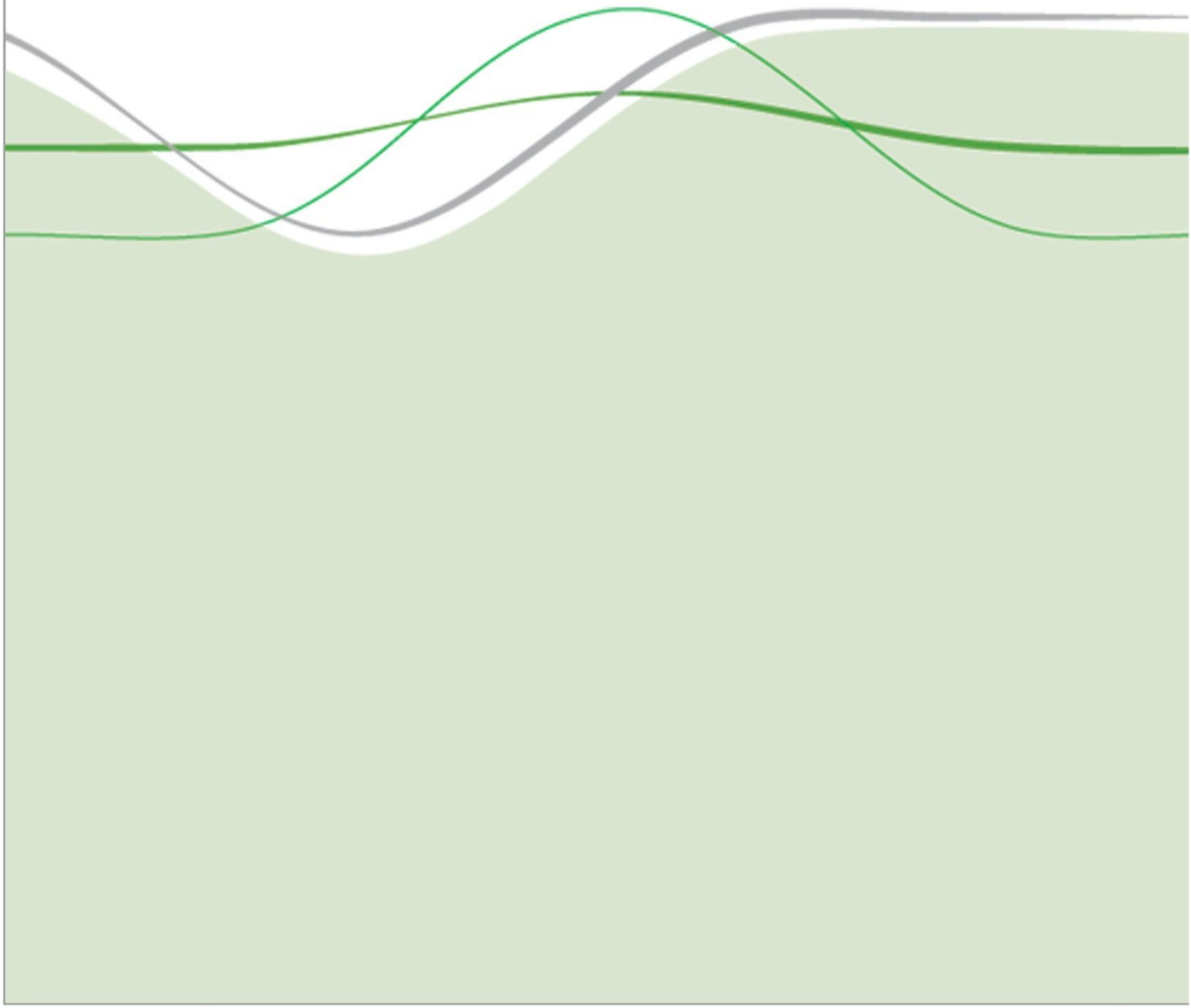
Office of Health Economics Research

Research Paper 17/01

What is the Normative Basis for Selecting the Measure of 'Average' Preferences for Use in Social Choices?

February 2017

Nancy Devlin, Koonal Shah and Ken Buckingham



What is the Normative Basis for Selecting the Measure of 'Average' Preferences for Use in Social Choices?

Nancy Devlin[1], Koonal Shah[1] and Ken Buckingham[2]

[1]Office of Health Economics

[2]Department of Preventive Medicine, University of Otago, New Zealand

February 2017

For further information please contact:

Professor Nancy Devlin

ndevlin@ohe.org

The Office of Health Economics
*(*a company limited by guarantee of registered number 09848965)

Southside, 105 Victoria Street

London SW1E 6QT

United Kingdom

Tel: +44 207 747 8858

About OHE Research Papers

OHE Research Papers are intended to provide information on and encourage discussion about a topic in advance of formal publication.

Any views expressed are those of the authors and do not necessarily reflect the views or approval of OHE, its Editorial or Research and Policy Committees, or its sponsors.

Once a version of the Research Paper's content is published in a peer reviewed journal, that supersedes the Research Paper and readers are invited to cite the published version in preference to the original version.

Acknowledgements

This paper is based in part on a study that was funded by a Department of Health Policy Research Programme grant (NIHR PRP 070/0073). Views expressed in the paper are those of the authors, and are not necessarily those of the Department of Health.

The authors are grateful to Tony Culyer, Paul Hansen and Gérard de Pouvourville for helpful comments received on an earlier draft.

Table of Contents

ABSTRACT

Value set studies for the EQ-5D aim to provide a single utility for each health state that reflects the *average* preferences of the general public. But what precisely do 'average preferences' mean in this context? There are a number of ways of combining numbers representing the utilities of individuals to achieve an 'average' reflection of society's preferences. EQ-5D valuation research typically relies on means, although the median has also been used. These reflect quite different things: the average of peoples' values, as opposed to the value of the average person. Still other approaches are possible, including the geometric mean, mode, or indeed any of the above taken together with various rules about the exclusion of outliers (in effect, a judgement about whose votes should count). Which approach to aggregation of individual preferences is chosen can have an important effect on conclusions about what 'society's' preferences are – with implications for decision making and the allocation of public funds. Which approach to calculating the average should be used is a *normative* question: it cannot be answered with recourse to empirical evidence alone. The choice of summary statistics is not merely a technical matter, but invokes ethical issues which need to be resolved.

The aim of this paper is to consider what normative arguments might exist for advocating the use of any given measure of the average in the context of health state values. We begin by providing examples of the importance and implications of the choice of the measure of central tendency in stated preference studies (including both EQ-5D values and corresponding issues in the willingness to pay literature). Then, drawing on the theory of social choice, voting models and welfare economics, we consider the criteria that are available for judging the 'goodness' of alternative approaches to aggregation, and evaluate their relevance to the selection of the measure of average EQ-5D values.

1. INTRODUCTION

Markets allocate resources based on individual preferences as reflected in consumers' observed willingness to pay (WTP) for goods. In many areas of the public sector such 'revealed preferences' data do not exist because there are no markets. This may be because the goods in question are public goods (defined as a good that it is not possible to exclude people from consuming once it has been produced (Culyer, 2010, pp.422-424)). Alternatively, they may be goods that the decision makers have decided to fund publicly for other market failure reasons or to meet social goals (e.g. the English NHS).

