The Impossibility of an Ideal Metric for Health Service Benefit Measurement

Professor Jeff Richardson
Foundation Director, Centre for Health Economics
Monash University

Dr John McKie
Research Fellow, Centre for Health Economics
Monash University

February, 2009

Centre for Health Economics
ISSN 1833-1173
ISBN 1 921187 32 8
Acknowledgements

The authors would like to acknowledge comments from Professor Reinhard Selten, Experimental Economics Laboratory, University of Bonn, Professor Peter Singer, Center for Human Values, Princeton University, and, in particular, the exhaustive and insightful comments of Dr Robin Pope, Center for European Integration Studies and Experimental Economics Laboratory, Bonn University.
# Table of Contents

1. Introduction .................................................................................................................. 1
2. EUT as the “Theoretically Correct Measurement” ......................................................... 3
3. The Impossibility of an Ideal Metric ............................................................................. 6
4. Methodology and Selection of Metrics ........................................................................ 12
5. Measurement without universal validity ..................................................................... 17
6. Discussion ..................................................................................................................... 19
7. Conclusion .................................................................................................................... 20

References ....................................................................................................................... 23
Abstract

Different policies in different contexts generate different winners and losers. As a result, the choice of any metric for measuring net benefits requires a normative judgement. However, there is no satisfactory way of making this judgement that allows one option to be considered "best" in a universal, timeless sense. The formulative approach to this problem in Arrow’s impossibility theorem is rejected as a framework for further progress. We conclude that the theoretical basis for selecting metrics should be sought from within an analytical paradigm that explicitly includes the quantification of public preferences and subsequent public consultation. The examples used in the development of the argument focus on health policies.
1 Introduction

The issue addressed in this article has simmered in the literature for a number of years in the context of the policies of international organisations such as the World Bank and the IMF, and the policies of national governments concerning economic progress, exchange rates, the environment and health policies. This is the matter of the correct metric for measuring net benefits. In the context of economic evaluation of health services, Torrance (1986) argued in an influential article that the theoretically correct metric was utility, and that the theoretically correct way to measure utility was within EUT (expected utility theory), via SG (a Standard Gamble procedure), because of its derivation from the axioms of von Neumann and Morgenstern. The SG was advocated by both von Neumann and Morgenstern (1944) and by Ramsey (1926) in his pre-axiomatisation of expected utility theory. For practical reasons other metrics were admitted as proxy measures of health benefits.

Subsequent authors have suggested that some of these latter metrics for health policy benefits may be more than just a proxy for the SG and might be, for different reasons, the one that best suits the purposes of economic evaluation in the health sector. Candidates include the PTO, Person Trade-off (Nord 1995), TTO, Time Trade-off (Richardson, 1994) and more recently the RS, Rating Scale (Parkin and Devlin 2006). A related debate concerned the “theoretical” merits of two of the 270 or so “composite metrics” which can be created from the metric’s constituents; namely, the Quality-Adjusted Life Year (QALY) versus the Healthy Year Equivalent (HYE) where these are differentiated by the time period of the measurement – one year versus the duration of the health state - and by the metric and procedure used to measure the quality of life – the SG or the TTO.

On the basis of its influence and the volume of empirical work in the area of health state measurement, a third metric to consider is the Disability Adjusted Life Year (DALY). This metric is based on expert opinion with respect to the quality of life and a measurement scale calibrated using

---

1 A non-exhaustive list results in 270 options from combinations of the following four components of choice: scaling instrument (whether TTO, SG, PTO, RS, ME, WTP); time frame (1 year, episode, whole of life); choice of judge (public, patients, potential patients, professionals, ); and perspective (ex ante, current, ex post - i.e. preferences before, during or after experiencing the health state) \( 6 \times 3 \times 5 \times 3 = 270 \) proposed for choosing the best metric. Not all these combinations are coherent. However, it is likely that the reason for the pre-eminence of the QALY and HYE was more related to historical chance (and the selection of an appealing label) than to an empirical demonstration that their elements corresponded with social preferences.
a limited number of PTO values for marker states (Murray and Lopez 1996, Stouthard et al. 1997, Begg et al. 2007).

Unlike the literature relating to the PTO and TTO, which used eclectic arguments and criteria for evaluating options, most of those engaged in the QALY-HYE controversy have sought, in the manner of Torrance, to demonstrate the validity of their arguments by appeal to the von Neumann/Morgenstern (NM) axioms - i.e. by assuming the validity of EUT.

Another approach, which also draws upon EUT, involves determining the social willingness to pay for a human life from the value placed upon the risk of death in different contexts (Schelling 1968) and it has been argued that is the only measure consistent with Paretian welfare theory Mishan (1971). As used by a leading exponent, Viscusi (1992) (Viscusi and Hersch 2008), the wage paid to workers exposed to the risk of death is econometrically related to various attributes of the work, the workforce and the level of risk, and can be used to determine the dollar compensation paid for the risk. Assuming the NM axiom of preference invariance under linear transformation, the value of risk is converted into the value of a life year: compensation of $2,000 for a 1 in 100,000 risk of death implies a value of life of 100,000 x $2000 = $2,000,000. These methods are currently being employed to determine the social value of a QALY by research groups funded in the UK by NICE.2

More recently there has been a general recognition of the shortcomings of expected utility as providing the theoretical basis for measuring social welfare. In the context of economic evaluation in the health sector, Culyer and others (Culyer 1989, Culyer 1990, Wagstaff 1991) have suggested that health should be the maximand, or at least a supplementary objective in economic evaluations (extra welfarism). More generally, Sen has nominated “capabilities” as the appropriate objective, arguing, inter alia, that those with expensive tastes may benefit (unfairly) from the use of utility, while those who have adapted to objectively poor circumstances may be (unfairly) disadvantaged (Sen 1982).

In this article we argue that there is no metric – the SG, TTO, WTP - whether utility-based, health-based, or capability-based that can be shown to be “correct” in some universal, timeless sense. Rather, we argue that the search for such an "ideal" metric for the purposes of economic evaluation is futile, and that the final choice of metric should depend upon social values that are potentially country-specific and mutable through time. Such a metric, provided certain conditions are met, can be seen as an acceptable compromise among individuals and groups who hold a plurality of (sometimes incompatible) values.

The argument proceeds as follows. In section 2 below we look at, and reject, EUT’s standard gamble and the SG-based QALY or HYE for the position of ideal metric and, by extension, risk based willingness to pay. In section 3 we develop the argument for the impossibility of a universal, timeless metric. The argument is based upon the inevitability of winners and losers with any method of evaluation, the inability to compensate losers, and the impossibility of aggregating the benefits and losses in a way that represents a universal measure of social welfare valid for all time periods.

2 The ‘EuroVaQ’ and ‘SVQ’ projects are both headed by Donaldson (cam.donaldson@ncl.ac.uk).
The last step of this argument enlists Arrow’s Impossibility Theorem. We explain why it renders futile any endeavour to demonstrate the existence of an ideal metric, but also argue that the theorem is not an impediment to the practice of economic evaluation and that its chief practical message is somewhat unsurprising. We argue that the methodology introduced (strictly, reintroduced) by AIT is unproductive in terms of likely progress towards socially relevant measurement. The non-existence of a universally ideal metric focuses attention upon the need to identify and quantify population (“social”) objectives, particularly in the neglected area of fairness, to identify the impact of health services upon social goals, and to provide a richer menu of options for decision makers.

