Estimating a WTP-based value of a QALY: the 'chained' approach

Angela Robinson[1], Dorte Gyrd Hansen[2], Philomena Bacon[1], Rachel Baker[3], Mark Pennington[4], Cam Donaldson[3] and the EuroVaQ Team[5]

1. University of East Anglia; 2. University of Southern Denmark; 3. Glasgow Caledonian University; 4. London School of Hygiene and Tropical Medicine; 5. Comprising 37 members across 10 countries

Paper presented to 3rd Joint CES/HESG Meeting, Aix en Provence, 11-13th January 2012.

1. INTRODUCTION

The question of the value to place on a quality adjusted life year (QALY) has existed for some time in the health economics literature, and has been subject to considerable conceptual debate since some authors sought to make the link between cost-effectiveness analysis (CEA) and cost-benefit analysis (CBA) (Phelps and Mushlin, 1991; Johannesson, 1995; Bleichrodt and Quiggin,1999; Dolan and Edlin, 2002). More recently, the same question has come to the fore in policy circles through the creation of health technology assessment (HTA) agencies around the world. In making one-off recommendations about adoption or otherwise of an evaluated therapy, such agencies are, in effect, placing a monetary value on health gains. This has spawned several innovative attempts at empirical estimation of the value a QALY (Gyrd Hansen, 2003; Pinto-Prades et al., 2009; Baker et al., 2010; Shiroiwa et al., 2010). The data reported here arise from another empirical study, the EuroVaQ (European Value of a QALY) Project. The part of EuroVaQ reported here took place across nine European countries, with internet-based surveys of members of the public in each country in order to elicit a monetary value of a QALY. In many of these countries, health technology assessment agencies or health ministries use the QALY as a health benefit metric, and, thus, must consider what value to place on it.

The valuation procedures used in EuroVaQ extend the work of Baker et al. (2010) and Pinto Prades et al. (2009), who used a 'chaining' procedure, whereby a willingness-to-pay (WTP)-based value of a QALY is estimated by combining respondents' answers to standard gamble (SG) and WTP questions. A major result, common to these studies was that, for a significant number of respondents, the value of a QALY resulting from combining their SG and WTP responses at the individual level, were extremely high, thus rendering any mean value implausible in policy terms. The aim of this paper is to report on the procedure devised within EuroVaQ in an attempt to address these issues.

The remainder of the paper is organised in five further sections as follows. We begin with a brief review of the literature on survey-based approaches to valuing health including a brief review of the results of the most relevant studies published to date. The survey design and methods used to estimate a value of QALY in the EurovaQ 'chained study' are then outlined along with a set of hypotheses tested, before presenting results and discussing the contributions of the EuroVaQ study and any outstanding issues still to be addressed.

2. LITERATURE REVIEW

2.1 Valuing QALYs

Through the 1990s, development of national-level HTA agencies, led to calls for monetary values of a QALY to aid decision making at a national level (Johannesson, 1995; Garber and Phelps, 1997). For example, in the UK, there has been significant debate about the empirical basis of the cost-per-QALY threshold above/below which the National Institute for Health and Clinical Excellence (NICE) would recommend rejection/adoption of a therapy by the National Health Service (NHS). The proceedings of the recent House of Commons Health Committee (in 2007) criticise the current NICE threshold on the bases that it "…is not based on empirical research and is not directly related to the NHS budget, nor is it at the same level as that used by PCTs [primary care trusts] in providing treatments not assessed by NICE."

Some estimates have been made of the value of a QALY based either on modelling approaches or survey methods. Modelling studies have been reviewed elsewhere and values of a QALY vary greatly depending on how the data are manipulated (Mason et al., 2009). Survey work on the value of a QALY has been limited. Typically, individuals have been asked about their WTP for health gains for which utility values exist (e.g. using EQ5D population tariffs); and uncertainty has not been fully accounted for (i.e. by presenting scenarios involving certain gains in quality of life) (Gyrd Hansen, 2003; Shiroiwa et al., 2010). Respondents have also been asked to assume they are suffering from a relatively serious illness (Shiroiwa, 2010), and, in some cases, values have been elicited from patients and not from members of the general public (Byrne et al., 2005; King et al., 2005).

In terms of methods used, the closest studies to EuroVaQ are those of Pinto Prades et al. (2009) and Baker et al. (2010) which, similarly, attempted to estimate a value of a QALY using a 'chained' approach, a method which has previously been used to estimate the value of a prevented fatality in the UK (Carthy et al., 1999).

2.2 The 'chained' approach'

The 'chained' approach referred to above was introduced to the value-of-life literature in order to overcome well-documented difficulties survey respondents have when asked to value very small risk reductions (Beattie et al,. 1998; Carthy et al., 1999). The approach was intended to break down such valuation tasks so that, in the QALY case, a respondent would be asked to value avoidance of a small, but certain, health detriment, and then provide a health state utility for the scenario using, for example, a SG procedure. The two are then combined to compute a value of a QALY. This is essentially the approach used by Pinto Prades et al. (2009) and Baker et al. (2010), and subsequently adapted for EuroVaQ.

Attempts to apply the chained approach to the value of a QALY have unearthed issues relating to methods of aggregation and measures of central tendency. For example, if a respondent in a SG exercise says s/he is only willing to take a 1 in 100,000 risk (or less) of death to avoid being permanently in a minor health state, his/her WTP to avoid 12 months in that state would then be multiplied by a hundred thousand, potentially giving an astronomical figure for the value of a QALY[1]. In the UK Social Value of a QALY (SVQ) Project, reported

[1] Someone who says they are only willing to take a 1 in 100,000 risk of death to avoid the chronic illness state is taken to be indexing that health state at 0.99999: i.e. a year spent in that state is taken to amount to the loss of

by Baker et al. (2010), over 25 per cent of respondents gave responses to a 12-month scenario which, in combination, imply values of a QALY of more than £1m. Some respondents generated values of thousands of millions of pounds, the mean values for a QALY being £3x10^8 for avoiding a minor stomach illness and £7x10^8 for a minor head condition. It has been widely accepted that when such values are being used to guide public policy, it is the mean figure which should be used as the best indicator of social welfare. But clearly, a mean value for a QALY of £5 x 10^8, or even a figure one-thousandth as big as that (i.e. £500,000), would be totally anomalous in a world where the UK value of a prevented fatality (VPF) is £1.7m (representing the prevention of a death which, on average, entails the loss of about 40 years of life expectancy) and the NICE cost per QALY threshold is £20,000-30,000.

Other ways of managing the data were therefore devised. In SVQ, rather than computing a ratio of WTP/QALY loss for each individual and then taking a mean, these alternative approaches involved taking a measure of central tendency (either mean or median) for WTP and the corresponding measure of central tendency for the QALY loss and then computing the ratio, rather more like a point estimate. Pinto Prades et al. (2009), having experienced the same problem when attempting to compute the mean value of individual values of a QALY, took the same approach. A recent paper by Gyrd-Hansen and Kjaer (2011) reports how sensitive estimates of WTP per QALY are to the analytical strategy used (e.g ratios of means vs means of ratios).

One factor possibly contributing to the extremely high values estimated previously is, as the duration and/or risk of the health gain presented in the WTP questions was 'fixed' in advance, respondents were potentially each valuing very different QALY gains. So, for example, respondents who had attached a very high utility value for the health state in the standard gamble component would be valuing a very much *smaller* fraction of a QALY than those who had attached a lower utility value to the health state. Given the well documented phenomena of insensitivity to scale and scope in WTP studies, it is not surprising that asking certain respondents to value very small gains indeed would - when combined with their WTP response - would generate very high WTP per QALY estimates.