In the public sector, the allocation of resources between alternative projects – such as choices about investments in roads, schools, and spending on new health care technologies – is determined by non-market decision processes, nested within a political system. High-level decisions about which types of services should be funded or provided publicly, and about the level of taxation funding to be devoted to them, are influenced by the preferences of the general public as reflected in their choices at the ballot box. As Arrow (1970) notes, "The methods of voting and the market are both methods of amalgamating the tastes of many individuals in the making of social choices."

More detailed decisions about which specific services to fund from public sector budgets often entail some form of cost benefit analysis. In the absence of readily available information about revealed preferences, these approaches often rely on the use of 'stated preferences' methods to estimate the value of each option under consideration. Stated preferences methods encompass a variety of approaches, which have in common that values are obtained by presenting people with hypothetical choices and assuming that the choices they state they would make provide an accurate representation of their preferences over the available options. Value may be expressed in various ways – for example, in monetary terms by establishing people's WTP via discrete choice experiments or contingent valuation methods; or in utility index terms, such as the quality of life weights ('utilities') applied in the estimation of quality-adjusted life years (QALYs). The latter may be elicited using time trade-off (TTO), standard gamble, discrete choice experiments, or visual analogue scale (Brazier et al., 2007).

In cost benefit analysis, the procedure for aggregating these stated preferences is straightforward in principle: the *sum* of the WTP (or willingness to accept; WTA) of each affected person for a given change in health provides an estimate of the compensating or equivalent variation as a measure of the increase (or decrease) in utility. Welfare economics provides a normative framework for aggregating these measures of individuals' utility to make judgements about the corresponding change in social welfare. However, in practice, frequently it will be too costly to obtain such information from all affected parties for all options under consideration. Instead the WTP values of a *sample* of people are used to infer something about societal benefits more generally. Put more simply, sample data are used to estimate points on the market demand curve, thereby allowing estimates of welfare changes. For example, the value of a statistical life year figure routinely used in transport decision making in the UK is based on estimates of the WTP to reduce risk of death obtained from a sample of the general public, and the *average* of these values is then applied to the valuation of reduced mortality benefits across multiple decisions affecting different people (Dionne and Lanoie, 2004; Viscusi and Aldy, 2003).

A similar approach is used to establish health state values. For the estimation of QALYs, as used in cost effectiveness analysis, decision makers such as the UK's National

Institute for Health and Care Excellence (NICE), and similar bodies internationally, require a *single* value for each health state described by a standardised questionnaire such as the EQ-5D (NICE, 2013). These values are used to provide a single numeric summary of the health profile that patients self-report on the EQ-5D (Parkin et al., 2010). By convention, these values do not come from persons who are affected by ill health or are candidates for treatment (i.e. they are not measures of the *experienced* utility of patients), but rather are the values assigned to health states by members of the general public, asked to imagine what it would be like to experience them. Stated preference methods are used to generate these values based on the preferences of a representative sample of the general public, and these are used to estimate the *average* preferences of the general public for each state[1]. The values are then used in cost effectiveness analysis to assess the health benefits of technologies.

But what precisely do *'average preferences'* mean in this context? There are a number of ways in which we can combine numbers representing health state values, or WTP, to measure the 'central tendency' (defined as "the single value that is most typical/representative of the collected data" (Manikandan, 2011)) of society's preferences. The most obvious way is to use the arithmetic mean: add all the observed values together and divide by the number (n) of values. This approach is very widely used as the basis for analysis of WTP and health state values, and underpins the econometric approaches usually used to model 'value sets' for health states. But the rationale for doing so is rarely considered, and there are alternatives – perhaps most obviously the median (arrange n values in an ordered sequence and identify the middle or $(n/2)^{th}$ value in that sequence). There has been some interest in the use of the median as an alternative way of modelling health state values – for example, see Shaw et al. (2010). Considered more broadly, the arithmetic mean and the median are just two of a wider set of ways that individuals' preferences data could be aggregated or represented. For example, the geometric mean (multiply all the numbers together and take the n^{th} root of their values); the mode (identify the most commonly chosen option); or indeed any of the above taken together with rules regarding the exclusion of outliers.

2. EXAMPLES OF THE ISSUE AND IMPLICATIONS FOR DECISION MAKING

Which measure of central tendency is chosen can have an important effect on conclusions about what society's preferences are – with implications for decisions about the allocation of public funds. For example, the question of which measure to use has been noted – but not resolved – in the context of contingent valuation studies of the value of a statistical life year. The skewedness in distributions of participants' responses typical in such studies means there is often a substantial difference in the results suggested by mean and median responses. Whilst many studies have reported such results, few have considered the implications for social choices. A rare exception is Jones-Lee et al. (1985), who note that "the appropriate value to place upon the avoidance of one 'statistical' death (or, more succinctly, the 'value of statistical life') is given by the *population mean* of the relevant marginal rates of substitution" (p.51) (emphasis added), implying the arithmetic mean to be the correct way to represent average preferences. However, they also cite a UK Department of Transport report which

[1] Each individual's stated utility (value) for a given health state is itself a measure of ***average*** utility per period of time - a measure, from their point of view, of the average 'flow' of utility over a specified duration of utility, obtained (in the case of TTO), by convention, using a 10 year time horizon (Buckingham and Devlin 2009).

states that "the general principles of cost benefit analysis . . . would suggest that the Department should aim to find the amount that an *average individual* would be willing to pay" (Leitch Committee Report, 1977, p.104, cited by Jones-Lee et al., 1985), instead implying the median would be the appropriate measure of central tendency. The results reported by Jones-Lee et al (1985) highlight the implications of this choice: the mean value of a statistical life year of £1.43m derived from one of the questions can be compared to a median value of £500k. The authors note that:

> "One simply has to face the fact that setting the value of a statistical life at the level implied by the mean would result in a situation in which *a minority of individuals with very high marginal rates of substitution would be 'dragging along' an unwilling majority*. Furthermore, since it would be virtually impossible to identify individuals with high rates of substitution, there would be no way, even in principle, of arranging for compensating taxes or transfers. *This might therefore be a case in which efficiency ought, to a degree, to be sacrificed in the interests of democracy, with the value of a statistical life being set at the median value*" (p.70) (emphasis added).

Issues caused by 'extreme' responses were also encountered in research to establish the WTP for QALY gains. Donaldson (2011) notes that in a study by Baker et al. (2010), the influence of outlier responses was such that estimates of the social value of a QALY based on the mean ran into several millions of pounds, leading the study team to explore other ways of analysing the data, including the exclusion of outlier responses and the use of medians.