2 EUT as the “Theoretically Correct Measurement”

In orthodox economics, people are assumed to obey EUT. Yet it has been known for a long time that EUT is empirically flawed (Allais 1979, Schoemaker 1982, Kahneman and Tversky 1979, Luce 2000). One reason for its continued use in policy and thus in prescriptive (as distinct from descriptive) economics is a desire to employ what is believed to be a rational decision procedure: EUT’s axioms are taken to encapsulate an ideal of rationality (e.g. in Marschak 1950, Schoemaker 1982, Thaler 1992, and in the health economics context, Torrance and Feeny 1989). Formally, this defence of EUT is unsatisfactory. Pope (2004) demonstrates that the conclusions about which particular health interventions yield greatest benefits may be reversed depending on whether the von Neumann and Morgenstern or the Ramsey version of EUT’s SG is used. Yet if EUT were descriptively valid, or an acceptable criterion of rationality, both procedures would yield identical conclusions as both are derived from EUT.

But the flaw in EUT is more fundamental. The theory omits considerations which are potentially so important that ignoring them would normally be considered a sign of irrationality. The error arose from EUT theorists focussing on simplistic accounts of social gambling instead of upon serious decisions concerning health, jobs and investment. Pope delineates the omissions in a succession of revealing publications3. Two of these papers relate to the use of EUT’s Standard Gamble (Pope 1989, 2004) in health economics. In this, a person’s NM utility is calculated from three pieces of information: (i) the utility of full health (conventionally put at 1.00), which is experienced if the outcome is favourable; (ii) the utility of death (0.00), which is experienced (so to speak) if the outcome is unfavourable, and (iii) the probability \( p^* \) that would leave a person indifferent to the option of a designated health state and a second option consisting of probabilities \( (p^*) \) of being dead, and \( (1-p^*) \) of being in full health. Under EUT, \( p^* \) provides a cardinal measure of the utility experienced in that designated health state. Note that under EUT, “experienced” refers to the psychological state after the outcome is known (ex post). This is because EUT requires choosers to evaluate each outcome “as if certain” (Friedman and Savage 1948), and an outcome can only be certain ex post. This restricts admissible emotions to those unrelated to risk and uncertainty. For instance, the emotions arising from a pain-induced bad mood can be included in EUT.

However, this information is insufficient for reasonable or rational choice. Risk and uncertainty are necessarily temporal concepts. At the time choices are made, outcomes are in the future. But this fundamental characteristic of the choice context is not reflected in the NM axioms, which are a-

---

temporal. In addition to ex post utility, to which EUT is restricted, reasonable choice requires people to consider their experiences in the “pre-outcome period” (or periods), i.e. in the period(s) from the time when it is first known that choice will be necessary to the time when the chooser knows which outcome will actually occur.

These can affect a person for at least two reasons. First, a person’s capacity to plan is affected by the level of their “knowledge ahead” (Pope et al. 2006a). A patient may elect to take a world trip while their prognosis is only a 50:50 chance of life. With a certainty of life they would spend this time engaged in other activities. Similarly, an AIDS patient with only a 10 percent chance of living for 20 years might adopt a five year old child immediately rather than wait some years before conceiving a child naturally. That is, with varying levels of “knowledge ahead” a person will change their behaviour in the pre-outcome period which is ignored by the axioms of EUT.

Secondly, and of particular relevance in health care decision making, the riskiness in the pre-outcome period generates risk-related emotions such as anxiety, hope, excitement or fear about the future unknown outcome. These are not associated with the ex post health states per se but with the risk per se. In the ex post health state “full health” these emotions will not be experienced; neither will they be experienced in the “health state” death. EUT cannot therefore capture the emotions of a person facing a probability of full health and a probability of death.

Consistent with this argument, recent research confirms the existence of a specialised synapse in the brain that is responsible for surprises and anticipation. It is the anticipation of reward and not the reward itself that stimulates the relevant neurons (Klein 2006). Following from this, the response of this part of the brain to the anticipation of outcomes cannot be reliably deduced from the effect of actual outcomes (in a neurologically distinct part of the brain). There is no evidence that these complex neurological systems are synchronised according to the rules of mathematical logic. More generally, anticipation cannot be plausibly modelled in an a-temporal framework.

Risk-based emotions have been referred to in the literature variously as the "specific utility of the gamble per se", “the love of gambling”, etc. However, until Pope’s pioneering work they were not analysed in a temporal context of pre- and post-outcome. In the SG typically used by health economists the two outcomes of the gamble are the extremes of death or full health. If the gamble is taken seriously these options are likely to evoke a strong response and it is these risk-based emotions, “g”, which are external to the health state, that are the wildcard in the calculation. They influence, and possibly dominate, choice. The probability p* is not, therefore, purely and simply an indication of the average utilities of later learning that one has full health or death (index numbers 1.00 and 0.00) or the health state under consideration (U_s). They also reflect a potentially powerful and kaleidoscopic mix of emotions experienced before the outcomes are known. Pope (2001) introduces the term “primary” satisfactions for those that EUT includes, and “secondary” satisfactions for those that EUT excludes, due to the theory’s exclusion of factors based on degree of knowledge ahead. Her description of a person’s evolving degree of knowledge ahead can be summarised as follows: Upon becoming aware of future uncertainty.

“He does not (initially) know his alternatives. Later he has a change in knowledge ahead having discovered at least two alternative acts ... still he does not know what act he will choose. Later he has a change in knowledge ahead, having chosen an act. Still, if the act is risky he still does not know the outcome of his act. He is in his pre outcome period. Later he
has a change in his knowledge ahead having learned the outcome of his chosen act. He is in his post outcome period. This division of the future by evolving stages of knowledge ahead allows the chooser to anticipate his secondary satisfactions ... and have those affect his choice of an act”. (Pope et al. 2007, p. 203)

Observed behaviour will reflect the individual’s decision rules through these changing stages. Their optimal behaviour should take account of present knowledge and knowledge of the future – knowledge ahead. At the (pre outcome) time of the key decision, knowledge ahead differs from the (ex post) knowledge included in the EU algorithm.

Neither is it plausible to posit that an associated emotion must be linked to a particular final outcomes in some universally applicable way when, in the initial stages of the sequence described above, outcomes are imperfectly known or not known at all, and all risk-based emotions depend not on a single outcome, but on the extent to which there are or were multiple different possible outcomes. The potency of the critique arises, in part, from the fact that it is not confined to EUT but to all a-temporal attempts to axiomatise behaviour under risk and uncertainty.

In view of this critique it is unsurprising that von Neumann and Morgenstern did not believe that their theory represented a general theory of choice under risk, a point reiterated by Morgenstern (1979, p. 181).

“...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 risk, etc. But what we did say and what I do feel I have to repeat even today after so many efforts have been made by so many learned men, is that the matter is still very elusive. I know of no axiomatic system worth its name that specifically incorporates a specific pleasure or utility of gambling together with a general theory of utility.... I am not saying that it is impossible to achieve it in a scientifically rigorous manner. I am only saying (as we did in 1944) that this is a very deep matter.”

But von Neumann and Morgenstern’s own view about their contribution to EUT has been largely ignored in mainstream economics. EUT’s SG is still widely and wrongly seen as the gold standard for measuring QoL. However, from a normative perspective, to argue that a person should ignore powerful emotions and behaviour change prior to the knowledge of final outcomes could indeed be described as to argue irrationally.

**Two Erroneous Rescue Attempts**

It has been argued that any emotions relevant to utility could be treated as outcomes - “elaborated outcomes” - and included in EUT. The case for this is strengthened by the emotion-charged nature of many health states. Emotions such as anxiety and reassurance are often important ingredients in health care decision making, for example in relation to breast cancer screening (Farmer 2000, Brain et al. 2000, Holloway et al. 2004) and pregnancy screening (Santalahti et al. 1998, Kornman et al. 1997, de Graff et al. 2002).

