GET MORE, PAY MORE? An elaborate test of the validity of willingness to pay per QALY estimates

Ana Bobinac, Job van Exel, Frans Rutten, Werner Brouwer

iBMG, Erasmus University Rotterdam

1. INTRODUCTION

Economic evaluations aim to inform allocation decisions in the healthcare sector by evaluating alternative interventions in terms of their costs and benefits (typically expressed in non-monetary terms as Quality-Adjusted Life-Years or QALYs). The results of such economic evaluations are commonly summarized in an incremental cost-effectiveness or cost-utility ratio (or ICER). Subsequently, common decision rules indicate that the intervention can be considered good value for money if the ICER falls below the relevant cost-effectiveness 'threshold'. This threshold represents the relevant monetary value of a health gain (commonly a QALY), appropriate in a specific decision making context. There are different normative views regarding the exact nature of the threshold (Claxton et al., 2010). It can be seen as a representation of the opportunity costs within the health care sector, i.e. the marginal cost-effectiveness of replaced activities in the context of a fixed exogenously set budget. Alternatively, and more commonly in the theoretical literature, the threshold can be viewed as the consumption value society places on marginal health gains, i.e. the monetary value society is willing to pay (or, the consumption it is willing to sacrifice) in order to obtain an extra unit of health. If the latter viewpoint is taken, the monetary value of the QALY can be empirically estimated in the population, using some preference elicitation method. The most prominent method in this context is contingent valuation (CV), i.e. the willingness to pay (WTP) method, which has been applied several times to estimate the monetary value of a QALY (e.g. Gyrd-Hansen, 2003; King et al, 2005; Bobinac et al, 2010). Given the far reaching consequences of choosing a particular threshold, and thus the potential relevance of the WTP for a QALY estimates, it is necessary to address the issue of validity[1] of WTP per QALY estimates. In the words of Bateman and Brouwer (2006), when WTP estimates are intended to inform decision-making in the healthcare sector, these need to be robust and therefore issues of validity and reliability are not only of academic interest.

Broadly speaking, validity of WTP estimates refers to whether the estimates concur with the underlying economic theory (e.g. the neoclassical theory of consumer behavior). This would predict that larger gains result in higher WTP, ceteris paribus. The validity of WTP for a QALY estimates can thus be judged by considering the robustness of WTP to changes in the QALY gains valued (for instance by varying the size of the quality improvement or the duration of the health gain). Theory, however, only predicts that WTP should increase with increasing QALY gains, but it does not predict the exact *size* of that increase (Fisher, 1996; Bateman and Brouwer, 2006). The relationship between WTP and QALY gains is expected to be increasing yet concave, such that an increase in the QALY gain yields a less then proportional increase in WTP (Bradford, 1972; Smith,

[1] Validity is concerned with whether the measure reflects what it intends to (i.e. the accuracy), as opposed to reliability that deals with whether the instrument measures something other than random noise (i.e. reproducibility) (Jorgensen et al., 2004). Validity of an instrument presupposes its reliability. Reliability is usually measured though test-retest or convergent validity exercises and is not the focus of this paper. For more discussion on the topic see e.g. Jorgensen et al. (2004), NOAA (1993).

2005; Olsen et al., 2004). This is mainly due to diminishing marginal utility of health and the 'income effect', i.e. an increasing WTP takes a higher proportion of income, consequently decreasing the ability to pay (Flores and Carson, 1997; Smith, 2005).