3. SURVEY DESIGN

3.1 Overview of methods

In EuroVaQ, two main approaches to elicit the value of a QALY were adopted- the 'chained' and 'direct' methods (see Final Report at the EuroVaQ website at **http://research.ncl.ac.uk/eurovaq/**). Briefly, the 'direct' questionnaire tested the notion of presenting health gains 'directly' using a simple graphical and textual description, and many of the health gains presented were of one whole QALY. In contrast, the chained approach - which is the focus of this paper - set out to build upon the use of the 'chained' approach used previously (see above). In previous WTP/QALY studies using a chained approach, respondents were asked to complete an utility assessment exercise in order that their utility value (between 0 and 1) for a given health state could be ascertained. Next they were asked their WTP to avoid a given duration/risk of that health state. Combining the respondent's answers to both components then allows that respondent's WTP per QALY gained to be estimated (essentially by 'multiplying up' their WTP for a known fraction of a QALY into one whole QALY). For example, if we know that a respondent is willing to pay £1000 to avoid one year with certainty in a health state with utility value of 0.95 (i.e. a loss of 0.05 or

0.00001 of a QALY. So if 100,000 such people were each willing to pay, say, £300 to avoid the 12-month illness, they would between them be paying £30 million and their combined benefit would add up to just one QALY.

1/20th of a QALY), we can estimate their WTP per QALY to be 20 * £1000 = £20,000 per QALY gained (assuming linearity).

The basic principle behind the chained approach is that the 'health losses' being considered in the WTP component of the exercise are not too large such that stated WTP values would likely be subject to 'budget constraints'. For example, most people would 'value' the prevention of a *certain* catastrophic health event (such as going blind or losing the use of their limbs) to be greater than the sum they could possibly afford to pay. Of course, in reality people are unaccustomed to being asked to pay to avoid the *certainty* of such large health losses. Rather, they are more accustomed to thinking about paying to insure themselves against the *risk* of some catastrophic event. Alternatively, people are more accustomed to considering the prospect of out-of-pocket payments to treat more minor ailments. Hence, the expected health losses people are used to considering in real purchasing decisions are relatively modest.

Whilst keeping the QALY gains being valued 'small' is a common feature of previous 'chained' approaches to estimating WTP per QALY, the novel aspect of the research reported here is that we also strove to keep the QALY gains *constant* across respondents. We did this by developing two different types of WTP questions: 'risk variant' and 'time variant' questions. Risk variant WTP questions ask respondents about their WTP to avoid some *risk* of a health state (that they had previously given a utility value for). The WTP questions presented to individual respondents were customised according to their utility valuation by varying either risk or time in order to keep the QALY gain constant across respondents. So for example, setting the QALY gain at 0.05 for all respondents would mean that a respondent whose utility value for a given health state was 0.90, would be asked their WTP to avoid a 50% chance of that health state for one year. Someone who gave a utility value of 0.80 for the same health state would be asked their WTP to avoid a 25% risk of the health state for a year. In time variant WTP questions, respondents are asked about their WTP to avoid some *duration* of a given health state (again that they have previously given a utility value for), this time with the duration varied in order to keep the QALY gain constant across respondents. For example, for the respondent with a utility value of the health state 0.90, avoiding 6 months in that state with certainty, would amount to a gain of 0.05 QALYs (0.10 * 6/12)[2]. We elected here to 'set' the QALY gains at 0.05 and 0.10 of a QALY. To the best of our knowledge, no other researchers have designed an interactive WTP study of this nature previously.

3.2 The utility assessment component

Every respondent completed a utility elicitation exercise on two EQ-5D health states, namely 21121 and 22222, referred to henceforth as the yellow and green states respectively. These health states were chosen as we wanted to include both mild and moderate health states and because yellow 'dominates' the green (in that it is better on 3 dimensions and no worse on the other 2). This provides us with a simple consistency test in that the utility value of the yellow state ought to be greater than that of green. Whilst all respondents valued *both* states, health states utilities were derived either by standard gamble (SG) *or* time trade off (TTO)[3].

[2] The respondent with a utility value for the health state of 0.8 would need only avoid 3 months in that health state with certainty to gain 0.05 QALYs.

[3] It is assumed that readers will be familiar with these methods, so we limit a general discussion of the techniques here and refer the reader to the full report for further details.

In the SG procedure, respondents were first taken through a 'risk introduction' exercise designed to familiarise them with the format of the SG questions and to get a feel for what a risk of 'X in 100' meant to them. The 'smiley face' format was then used for the SG questions themselves. Respondents were asked to consider two situations; with and without treatment. Without treatment they would be in the target health state for sure for the rest of their lives. With treatment they faced some probability 'p' of full health (EQ-5D state 11111) and some probability (1-p) of immediate death. In all cases, the starting values of 'p' and '1-p' were 0.60 and 0.40 respectively. That is the 'with treatment' option presented respondents with the prospect of a 60% chance of full health and a 40% chance of death. The full health state (EQ-5D 11111) was denoted by the 'pink health state'.

The SG followed an iterative procedure which generally presented respondents with a maximum of four questions before their utility value could be estimated. If, however, respondents preferred the 'without treatment' option even when the chances of full health with treatment were 99% (and hence chances of death 1%) they were then asked a supplementary question to try and accurately estimate their utility value for the target health state. Those respondents who did not accept a 1 in 10,000 risk of death were considered to have a utility value of '1' for that health state and were classified as 'non –traders'. We return to the issue of non traders below.

In the TTO, respondents were first asked their age and a rough life expectancy was then estimated for them (of around 80 years)[4]. They were then asked to consider two different 'lives'; A and B. Life A was remaining life expectancy, Y, (in the example in footnote 4, 30 years) in the target health state and life B was X years (where X< Y) years in full health. In all cases, the starting value of 'X' was set at 60% of Y, the respondent's remaining life expectancy. They were then asked to consider whether they would prefer life A, prefer life B or they considered both lives to be equally preferable.

The TTO procedure replicated exactly that used in the SG procedure (by simply swapping chances of full health in the SG with proportion of life expectancy in the TTO). If respondents preferred 'Life A' even when the years spent in full health in life B were 99% of life expectancy they were asked whether they would be willing to give up *any* time in order to avoid the target health state. Those who said they would not be willing to give up any time at all were considered to have a utility value of '1' for the target health state and are classified below as 'non traders'.

3.3 The WTP component

In the WTP component, respondents were presented with some risk/duration of the health states they had already seen in the health state utility elicitation component. Respondents were first presented with a screen introducing the WTP exercise. Respondents were then either asked risk variant WTP questions or time variant WTP questions which are explained in detail below[5]. Henceforth, all WTP questions that relate to the green health state (EQ-5D