More generally, the success of health services is subject to risk. However, the elaboration of outcomes encounters practical and theoretical problems. In the case of the SG the risk of life-death is introduced as a measurement instrument and the emotions associated with the likelihood of
Instant death in the pre-outcome period needs to be anticipated, measured, and subtracted from the overall expected utility of that branch of the gamble. Pragmatically, a methodology for achieving this and the identification and measurement of other emotions has not been devised and is probably intractable. It would be necessary to elicit from a person their assessment of the disutility of death while purging their emotions of anxiety, fear, hope, anticipation, etc, a task particularly difficult for the majority of the population that believe in an uncertain afterlife. In view of the evidence cited from Klein, above, the task is likely to be psychologically impossible. The theoretical objection to the elaborated outcome solution is that it ignores the constraints that the EUT axioms impose; constraints that require each outcome to be evaluated identically and have the same emotions, regardless of whether there is a 1 in 10,000 or 95 percent chance of death. Pope (2000) has shown that this is not simply implausible, but introduces contradictions in the axiom’s plan from which EUT derives its authority.

The second, more general rescue attempt is to argue that as some level of risk is associated with all decisions, it would be necessary that decisions should be made “under risk” and that this is achieved by EUT; that the risk element of EUT will reflect, in some sense, the person’s “risk attitude”. This view was probably reinforced by Marschak (1950) who redefined risk to mean a riskless concave utility function. But “any risk” is not “all risk” and Marschak’s definition is seriously misleading. The commitments, behaviour change and emotions affected as a person passes through the various stages of knowledge ahead are potentially highly variable with the level of risk. Small risks may be a source of excitement (poker machines, white-water rafting, etc). Risks to life may be the cause of paralysing fear and dread. The notion that all risk may be captured by any risk borders upon mysticism and the world of “under risk” is a misleading construct.

3 The Impossibility of an Ideal Metric

Disproving an existential proposition is impossible. The assertion in the title of this paper is therefore strictly incorrect as, in principle, a supposed non-existent may always be found. (It is always possible that there are fairies at the bottom of the garden). At best we may argue, by inference, that numerous attempts to identify the best objective metric have failed to persuade others of its superiority and that the most plausible explanation for this is that there is no argument which could achieve this. Nevertheless, a number of famous attempts have been made to find criteria which will allow the identification of the best metric, or the conditions under which a metric may be deemed to have demonstrated a superior outcome. These have been suggested in both the economics and ethics literatures. We briefly consider some of these criteria below and then suggest an overarching reason why no argument may be persuasive. This draws upon public choice theory.

As noted earlier, the health economics literature has focussed upon the QALY, HYE and DALY. The criterion used to judge between the former two has been the correspondence between these metrics and the axioms of EUT. This criterion, however, cannot establish either of these as an objective gold standard, *inter alia*, for the reasons discussed above. However, other more general criteria have been suggested.
The Pareto Criterion, Winners and Losers

In welfare economic theory, the Pareto criterion attempts to circumvent the need for ethical or social judgement at least in certain contexts. A change is a “Pareto improvement” if it makes no one worse off and at least one person better off as defined by their personal utility. A state of affairs is “Pareto optimal” if it is not possible to increase one person’s utility without decreasing the utility of another person. Much of the appeal of WT (Welfare Theory) stems from the fact that in a perfectly competitive model every competitive equilibrium results in a Pareto optimal outcome (Debreu, 1959; Arrow and Hahn, 1972) and, consequently, the criterion appears to be highly relevant for practical economic policy: the change that benefits at least one person without making anyone else worse off should command universal assent.

However, one problem with using the Pareto Criterion to select an ideal metric is that none of the metrics produces only winners; each produces losers as judged by the other metrics. Other problems cast doubt upon the criterion’s practical usefulness and conceptual validity but these are of less immediate relevance here. This is illustrated in Table 1, with 3 elements of a metric, namely, the time frame (life, life year); the measurement of central tendency (mean, median); and the quality of life measure (TTO, SG etc). As indicated under “Group favoured” in Table 1, denoting benefits in "lives saved" advantages projects which save many lives, even if they are only extended briefly – for example, life saving for the elderly. Use of "life years" advantages projects that significantly increase life expectancy – e.g. services for the young. Life expectancy – the metric incorporated in all of the common composite metrics - advantages those who might die early as the relatively large increase in life years for one person would have a relatively large impact upon the mean but not median gain. In contrast, the median gain would favour those who might die close to but greater than the median increase in life years. Use of the mean would favour the provision of neo-natal ICUs; use of the median might favour ICUs for coronary care amongst the middle to old age cohorts. A population with low risk aversion would select a higher probability of death (lower apparent utility score) if the SG was used to measure utility and disadvantage projects promoting quality relative to a methodology where risk aversion was unimportant. Use of the person trade-off (PTO) would (probably) advantage individuals requiring therapies where the benefits were fairly distributed relative to those where only a few gained an advantage.

In sum, choice of metric will have a complex distributive effect on the pattern of benefits, and in particular will produce a profile of winners and losers. Since the Pareto criterion does not apply when there are losers, it cannot provide a criterion for choosing between the metrics and the distribution of benefits becomes a pivotal issue.

4. The practical problem is that the criterion requires the measurement of utility order to ensure that it is not reduced for the person who is not made better off – for example, as a result of envy. When utility is replaced by income in the definition the criterion is rejected; the majority of a sample of Australians did not believe there was social improvement when the wealthy became wealthier and the poor did not have their income reduced (Anderson et al. 2007). The conceptual problem with the Pareto criterion is that it is applicable only if there is a welfarist social welfare function. Sen (1982), for example, notes in his “Impossibility of a Paretian Liberal”, that the criterion may be inconsistent with the principles of liberalism. However, this is only one example of a much more damaging criticism. If social welfare is viewed in anything but welfarist terms – for example, egalitarian, communitarian, proceduralist or any of a number of deontological views – the criterion does not result in an improvement in social welfare.
Four traditions are relevant to this problem of winners and losers - viz, a potential compensation, use of a social welfare function, normative ethics and public choice theory. “The first approach seeks to avoid the problem; the second and third tackle it head on, and the fourth suggests that it is intractable.”

Table 1. Advantage bestowed by Different Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Group favoured</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Life</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Frame</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lives</td>
<td>Elderly and Short LE</td>
<td>Maximised by large number of short lives</td>
</tr>
<tr>
<td>Life Years</td>
<td>Young and Long LE</td>
<td>No diminution in benefit as one person gains more</td>
</tr>
<tr>
<td>Measure of central tendency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>Young, persons with large gain</td>
<td>As for Life Year: no diminution in benefit to one person</td>
</tr>
<tr>
<td>Median</td>
<td>Persons just above median (middle aged)</td>
<td>No disadvantage if LE is small (above the median)</td>
</tr>
<tr>
<td>Valuation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Stated Preference, WTP</td>
<td>Wealthy patients capable of paying more</td>
<td>Based upon Welfare Theory, WTP must apply to the individual paying for themself. It does not apply to others with different income (Richardson and Smith 2004)</td>
</tr>
<tr>
<td>(b) Extrapolated risk, WTP</td>
<td>Patients whose raters have high risk aversion</td>
<td>Aversion to risk is interpreted as (extrapolated to) a high value of life</td>
</tr>
<tr>
<td>B. Quality of Life</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Utility</td>
<td>Patients with poor QoL evaluated by those with low risk aversion</td>
<td>Risk related pre outcome emotions do not result in high P implying a large quality of life effect (1-P)</td>
</tr>
<tr>
<td>SG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTO</td>
<td>Patients with poor QoL evaluated by those with low rate of time preference (RTP)</td>
<td>Discounting impacts more upon longer LE in poor QoL in the TTO instrument. Low RTP discounts future years of poor QoL less</td>
</tr>
<tr>
<td>PTO</td>
<td>Patients with less severe illness and limited potential benefits if raters values equality of benefits</td>
<td>PTO Raters will prefer the option benefiting the most patients because of a preference for equity per se. The instrument is calculated by the number of beneficiaries which must be greater when benefits are smaller.</td>
</tr>
<tr>
<td>WTP</td>
<td>See 'Stated Preference’ above</td>
<td></td>
</tr>
<tr>
<td>(b) Non Utility</td>
<td>Patients with high “Set Point”/ non depressive disease</td>
<td>Greater potential gain in QoL</td>
</tr>
<tr>
<td>Happiness (SWB)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capabilities</td>
<td>Patients with low capabilities; high utility and SWB</td>
<td>Patients receive unsought benefits</td>
</tr>
</tbody>
</table>