A problem here arises since WTP estimates have been criticized for their insensitivity to scale, implying that willingness to pay does not vary 'meaningfully' with the quantity of the good on offer. Given that a non-proportional increase is expected theoretically, the *degree* of robustness must also be evaluated in order to allow more general claims about the validity of WTP estimates. In other words, finding a significant (and positive) coefficient of the health gain (or income) in a linear regression explaining the variance in WTP, usually termed "theoretical validity" (e.g. Ryan, 2004; Lienhoop and MacMillian, 2007), is a necessary yet not a sufficient condition for a more general claim about validity. Indeed, the appropriate *sign* does not necessarily imply that the associated variation is 'practically meaningful' or, as it also has been labeled, 'theoretically plausible' (Olsen et al., 2004), let alone that the estimates can be directly applied in decision making[2]. Results that cannot be shown to be theoretically invalid may still be considered practically irrelevant, therefore. Judging whether results are practically meaningful, i.e. whether the *size* of the coefficient is deemed appropriate, requires a normative judgment, one that is not directly informed by theory (Hammitt and Graham, 1999). This issue has not (yet) received much attention in the literature. It has, for instance, been suggested that WTP estimates for small risk reductions need to be near-proportional (increasing and strictly concave) to the size of the risk reduction (NOAA, 1993; Hammitt and Graham, 1999). Although perhaps somewhat restrictive, the condition of near-proportionality might thus be appropriate in establishing what practically meaningful (i.e. 'theoretically plausible') refers to.

Although the validity of WTP estimates for goods other than health gains attracted considerable attention[3], there has been only a limited in-depth empirical interest in the validity of WTP for changes in health *per se* (Olsen et al., 2004; Smith, 2005; Yeung et a, 2003), in particular when health is expressed in terms of QALYs. (A notable exception is the recent paper by Pinto Prades et al., 2009.) However, given the commonness of using the QALY as a measure of health gains, and the increased interest in the monetary value of QALYs, such studies appear warranted.

Our study contributes to the literature by extensively exploring the validity of WTP per QALY estimates, using a data set explicitly designed both to estimate the WTP for a QALY *and* to test the various aspects of the validity of these estimates (Bobinac et al., 2010). We define validity in terms of "construct validity" (Jakobsson and Dragun, 1996) which encompasses scale sensitivity of WTP[4] and the related sub-additive impartiality (or 'part-whole' bias). The latter refers to a situation in which two goods are valued higher when valued separately than when valued jointly. The design of the study allowed validity testing along both dimensions of the QALY measure (i.e. length and quality of life), both across and between samples and on aggregate and sub-group level, thus allowing accounting for the underlying heterogeneity in preferences. Particular hypothesis are described in more detail in section 2 and tested in section 3. The implications of the results, also for the

[2] Although the issue of validity is not the only issue precluding WTP estimates to be directly applied in policy making.

[3] See for example Desvousges et al., 1993; McFadden and Leonard, 1993; Jones-Lee et al., 1995; Carson and Mitchell, 1995; Frederick and Fischoff, 1998; Kahneman et al., 1999; Hammitt and Graham, 1999; Norinder et al., 2001; Smith, 2001; Van Exel et al., 2006; Van Houtven et al., 2006; Smith and Sach, 2009; Baker et al., 2010.

[4] Since the QALY is the concept of interest, we will not address the sensitivity to scope in this paper (i.e. sensitivity to a range of goods on offer).

application of WTP studies in estimating the monetary value of QALY gains in the context of determining an appropriate cost-effectiveness threshold, are discussed in section 4.

2. THE TEST

This study uses a data set obtained in a representative sample of the Dutch population, designed to estimate the WTP for a QALY from the individual perspective under certainty (Bobinac et al., 2010) and to test the construct validity of WTP per QALY estimates. The web-based questionnaire was administered in October 2008. Participants did not receive direct monetary compensation, but a small sum was donated to a charity of their choice upon completion of the questionnaire. More information about the design, the data collection and the main findings in terms of average WTP per QALY estimates can be found in Bobinac et al. (2010).

The construct validity of WTP per QALY estimates was tested through the following hypotheses:

Hypothesis I: WTP is sensitive to scale in terms of quality of life. That is, for a given duration, a larger improvement on the QALY scale should result in an increase in WTP, both between and within samples. This increase will be evaluated both in terms of theoretical validity and practical meaningfulness.