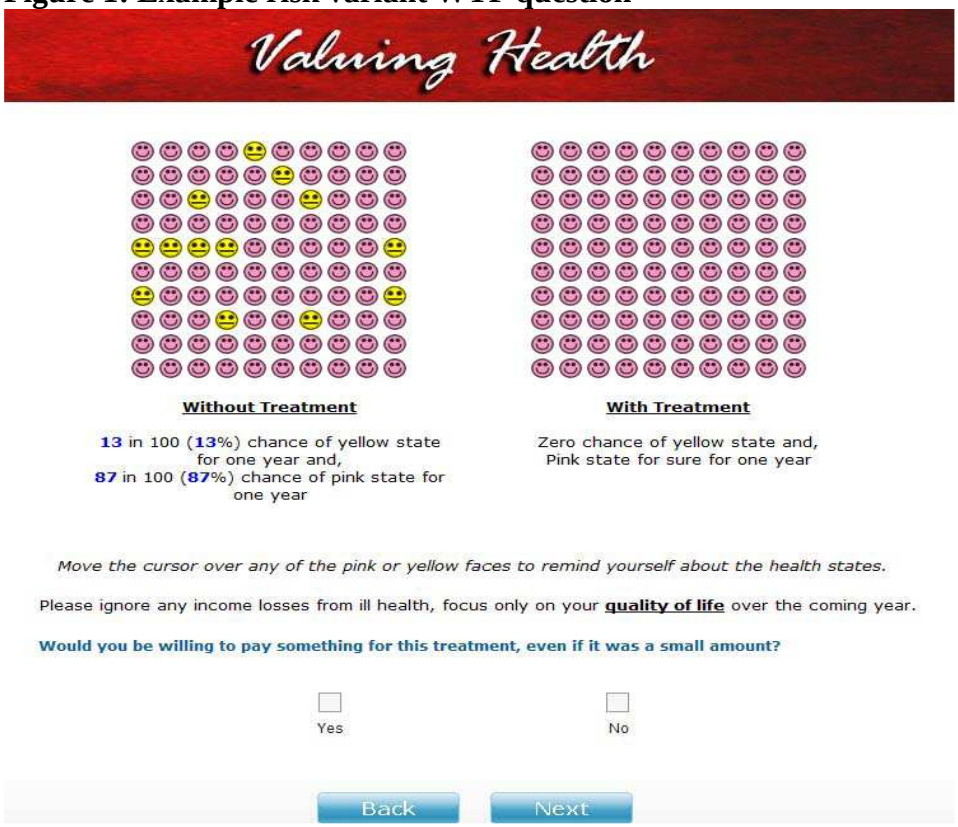
[4] So for example, respondents aged 52 were then told; 'very roughly then we can say that people of your age are likely to live for another 30 years or so…' .

[5] Additional questions were adopted in order to achieve overlap with the direct approach whereby respondents were asked about one year in the health state with certainty. As the magnitude of the health gains involved in these questions are potentially large. Hence, as these are not in keeping with the rationale for the 'chained' design outlined above, they are not discussed further here. See the EuroVaq report for the whole set of questions deployed in the chained approach.

state 22222) are prefixed by 'G' and those that relate to the yellow health state (EQ-5D state 21121) are prefixed by 'Y.'

In the risk variant WTP questions respondents were asked to think about the value they would put on avoiding some *chance* of suffering the target health state for one year. The 'chances' were set according to a) the respondents' utility value derived in the utility assessment component and b) whether the QALY gain was 'set' at 0.05 or 0.10. So, for example, a respondent who attached a utility value of 0.60 to the yellow health state would be presented with a 13% chance of that health state in a 0.05 QALY gain question and 26% chance of that state in a 0.10 QALY gain question. As the QALY gain for one year with certainty is 0.4 QALYs (from 0.6 up to 1), multiplying by 13% and 26% yields 0.05 and 0.10 respectively (after rounding to whole numbers). An example risk variant WTP question is given in Figure 1.

Figure 1: Example risk variant WTP question

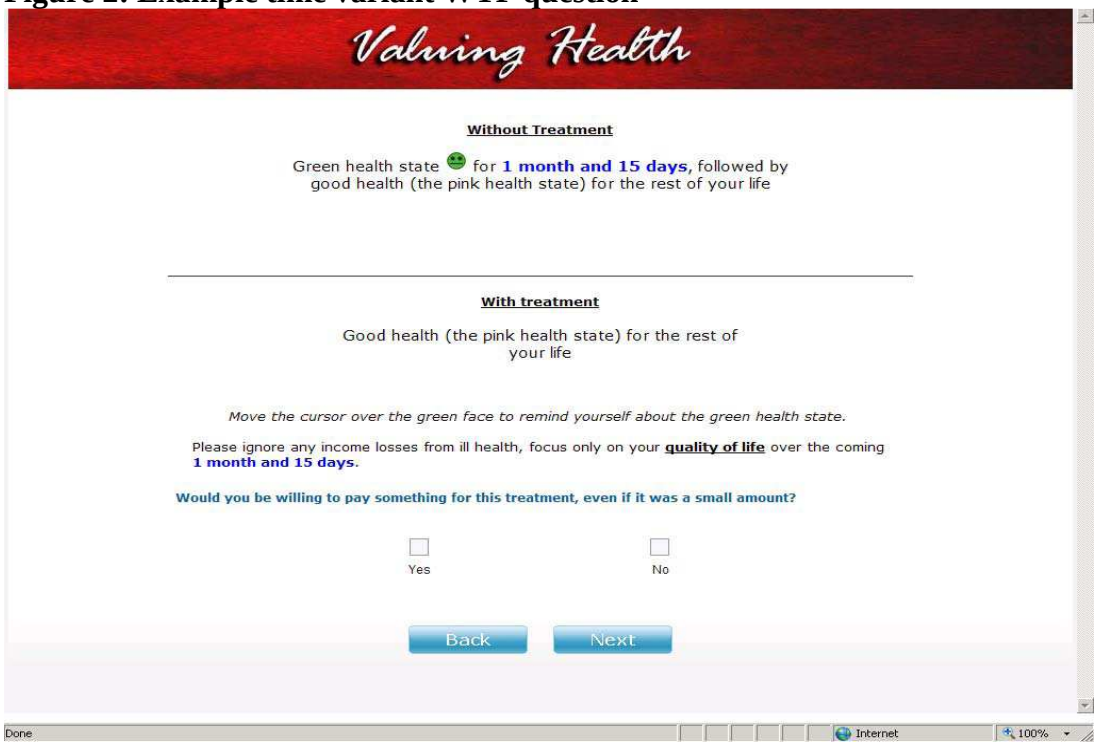


In the time variant WTP questions respondents were asked about their WTP to avoid some *duration* of a health state with certainty. Durations were set according to a) the respondents' utility value derived in the utility assessment component and b) whether the QALY gain was 'set' at 0.05 or 0.10 QALY. So, for example, a respondent who attached a utility value of 0.60 to the green health state would be presented with a 3 month duration of that health state in a 0.10 QALY gain question and 1.5 month duration (expressed as 1 month and 15 days) in a 0.05 QALY gain question. An example time variant WTP question is given in Figure 2.

It is important to stress that respondents could be allocated to either risk or time variant WTP questions irrespective of whether they completed SG or TTO in the utility assessment component. The specific nature of the WTP questions will be outlined in the following

section where we set out the related hypotheses to be tested. In all cases, respondents were first asked whether they would be willing to pay anything at all, even a small amount to avoid the health loss. Respondents who were willing to pay at least something were then taken through a 'card sort' procedure. Respondents were presented with a series of 15 money amounts in random order and asked to decide whether they definitely would pay that amount, definitely would not pay that amount or whether they were unsure whether or not that would pay that amount. For each amount in turn, respondents were invited to 'click and drag' the card into one of 3 boxes; 'definitely would pay', 'definitely wouldn't pay' or 'unsure'.