* “Raters” refers to the group carrying out the QoL exercise upon which results are based.
The Kaldor-Hicks potential compensation principle for losers

The Kaldor-Hicks potential compensation principle is a corner-stone of Cost-Benefit Analysis and is commonly evoked to overcome the problem of interpersonal comparisons and to create the necessary logical nexus between changed states and social welfare (Kaldor 1939, Hicks 1939). According to this principle, one state of the world is better than another if those who are made better off could potentially compensate those who are made worse off. As currently used in cost-benefit analysis, it involves consumer surplus calculations that are inconsistent for different health interventions. In theory, that inconsistency problem can be overcome by an alternative approach (Selten and Winter 1994).

However, there are other difficulties. The Kaldor-Kicks principle can only apply if it is possible to compensate those who are made worse off. For patients left with a very poor quality of life this may be impossible, and it is unequivocally impossible when a patient dies. This implies that the attempt to base policy decisions on the achievement of Pareto efficiency fails in the health sector, and with it the attempt to by-pass ethical issues in the formulation of health policy. While the economics literature usually acknowledges the unreality of much of WT, this final conclusion is less commonly acknowledged.

A more fundamental criticism is that the Kaldor-Hicks criterion simply inserts an additional and controversial normative assumption: potentially superior states should be treated as if they are superior. However, it is questionable if losers would agree that a situation was better, rather than simply potentially better, in the absence of compensation. This highlights a final anomaly. In an NHS, losers (those who do not receive health services) include wealthy, healthy taxpayers. Compensation would require the taxation of poor, unhealthy recipients of public health services and a redistribution back to healthy, wealthy losers. But this would offset one of the explicit objectives of all national services and the option cannot, therefore, resolve the present problem, even in principle.

Social Welfare Functions

The second school of thought constructs social welfare functions and seeks to incorporate ethical judgements into a "function" which, in principle, allows a comparison of winners and losers. The Bergson-Samuelson social welfare functions pioneer this genre (Bergson 1938, Samuelson 1947). In practice this is formally equivalent to agreement upon the criteria that are to be used to select social objectives and the metrics which seek to measure them. This was the approach adopted by Richardson (1994) to distinguish between the use of the Standard Gamble, Time Trade Off and Person Trade Off, and further developed in Richardson (2002a). However, this approach also encounters a problem, this time in the form of the logical argument foreshadowed earlier. There are, potentially, a very large number of sets of criteria reflecting the different purposes (social goals) individuals and societies seek to achieve. Following this second approach the competing criteria would need to be evaluated. But this requires meta criteria in terms of which the contenders must be evaluated, which themselves require meta meta-criteria. The infinite series generated in this way indicates that logic alone cannot determine the superiority of one set of criteria with which to evaluate the competing metrics.
A second more damaging criticism is that the social welfare function does little more than obscure the problem by abstraction and by applying a technical label to a non-existent solution. A “function” per se does not represent a logical or ethical argument why variables should be included or arranged in a particular way. The arrangement may appeal to some or even to a majority of the population; but the task is then to justify the adoption of this as a criterion. A “function” implies a consistency in the ordering of states according to the variables that are included in the function, but as context changes geographically, temporarily and socially, numerous variables are omitted and inconsistency could generally be explained in terms of some ad hoc event or circumstance omitted from the function. In sum, the social welfare function generally represents an avoidance of the issue in its abstract form and a (generally) unsupported assertion in its concrete manifestations.

**Normative ethics**

The third school of thought recognises that choosing a gold standard requires identifying who should be the winners and losers, which is unavoidably an ethical problem. Several philosophers have argued that this is possible:

Firth (1952) suggests that agreement on ethical matters can be achieved, or approached, by imagining what an "ideal observer" would do when confronted with a moral choice. By hypothesis, the ideal observer is omniscient, disinterested, dispassionate and benevolent. He or she has faultless powers of prediction, is fully aware of the needs of others, knows the economic and political conditions that obtain in society, and so on, but is not influenced by his or her own interests or passions.

Brandt (1979) argues that agreement on ethical issues can be achieved by attending to the moral code that a "fully rational person" would support (for a society in which he or she expected to live a lifetime). A completely rational person is not necessarily omniscient and disinterested, but is one who has undergone a thorough and effective course of "cognitive psychotherapy", during which he or she has examined, criticised and ultimately replaced all ignorant and prejudiced desires with rational ones.

Hare (1981) suggests that deliberators will converge on the correct answer to moral questions by following the dictates of an "archangel" who is devoid of human imperfections. The archangel has "superhuman powers of thought, superhuman knowledge and no human weaknesses". According to Hare: "archangels, at the end of their critical thinking, will all say the same thing, on all questions on which moral argument is possible; and so shall we, to the extent that we manage to think like archangels".

Habermas (1979) suggests that agreement on ethical issues can be achieved, or progress towards it made (1992), by placing individuals in a hypothetical "ideal speech situation". In this situation participants are not limited in their knowledge or reasoning ability, nor are they subject to the distorting influence of oppressive social relations. Knowledgeable, rational, unbiased participants in un-coerced "ethical discourse" aimed at clarification, justification and understanding, will approach the "truth" on ethical matters.

Rawls (1971) suggests that ethical agreement can be reached on matters of justice by considering what a "rational egoist" would choose, or what a group of rational egoists would agree upon, from
behind an imaginary "veil or ignorance". The rational egoist is motivated by self-interest, but does not know "his conception of the good, the particulars of his rational plan of life, or even the special features of his psychology such as his aversion to risk or liability to optimism or pessimism" (p. 118).

These approaches hold out the hope that ideal ethical agents placed in ideal circumstances, might arrive at uniquely correct answers to ethical questions. But there is no reason to believe this could translate into practice. The conclusions that I believe would be reached by an "ideal observer" not influenced by his or her own interests or passions, or by a "fully rational person" after thorough and effective cognitive psychotherapy, or by an "archangel" with superhuman powers of thought and knowledge, or by a member of an "ideal speech community", or as a "rational egoist" behind a veil of ignorance, may be different from the conclusions you believe such an agent would reach. While ideal ethical agents can leave aside their own interests and transcend their own empirical circumstances, in attempting to simulate the reasoning of such agents, or discern the conclusions they would reach, we cannot. Each individual’s assessment of how much weight an ideal agent will give to competing moral principles, their assessment of what facts he or she will consider relevant, or what importance he or she will place upon the keeping of promises, the avoidance of suffering, and so on, will be shaped by each individual’s own sympathies, commitments, and so on, which are in turn the products of their character, history and genetics.