Hypothesis II: WTP is sensitive to scale in terms of duration. That is, for a given gain on the QALY scale, a longer duration of this gain should result in an increase in the WTP. This increase will be evaluated both in terms of theoretical validity and practical meaningfulness.

Hypothesis III: Average-level data hide the underlying heterogeneity in preferences in different, relevant subgroups and thus the underlying sensitivity in WTP.

Hypothesis IV: WTP per QALY estimates are affected by the sub-additivity bias. That is, the value of two smaller QALY gains valued separately is expected to add up to more than when these gains are valued jointly.

A significant difference between the WTP for smaller and larger gains is a necessary condition for establishing construct validity, but it need not be a sufficient one. A more definite test would be disproving the sub-additivity bias, i.e. finding a “near-proportional” relationship between the WTP estimates and the size of the health gains on offer and thus establishing a (near)additive relationship between them.

Survey instrument

In the survey, respondents were randomly assigned to one of 10 blocks consisting of five scenarios. Each scenario presented two health states described using the EuroQoL-5D descriptive system (or EQ-5D; The EuroQol Group, 1999). A total of 42 different health states were paired into 29 scenarios that were, with some overlap, assigned to one of the 10 blocks. These health states were chosen in such a way that they represented a fair spread of QALY gains across the utility range. The majority of the pairs were originally applied when deriving the UK tariffs for the EQ-5D (Kind et al., 1998) and 16 out of the 29 pairs were applied in deriving the Dutch tariffs (Lamers et al., 2006). Four of the five scenarios from each block were relevant for this study as they were specifically designed and combined to test the construct validity of WTP per QALY estimates. The blocks of scenarios were constructed such that the first two scenarios in each block represented smaller health gains that, according to Dutch EQ-5D tariffs, added up to a larger health gain presented in the third scenario

(Table 1). Health gains in different scenarios purposefully started either low or in the middle of the QALY scale, therefore ending either in the middle or high on the scale. In each block, one of the scenarios was repeated as the fourth scenario, but then with a longer duration (i.e., 3 or 5 years instead of 1 year; see the right-hand side of Table 1). The combinations of health states and duration were chosen to ensure comparability across blocks.

Table 1 here

Given that the difference in valuation between the two states per scenario indicated the size of the QALY gain on offer, it was important to obtain respondents' health state valuations, using a VAS scale. In the survey, respondents were first asked to rate their current health, death and perfect health on the VAS. Next, they were assigned to one of the 10 blocks, randomly offered the scenarios from this block and in each scenario asked to indicate which of the two health states they thought was better. The two health states had to be rated on the VAS showing previous valuations of current health, death and perfect health, providing a context to VAS valuations. Respondents were then instructed to imagine being in the better health state and to indicate their WTP for avoiding spending one year in the health state they had chosen as the worse. This health loss, i.e. the difference between the better and the worse health state, could be avoided by taking a painless medicine of unspecified properties[5] once a month, for which one had to pay out-of-pocket in 12 monthly installments. WTP was elicited in a two-step procedure: a payment scale (PS) (Donaldson et al., 1995; Donaldson et al., 1997; Olsen and Donaldson, 1998) was offered, followed by a bounded 'open-ended' (OE) question. In the first step, respondents were presented with an ordered low-to-high payment scale of monthly installments[6] and asked to indicate the amount they would certainly pay and the amount they would certainly not pay (Donaldson et al., 1997). Together, the two answers provide information regarding the range of values for which people were uncertain (Dubourg et al., 1997). In the second step, respondents were given a bounded 'open-ended' follow-up question, asking them to indicate the maximum amount they would pay if asked to do so right now, within the boundaries determined by the amounts they had indicated to certainly pay and certainly not pay in the first step. The two step approach was applied to arrive at a more precise estimate of the maximum WTP and add information and potentially robustness to our findings since the respondents used two different valuation techniques within one questionnaire. Such a combination of two WTP questions, although in the context of a bidding game, was applied before (e.g. Bhatia and Fox-Rushby, 2003; Cameron and Quiggin, 1994). The benefit of employing two different WTP formats, although in a context of separate WTP questions, was investigated by Johnson et al. (2000).