Figure 2: Example time variant WTP question



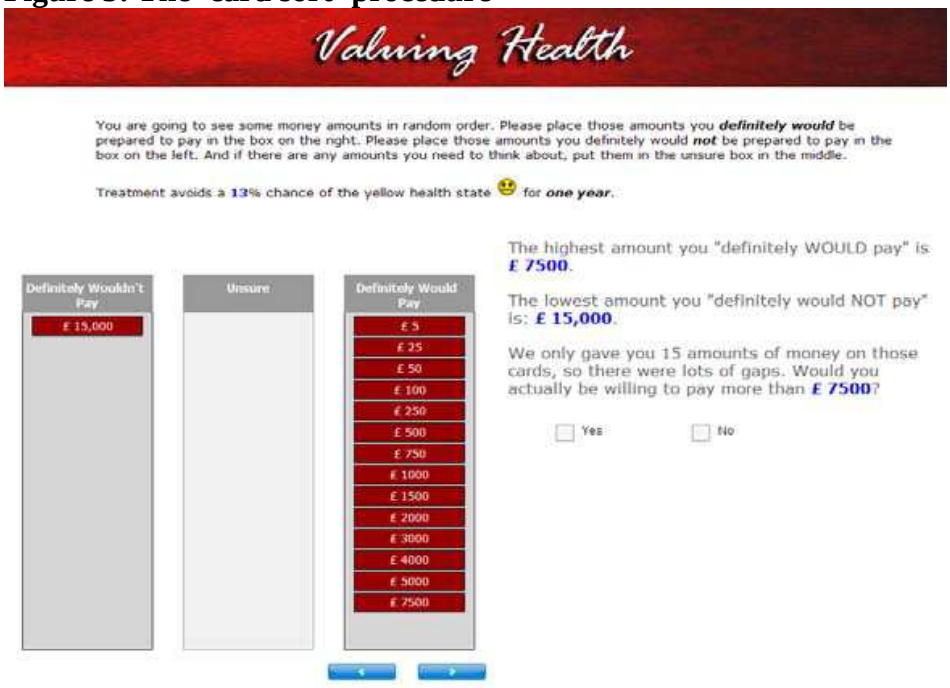
Once respondents had clicked and dragged all 15 amounts into the boxes the interactive programme then presented back to them their highest 'definitely would pay' amount and lowest 'definitely would not pay' amounts. They were then asked whether they would be willing to pay more than their highest 'definitely would pay' amount and, if so, were invited to write in that amount (in the example given in Figure 3, somewhere between £7,500 and £15,000) and that amount was taken to be the best estimate of their maximum WTP for the treatment. If they were unwilling to pay more than their highest 'definitely would pay' card, then *that* amount was taken to be the best estimate of their maximum WTP for the treatment.

An important design feature of the card sort procedure was that the set of cards presented to respondents aimed to keep the 'range' constant in terms of *implied WTP per QALY*. This was done in order to try and reduce any 'framing' effects on WTP per QALY introduced solely as a result of the chosen card sort amounts and, as argued above, gave the values the 'best shot' at uncovering robust WTP per QALY estimates. Hence, in the UK, the amounts used in the 0.05 QALY gain questions, ranged from £5 to £15,000 (i.e. from £100 to £300,000 per QALY) whilst the amounts used in the 0.10 QALY gain questions ranged from £10 to £30,000 (i.e. again from £100 to £300,000 per QALY). We believe that most readers will consider the range wide enough not to unduly constrain respondents to any particular valuation (and they had the option of stating values outside this range). The card sort amounts

were converted from UK currency to the prevailing local currency using Purchasing Power Parity conversion rates for 2008 and rounded to two significant figures.

Those respondents who were unwilling to pay anything at all, *even a small amount*, were instead presented with a supplementary question to ascertain the reason behind such a response. Respondents were presented with a series of possible reasons for not paying anything at all, and asked to tick *all* statements that applied to them. Whilst 4 of the 5 options could be taken to be 'legitimate' reasons for being unwilling to pay anything at all (i.e. 'the risk is too low to worry about' or 'I would get better anyway, so it is not worth paying'), respondents ticking only 'I do value the treatment, but do not want to pay because the government should provide health care' were considered to be 'protestors'[6]. The issue of how 'protestors' were dealt with is included in the 'analytical strategy' section in the Appendix.

Figure 3: The 'card sort' procedure



3.4 The chained questionnaires

The chained questionnaires were piloted in several countries using an interactive computer-based 'prototype' of the final internet-based survey. In order to cover the range of questions without over-burdening respondents, 8 different versions of the survey were subsequently programmed and hosted on the internet. Respondents were initially randomised to one of 8 versions[7] and were presented with a maximum of 5 WTP questions. Then, within each version, respondents were further randomised *to begin* by answering questions relating to either the green or yellow health state. Those randomised to yellow *first* would complete an utility assessment exercise (either SG or TTO) involving the yellow health state and then go

[6] As respondents had the option of ticking multiple options, we elected here to define respondents ticking *only* 'government should pay', and no other 'legitimate' reason, to be protestors. This was subject to sensitivity analysis in the full report.

[7] Familiarity with the versions is *not* essential to understanding the results. Details of how the questions were 'blocked' into the 8 versions and the questionnaires themselves can be found on the EuroVaQ website; http://research.ncl.ac.uk/eurovaq/questionnaires.

on to answer the set of WTP questions relating to the yellow health state (the 'set' depending on which of the 8 versions the respondent was randomised to). They would then go on to complete an utility assessment exercise involving the green health state before answering the set of WTP questions relating to the green health state (again the set depending on version). Those randomised to green first did the opposite. Further, the order in which the WTP questions were presented was randomised wherever possible in order to avoid introducing any systematic 'ordering' bias. We set out to achieve a sample in each country that was representative in terms of age, gender, region and social class/income group.

4. HYPOTHESES TO BE TESTED

As above, the risk variant questions asked the respondent about their WTP to avoid some chance of the target health state for *one year*. These questions are referred to below as **Gj,** and **Gd** (relating to the green health state and 0.05 and 0.10 QALY gains respectively) and **Yh** and **Yi** (relating to the yellow health state and 0.05 and 0.10 QALY gains respectively). A comparison of responses to a) Gj and Gd and b) Yh and Yi, then offer straightforward sensitivity to scope tests. If the standard QALY model holds and WTP is proportional to QALYs gained, we would expect the responses to the 0.10 QALY gain question to be roughly double those to the 0.05 QALY gain question and, hence, estimated *WTP per QALY* to be roughly equivalent. If, on the other hand, respondents were, on average, willing to pay the same for the 0.05 and 0.10 QALY gains, the estimated WTP per QALY derived via the 0,05 QALY gain question would be roughly twice that estimated via the 0.10 QALY gain question.

The time variant WTP questions asked the respondent about their WTP to avoid some *duration* of a health state with certainty. The time variant WTP questions are referred to below as **Gm,** and **Gn** (relating to the green health state and 0.05 and 0.10 QALY gains respectively) and **Yq** and **Yr** (relating to the yellow health state and 0.05 and 0.10 QALY gains respectively). Here, comparing responses to a) Gm and Gn and b) Yq and Yr offers straightforward sensitivity to scope tests. Again, assuming the standard QALY model holds and WTP is proportional to QALYs gained, we would expect the responses to the 0.10 QALY gain question to be roughly double those to the 0.05 QALY gain question and, hence, estimates of *WTP per QALY* to be roughly equivalent.

It should also be obvious that Gm and Gn are the time variant counterparts of the risk variant questions Gj and Gd. Likewise, Yq and Yr are the time variant counterparts of the risk variant questions Yh and Yi. A comparison of responses to a) Gj and Gm, b) Gd and Gn, c) Yh and Yq, and d) Yi and Yr, then offer a test of sensitivity of WTP responses to the 'framing' of the WTP question i.e. whether a risk variant or time variant framing was used. If the 'framing' of the question had no influence on responses, then, for each magnitude of QALY gain, we would expect estimates of *WTP per QALY* to be roughly equivalent across 'frames'. Table 1 describes the 8 WTP questions described above in terms of the QALY gain and the 'framing' of the question and forms the basis of three of our four hypotheses.