These same elements would also influence the answer to two related questions; first, knowing the answer that would be given by an ideal ethical agent, would each individual or even the majority of individuals accept it? If not, the practical relevance of the solution must be questioned. Secondly, would people accept a methodology based upon the transfer of their own (possibly self-interested) influence to some higher moral authority?

A simple thought experiment suggests that the answer to both questions would be "No". Suppose – as is likely – that ideal ethical agents rejected the importance of national boundaries, adopted a world perspective, and suggested a major redistribution between nations; or an inter-temporal perspective which resulted in a punitively high worldwide carbon tax. It is possible that many would accept the advice in principle but it is also probable that it would have little potential impact on policy. Many, or the majority of those consulted, would argue that their reasoning, sympathies, commitments, etc are as legitimate as those of any other moral agents. Solutions that disenfranchise individuals because of their individual characteristics do not solve the current problem, which is to indentify an ideal metric for use in the real world.

Public Choice Theory

From one perspective the conclusions reached above are unsurprising, even trivial. Both welfare and ethical approaches are based implicitly or explicitly upon the assumption that certain arguments, axioms or principles are so plausible that they would attract universal or widespread agreement. While such a principle may exist, there is no reason why this is necessarily true, and after many centuries of unsuccessful searching, there are good grounds for doubting this. The assumptions underlying our reasoning are not "technically correct" - they are normative – and the plurality of conflicting experiences in society suggests that there is no realistic prospect of agreement on any such principle.
The alternative approach to the problem of winners and losers is therefore to seek some form of ideal social decision making, drawing upon social choice theory. However, within democratic countries the methods of social choice rely directly or indirectly upon voting, and Arrow’s impossibility theorem demonstrates that these cannot satisfy a minimum set of conditions – “the axioms of democratic legitimacy” - which are needed for ideal decision making.

The theorem was a turning point in terms of Public Choice Theory and its analytical methods. In the words of Suzumura, one of the editors of The Handbook of Social Choice and Welfare:

“Arrow pioneered the axiomatic approach in social choice theory, which enabled him to analyse (all) Arrovian social welfare functions all at once; by imposing a set of axioms which are deemed necessary for the Arrovian social welfare functions to be reasonable, hence acceptable. It is this novel methodology which enabled him to analyse all the relevant social welfare functions at once stroke and led him to the celebrated General Possibility Theorem, or the Arrovian Impossibility Theorem” (Suzumura 2002 p. 11)

The methodology is not, in fact, novel, as it is the methodology of (philosophic) rationalism which has a 2,500 year heritage (Tarnas 1991). It was, however, new in this context. Arrow’s axioms, or more correctly, criteria, in the 1963 revised version of his theorem (1963, pp. 96-97), were as follows: (1) every individual may have any preference order, and it must be possible to aggregate these individual preference orders into a social preference order which has, inter alia, the property of transitivity, (2) unanimous preferences must result in social preference, (3) it must be (a) necessary and (b) sufficient to know how individuals rank two states to determine social order and (4) there can be no dictator. Arrow proved that these conditions are mutually inconsistent at least some of the time. For a start, preference orderings may be intransitive. This is sufficient to demonstrate that there is no method of voting that can identify an unambiguous concept of social preference based upon ordinal preferences. This in turn implies that there is no way of determining which of the competing metrics (inter alia, those in Table 1) best measures “social choice”. Any metric selected results in winners and losers, and AIT, the Arrovian Impossibility Theorem, implies there is no legitimate, logical satisfactory way of judging between the options.

4 Methodology and Selection of Metrics

This latter conclusion appears to leave the theoretical basis for measurement in disarray. How do we justify the use of a particular metric? The axiomatic approach of welfare theory is unsatisfactory as it avoids the issue of interpersonal comparison. The differences in individual characteristics imply the need for public choice and a social contract but AIT indicates that voting systems are intellectually unsatisfactory in violating one or more apparently reasonable conditions.

Unsurprisingly we do not resolve this apparent dilemma below but make two suggestions. The first concerns methodology: we argue that, for the purposes of choosing a metric, subsequent analysis should abandon rather than build upon the analytic foundations of AIT. Secondly, we argue that the “solution” to the dilemma posed above lies in determining what may reasonably be expected of economic analysis. Rearranging the goal posts is justified if it allows subsequent play without the need for arbitrary rules.
Methodology

One approach to the problem consistent with the literature is to seek the ultimate justification for a metric from within social choice theory through some modification or extension of Arrow’s framework. It is argued below that while the AIT serves a limited purpose – demonstrating the impossibility of an “ideal” metric - the theorem should be viewed as having little practical or intellectual relevance beyond liberation of the analytical framework from the flawed notion that there is such an ideal. We argue that subsequent analysis should not build upon the analytical foundations of AIT; that the theorem is useful for demolition not construction.

The Limited Importance of Arrow’s Impartiality Theorem

Evaluation of the potential contribution of AIT for constructing a framework requires criteria. Three are considered below: first, the intuitive “insight” AIT gives into social decision making; second, the explanatory power of the theorem in terms of social decision making; and third, the usefulness of AIT for building a foundation for applied evaluation studies and specifically the selection of a metric for evaluation studies. It is argued below that the intuitive “novelty” of AIT with respect to understanding behaviour is primarily an artefact of the abstraction of assumptions and the complexity of the subject matter at this level of abstraction and that, to the extent that it demonstrates something counter-intuitive, we should not be too surprised. Negative assessments using the second and third criteria follow from this conclusion.

The best known example of AIT predates Arrow and is attributable to Concorcet (1785). It applies to majority voting. Assume three voters A, B, C with preferences for outcomes X, Y and Z, as follows: A (X>Y>Z); B(Y>Z>X); C(Z>X>Y), where “>” means “preferred to” and there is first-past-the-post voting. Pair-wise voting between X and Y, and Y and Z results in (X>Y) and (Y>Z), each with a 2:1 majority. Assuming transitivity this implies X>Z, but pair-wise comparison results in Z>X also with a 2:1 majority.

Now contextualise the paradox. Suppose that person A is a religious aesthetic who prefers church attendance (X) to charity work (Y) to an afternoon at McDonalds (Z). Person (B) is a moral atheist who prefers charity (Y) to hamburgers (Z) to attending church (X). (C) is a gluttonous hedonist who prefers McDonalds (Z) to church attendance (X) to charity (Y) (which involves some work). From the paradox, church is preferred to charity, charity is preferred to McDonalds but, “paradoxically” McDonalds is preferred to church.

In this example, the AIT is intuitively simple. Each of the pair-wise comparisons uses a different set of criteria and the result of the different criteria is not tidy – i.e. transitive - at the aggregate level. The first comparison, church and charity, “disenfranchises B” - i.e. her vote does not affect the decision. The effective criteria determining the outcome are the religious beliefs of A and the hedonism of C. In the comparison of charity and hamburgers, it is C that is disenfranchised and the effective criteria are the aesthetism of A and the moral behaviour of B. In the final comparison hamburgers are preferred to church because of the atheism of B and the gluttony of C.

In sum, the AIT is not paradoxical when it is contextualised. Different decisions involve different criteria and when criteria change social ranking changes. Sen (2002) makes a similar point when he argues that “health equity is inescapably multi-dimensional...”. This militates against the
expectation that in every comparison of social states there must be a full ranking that places all alternative states in a simple ordering (Sen 2002 p. 660). An alternative intuition in this example is that the choice of question creates an effective coalition between two of the three voters, at the expense of the third, and that with changing coalitions there is no reason to expect transitivity of choice.