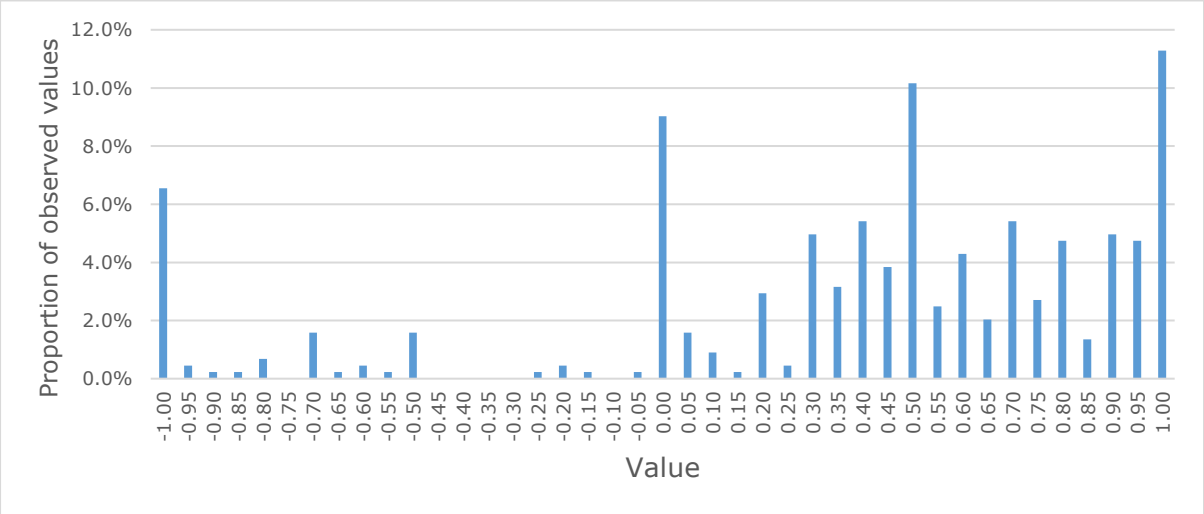
The same issues are also encountered in studies to elicit health state values. For example, Devlin et al. (2013) report an exploration of different variants of TTO task that might be used to elicit values for states 'worse than dead'. They find that, for a small but stubborn sub-sample of participants, extreme distaste regarding very severe states of health was such that their values approached negative infinity – clearly problematic where one wishes to calculate arithmetic means across the sample.

More generally, other sorts of extreme responses to the TTO task (such as non-trading behaviour – in effect, valuing all health states as equivalent to full health) are very common, and affect the distribution of responses which are observed. This means that the decision about how to characterise the 'average' preferences for EQ-5D states has important repercussions. For example, NICE decisions about new technologies currently rely on EQ-5D health state utilities based on means of TTO values elicited from a sample of approximately 3,000 members of the general public (Dolan, 1997). A different choice about how to represent the average of those values would have led to quite different estimates of QALYs and incremental cost effectiveness ratios, and – potentially in some cases – different decisions about whether to fund the technologies under evaluation.

A (fairly typical) example of a TTO value distribution is shown in Figure 1. This shows the distribution of values given to EQ-5D-5L health state 32442 by 443 UK respondents in a study by Shah et al. (2016). There is clustering at -1, 0, 0.5 and 1 – these four values account for 37% of all observations. The mean, median and mode values are 0.38, 0.50 and 1.00, respectively. The mean is driven downwards by a small group of respondents who gave an extremely low value of -1 to this health state. In fact, in the particular variant of TTO used in the Shah et al. (2016) study, the minimum value was bounded at -1 by design – it was not possible for respondents to trade any more time in order to provide even lower values. But the large number of respondents reporting a value of -1

is likely to contain some individuals who would have given a value lower than that if the valuation protocol had permitted them to do so. This issue is discussed further in section 7.

Figure 1. Value distribution of EQ-5D-5L health state 32442 in Shah et al. (2016) study



Empirically, there have been a number of attempts to model EQ-5D values based on medians – see, for example, Shaw et al. (2010) – and some statistical advantages are claimed of those models. However, which approach to calculating the average *should* be used is ultimately a *normative* question: it cannot be answered by recourse to empirical evidence alone (though empirical investigations can help address questions about the characteristics of the individuals giving extreme responses). The aphorism "society should be the master of mathematics, not its slave" is relevant here; as Nicholl (1989) notes, "the choice of summary statistics is not merely a technical matter for statisticians, but involves an ethical issue which should be resolved". A number of authors have pointed to the need for great care in the selection of central tendency measures. However, in spite of the considerable importance of this issue and its implications for decision making, we have been surprised to find almost nothing in the literature directly addressing this as a normative issue.

The aim of this paper is therefore to consider the normative arguments for preferring one measure of central tendency over another in the context of health state values. Drawing on the theory of social choice and welfare economics, we review the criteria available for judging the 'goodness' of alternative approaches to aggregating individual preferences, and evaluate their relevance to the selection of the measure of 'average' health state preferences.

3. WHAT CONSIDERATIONS ARE RELEVANT?

We take as our starting point the assumption that it is the general public's preferences for health that are relevant, rather than those of patients. While there is some debate about whether patients' values are relevant to the cost effectiveness analysis of health care technologies (e.g. Brazier et al 2009), the prevailing orthodox approach to health technology assessment uses 'value sets' obtained from the general public, on the

grounds that it is the view of public, as taxpayers (and therefore funders) and potential users of the health care system, that are relevant to decisions about health care.

The valuation of health states can be regarded as a voting system for preferences over those states. For example, just as the recent UK European Union membership referendum required voters to indicate which of two outcomes they preferred, pairwise choice tasks (as used in discrete choice experiments, for example), ask study participants which of two heath states they prefer. We should say, at the outset, that there is no perfect way of answering the question of how we use those votes to construct a measure of 'average' social preferences. Indeed the question itself is recursive. We cannot agree on a voting system without having a method of agreeing on a voting system. In more tangible terms, without a way of determining social preferences via a political system, we have no unchallengeable way of determining how individual preferences should be combined to arrive at a social preference.

We therefore need to construct, from first principles, the criteria to use in selecting the particular approach to aggregation. For example, as a starting point, and assuming that social choices should be informed by evidence of people's preferences, it might be argued that the method of calculating an average should obey some of the properties of the democratic systems within which publicly-funded health care systems are contained:

(i) It should respect the majority view.
(ii) It should give every person's preferences an equal weight in the final outcome.
(iii) It should ensure representativeness, and therefore minimise the number whose preferences are excluded from counting in the voting.

There are two elements of the issue of representativeness: *ex ante* choices about who is invited to take part in stated preference studies and *ex post* decisions about which of the data generated should be included in the analysis to inform decision making. With respect to the *ex ante* choice, we note that there are already some, relatively non-contentious, exclusions which are common to both political voting and stated preferences for EQ-5D, e.g. in both cases, there is an *ex ante* rule that only those older than 18 years of age are eligible to participate[2]. While those under the voting age may have preferences, the restriction reflects an apparently widely accepted judgement that those below a certain age are unable to form sufficiently rational or well informed views. Other sorts of exclusions are also common, but more contentious, such as whether the preferences of criminals should be excluded. Some countries, such as the UK, do not allow prisoners to vote (although elsewhere in Europe prisoners are allowed to vote). In the US, this is at the discretion of the state: 14 states permanently bar ex-felons from voting and 29 states prevent criminals from voting whilst on probation. Only two states follow the European approach of allowing all inmates and ex-convicts to vote. Similarly, it is common for samples recruited for stated preference studies to exclude those in institutions such as prisons (Shaw et al., 2010[3]).

Still more contentious, and highly relevant to the issues around constructing average preferences, is the extent to which those exclusion criteria should be extended to making *ex post* judgements about whether participants' responses to stated preferences tasks

[2] Also common to both democratic voting systems and EQ-5D valuation studies is that there is generally no upper age limit on whose preferences count.
[3] This is also the case in the provisional EQ-5D-5L value set for England study (Devlin et al., 2016) although there (as is probably the case in similar studies) this does not reflect a judgement about whether criminals' preferences should count, but rather is a reflection of the sampling approach used, which relies upon a random draw from residential addresses.

are 'valid' or 'rational'. Exclusion rules are sometimes used to handle cases where data are clearly of poor quality – such as where the time taken in responding to stated preference tasks is implausibly low, suggesting lack of engagement with the choices. However, some people's responses may simply be 'uncodeable' – people may respond in ways that are meaningful for them (truly reflect their values and preferences) but are outside the expectations and norms of researchers. For example, Devlin et al. (2004) encountered people who valued 'dead' at both 100 and 0 ('dead in heaven, dead in hell') in a visual analogue scale rating exercise, rendering their data unusable. The exclusion of such data is perhaps the procedural equivalent of discarding 'spoiled' votes at the ballot box.