Table 1: Summary of the WTP questions

		Framing	
		Risk	**Time**
QALY gain	**0.05**	**Gj, Yh**	**Gm, Yq**
	0.10	**Gd, Yi**	**Gn, Yr**

4.1 Hypothesis one

Our first hypothesis is the straightforward sensitivity to scope tests discussed above when only the magnitude of the QALY gain differed. Formally, this tests;

Ho: Gj = Gd vs H1: Gj ≠Gd. Ho: Yh = Yi vs H1: Yh ≠Yi.
Ho: Gm = Gn vs H1: Gm ≠Gn. Ho: Yq = Yr vs H1: Yq ≠ Yr.

Whilst both parametric and non-parametric tests were undertaken, we report only the difference in means tests here. Given the large sample size and the number of tests to be carried out, we considered it reasonable to adopt a stringent test of statistical difference and, hence, have tested at the 1% significance level. It is important to stress that all hypothesis tests were conducted on the implied WTP *per QALY* results, rather than on the WTP responses.

4.2 Hypothesis two

Our second hypothesis is that, for a given magnitude of QALY gain, whether the use of a time variant or risk variant 'frame' affects responses. Formally, this tests:

Ho: Gj = Gm vs H1: Gj ≠Gm. Ho: Yh = Yq vs H1: Yh ≠ Yq.
Ho: Gd = Gn vs H1: Gd ≠Gn. Ho: Yi = Yr vs H1: Yi ≠Yr.

4.3 Hypothesis three

Our third hypothesis is that derived WTP per QALY estimates will be robust to the severity of the health state that the 'chaining' process was based upon. That is, the WTP for a QALY gains of 0.05 or 0.10 are independent of the health state used as the starting point. This essentially involves comparing responses to the green health state (prefixed with 'G' in table 1) with their yellow counterparts (prefixed with 'Y' in Table 1). Formally, this tests;

Ho: Gj = Yh, vs H1: Gj ≠Yh. Ho: Gm = Yq vs H1: Gm ≠Yq.
Ho: Gn = Yr, vs H1: Gn ≠Yr. Ho: Gd = Yi vs H1: Gd ≠Yi.

4.4 Hypothesis four

Our fourth hypotheses (not highlighted in Table 1) relates to whether the SG or TTO was used in the utility assessment component. Respondents in versions 1-4 completed SG whilst respondents in versions 5-8 completed TTO in the utility component. Irrespective of whether respondents had completed an SG or TTO exercise, respondents could be presented with either a time variant or risk variant WTP question depending on which of the 8 versions they were randomised to. In this way, each WTP question listed in Table 1 can be further disaggregated into whether respondents were randomised to versions 1-4 (and completed SG) or to versions 5-8 (and completed TTO). Whilst hypothesis four may be tested on all 8 questions that appear in Table 1, we limit ourselves here to testing one question from each of the 4 cells. Formally this tests;

Ho: $Gj_{(SG)} = Gj_{(TTO)}$ vs H1: $Gj_{(SG)} \neq Gj_{(TTO)}$. Ho: $Yq_{(SG)} = Yq_{(TTO)}$ vs H1: $Yq_{(SG)} \neq Yq_{(TTO)}$.
Ho: $Gn_{(SG)} = Gn_{(TTO)}$ vs H1: $Gn_{(SG)} \neq Gn_{(TTO)}$. Ho: $Yi_{(SG)} = Yi_{(TTO)}$ vs H1: $Yi_{(SG)} \neq Yi_{(TTO)}$.

5. RESULTS

5.1 Achieved sample

A total of 21,965 completed interviews were achieved of which 2,312 were conducted in the UK and 2,673 in France. The final sample was broadly representative but with significant under-representation of elderly females in most countries including France. Table 2 shows the target and achieved samples by age and gender in the UK and France (the full sample information is shown in Table 3.3 in the EurovaQ report).

Table 2: Sample by age and gender: UK and France.

	UK		France	
	Target	Sample	Target	Sample
Male 18-25	6.4%	6.2%	6.4%	5.6%
Male 26-35	8.8%	8.3%	8.8%	9.1%
Male 36-45	9.6%	9.3%	9.6%	9.8%
Male 46-54	7.2%	7.6%	8.0%	7.4%
Male 55-64	6.4%	7.4%	5.6%	7.8%
Male 65+	8.8%	8.4%	8.8%	6.7%
Female 18-25	6.4%	7.0%	6.4%	6.8%
Female 26-35	9.6%	9.0%	9.6%	12.4%
Female 36-45	9.6%	9.6%	9.6%	12.2%
Female 46-54	8.0%	9.0%	8.0%	10.0%
Female 55-64	7.2%	7.4%	6.4%	8.8%
Female 65+	12.0%	10.9%	12.8%	3.4%

5.2 Estimated WTP per QALY

Table 3 shows mean and median estimates of *WTP per QALY* by country for the 8 questions outlined in Table 1. Means have been top trimmed at 1% by country[8] and estimates are in $US dollars PPP[9]. Other issues relating to the data are discussed in the 'analytical strategy' section in the appendix. Estimates of the value of a QALY clearly varied across question and country with (trimmed) mean estimates ranging from $13,228 to $29,308 in the UK and from $11,317 to $26,890 in France. Untrimmed means were considerably higher in both countries. The corresponding median value of a QALY estimates ranged from $3,064 to $6,775 in the UK and from $2,745 to $4,574 in France. Table 3 shows that in many cases the mean and median value of a QALY estimates for UK and France tend to be among the lower estimates across countries. For example, for question Gj the median values of $6,317 and $4,574 in the UK and France respectively compare to estimates of $10,748 and $7,671 in Poland and Hungary respectively.

5.3 Hypothesis one: sensitivity to 0.05 or 0.1 QALY gain

The risk variant questions **Gj,** and **Gd** relate to the green health state and 0.05 and 0.10 QALY gains respectively whilst **Yh** and **Yi** relate to the yellow health state and 0.05 and 0.10 QALY gains respectively. Looking at the overall results, the last row of Table 3 shows that the ratio of mean (medians) for WTP per QALY derived via Gj and Gd is 1.14:1 (1.10:1), indicating that, whilst the general direction of the aggregate results indicates insensitivity to scope, *some* degree of sensitivity exists (as neither ratio is close to 2). The corresponding ratio of means (medians) of WTP per QALY derived via Yh and Yi at 1.38:1 (1.59:1), is, however, closer to two, indicating there is less sensitivity to the magnitude of the QALY gain in those WTP questions involving the yellow health state.

The time variant WTP questions **Gm,** and **Gn** relate to the green health state and 0.05 and 0.10 QALY gains whilst **Yq** and **Yr** relate to the yellow health state and 0.05 and 0.10 QALY gains respectively. Again looking at the overall results, the last row of Table 3 shows that the ratio of mean (medians) for WTP per QALY derived via Gm and Gn is 1.50:1 (1.27:1). The ratio of means (medians) of WTP per QALY derived via Yq and Yr is 1.62 (1.25). So again,

[8] That is the top 1% of WTP per QALY estimates *in each country* were trimmed- not the top 1% overall.

[9] Untrimmed means were considerably higher in all cases due to the presence of extreme outliers - see full report for untrimmed means.

some degree of sensitivity has been shown in the aggregate results. Focusing on means, for both health states the ratios are closer to 2 than in their risk variant counterparts, indicating that there was less sensitivity to the magnitude of the QALY gain when duration was varied than when risk was varied.