A second example indicates that under parallel contextual conditions non-transitive preferences are a familiar and unsurprising problem for an individual. Consider a person who seeks to order his preferences regarding gluttony, G, parsimony, P, and immediate charity C (street beggar/door knock appeal). Upon introspection or recollecting past preferences, the person identifies three patterns arising from separate contexts. In the first he has recently and directly observed deprivation and is motivated to provide immediate assistance. Gluttony appears self-indulgent and distasteful. The resulting preference ordering is C>P>G. The second pattern recognised is that following exposure to information relating to his personal financial situation, he becomes self-interested but cautious. His resulting pattern is P>G>C. Finally, he recognises that when he is carefree he has a desire for instant gratification – primarily via consumption but also from a feeling of virtue. This results in G>C>P.

In sum, the person recognises that he is, effectively, three people about whom he has complete information and whose preferences he perceives (analytically) to be context specific. In order to be rational and impose consistent preferences he grants each person one vote. This replicates exactly the conditions and consequences of the Condorcet paradox. The individual cannot rank his preferences in a satisfactory way. In a particular context, pair-wise comparisons will be made but the equivalent to the aggregation problem in the AIT is that she will recognise the existence of the other two contexts, the preference patterns they generate and the conflicting ranking which the remaining two contexts imply for any pair-wise comparison.

The chief difference between this case and AIT is that this “paradox” is well known and embodied in our vocabulary. People “vacillate”. They “waver and dither”. They “change their mind”, “chop and change”, etc. The absence of corresponding words that apply at the social level is probably indicative of the relative importance of the “paradox” at the two levels. Of course, at the individual level, decisions will be made. It is possible, though unlikely, that the individual will attach stable preference weights to decisions and subsequently act consistently (a solution unavailable in any existing voting system). More likely, however, is that in practice decisions will reflect the preferences of the most recently dominant context. It is unlikely that even a reflective and analytical individual would consider the problem unusual or paradoxical. In sum, as with the previous example, the paradox is simply a function of the existence of different criteria, this time changing through time with context.

Evidence that individuals may sometimes have intransitive revealed and stated preferences was noted by May as early as the 1950s (May, 1954). Experimentally derived stated preferences resulted in intransitivities by the use of three different ranking criteria for partners, with pair-wise choice based upon good looks, intelligence and wealth. Group pair-wise choices were circular and intransitivities occurred with 36% of individuals.5

5 May also notes the intransitive revealed preferences of hungry rats that prefer food to sex, sex to pain avoidance, but pain avoidance to food. (He also notes that all people do not always behave like rats). In a
Prima facie it does not appear surprising that if individuals are unable to reach clear preference orderings then social groups formed by the aggregation of these individuals might likewise be unable to reach clear social rankings. However, economic theory creates an intellectual environment that inhibits the transference of intuitive understanding of the individual to the social level by assuming that the well-known attribute of human inconsistency does not exist and by defining “rationality” itself in a way that discounts this attribute. (It is not rational to define “rationality” in a way that discounts a fundamental human attribute - in this case the effects of inconsistency - and then expect the theory to explain behaviour.) Without this denial of human nature it would be intuitively less surprising to find individual traits replicated at the social level.

The AIT does not depend, as in our examples, upon first-past-the-post majority voting. Nevertheless, the general point remains valid. In any voting system there is the likelihood of multiple criteria that differ in importance between people, and permitting this is the first of Arrow's axioms of democratic legitimacy. Voting necessarily "disenfranchises" some voters in the sense illustrated above – these people’s votes cease to be important because their preferences become those of the minority. Overall, therefore, the importance of different criteria shift and the "average criteria" changes: a different de facto coalition has different criteria. Thus, different effective criteria may apply when a conclusion is reached from two or more pair-wise comparisons and the assumption of transitivity on the one hand, and from a direct pair-wise comparison of the first and last options on the other. It is not paradoxical but a logical consequence of the fact that different criteria result in different outcomes.

Similar comments to these apply to Sen’s famous paradox which demonstrated the possibility of conflict between the principles of liberalism (individual freedom of choice) and Pareto improvement, requiring coercion to increase the utility of two individuals (Sen 1970). Like the AIT, Sens “paradox” arises from the (unsurprising) possibility of a conflicting outcome when two different criteria are employed. Similarly, conflict may arise if an optima is to be simultaneously judged by Pareto and egalitarian criteria, equality of opportunity and of outcome, procedural and distributive justice, equity and efficiency or, generally, between any two distinct normative principles.

Nevertheless, it may appear that the very generality of AIT is the reason for its counter-intuitiveness. Surely there must be some voting system that satisfies the axioms? Our intuition surely suggests some form of democratic procedure is reasonable. However, in general terms we should be deeply suspicious of our superficial intuitions, which we know to be very limited outside the context of what is immediately needed for day by day activities, and the abstraction of Arrow’s general argument puts it far outside this range. We have little capacity to appreciate the infinitesimally small magnitudes that exist in the quantum world or the consequences of the simple mathematical relationships in chaos theory that explode into the extraordinary complexity of fractals. The desire for transitivity is probably best explained as a manifestation of the “Sunday School Effect” - the desire for the world to be simple and intuitive.

Paradoxical or not, AIT could be an important element in the framework for evaluation if it provided insight and explanatory power. However, we are unaware of any example in the history of social different context, the reported preferences of desperate pilots in World War II were for death by flames, to death by hot metal, to death by falling, to death by hot flames (May 1954 p. 7). (Results were not, presumably, based upon a random control trial.)
decision making when the AIT has provided such insight. Examples may exist but they are not obvious or commonplace in the literature. This is unsurprising, as the idealised axioms of AIT do not capture real world conditions. Individuals do not have well-ordered preferences over a wide range of issues; there is preference failure (Goodin 2000), preference pollution (George 2004). The doctrine of the “immaculate conception of preferences” is false (Rice 2003). People rarely vote for issues but rather for parties, and people commonly alter their views with party policy or with the views of friends, radio, television or newspaper columnists: individual preferences are unstable and often quickly mutable.

Within Parliament, similar “imperfections” exist. It does not approximate the assumptions of AIT. In the short run there is commonly a near dictatorship of the central Cabinet plus senior public servants. Members of Parliament follow Party discipline. Parliament does not rank options so much as assess the cost implication for other options, something that is often opaque. The legitimacy of policies is commonly equated with their presentation and debate during elections and discussion with stakeholders. It is again doubtful if, in the cumulative history of Parliamentary democracy, legitimacy has ever been challenged because of the non-transitivity of decisions or other elements associated with the AIT.

The rejoinder to these observations, of course, is that AIT was never intended to be descriptively accurate but to indicate the difficulty of determining a satisfactory basis for legitimate choice. But there has never been difficulty. As an empirical fact decisions have been made as long as social units have existed, and such decisions have been considered “legitimate” (Daniels and Sabin 1998) and this also applies to democratic, voter-based polities. Legitimacy does not necessarily imply consistency but varies with the social history of the Polity. Even faced with an AIT type paradox, few would question the right of Parliament to make a choice, albeit one inconsistent with other decisions. Rather, AIT commences with the implicit assumption that there should be a fully consistent intellectual edifice and demonstrates that this is problematical. This negative conclusion does not provide the basis for a subsequent theoretical framework offering practical guidance (nor, we suspect, do the subsequent efforts to create one).