Such exclusions are of concern, and may be difficult to defend, if preferences that may seem implausible to researchers are nevertheless consistent with some underlying beliefs – for example, a refusal to trade off any time for any health state in a TTO may be motivated by religious beliefs that 'life is precious' and that poor health is part of God's plan (Papadimitropoulos et al., 2015). The use of exclusion rules has a non-trivial effect on the value sets modelled from the data (Devlin et al., 2003). Since these exclusions tend to pertain to extreme values (outliers) in stated preferences data sets, judgements about them have an important influence on mean values (and the extent to which there is concordance between means and medians).

In general terms, minimising exclusions is important for the extent to which value sets can legitimately claim to provide a 'representative' average of society's preferences; while transparency about decisions made about exclusion criteria is required for accountability and to help others judge the legitimacy of the exclusions. Representativeness is also a consideration in the choice of the measure of that average, and the extent to which social preferences reflect information on the preferences of all members of the sample. For example, using the arithmetic mean arguably includes information on the utilities of all members of the sample, and is therefore influenced by any heterogeneity in the underlying preferences of diverse types of people. Decision making bodies, such as NICE, may consider that the legitimacy of its decision process, and its ability to defend those decisions to taxpayers, is better served by the mean on this basis. In contrast, the median is concerned just with the preferences of the 'average' person. If extreme values are particularly associated with a particular sub-group[4] (perhaps defined by ethnicity, religion, or age), then the exclusion of that group's vote on the 'average' value of health states might contravene equalities legislation that Governments are required to respect.

A further consideration is the extent to which any given measure of average value allows an approximation of the total changes in welfare associated with a given option or decision. As noted in section 1, welfare economics is concerned with the *sum* of WTP (or WTA) of affected persons, in order to determine whether society is, overall, better off as a result of that option (including both Pareto improvements and potential Pareto improvements). The arithmetic mean has the appeal that, multiplied by the number of persons, it provides an accurate prediction[5] of the sum of the effects. That would not be so for the median or mode (except in special cases). Pragmatic arguments might favour

[4]One example of this is the tendency for income to exert an effect on WTP responses. Another example might be the effect of religious beliefs (for example, about the sanctity of human life; or about the existence of an afterlife) on people's willingness to trade time in a TTO (Jakubczyk et al., 2016).

[5] The accuracy of the prediction is dependent on the extent to which the sample from which stated preferences are elicited is representative of the general public, both with respect to socio-demographic characteristics, and also with respect to their preferences.

the median, as we only need to locate the median preference, or indeed the mode, as we need only go so far as finding the most popular choice or most commonly expressed preference.

In the following sections, we consider further what guidance is available from politics, social choice and welfare economics, on the question of how to represent the average in stated preference studies.

4. WHAT GUIDANCE IS AVAILABLE FROM POLITICS?

We might consider whether there are lessons for health state valuations that might be learned from the political system within which the health care system is embedded. If that system at least satisfies the electorate as a means of making policy, it might also satisfy them in deciding on health state values. Most obviously, a democratic political system attempts to assign equal weights to all voters. It then combines choices by mode and choices by median. Political representatives are selected as the candidate receiving the largest single vote (the mode). The decisions made by those representatives are formally determined by acceptability to more than 50% of the representatives (the median). In practice, representatives receiving the modal vote from their constituencies will exert a stronger influence if they form a group which is itself a mode (the largest party usually takes the lead in forming the government).

The mode is simple to apply to qualitative options, as it does not rely on any additional numerical information and probably accounts for its use in the selection of members of parliament. The median can be used directly for qualitative choices between two options (in which case it degenerates to the mode). Where there are more than two options, further information is required to be able to sequence the choices (recall that the median involves identifying the central position in an ordered sequence). The arithmetic mean is well suited for numerical choices. In a political context we would need choices to be numerically quantified, for example by each individual casting a vote being asked to assign points to various candidates. Another example might be where decision makers assign points to various policies – as in option appraisal and multi-criteria decision analysis (Devlin and Sussex, 2011). Note that the number of points assigned by any individual cannot be infinite or indeed very large if it is not to swamp other responses.

An extensive political theory literature exists on the methods by which votes may be aggregated. For example, various rules exist by which preference ranking data can be aggregated, such as Borda counts and the Condercet method. However, those approaches are specific to the aggregation of *ordinal* information, where there is no information on the *strength* of the preferences. In contrast, a fundamental property of the outputs of health state valuation studies is that they are (or are intended to be) cardinal measures of value. For example, the methods for valuing health states on the 0-1 utility scale are intended to convey interval and ratio scale measurements. The aforementioned aggregation methods would fail to make use of the information that techniques such as WTP and TTO convey about the strength of people's preferences.

5. WHAT GUIDANCE IS PROVIDED BY WELFARE ECONOMICS AND EXTRA WELFARISM?

Welfare economics provides the normative foundations of economics, and is concerned with fundamental issues about how society's welfare might be measured and maximised. It typically has as its basis the assumption that individuals are the best judge of their

own welfare, and that society's welfare should be a reflection of the utilities of its members. However, where utility can only be measured in ordinal terms, there is a well-known problem with the aggregation of individual utilities. Arrow's Impossibility Theorem (AIP) concludes that where individual preferences are observable only in ordinal terms – and assuming we disallow the possibility of a 'dictator' – no voting system can yield a complete and transitive set of social preferences. Specifically, the criteria considered in AIP include:

(a) Unrestricted domain: all preferences of all voters must be allowed, and must yield a complete ranking of preferences
(b) Non-dictatorship: the preferences of any one individual cannot dominate all others
(c) Pareto efficiency: if every individual prefers an option to another, then that option should be preferred
(d) Independence of Irrelevant Alternatives

Indeed, beyond the early contributions of Mill and Bentham, who concerned themselves with the explicit weighing up of pleasures and pains, later developments in welfare economics and microeconomics, including the contributions of Pareto, Edgeworth, Jevons and Arrow, largely focussed on the *ordinal* treatment of preferences and the implications both for social choice and general equilibrium. This body of theory therefore has little to say on the specific issue that is the concern of this paper: modern welfare economics (as also noted above with respect to the various methods of voting 'counts') is largely concerned with the means of aggregating *ordinal* preferences data; whereas EQ-5D values are generally treated as having cardinal properties.

5.1. Extra welfarism

The estimation of QALYs and their use in cost effectiveness analysis rests on extra welfarism. Extra welfarism rejects utility as the (sole) basis for social choices and instead, following Sen's (1979) critiques of utility and welfarism, allows for non-utility information about individuals to be admitted into the process of comparing social states. Culyer (2012) summarises the key distinctions between welfarism and extra-welfarism as follows:

> 'The extra welfarist approach differs from the welfarist in four general ways: (1) it permits the use of outcomes other than utility (2) it permits the use of sources of valuation other than the affected individuals (3) It permits the weighting of outcomes (whether utility or other) according to principles that need not be preference based (4) It permits interpersonal comparisons of wellbeing in a variety of dimensions, thus enabling movements beyond Paretian economics. (Culyer, 2012, p.72).