Table 3: WTP per QALY estimates in $US (top trimmed at 1% by country)*

		Gj	Gd	Gm	Gn	Yh	Yi	Yq	Yr
Netherlands	**N**	504	1,017	1,119	1,111	439	759	1,124	1,089
	Mean	24,892	22,211	27,418	18,623	23,133	16,819	27,345	15,738
	Median	7,904	7,398	4,554	4,556	3,557	4,378	4,553	3,415
UK	**N**	459	882	953	954	394	646	960	942
	Mean	23,267	21,182	29,308	15,897	23,285	14,848	20,525	13,228
	Median	6,317	6,775	3,064	3,065	3,064	3,064	3,064	3,256
France	**N**	569	1,129	1,163	1,146	474	824	1,143	1,091
	Mean	26,890	24,852	25,965	16,613	20,115	13,927	19,952	11,317
	Median	4,574	4,391	3,294	3,293	3,327	3,245	3,293	2,745
Spain	**N**	496	1,050	1,130	1,140	441	799	1,102	1,093
	Mean	38,162	50,083	52,876	33,789	25,629	28,049	39,356	26,299
	Median	12,669	7,347	8,005	7,338	8,005	6,671	8,005	6,671
Sweden	**N**	500	1,056	1,107	1,122	428	767	1,100	1,095
	Mean	35,200	34,824	28,805	19,287	27,696	16,908	33,142	18,292
	Median	7,842	6,862	4,313	3,235	5,392	3,235	4,313	3,235
Norway	**N**	365	760	807	818	288	523	801	794
	Mean	37,427	33,685	41,298	26,399	29,474	21,602	41,003	24,757
	Median	15,472	10,637	7,659	7,659	10,745	7,621	7,659	7,659
Denmark	**N**	473	978	1,131	1,149	393	762	1,127	1,142
	Mean	57,389	34,063	42,118	31,456	37,309	25,328	41,316	24,796
	Median	15,409	11,499	6,899	7,704	11,499	7,704	7,589	5,749
Poland	**N**	510	898	959	973	460	750	955	949
	Mean	40,023	35,773	29,188	22,434	32,104	23,104	25,716	18,601
	Median	10,748	7,738	5,171	5,159	9,921	6,798	5,159	3,611
Hungary	**N**	516	948	946	952	464	715	938	932
	Mean	26,132	19,617	21,791	13,222	21,181	11,565	19,117	10,938
	Median	7,671	4,299	7,176	4,285	4,478	3,081	5,731	3,081
Total	**N**	4,392	8,717	9,314	9,364	3,781	6,544	9,250	9,127
	Mean	34,097	30,581	33,236	22,057	26,386	19,129	29,726	18,247
	Median	8,211	7,473	6,382	5,015	7,295	4,584	4,687	3,723

*as outlined in the analytical strategy section, 'non–traders' in the SG/TTO and 'protests' in the WTP are omitted

Table 4 gives the results of the formal hypotheses tests undertaken and shows that, at the 1% significance level, three of the four tests undertaken under hypothesis one rejected the null - the exception being the Gd vs Gj comparison (where p= 0.017). A similar picture emerges for both France and UK when data are analysed at the level of the individual country, suggesting that respondents in both countries were generally insensitive to the size of the QALY gain.

Table 4: Hypothesis One Summary data (based on WTP *per QALY*)

H0:	Difference in means	Std err	t stat	P-value
Gj = Gd	3516.8	1476.77	2.38	0.0173
Yh = Yi	7257.7	1081.41	6.71	0.0000
Gm = Gn	11178.6	1103.77	10.13	0.0000
Yq = Yr	11479.4	999.53	11.48	0.0000

5.4 Hypothesis two: time versus risk 'frames'

Our second hypothesis is that, for a given magnitude of QALY gain, whether the use of a time variant or risk variant 'frame' affects responses. The last row of Table 3 shows that the median WTP per QALY estimated via the risk variant questions Gj, Gd, Yh and Yi of $8211, $7473, $7295, and $4584 respectively are in each case higher than their time variant counterparts Gm, Gn, Yq and Yr of $6,382, $5015, $4687 and $3723 respectively. Thus, it would appear as if estimated WTP per QALY is not independent of the framing used in the WTP question in that, all else equal, using a risk variant format appears to yield higher estimates than a time variant format. Table 5 gives the results of the formal hypotheses tests undertaken and shows that, at the 1% significance level, two of the four tests undertaken under hypothesis two rejected the null hypothesis whilst it was accepted in the remaining two cases. At the individual country level, in both UK and France, only one of the four tests undertaken rejected the null (Gd=Gn in both cases).

Table 5: Hypothesis Two Summary data (based on WTP *per QALY*)

H0:	Difference in means	Std err	t stat	P-value
Gj = Gm	861.85	1490.97	0.58	0.5632
Yh = Yq	-3339.9	1296.46	-2.60	0.0094
Gd = Gn	8523.6	1084.5	7.86	0.0000
Yi = Yr	881.7	716.61	1.23	0.2186

5.5 Hypothesis three: robustness to severity

This essentially involves comparing responses to the green health state (prefixed with 'G' in table 1) with their yellow counterparts (prefixed with 'Y' in Table 1). Again looking at the last row of Table 3, the mean WTP per QALY from the green health states are in each case higher than their yellow counterparts. The means of Gj, Gd, Gm and Gn are $34,097, $30,581, $33,236 and $22,057 respectively, and, in each case, is higher than the yellow counterpart, as Yh, Yi, Yq and Yr means are $26,386, $29,726, $19,129 and $18,247 respectively. Observing the medians shows the same general pattern in the aggregate results.
Table 6 gives the results of the formal hypotheses under three and shows that, at the 1% significance level, the null is rejected in all four cases. Hence, the severity of the health state that the 'chaining' process is based on does appear to matter in each case. At the individual country level, however, in both UK and France in only two of the four tests undertaken was the null rejected.

Table 6: Hypothesis Three Summary data (based on WTP *per QALY)*

H0:	Difference in means	Std err	t stat	P-value
Gj = Yh	7710.9	1485.83	5.19	0.0000
Gm = Yq	3509.1	1292.4	2.72	0.0066
Gn = Yr	3809.9	739.64	5.15	0.0000
Gd =Yi	11451.8	1068.93	10.71	0.0000

5.6 Hypothesis four: SG versus TTO

Our fourth hypotheses related to whether the SG or TTO was used in the utility assessment component. Respondents in versions 1-4 completed SG whilst respondents in versions 5-8 completed TTO in the utility component. Whilst this data is not shown in Table 3 alongside the other aggregate results, Table 7 shows the mean estimates for 4 of the 8 questions broken down by either SG or TTO. For example, mean WTP per QALY estimates for Gj starting with the SG (versions 1-4) and TTO (versions 5-8) are $32,248 and $36,267 respectively. In each case, the estimated WTP per QALY starting with the SG procedure are lower than those starting with the TTO procedure, although in only one of the four cases highlighted is the difference statistically significant (for Gn). At the individual country level, in both France and the UK none of the tests undertaken here rejected the null (even at higher significance levels). Indeed, the country level results show that, even for Gn, it is only in 2 countries: Norway and Poland that the null is rejected.

