector Management Institute, Monash University.

Robinson, Jan, 1965, *The Accumulation of Capital* (2nd E.), London:Macmillan.

Rosser, R.M., and Kind, D.P., 1978, "A Scale of Valuations of States of Illness: Is There a Social consensus?", *International Journal of Epidemiology*, 7, 347.

Savage, L.J., 1954, *The Foundations of Statistics*, second revised ed., Dover Pub., New York.

Shoemaker, P., 1982, "The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations", *Journal of Economic Literature*, 20, 529—563.

Torrance, G.W., 1976, "Social Preferences for Health States, An Empirical Evaluation of Three Measurement Techniques", *Socio Economic Planning Science*, 10, 129—136.

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

Torrance, G.W. and Feeny, D., 1989, `Utilities and Quality Adjusted Life Years', *International Journal of Technology Assessment in Health Care*, (forthcoming).

Viner, J., 1925, "Utility Concept in Value Theory and its Critics", in Page (ed.), 1968, *Utility Theory: A Book of Readings*, John Wiley & Sons.

Von Neumann, J. and Morgenstern, O., 1947, " *Theory of Games and Economic Behaviour"*, Princeton University Press.

Watkins, J.W.N., 1977, `Towards a Unified Decision Theory: A Non-Bayesian Approach', in Butts and Hintikka (eds), *Fundamental Problems in the Special Sciences*, Reidel Dordrecht pp. 347—379.

Williams, A., 1987, "Response: QALYfying the Value of Life", *Journal of Medical Ethics*, 13, 123.

Wright, S.J., 1986, "Age, Sex and Health: A Summary of Findings from the York Health Evaluation Survey", Discussion Paper 15, *Centre for Health Economics*, University of York.

APPENDIX 1

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 undesirablity of pairs of health states. For example, is one state two or three times worse than the other state? 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 healthy 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 then 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 [41]. See also [50].