The arguments set out in this paper have assumed that: (a) social choices and resource allocation decisions should be informed by evidence on preferences; and (b) some measure of the average of those preferences is required. However, the extra welfarist approach allows for consideration of the rationality, reasonableness, acceptability and heterogeneity of preferences. This is particularly pertinent when dealing with outliers and extreme values. It is our view that extra welfarists would be interested not only in the average value (which will usually mask any underlying heterogeneity) but also in the variance and skewness of the distribution of values. It may be acceptable within the extra welfarist account to place relatively less weight on extreme values based on

judgements that those values reflect preferences that should not be reflected in public decisions.

Extra welfarism, as it has been applied in the economics of health care, has come to be associated with an approach which focuses on 'health' as the principal maximand of the health care system, with health being measured in terms of QALYs. Extra welfarism therefore provides a relevant theoretical foundation for the use of the health state values produced by stated preference studies e.g. EQ-5D value sets based on TTO data.

We reviewed the literature on extra welfarism to check whether it offered any guidance on the selection of the appropriate measure of central tendency for health state values. Culyer (1989) arguably provides the seminal account of extra welfarism in health economics. That work touches only very briefly on the issue of where values come from and how they might be aggregated, in response to critiques of the stated preference approaches to the value of a statistical life year, vis:

> "Broome (1978) argued that the only appropriate value for a 'statistical life' was infinity on the grounds that, eventually, statistical probabilities of death (or of opportunities for life extensions not taken advantage of) translate into deaths of actual individuals who might reasonably be expected to exercise a veto. This clearly poses something of a challenge to Paretian welfarists though less of one, of course, to non-welfarists. A welfarist may find something of a defence in the reflection that he might himself agree to an option offering some benefit but with a very small prospect of its entailing his own death, so why should a society of like-minded folk feel differently? *An extra welfarist might take the view that she would be guided by the majority view on the value of (or differential values of) life".* (p.54). (emphasis added)

An extra welfarist might also wish to be guided by the majority view as to *whose* preferences should count (Culyer, A.J., personal communication, 2 July 2015).

We noted earlier the point that the methods of applied welfare economics aim to identify potential Pareto improvements i.e. where the sum of the gains is greater than the sum of the losses, such that (in principle) compensation could take place. That logic applies to the way WTP is assessed in stated preference studies, since WTP/WTA is intended as a measure (or at least a proxy) of the corresponding total compensating variation or equivalent variation. However, that same logic does not necessarily apply to health state valuations being applied to cost effectiveness analyses within an extra welfarist framework. The purpose is not to identify net welfare-increasing projects by establishing whether compensation could be made in principle. Rather, it is simply to establish a reasonable common denominator for assessing health gains against health losses, the latter being defined as the opportunity costs that are unavoidable, given the assumption of a fixed health care budget. The implications of this for the choice of the way average weights are constructed are not obvious.

It is worth noting that extra welfarism is also much more permissive than welfare economics about the sources of utility. While in welfare economics the affected group of individuals is the primary source of valuation, in extra welfarism, "any number of stakeholders might be regarded as the appropriate source of different values" (Culyer, 2012). Sources of values might appropriately come from "an authority (decision makers, wise women, the general public, an elected or appointed committee, a citizen's jury, or some other organ)" and whilst "economists may be able to derive values from experimental groups or samples of the relevant population through modern methods for

eliciting preferences . . . the choice about which groups to sample are not normally for the analyst to make but for the ultimate decision maker, advised by the analyst" (Culyer, 2012, p.77).

While this does not offer any specific guidance on how to aggregate values, it does serve as a useful reminder of the role of elicited preferences in extra welfarism, and the paramountcy of the decision maker. Alan Williams' early work (albeit on cost benefit analysis) emphasised what he called 'the decision makers' approach'. Under this pragmatic approach, it is possible for values (such as weights) to come from the decision makers themselves ('postulated values').

> 'You cannot ascribe values without making value judgements. Market prices are acceptable if the value premises underlying market behaviour are acceptable… Postulated values are acceptable if it is believed that the value premises underlying both market and imputed prices are misconceived for the purpose in hand (e.g. if one accepts the propriety of a paternalistic or collectivist basis for valuation'). (Williams, 1972, cited by Sugden, 2008, pp.7-8).

Under this perspective, the emphasis is on the role of the decision maker in making judgements on behalf of 'the citizenry as a whole'. Sugden (2008) notes that, in William's later work, "much more than in his earlier work on CBA . . . Alan the health economist wants to draw the citizenry into the decision making process. Much of his work in health economics was concerned with eliciting, from representative samples of the general public, citizen-perspective judgements about marginal trade-offs, on the grounds that this is relevant evidence for decision makers". However, "Alan's decision maker appears to be reserving to himself the final decision about what to do in the light of the results. *It seems that the survey of citizen judgements is intended to inform the decision, not to make it*". (Sugden, 2008, p.17) (emphasis added)

While there is mention of 'collective averages' (Culyer, 2012) and of 'median trade off rates from the representative sample of the population' (Williams, 2003, cited by Sugden, 2008), the theory of extra welfarism offers no specific guidance on the way decision makers should aggregate the information collected in stated preference surveys – other than to remind us that it is the decision makers themselves (*not* economists) who should make that decision!

6. WHAT GUIDANCE IS AVAILABLE FROM MATHEMATICS?

We suggested in section 2 that mathematics should be the servant of policy rather than the master; however, it is well to remind ourselves what numbers can do. The information that a number can embody is hierarchical and the nature of a number can constrain what sort of measure of central tendency is possible.

Numbers can categorise phenomena; as in: call the red team '1', call the blue team '2' and call the yellow team '3' (categorical). Numbers can be used to arrange phenomena in order; as in: red team took position 1, blue team took position 2 and yellow team took position 3 (ordinal). Numbers can be used to quantifying the size of phenomena; as in red team scored 5, blue team scored 3, yellow team scored 1 (cardinal). We can refer to this as a hierarchy because in moving from categorising to sequencing to quantifying involves an increase in the information that the numbers contain. Sequencing also contains sufficient information to allow categorising. Quantifying also contains sufficient information to allow categorising and sequencing.

Categorical data can only be used to determine a mode. Ordinal data can be used to determine a mode and a median. Cardinal data can be used to determine a mode, a median and a mean. Because it requires all the information that is contained within cardinal data, the mean embodies more information than the mode or the median. There is an argument which says that, if you have a statistic with which you are fully confident, you should make full use of all the information that it contains. The mean incorporates the 'weights' that quantify the phenomena of interest. It represents it in a way that other measures of central tendency do not. Accordingly, there is a mathematical argument which says that, if we wish to make the maximum use of the data we have, we should calculate the mean if we have cardinal data. However, it is still our decision as to whether that information is useful for our purposes.

We should note that infinity is not a number. If we have a data set that includes infinity, we have a categorical data set rather than a cardinal data set. The choice of the measure of central tendency is limited to the mode and the median. The median utilises more information from the data set than the mode and might therefore be preferred. This does not imply that, if the data contains observations that include infinite values, we should necessarily change the data, or the way it is collected to rid ourselves of those awkward observations. There is a choice to be made: we might decide that the infinite values are necessary and restrict our choice of the measure of central tendency accordingly. Alternatively, we might decide that the infinite values are simply artefacts of the way the data are collected and consider collecting the information in a different way.