Table 7 Hypothesis Four Summary data (based on WTP *per QALY*)

Variable	Mean	Std err.	t stat	p value
Gj_sg	32247.7	1524.69		
Gj_tto	36267.4	1764.85		
Combined	34097.4	1156.53		
Diff	-4019.7	2332.20	-1.72	0.0849
Yq_sg	28874.8	1302.53		
Yq_tto	30604.4	1197.51		
combined	29726.4	885.92		
Diff	-1729.6	1769.35	-0.98	0.3283
Gn_sg	19616.5	759.07		
Gn_tto	24462.1	866.19		
combined	22056.9	576.94		
Diff	-4845.6	1151.72	-4.20	0.0000
Yi_sg	18850.9	753.23		
Yi_tto	19477.3	792.85		
Combined	19128.7	547.09		
Diff	-626.4	1093.60	-0.57	0.5668

6. DISCUSSION

We have reported a novel method of estimating WTP per QALY that has largely overcome many of the problems associated with previous applications of the 'chained' approach. Whilst previous approaches have estimated extremely high mean WTP per QALY values, the means reported here do not appear to be unreasonably high and, if anything, many readers may consider the means – and medians - to be somewhat on the low side. For example, the mean WTP per QALY estimates reported for the UK are all within a range that NICE generally considers interventions to be cost effective. Recall that the values used in the card sort procedure allowed the WTP per QALY estimates to range from £100 to £300,000, so we do not believe that the procedure itself would 'lead' respondents to give low values. Rather, it is more likely that the magnitude of the QALY gains valued here – which are perhaps larger than those valued in other studies[10] - is driving the valuations. Whilst there is clearly no 'correct' magnitude of QALY gain to ask respondents to value, valuing smaller- rather than larger gains is more in keeping with the spirit of the chained approach. This could be a reason to favour the WTP per QALY estimates based on the 0.05 QALY gains, but that is clearly a subjective assessment. That we have uncovered a 'range' of valuations, rather than a single point estimate, is perhaps not surprising and we have identified a number of factors that appear to influence estimates of WTP per QALY derived in this manner.

First, there are straightforward insensitivity to scope effects in that WTP per QALY estimates based on smaller QALY gains (0.05 QALY) are higher than those based on larger QALY gains (0.10 QALY). This is despite the values used in the 'card sort' valuation procedure being tailored to the size of the QALY gain so that the WTP amounts respondents were presented with when valuing 0.10 of a QALY were double those in the 0.05 QALY questions. It could be argued that this was giving the method it's 'best shot' at showing sensitivity to scale- yet the classic 'embedding' effects are still evident in the data. As embedding is an already well documented phenomenon, we do not discuss this finding further here.

Other findings are perhaps somewhat harder to explain and we would welcome discussion on the results of our second and third hypotheses in particular. The results reported under hypothesis two showed that estimated WTP per QALY is *not* independent of the framing used in the WTP question in that the risk variant format appears to yield higher estimates than a time variant format. This is perhaps counter intuitive as it may seem more likely that respondents will pay more to avoid the certainty of some duration of a health state than to eliminate the risk of that health state. It may indicate, however, that introducing risk into the WTP questions has resulted in risk aversion influencing the amount respondents are WTP. Interestingly, the results reported under hypothesis one indicated that greater sensitivity was shown when risk - rather than duration - was varied in the WTP questions.

The results reported under hypothesis three showed that estimated WTP per QALY is *not* independent of the severity of the health state used in the chaining process. In particular, estimates of mean WTP per QALY were higher the *more* severe was the health state used in the chaining process. Again, it is not immediately obvious why this might be the case bearing in mind that the utility value attached to the health state has already been taken into consideration in keeping the QALY gain valued in the WTP component the same (such that, if the utility value for one state was higher than for the other- they were presented with a higher risk – or longer duration- of that state in the WTP component). It may, however be

[10] For example in the SVQ project respondents were asked to value the avoidance of minor stomach illnesses or minor head conditions.

that, as there were more respondents already in a health state equivalent to- or worse than- the yellow state than is the case for the green health state-this may have had a downward influence on valuations. This brings us to an important issue in that, the results reported above all assume implicitly that the basic QALY model is 'correct'. In attempting to present respondents with an equivalent QALY gain - and subsequently in estimating their WTP per QALY - we have thus far applied the strong assumptions inherent in the QALY model (for example linearity with risk and duration).

Hence, it may not be the case that any 'problem'[11] uncovered here is necessarily a problem with the WTP method; it could equally be that the strict QALY assumptions applied here simply do not hold. For example, it may not be the case that the model is linear in duration and that doubling the duration of a health state does not double the perceived loss of health. It is beyond the scope of the current paper to discuss here, but plans for further data analysis include those to model WTP responses against utility, risk and duration and estimate a model that explains WTP given the components of the QALY - but without imposing the standard QALY model in advance. This will allow us to examine different functional forms of alternative QALY models and examine the robustness of WTP values in light of those alternative models.

We turn now to issues that have not been explored in detail in this paper, but which we believe are nevertheless worthy of mention here. As outlined above, respondents who accepted no risk of death in the SG -or traded off no time at all in the TTO - were classified as 'non traders'. For SG, the percentages of 'non-traders' were 2.9% and 5.2% for the green and yellow health states respectively. The corresponding percentages for TTO were 7.1% and 10.6%. According to the SG/TTO methodology, they were then considered to have attached a utility value of '1' to the health state. Hence, when faced with a WTP question to avoid some risk/duration of that health state, according to the methodology underlying the chained approach, were being asked to value a zero QALY gain. Whilst one approach would have been to exclude such respondents from the WTP questions, we elected instead to ask *all* respondents to answer the WTP questions, irrespective of their prior response in the utility assessment component. This raised the possibility that respondents could state a positive WTP value for a zero QALY gain and, hence, effectively have an infinite WTP per QALY. By way of illustration, the figures for one question (Yq) show that 66% of non–traders were WTP at least something to avoid some duration of that state. As outlined in the 'analytical strategy' section in the appendix, the results presented here have excluded these respondents (rather than assign an arbitrary small utility loss to the health state and then estimate an arbitrary high WTP per QALY). But the fact that many 'non traders' *did* go on to state a positive WTP to avoiding that health state clearly calls into question some of the underlying assumptions behind the 'chained' approach.

If we consider that a positive WTP value denotes 'trading' on the wealth dimension, then other respondents traded on the utility scale (i.e in SG or TTO), but did not trade on the wealth scale (i.e. the WTP component). Whilst these respondents present no particular challenge analytically (they simply have a zero WTP per QALY), the issue of trading on one dimension- but not on another may suggest that the two scales are treated very differently. Indeed it seems quite intuitive that someone would be willing to sacrifice wealth to avoid some situations, but would be unwilling to sacrifice life years (or chances of survival) to do so

[11] We do not regard producing a range of estimates, in itself, to be a problem and believe that any method of estimating the monetary value of a QALY – whether that be using WTP or very different methods-will likely come up with a range, rather than a single estimate.

(perhaps for religious reasons). It is, however, more difficult to understand why other individuals would be willing to trade life years (or chances of survival) but not even a small amount of wealth. Again illustrating this with the data for one question (Yq), 29% of those respondents who *did* trade on SG/TTO were unwilling to pay anything at all to avoid some duration of that health state[12]. Such behaviour does not accord with economic theory according to which goods are only valued if individuals are willing to pay a positive amount for them. The implication of these findings for the chained methodology needs further consideration.

Another issue highlighted in the 'analytical strategy' section is the issue of 'protest' zeros. After much discussion amongst the research team, the decision was made that, for the purpose of reporting the 'base case' results above, respondents who had given 'I do value the health gain, but think that the government should provide health care' as the sole reason for being unwilling to pay anything at all, were deemed to be a 'protestor' and their zero WTP value was set to 'missing'. Although not shown here, sensitivity analysis was then done around different decisions regarding the 'zero' valuations (obviously the mean WTP per QALY rises the fewer zero valuations are 'allowed' in the data). The issue of 'protestors' and the treatment of zero valuations in general is obviously subject to debate and we look forward to hearing the views of HESG/CES members on this issue.