The suggested conclusion to be drawn from the AIT is that there is, in fact, no fully coherent concept of “social preference” and that phrases like “society is better off”/ “social improvement” have no unambiguous meaning. That is, it is probably true that the theorem demonstrates that there is no “technically correct” meaning of these terms, although, like many phrases, they can be usefully employed to signal an area of interest for discussion6. They are “essentially contestable” concepts (Grafstein 1988; D’Agostino 1996). That is, while there is sufficient agreement on their basic meaning to enable discussion and debate, there is disagreement about the application of these concepts to concrete phenomena (actions, policies, institutions etc). Meaning could be bestowed on these phrases by defining “better off” in utilitarian terms, for example, but those ascribing to this doctrine do not have the authority to impose it upon others.

In support of the “essential contestability” thesis, it might also be argued that the existence of intuitively appealing phrases is no guarantee of a corresponding coherent concept and particularly

6. Similarly, the disciplines of ethics, metaphysics and aesthetics, amongst others, have a long history of disagreement about the precise meaning of numerous words that are usefully employed in everyday communication by the majority of the population.
one that might be operationalised. Words are used in a context where they are meaningful and where they might become familiar in this context. They may then be transferred to another context (often metaphorically) and the familiarity of the phrase may then suggest the existence of a useful, coherent concept which does not, in fact, exist. In the present case, there may be a form of social anthropomorphism in which the characteristics of an individual are transferred, unhelpfully, to the society as a whole. The nihilist conclusion from this line of reasoning might be that, in the absence of “coherent” concepts of “social improvement” etc, there can be no useful metric which might justifiably be used in economic evaluation (or elsewhere).

5 Measurement without universal validity

The previous conclusion is unnecessarily pessimistic, and reflects a view that is implicit in much of the orthodox treatment of benefits - i.e. that there must be a universal “correct” metric which can gain its authority from axioms such as those proposed by von Newman/Morgenstern at the individual level, or Arrow’s conditions for social aggregation. However, it is not necessary to pursue these paradigms. Alternatives are readily available in the physical sciences and, indeed, in other areas of economics. In these arenas, it is normal to construct a descriptively accurate theory which explains what is known within a particular (psychological, economic etc) domain, and to use this as the basis for subsequent theoretical or practical predictions. In the case of normative analyses, the descriptively accurate theory is a constraint upon independently determined objectives (the normative objective of preserving life results in policies constrained by the theories of anatomy and medical science).

The two suggestions in this section are, first, that there are concepts of "social preferences" and "social improvement" which, although ambiguous, are nevertheless useful, and that do not obtain their authority from the orthodox paradigm, and, secondly, that the use of a particular metric in economic evaluation does not require a rigorous justification in terms of its compatibility with a universal theorem of optimal individual and social choice: much less stringent justification is needed so long as the limitations of a metric are understood.

With respect to the first question there is a useful analogy between the concepts in question and the “social welfare function (SWF).” There is no universal, technically correct function and the SWF may incorporate variable and eclectically selected values. The concept is vague but the label denotes the need for a collective judgement to be made that incorporates inter-personal comparisons. The role of the term “social welfare function” (SWF) is useful even if the “function” does not indicate consistent ranking or other particular characteristics. The properties – Welfarist, Rawlsian, Libertarian etc – may be added in particular contexts. Analogously, the concepts of “social preference”, “social improvement” can – and are – used flexibly and are generally understood. The assertion that social improvement has occurred will usually be translated as meaning that according to the values of the person making the assertion or the values of a particular group, improvement has occurred. A “social preference” for an option will be translated as a preference by some group and possibly a majority of the population for that option. The phrase may be endowed with much greater precision if, like a particular SWF, the arguments in the function are precisely defined.

The argument here defends the semantics but not the substantive role of the “concepts”, which is to help determine whether or not society is “better off” because an option satisfies “social
preferences” and in this sense represents “social improvement”. The negative message of the AIT is that such universal concepts are incoherent and their appeal probably relates to the Sunday School effect.

While this latter conclusion implies that a fully integrated intellectual edifice which may be used to demonstrate the universal superiority of a particular metric in monitoring and evaluating social improvement does not exist, this is probably a reflection of the underlying reality. As post-Platonists had to abandon such an edifice in the physical sciences, economists may likewise have to abandon this aspiration.

Nevertheless, useful economic evaluation can be undertaken and the assessment of alternative metrics is a useful venture even if there is no fully satisfactory universal edifice. Economists do not make ultimate decisions in an NHS or, more generally, in the economy in their capacity as professional economists. Their role is to provide advice concerning options and the ranking of options and rankings may be assessed with different methodologies and metrics. Economic evaluation and analysis may potentially be justified by the (positive) hypothesis that whatever the prevailing system of voting or (normative) decision making, there is a greater probability of progress – as defined within a particular polity – when the number of options and evaluations increases. Conversely, the value of the options and evaluations (as judged by their ability to enhance progress defined in this way) is increased when they are based upon concepts and metrics suitable for the questions of interest to decision makers and to those influencing decision makers. Options and evaluations may be undertaken, not because they are implied by an edifice, but because they are relevant to (1) the government; (2) the government bureaucracy; (3) the values of a significant proportion of the public; (4) the values of patients; (5) increasing GDP with or without the possibility of compensation for losers. Work may be usefully undertaken in the belief that it will attract the attention and possibly alter the values of any of these players. The technical needs of these analyses in turn create the criteria for evaluating various metrics.

The previous paragraph subsumes what has been described as "empirical ethics" (Richardson 2002b; Richardson and McKie 2005) - the identification of population values by empirical enquiry. These represent country-specific "social preferences". No normative ethical theory can be “confirmed” in terms of popular opinion. Nevertheless, knowledge of majority values and their strength can be an important input for decision makers charged with making practical decisions. When values are widely held it is sensible and appropriate for economists to adopt them provisionally, until they are explicitly addressed and possibly replaced.

This approach is consistent with the more empirical branches of economics, which avoid questions of ultimate justification by adopting widely accepted goals – full employment, low inflation, economic growth – investigating their determinants and formulating policy for their achievement. "Empiricism" in this form does not imply immutable goals. Through the interaction of research, evolving public opinion and the response of decision makers, objectives may mutate to include leisure, environmental protection and, possibly, subjective well-being. As they do so, the metrics used in the social maximand will be redefined, with the criteria for "best metric" being the reliable and valid measure of the social goals. We are suggesting a similar relationship between metrics and objectives in the area of economic evaluation.
Discussion

In 2500 years the discipline of ethics has been unable to identify a universally acceptable criterion for judging the superiority of alternative social states; but this has not held back practical policies which, on reflection, most (but not all) would believe led to “social improvement”. The policies were based upon social, scientific, political and ethical theories which, in turn, dictated the nature of various metrics. It was not the case that the metrics were created a priori from axioms defining improvement in an overarching intellectual edifice. The progress was much more ad hoc and the relationship between theory and values much more interactive. We have suggested here a similar relationship exists in economics and specifically health economics between measurement, individual welfare, social choice theory and social progress. This is consistent with the textbook lesson to students that economics generally seeks marginal improvement, local not global optima.

We have suggested an impossibility theorem; every metric used for evaluation purposes favours one group over another, and there is no objective criterion for choosing between them. In the context of an NHS, compensation cannot be given even in principle, so an overall judgement of the social benefits of different options which compares winners and losers is inevitable. But in the existing framework of democratic countries in which decisions are made directly or indirectly by voting, it is impossible to do this in a way that satisfies certain minimum notions of coherence in the decision process, implying that there is no clear and unambiguous notion of “social improvement” which, consequently, renders impossible the task of identifying a “best” metric – the metric which will best identify social progress defined by quasi-technical criteria. We have further suggested that this need not be seen as a fundamental problem for economic evaluation or analysis but as an opening up of the question of the choice of metric.