7. THE NATURE OF HEALTH STATE VALUES

We have already noted that the valuation of EQ-5D health states is based on stated preference methods – that is, people's responses to hypothetical rather than real choices.

There are two ways of thinking about the preference data yielded by such exercises. One view is that people have reasonably well formed, comprehensive, pre-existing preferences; and that, provided they are well designed, stated preference exercises are capable of tapping into and retrieving this information. Fischoff (1991) refers to this as 'the philosophy of articulated values'. In contrast, the 'philosophy of basic values' suggests that people lack clearly formulated preferences for all but the most familiar of evaluation questions. In responding to stated preference tasks, individuals are therefore engaging in a 'mental production process' to create their response. Preferences are 'constructed' in response to a particular choice or decision problem which is presented (Jones-Lee et al., 1995).

The possibility that preferences are 'constructed' has particular relevance to the valuation of generic health states. For example, in valuing EQ-5D states, participants are being asked to 'think' about health in a highly abstract way that is likely to be unfamiliar to them; and to imagine what it would be like to experience states described in a generic manner. For many participants, EQ-5D valuation tasks will be acting to help participants to *create* their values for these unfamiliar concepts, not simply *eliciting* them. This view of health state valuation may help to explain the differences between the values yielded by different methods. It further suggests that the values we obtain in such studies are *in part* a reflection of something 'real' about people's underlying preferences, and in part a reaction to the particular way questions are asked, via framing effects or decision heuristics adopted by participants.

This in turn is relevant to the way we interpret 'extreme' values. For example, Devlin et al. (2013) observed a sub-sample of participants providing values for very severe health states (using 'extended' TTO tasks) that approached minus infinity. The authors speculate that these responses may arise because the participant is simply trying to express a (qualitative) view that the state 'is very bad indeed'; or that their previous responses had 'boxed them into a corner'.

Similarly, other characteristics of distribution of EQ-5D valuation data may be understood as a direct consequence of the way preferences have been elicited. For example, in the research protocol used by the EuroQol Group (Oppe et al., 2014) the particular TTO approach used (a 'composite' TTO, comprising use of both 'conventional' TTO and lead time TTO) generates values between -1 and 1, hence the minimum value is -1. Where participants' responses indicate a value of -1, we know their value is *at most* -1, since that observation is censored. This suggests that using these values to establish average preferences needs to take account of this censoring in some way (see Feng et al., 2016).

Other characteristics of valuation data also reflect the constructed nature of the data. For example in the England EQ-5D-5L value set study, participants were far more likely to value a state at 0, 0.5 or 1, rather than any intermediate number; there were very few values in the range -0.5 to 0 (Mulhern et al., 2013). These discontinuities suggest participants are providing quite crude responses – indeed some participants valued all health states encountered using either a single high value (such as 1) or a single low value (such as 0), possibly providing a broad signal that states are either 'good' or 'bad', rather than providing preferences that have any cardinal meaning. The presence of significant interviewer effects (Mulhern et al., 2013) gives further reason to doubt the straightforward interpretation of valuation data as 'true' representations of individuals' preferences about health states. Put another way: we have reason to suspect that many of those participating in such tasks do not understand them, are not fully engaging with the tasks at hand, or simply want to finish the interview as quickly as possible; and that even the preferences data provided by those who *do* understand and are engaged are inevitably coloured by the framing of the tasks and the interviewer's approach.

These problems with data quality may have a bearing on the way we aggregate preferences data, with respect to the treatment of extremes; the way 'crude' data are best modelled; and whether some data are of such poor quality (e.g. logically inconsistent or 'unexpected' responses; tasks being completed in implausibly short times) that they should be excluded from value sets on the grounds that they do not represent the considered views of 'informed citizenry'.

This suggests a further principle is justified in aggregating individuals' preferences: that the final value set which it produces should show that a better (objectively judged or logically defined) health state is preferred to a worse (objectively judged or logically defined) health state. With respect to EQ-5D-5L, for example, the aggregate value for the worst state defined by it, 55555, should be lower than that of any other state. From the 3,125 states there are a number of pairs of states which can be logically ordered with respect to the dimensions and levels of the descriptive system – e.g. 13254 is worse than 12154 (it is worse on two dimensions, and no better on the other three dimensions). Any measure of central tendency (and any set of exclusion criteria used in conjunction with that) needs to be able to discriminate between these logically better and worse states in order to be of use to decision makers e.g. in cost effectiveness analysis. While this may seem self-evident, such a criterion may lead to rejecting the

mode and maybe even the median as the best representation of average preferences for EQ-5D-5L. The mode is probably 1, 0 or -1 for most states – and bi- (or multi-) modality is also common in valuation data. The median probably suffers from the same problem: many EQ-5D-5L health states will have the same median value. In the Shah et al. (2016) study on which Figure 1 was based, for example, four EQ-5D-5L health states – 11111, 21111, 11121 and 11112 – all shared a median (and mode) value of 1.

Let us briefly consider how the principal valuation procedures used by the EuroQol Group measure up to these criteria.

7.1. The time trade-off (TTO)

The TTO for states better than dead effectively gives everyone a time budget to 'spend' in exchange for an improvement in health. Conventionally, the individual values derived in this way are averaged using the arithmetic mean. For normally distributed preferences the method would reflect the majority view. The fact that all people are given an equal amount of time to spend might imply that all preferences can legitimately carry an equal weight. Usually we try to minimise the people excluded, although practice varies across previous studies (Szende et al., 2007; Engel et al., 2016) and there is currently no EuroQol Group consensus on or guidance this issue.

Various alternative TTO variants have been proposed to elicit values for states worse than dead (Oppe et al., 2016). The method used in the Measurement and Valuation of Health study (Dolan, 1997) was as follows. The respondent is asked to choose between immediate death, and spending a length of time (10-x) years in health state H_i, followed by x years in full health. x is varied until the participant is indifferent between the two options. The value of H_i is given by $U(H_i) = -x/(10-x)$. Remembering that these respondents consider the state to be worse than dead, the respondent would prefer to be dead than to endure the health state. The worse they consider the poor health to be, the more they must be compensated for enduring it, with increasing amounts of full health. Looked at from the opposite point of view, the price they pay for a life (of unspecified duration) is increasing amounts of time in poor health or decreasing amounts of time in good health. This resulted in a lack of consistency in what was being valued. Essentially the amount of time in poor health was allowed to take different values between individuals. At the extreme, the method allowed respondents to decline to accept *any* time in poor health, implying a value of negative infinity as their valuation of poor health. The impact of this method is to allow the extreme preferences of some respondents to completely overwhelm the preference of those with less extreme values (no amount of finite positive values can outweigh negative infinity).

It is at this point that choice of numerical methodologies becomes particularly important. (We may note in passing, that the same issue arises in the political context. In a democracy, we cannot allow an individual with infinitely strong preferences, save for their instatement as monarch or dictator, to outweigh the preferences of the majority.) The arithmetic mean is undefined over a set that includes infinite numbers, under which circumstances even if we wished to use the arithmetic mean to represent the view of society we should not.