Attention was also given to reducing the hypothetical bias inherent to contingent valuation exercises, through *ex ante* and *ex post* mitigation (Blomquist et al., 2009). *Ex ante*, respondents were reminded to take their household income into consideration when solving the exercise (NOAA, 1993). Moreover, the visual image of health states rated on the VAS remained present on the right-hand side of the survey screen, as a reminder of the size of the health gain being valued. *Ex post*, respondents were asked on which element of household spending they would economize in order to be able to pay for the health gain[7] (NOAA, 1993) and to indicate

5 The vehicle of health improvement was only described as "painless medicine" in order to remove any possible contamination of the health gain evaluation according to the means by which that improvement would be brought about (Smith, 2001).

6 In €: 0, 10, 15, 25, 50, 75, 100, 125, 150, 250, 300, 500, 750, 1,000, 1,500, 2,500

7 Answer options included: (i) food; (ii) clothing; (iii) entertainment; (iv) sport; (v) savings; (vi) charity; (vii) other) (Smith, 2006)

how sure they were about their indicated WTP[8]. Finally, when respondents choose €0 as their maximum WTP, they were asked to indicate the reason behind this choice[9].

The questionnaire was pilot-tested in a random sample of 100 respondents in order to determine the plausibility and clarity of the tasks, the feasibility of the full questionnaire and the appropriateness of the range of the payment scale. The respondents had several opportunities to express their opinion about the tasks at hand. The pilot showed that the spread of the payment scale was not optimal; the initial scale encompassed three value categories above €2,500 (i.e. €5,000, €7,500 and €10,000), that were never chosen. To avoid loss of information and possible anchoring to exaggeratedly high values, the maximum was changed to €2,500 and additional value categories were added to the scale around the most frequently chosen values.

Analyses of hypothesis

The design of the scenarios (i.e. health state descriptions) was based on the EQ-5D descriptive system. The utility weights attached to each health state originated either from the available Dutch tariffs (Lamers et al., 2006) or the sample-specific VAS scores obtained from the valuations in the questionnaire (labeled "raw" scores). VAS scores were calculated in a rescaling procedure that included the mean scores of perfect health and death and the gains in QALYs, customarily, as the difference between the weights of the two health states.

The distributional properties of WTP estimates were analyzed using Kurtosis and Shapiro-Wilk tests for normality. The hypotheses were tested using both the parametric t-test on log-transformed WTP estimates and the non-parametric Mann-Whitney u-test (Yeung et al., 2003). Particular attention was paid to testing of the income variable and its effect on the WTP per QALY estimates. Statistical analyses were performed in STATA version 10.

Sensitivity to scale in terms of quality of life (*Hypothesis I*) was examined both between and within blocks. Between blocks, the statistical differences between WTP estimates for differently sized gains was tested under the premise that different samples elicit similar WTP for similar gains and statistically different WTP estimates for gains of different sizes. (An example of such a test is the comparison between scenarios 1 across blocks, as shown in Table 1.) The within-block tests were performed to reveal whether respondents, when faced with consecutive health gains differing in size, assigned a (meaningfully) higher WTP to the higher gain. In terms of Table 1, it was tested whether: scenarios 1 and 2 yielded similar WTP estimates for similar-sized health gains; scenarios 1 and 2 yielded lower and statistically different WTP estimates than those obtained in scenario 3; and whether possible differences between WTP for smaller gains (in scenarios 1 and 2) and larger gains (in scenario 3) were more pronounced on subgroup levels. Because the scenarios were presented consecutively to respondents - thus potentially drawing attention to the size of the health gain - one could expect the within-sample tests to be more likely to detect a (meaningful) increase in WTP estimates, given the increase in health gains, than the between sample tests (Kahneman et al., 1999; Hammitt and Graham, 1999; Olsen et al., 2004).