A third issue is that inconsistent respondents have been retained in the data set. Again this is a subjective decision and one that is open to debate. Whilst there were a number of consistency tests that may be applied to the WTP data, the most straightforward test of consistency is that respondents clearly ought to attach a higher utility value to EQ-5D state 21121 (the yellow state) than to 22222 (the green state). A worryingly large number of respondents (approx 19% who answered SG and 15% who answered TTO) did precisely the opposite. This may indicate that respondents misunderstood the task and it is clearly an open question whether policy decisions ought to be based, at least in part, on the views of those who may have had cognitive problems with the exercise. A quick analysis of the data on the 'inconsistent' respondents shows that they are more likely to be older and have lower educations levels than 'consistent' respondents. Hence, removing inconsistent respondents form the data set will affect the demographics of the sample and the implications of this needs to be considered further.

The number of inconsistent responses in the SG/TTO uncovered here does raise an additional issue that has potentially important implications for other researchers setting out to do value elicitation surveys via the internet. The proportion of strictly inconsistent responses uncovered here are considerably higher than those found before using other response formats. This was despite intensive piloting of the survey and making it obvious that the yellow state 'dominated the green' by presenting them side by side at the start of the exercise. This raises questions about the 'quality' of internet responses vis-a-vis other response modes such as computerised assisted interviews (CAPIs). Whilst becoming an increasingly popular means of collecting large amounts of data in a relatively short space of time, there may be a need for some caution before the wholesale adoption of the format. Again, we welcome hearing the views of HESG/CES members on this issue.

[12] Only a third of these would be considered 'protest' zeros according to the definition of 'protestor' used in this study..

References

Baker R, Bateman I, Donaldson C, Jones-Lee M, Lancsar E, Loomes G, et al. (2010) Weighting and valuing quality-adjusted life-years using statedpreference methods: preliminary results from the Social Value of a QALY project. *Health Technol Assess* 14

Beattie J, Covey J, Dolan P, Hopkins, L, Jones-Lee M, Loomes G, Pidgeon N, Robinson A and Spencer A (1998) On the contingent valuation of safety and the safety of contingent valuation: part 1 – caveat investigator. *Journal of Risk and Uncertainty* 17: 5-25.

Bleichrodt H and Quiggin J (1999) Life-cycle preferences over consumption and health: when is cost-effectiveness equivalent to cost-benefit analysis? *Journal of Health Economics* 18: 681-708.

Byrne MM, O'Malley K and Suarez-Almazor (2005) Willingness to pay per quality adjusted life year in a study of knee osteoarthritis. *Medical Decision Making* 25: 655-666.

Carthy T, Chilton S, Covey J, Hopkins L, Jones-Lee M, Loomes G, et al. (1999) On the contingent valuation of safety and the safety of contingent valuation: Part 2 - the CV/SG 'chained' approach. *Journal of Risk and Uncertainty* 17(3):187-213.

Dolan P and Edlin R (2002) is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis? *Journal of Health Economics* 21: 827-843.

Garber AM, Phelps CE (1997) Economic foundations of cost-effectiveness analysis. *Journal of Health Economics* 16:1-31.

Gyrd-Hansen D (2003) Willingness to pay for a QALY. *Health Economics* 12:1049-60.

Gyrd-Hansen D and Kjaer T (2011) Disentangling WTP per QALY data: Different analytical approaches, different answers. *Health Economics*, published online Jan 2011.

Johannesson M (1995) The relationship between cost-effectiveness analysis and cost-benefit analysis. *Social Science and Medicine* 41(4):483-9.

King JT, Tsevat J, Lave JR and Roberts M (2005) Willingness to pay for a quality adjusted life year: implications for societal health care resource allocation. *Medical Decision Making* 25: 667-677.

Mason H, Jones-Lee M, Donaldson C (2009) *Modelling the monetary value of a QALY: a new approach based on UK data. Health Economics*, 18: 933-950.Forthcoming.

Phelps CE and Mushlin AI (1991) On the (near) equivalence of cost-effectiveness and cost-benefit analyses. *International Journal of Technology Assessment in Health Care* 7: 12-21.

Pinto-Prades JL, Loomes G and Brey R (2009) Trying to estimate a monetary value for the QALY. *Journal of Health Economics* 28: 553-562.

Shiroiwa T, Sung Y-K, Fukuda T, Lang H-C, Bae S-C and Tsutani K (2010) International survey on willingness-to-pay (WTP) for one additional QALY gained: what is the threshold of cost-effectiveness? *Health Economics* 19: 422-437.

Appendix: The analytical strategy

Capping

Whilst at first glance it may appear that we need simply multiply the respondent's WTP responses to the 0.05 and 0.10 QALY questions by 20 and 10 respectively in order to estimate their WTP per QALY, the analysis is somewhat more complicated as it was not always possible to keep the QALY loss constant across respondents due to 'capping' when respondents gave very high utility values for the health state. For example, for a respondent with a utility value of the yellow health state of 0.999, there is *no* risk of that health state (not greater than 100%) for one year they could be faced with that would amount to 0.05 (or 0.10) QALYs. Rather, the risk would be capped at 100% with an expected QALY gain of 0.001. Similarly, in a time variant WTP question, a respondent with a utility value of 0.999 would need to avoid spending 100 years in that state with certainty in order to gain 0.10 QALYs, which is clearly implausible. Risk variant WTP questions were then capped at 100%, time variant questions at life expectancy and the QALY gain respondents had actually been presented with had to be calculated in each case.

Non- traders in SG/TTO

As detailed in section 3.3 above, respondents who accepted no risk of death in the SG or traded off no time at all in the TTO were classified as 'non traders'. According to the principles underlying the SG and TTO approaches, respondents are then considered to have attached a utility value of '1' to that health state. Whilst we could have elected not to ask such respondents the subsequent WTP questions (as they had indicated that there was no utility loss associated with that state), we decided to elicit WTP values from *all* respondents including 'non traders'. The problem this poses is that, if 'non-traders' go on to pay a positive amount to avoid that health state, they effectively have an infinite WTP per QALY as they are paying to avoid something that involves zero utility loss. This is obviously problematic for the analyses, so the aggregate results presented here exclude 'non traders' in the utility assessment component.

'Protest' WTP responses

As outlined in section 3.3, respondents who were unwilling to pay anything at all for the treatment were asked to give their reasoning and one response category was; 'I do value the treatment, but do not want to pay because the government should provide health care'. Respondents could, however, tick multiple statements. This raised the issue of whether a respondent who ticks 'government should provide' *and at least one* other 'legitimate' reason ought to be classified as a protestor or not- again this is subjective. We elected to classify a 'protestor' in our 'preliminary case' for analyses as a respondent who ticked 'government should pay' *only* and to explore alternative assumptions in our sensitivity analyses. As we considered that a 'protest' zero was not a true zero WTP response, we excluded 'protest' zeros from the analysis reported here.

Inconsistent respondents.

There are a number of within-respondent consistency tests that may be carried out. For example, as the yellow heath state (EQ-5D 21121) dominates the green (EQ-5D 22222) then a respondent ought not to attach a lower utility value to yellow state than to the green. In addition, there are various checks for consistency in the WTP component of the chained approach. Whilst there is clearly an issue of whether or not inconsistent respondents should remain in the data, the results presented here retain all inconsistent respondents.

Trimming

As the untrimmed data contained a number of high outliers, the data presented here has been top trimmed at 1%. This was done by country so that the top 1% WTP per QALY estimates in each country have been removed before the means were calculated. The untrimmed results are given in the final EurovaQ report.