7.1.1. The lead time TTO

The lead time TTO is a variant of the TTO that has a different way of dealing with states worse than dead. As with the TTO for states better than dead, everyone is given the same 'time budget' to spend on health improvements. However, in this case, the time

budget *exceeds* the time in poor health that is being valued. This allows individuals to trade more time than the time spent in the health state being valued and effectively allows them to place a value on states that are worse than being dead.

In this case the question arises of what to do with people who would trade off all the time that they are allowed. Arguably these people have not been able to achieve a point of indifference and hence the value of the health state has not been determined. From an economist's point of view, we might increase the amount of time that they would be allowed to trade until they have sufficient to achieve an equilibrium. Unfortunately this gives rise to some problems. In effect, this puts more weight on the values of the people whose time budget is extended compared to those whose budget is not extended. Furthermore, we have determined experimentally (Devlin et al., 2013) that some people would not accept poor health regardless of how much extra time in full health they were given to trade. Such people hold, it appears, preferences that imply a value of negative infinity for very poor health states. This gives rise to the same issues as discussed in the preceding section, i.e. the impossibility of deriving a mean from a set of values including infinity. For example, in the case of the EuroQol Group's international protocol for valuing EQ-5D-5L (Oppe et al., 2014) , the use of the lead time TTO to elicit negative values entails respondents being given a time budget of b = 20 years to value a health state lasting 10 years, such that the minimum value assigned is (t-b)/t = (10-20)/10 = -1. This satisfies the democratic criterion that all 'voters' should be treated equally – but contradicts the economist's model that valuations are derived when the respondent is in equilibrium between two options. The EQ-5D-5L value set for England finds a neat compromise between these issues: the amount of available trading time is the same for all respondents, but the responses of those who provide a value of -1 are treated as 'censored', allowing us to model the possibility of a distribution of (non-infinite) values below the minimum (Feng et al., 2016).

7.2. Conjoint analysis

Conjoint analysis comprises quite a large family of techniques, with much of health care valuation based on pairwise choices tasks forming part of a discrete choice experiment. Preferences are calculated by examining the choices which are made. Strength of preference is inferred from consistency of choices. However the method fails where it might be at its strongest. Where individuals always select based on a single criterion (the case known as lexicographic preference) no scale of preference between alternatives is calculable. In analysing consistency in response and in restricting choices to categorical choices, conjoint analysis bears some similarity to an analysis of modal responses. Valuations can be derived by examining consistency across a sample, in which case there is no further need to aggregate values over the sample. Alternatively, valuations can be estimated at the level of the individual, looking for consistency within respondents' answers. In either case, the issue of whether all voters carry an equal weight does not appear to be a problem, except in so far as there may be exclusions, either of those with strong lexicographic preferences, or of those who may find the questions too difficult to answer and accordingly self-exclude.

7.2.1. Extreme values and conjoint analysis

In a two party system we might infer strong preference for party A over party B if party A were to receive 100% of the vote. However, we might be wrong to interpret consensus as strength of preference. If the two parties were distinguished only by party A being

expected to grow the economy by 5.1% compared with 5% for party B, voters might have a uniform but weak preference for party A.

The theoretical justification for conjoint analysis is that, where people have weak preferences over alternatives, they will sometimes report preference for choice A and sometimes report preference for choice B. The argument is extended to groups and inverted such that, if a group is equally divided between choice A and choice B, they have no strong preference between them. However, we cannot discount the possibility that a group may be equally divided between opposing strong opinions. Conjoint analysis as applied in health state valuation can be interpreted as a complex voting system with consensus being used to imply strength of preference. It is fundamentally a modal form of processing.

By asking only categorical questions (of the form 'choose A or B') the problems that arise from extremes in strength of preference are side stepped.

Although some have argued that pairwise comparisons offer a simple binary choice, care needs to be taken to ensure that the binary choice is comprehensible. This may not be the case where respondents are asked to consider scenarios each comprising five or more dimensions being compared at multiple levels. In the latter case the question of how to determine a representative value remains to be addressed.

8. CONCLUSIONS

Table 1 summarises the criteria identified in the previous sections, and the extent to which alternative measures of central tendency satisfy these with respect to health state values.

We summarise our main conclusions as follows. First, despite this being an issue with considerable importance to the way health state values (e.g. EQ-5D value sets) are presented and used in decision making, the normative basis for selecting the measure of average preferences has received almost no attention in the literature to date.

Second, existing theory offers very little guidance on this issue. Welfare economics is concerned with the aggregation issues in ordinal preferences. Its application (e.g. in cost benefit analysis) frequently does, of course, in practice, make use of estimates of WTP from stated preference surveys, and skewed distributions of WTP that are commonly observed result in important differences between means and medians. Where the aim is to establish potential Pareto improvements by proxying the sum of the compensating or equivalent variations of those affected by the options under consideration, then in that context the arithmetic mean has the important advantage that (by definition), multiplied by the number of people affected, it provides the total change in welfare.

However, there appears to be no clear normative rationale for the current use of the arithmetic mean to represent average preferences in the context of health state valuation. The literature on extra welfarism – the theoretical foundation for cost effectiveness analysis in health care – does not appear to address the issue of which approach to average preferences should be used in the estimation of QALYs. Here, the role of health state values is to provide a consistent basis for estimating changes in QALYs, so that improvements in health from new technologies can be weighed up alongside opportunity costs, given fixed budgets.

Table 1. Normative criteria for selecting a measure of average preferences, and characteristics of the three principal approaches

Normative criteria	*Arithmetic mean*	*Median*	*Mode*
(i) Respects the majority view	*No – extreme values held by a few people have a big effect on the mean*	*No – at the median value, by definition 50% people have a value higher than that*	*Not necessarily – the mode tells us the value that most people opted for, but this might not be the majority of people, and bi-modality may be an issue*
(ii) Each person's preferences carry an equal weight	*Yes, in the sense that everyone's 'vote' counts – but extreme preferences by a small number of people can drive results*	*No – the focus is only on the preferences of the average person*	*Yes*
(iii) Representativeness – exclusions are minimised	*No – because of the problem of 'extreme' values, exclusion rules are often applied*	*Yes – extreme values and censored data are less problematic*	*Yes*
(iv) Differentiates between logically better and worse states	*Yes – although in part simply because the mean can take more unique values than the median or mode (which are constrained to the minimum trading units in the way in which values are elicited)*	*Possibly – although the median is likely to be considerably 'blunter' than the mean, and in practice may lead to many health states having the same values*	*No – for example, the modal value is often 1 or 0 for many EQ-5D states*

In the case of 'well behaved' distributions of preference scores, we see no problem using the arithmetic mean as a representative value for social preferences. The case of less well behaved distributions – such as distributions characterised by 'spikes' and multi-modality which are common in health state valuation – is, however, more problematic.

Where some extreme preferences preclude the use of the mean, and where bi-modality is not a problem, the mode may be regarded as a valid measure of central tendency for health state valuation. However, where bi-modality does exist, it can present a problem for health state valuation as it can for democratic processes more generally. (Bi-modality in the political system, where it is associated with very strongly held beliefs, can present the most intractable of problems, as in the case of ethnically or religiously-divided societies. See McHugh et al. (2015) for an example of a study in which a small number of strong, opposing societal viewpoints are identified in the context of health care priority setting – the authors of that study warn against reporting the average preference or adopting majoritarian decision rules). Accordingly it may be that median values provide a reasonable compromise in determining the representative view where extremes preclude the use of arithmetic means.