In more concrete terms we have argued that each of the metrics commonly used (and the very large number which could, but have not, been considered) have an inherent bias relative to other metrics. Sometimes this is explicitly noted and defended. For example, while metrics with an ex ante and ex post perspective are likely to produce different magnitudes, one of these perspectives may be desired: according to one set of values ex ante (or decision) utility may be a better indicator of what is sought than experienced utility. The reverse is also possible. The theme of this article is that while criteria may and should be proposed to evaluate these options, there is no objective, universal answer - either option might better suit a particular society with its potentially different goals and values. This does not mean that ethics is relative to a culture (on “cultural relativism” see Brandt 1959 and Williams 1972), but that practical, political decisions must take account of local conditions. The same comment applied to comparison of the PTO, TTO, SG and various attempts at defining (social) willingness to pay.

The most obvious target in our critique has been the special position afforded the SG. The axioms that bestow its status omit consideration of the complex and possibly dominating emotions and the stages of information ahead that accompany ex ante decision making. More fundamentally, they presume a particular social goal, which is a function of individual utilities, and which evidence suggests is far from universal. National health schemes appear to fit the communitarian, extra-welfarist (or non-welfarist) description more readily than the welfarist model.
We have argued that even within an imperfect intellectual edifice progress can be achieved by generating options and providing alternative answers. Despite our emphasis on the differences between metrics it remains true that there is generally a moderate to large correlation between them. Even if, in retrospect, a metric was superseded, earlier evaluation based upon it would almost certainly be a more reliable guide to social objectives than idiosyncratic, clinician-driven decision making. The implied need for ongoing, culture-specific research into metrics is likely to produce second-order improvement, not a complete overthrow of present methods.

In lieu of “technically correct” methods we have argued that the rationale for economic evaluations should be seen as the provision of relevant options as this increases the likelihood of progress, however defined and operationalised within a particular polity. The process of progress is self-evidently erratic. It is mediated by politics, the press, vested interests and public opinion. New ideas will be "filtered" by this process and accepted or rejected according to the mores of the intellectual market place. Intellectual "natural selection" may be highly imperfect but it exists and justifies the production of new ideas (cf Dawkins (1989) on “memes”).

The arguments here imply the need for feedback from policy makers to help inform economists about the nature of the social goals. A dilemma which arises is that, presently, policy makers seek, rather than provide, advice on values from economists; that is, they commonly accept the authority that orthodox economics claims in these matters. This places an obligation on economists to provide, minimally, clear advice on the impact of options upon possible social goals; for example, whether the choice of one method of evaluation and choice of metric has consequences for different dimensions of fairness – access by socio-economic group, equity, distribution, severity of disease. Current evidence suggests this does not occur (Richardson and McKie, 2007). The conclusions here also suggests a more modest role for economics than the present self-appointed arbiters of social values, a role implied by J M Keynes when he famously argued that “if economists could manage to get themselves thought of as humble competent people, on a level with dentists, that would be splendid” (Keynes, 1989, p.332.)

More significantly, this problem suggests the need for a culture-specific research program to determine public values with respect to the possible objectives of health policy; whether ex ante or ex post utility is the preferred target; whether subjective well-being should be more explicitly included in measurements; whether utility, capabilities, health or some mix of these should be a metric; whether particular social groups should receive priority etc. This is the process we describe as “Empirical Ethics”. Importantly, it seeks to identify population values and to inform both policy makers and individual economic evaluators.

7 Conclusion

Can we say policy A is superior to policy B; that there is a “social preference” for A in some fundamental sense and that there is a particular metric that is universally and eternally correct in the sense that its maximisation indicates higher social welfare? The answer is probably “no”, as there is no uncontroversial interpretation of “superiority”, “social preference”, “social welfare”. The intuition that there should be such concepts is, as we have suggested, a manifestation of the “Sunday School effect” or social anthropomorphism, the belief that social groups should share values, world views, and so on. We have argued, in contrast, that intuitions in complex contexts are
The Impossibility of an Ideal Metric for Health Service Benefit Measurement

institutions should operate. Where different but defensible metrics exist that alter rankings, explicitly advantageous policy if policy options are increased and their consequences clarified and addressed explicitly. It is not the role of economists to dictate social values or to dictate how political institutions should operate. Where different but defensible metrics exist that alter rankings, the

The broad approach advocated here is not new. It follows the tradition of Bergson and Samuelson noted earlier in which social judgements are incorporated in a SWF which, in principle, could incorporate a range of ethical views. It is also consistent, though not uniquely so, with “William’s decision making school” (Sugden and Williams 1978) which seeks to limit the role of (health) economics to carrying out the wishes of decision makers. The additional element here – empirical ethics – is to advocate a program of empirical inquiry into population values to inform social debate, to evaluate the implications of these values as judged by alternative ethical criteria. There is, additionally, an important role for advocacy, and historically the ideas of economists have been highly influential. At the macroeconomic level, for example, advocates have suggested the replacement or supplementation of GDP as the dominant concept and metric by various other goals and their corresponding metrics. The initial suggestion by Nordhaus and Tobin (1972) for a measure of “net economic welfare”, including informal work, leisure and environmental damage, was followed by a suggested “Index of Sustainable Economic Welfare” (Cobb and Cobb 1994) which became the “Genuine Progress Indicator” (Hamilton and Denniss 2000). More recently the United Nations Development Program combined life expectancy, literacy and per capita GDP into its Human Development Index (United Nations Development Program 1990) and National Measures of Subjective Well-being have been developed (Cummins 2003). The OECD publishes 33 indicators of social well-being. Such metrics may or may not eventually prove to be useful but economists play a valuable role in creating such options for consideration. Our argument here is that the value of these metrics does not depend upon their theoretical “correctness” and, in particular, their derivation from a set of “self-evident axioms”. Rather they derive their value from their contribution to problem solving in a particular context.

Our argument is also in the tradition of moderate “empiricism”. Goals are informed by empirical research into problems and metrics designed pragmatically to assist with this, and the subsequent search for solutions. In contrast, the Arrovian Impossibility Theorem is in the tradition of the “a priori method”. Axioms – by implication, universally applicable axioms – are based on “introspection” and “simple and indisputable experience” (Robbins 1935, pp. 77-8), and the analytical focus is upon the construction of a universal edifice. The Bergson-Samuelson tradition is clearly more sympathetic to cultural and value differences and increases the probability of affecting incremental social improvement as defined within a polity. It is also more consistent with the branches of economics that focus on incremental improvement rather than the achievement of a global optimum.

We have argued that the role of economists is to generate rankings of options that may be accepted or rejected when evaluated within existing institutions. Suggestions will be filtered by any political system. The real world imperfections of the institutions do not invalidate our suggested rationale for evaluation studies under these conditions - viz, that there is an increased probability of advantageous policy if policy options are increased and their consequences clarified and addressed explicitly. It is not the role of economists to dictate social values or to dictate how political institutions should operate. Where different but defensible metrics exist that alter rankings, the
alternative results should be presented. None should be censored because “economic theory” demonstrates the superiority of the orthodox solution. Alerting policy makers and the public to the full set of options and underlying values provides more choice and this increases the likelihood of decisions which increase social welfare as defined within a particular polity.
References


The Impossibility of an Ideal Metric for Health Service Benefit Measurement


Richardson J (1994), Cost utility analysis: what should we measure, Social Science and Medicine, vol 39. No.1, 7-21


Richardson J, McKie J (2005), ‘Empiricism, ethics and orthodox economic theory: what is the appropriate basis for decision making in the health sector’, Social Science and Medicine, vol 60, pp 265-75.