Finally, given: (a) that as Alan Williams' early work reminds us, it is the *decision maker* who has the responsibility to decide on what values are legitimate to use in health care decisions; and (b) the apparent lack of any clear normative grounds for favouring any one approach to aggregation over another; arguably the role of health economists is not

to prescribe any one approach, but rather to take responsibility for providing decision makers with as much information as possible from stated preference studies. This might include generating a suite of value sets based on alternative aggregation approaches, alongside information on the nature of the distribution of values surrounding each measure of central tendency. It also suggests complete transparency with respect to researcher judgements in data handling, such as the reasons for excluding data points. Researchers have a responsibility to highlight the sensitivity of values – and therefore the social choices based upon them – to the selection of aggregation methods.

REFERENCES

Arrow, K.J., 1970. *Social choices and individual values.* Yale University Press.

Baker, R., Bateman, I., Donaldson, C., Jones-Lee, M., Lancsar, E., Loomes, G., Mason, H., Odejar, M., Pinto Prades, J.L., Robinson, A., Ryan, M., Shackley, P., Smith, R., Sugden, R. and Wildman, J., 2010. Weighting and valuing quality-adjusted life-years using stated preference methods: preliminary results from the Social Value of a QALY Project. *Health Technology Assessment*, 14(27).

Brazier, J., Dixon, S. and Ratcliffe, J., 2009. The role of patient preferences in CEA: a conflict of values? *Pharmacoeconomics*, 27(9), pp.705-12.

Brazier, J., Ratcliffe, J., Salomon, J.A. and Tsuchiya, A., 2007. *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University Press.

Buckingham, K. and Devlin, N., 2009. An exploration of the marginal utility of time in health. *Social Science & Medicine*, 68, pp.362-367.

Culyer, A.J., 1979. The normative economics of health care finance and provision. *Oxford Review of Economic Policy*, 5(1), pp.34-58.

Culyer, A.J., 2010. *The dictionary of health economics, second edition*. Cheltenham: Edward Elgar Publishing.

Culyer, A., 2012. Extra welfarism. Chapter 2 in: Cookson, R. and Claxton, K. (eds.) *The humble economist.* London: Office of Health Economics.

Devlin, N., Buckingham, K., Shah, K., Tsuchiya, A., Tilling, C., Wilkinson, G. and van Hout, B. (2013), A comparison of alternative variants of the lead and lag time TTO. *Health Economics*, 22(5), pp.517-532.

Devlin, N., Hansen, P., Kind, P. and Williams, A., 2003. Logical inconsistencies in survey respondents' health state valuations – a methodological challenge for estimating social tariffs. *Health Economics*, 12(7), pp.529-544.

Devlin, N., Hansen, P., Selai, C. (2004) Understanding health state valuations: a qualitative analysis of respondents' comments. Quality of Life Research 13(7), pp.1265-77.

Devlin, N., Shah, K., Feng, Y., Muhern, B. and van Hout, B., 2016. *Valuing health-related quality of life: an EQ-5D-5L value set for England*. Research Paper. London: Office of Health Economics.

Devlin, N. and Sussex, J., 2011. Incorporating multiple criteria in HTA: Methods and processes. Monograph. London: Office of Health Economics.

Dionne, G. and Lanoie, P., 2004. Public choice about the value of a statistical life for cost-benefit analyses: The case of road safety. *Journal of Transport Economics and Policy*, 38(2), pp.247-274.

Dolan, P., 1997. Modeling valuations for EuroQol health states. *Medical Care*, 35(11), pp.1095-1108.

Donaldson, C., 2011. *Willingness to pay and publicly funded health care: contradictions in terms?* Seminar Briefing. London: Office of Health Economics.

Engel, L., Bansback, N., Bryan, S., Doyle-Waters, M.M. and Whitehurst, D.G., 2016. Exclusion Criteria in National Health State Valuation Studies A Systematic Review. *Medical Decision Making*, 36(7), pp.798-810.

Feng, Y., Devlin, N., Shah, K., Mulhern, B. and van Hout, B., 2016. *New methods for modelling EQ-5D-5L value sets: an application to English data*. Research Paper. London: Office of Health Economics.

Fischoff, B., 1991. Value elicitation: is there anything in there? *American Psychologist*, 46, pp.835-847.

Jakubczyk, M., Golicki, D. and Niewada, M., 2016. The impact of a belief in life after death on health-state preferences: True difference or artifact? *Quality of Life Research*, 25, pp.2997–3008

Jones-Lee, M., Hammerton, M. and Philips, P.R., 1985. The value of safety: results of a national sample survey. *The Economic Journal*, 95(377), pp.49-72.

Jones-Lee, M., Loomes, G. and Robinson, A., 1995. Why did two theoretically equivalent methods produce two very different values? In: Schwab Christe, N.G. and Soguel, N.C. (eds.) *Contingent valuation of transport safety and the value of life*. Springer.

Manikandan, S., 2011. Measures of central tendency: the mean. *Journal of Pharmacology and Pharmacotherapeutics*, 2(2), pp.140–142.

McHugh, N., Baker, R.M., Mason, H., Williamson, L., van Exel, J., Deogaonkar, R., Collins, M. and Donaldson, C., 2015. Extending life for people with a terminal illness: a moral right and an expensive death? Exploring societal perspectives. *BMC Medical Ethics*, 16(14).

NICE, 2008. *Guide to the methods of technology appraisal 2013*. London: National Institute of Health and Care Excellence.

Nicholl, J.P., 1989. *Conceptual and technical considerations about using medians to evaluate the outcome of pragmatic clinical trials*. Paper presented at ISCB-10, Maastricht.

Oppe, M., Devlin, N., van Hout, B., Krabbe, P. and de Charro, F., 2014. A programme of methodological research to arrive at a new international EQ-5D-5L valuation protocol. *Value in Health*, 17, pp.445-453.

Oppe, M., Rand-Hendriksen, K., Shah, K., Ramos-Goñi, J.M., Luo, N., 2016. EuroQol protocols for time trade-off valuation of health outcomes. *Pharmacoeconomics*, 34, pp.993-1004.

Papadimitropoulos, M., ElBarazi, I., Blair, I., Katsaiti, M.S., Shah, K.K. and Devlin, N., 2015. An investigation of the feasibility and cultural appropriateness of stated preference methods to generate health state values in the United Arab Emirates. *Value in Health Regional Issues*, 7C, pp.34-41.

Parkin, D., Rice, N. and Devlin, N. (2010) Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? *Medical Decision Making*, 30(5), pp.556-565.

Shah, K., Mulhern, B., Longworth, L. and Janssen, M.F., 2016. An empirical study of two alternative comparators for use in time trade-off studies. *Value in Health*, 19, pp.53-59.

Shaw, J.W., Pickard, A.S., Yu, S., Chen, S., Iannacchione, V.G., Johnson, J.A. and Coons, S.J., 2010. A median model for predicting United States population-based EQ-5D health state preferences. *Value in Health* 13(2), pp.278-288.

Szende, A., Oppe, M. and Devlin, N., 2007. EQ-*5D valuation sets: an inventory, comparative review and users' guide*. Springer.

Sugden, R., 2008. Citizens, consumers and clients: Alan Williams and political economy of cost benefit analysis. Chapter 2 in: Mason, A. and Towse, A. (eds.) *The ideas and influence of Alan Williams*. Radcliffe Press/Office of Health Economics.

Viscusi, W.K. and Aldy, J.E., 2003. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty*, 27(1), pp.5-76.