8 Answer options included: (i) totally sure; (ii) pretty sure; (iii) neither sure nor unsure; (iv) not very sure; (v) unsure.

9 Answer options included: (i) I am unable to pay more than €0 (ii) avoiding the worse health state and remaining in the better health state in not worth more than €0 to me; (iii) I am not willing to pay out of ethical considerations; and (iv) something else (open text field for explanation); options (i) and (ii) were considered as a "true zero" WTP, options (iii) and (iv) as a "protest zero" answer.

Sensitivity to scale in terms of duration (*Hypothesis II*) was examined within blocks. Health gains of longer duration were expected to be valued (meaningfully) higher (Table 1).

The sensitivity to scale was further examined in subgroups of respondents with different (i) levels of household income[10] and (ii) reported levels of certainty in the WTP answers (Johannesson et al., 1999), thus addressing *Hypothesis III*. Subgroup analysis was performed both in within-sample and between-sample tests. With respect to income, the sensitivity to scale was expected to increase with the absolute level of income but decrease with the proportion of income sacrificed, regardless of its absolute level, thus disclosing the "income effect". We tested whether either explanation was significant in our study. In terms of expressed certainty, we expected that higher levels of certainty would yield more sensitivity to scale, due to more reasoned responses.

Sub-additivity (*Hypothesis IV*) was examined by comparing WTP for the health gain in scenario 3 of each block with the sum of WTP estimates for the health gains in scenarios 1 and 2 (Table 1). (Recall that the two smaller gains added up to the larger gain in terms of Dutch EQ-5D tariffs.) We expected the sum of the value assigned to two smaller gains (i.e. the first two scenarios) to exceed the value assigned to the summed gain in scenario 3. In additional analyses, the individuals were assigned to one of three sub-additivity categories: positive "scope" (the sum of the values assigned to the parts exceeded the value assigned to the whole), negative "scope" (the sum of the values assigned to the parts was lower than the value assigned to the whole), and neutral "scope" (the sum of the values assigned to the parts equaled the value assigned to the whole).

The validity of the WTP estimates were further explored with multivariate regressions, using the QALY gain as the independent variable, along with the most common socio-economic factors, such as income.

Finally, the results were tested for specific framing effects. First, we examined whether an order effect was present, by considering the strength of correlation between the mean WTP estimates across all scenarios and the WTP assigned in the scenario first presented to a respondent. Second, a simple test for the presence of a learning effect was conducted by inspecting if sensitivity to scale improved in scenarios presented later in the questionnaire, i.e. after the respondents had already gathered some experience by answering to the first scenarios.

3. RESULTS

In total, 1,091 respondents, representative of the Dutch population according to age (18-65 years), gender and education, participated in the survey. The participants were predominately married, employed and in very good health. On average, 2.44 people shared an average net monthly household income of €2,564, adequately representing the Dutch national figures for 2008 (CBS, 2009). Descriptive statistics for the population are given in Table 2.

Table 2 here

[10] The income groups were defined according to the distribution of income, such that the poorest (below 1,000 € of household income) and the richest group (above 3,500 € of household income) made 13 % and 12% of the sample, respectively, while two middle groups made 35 and 40% of the sample.

Overall, the QALY gains based on sample-specific VAS scores were somewhat lower than those based on Dutch EQ-5D tariffs, with one notable exception (scenario 2 in blocks 5-7, Table 3). Most scenarios were designed such that one health state was unambiguously better than the other. It was tested and confirmed that the better health states received higher average VAS valuations. Respondents reversed the ranking (i.e. valuing the worse health state higher on the VAS) in less than five percent of scenarios. However, the correlation between EQ-5D tariffs and sample-specific VAS scores was relatively low (r=0.24, p=0.02). Although the average ratio between the QALY gains based either on existing tariffs or the VAS scores was 0.97, Table 3 shows that the dispersion of estimates around the ratio of 1 is considerable. Since the VAS QALY gains represent the estimates that respondents themselves provided and subsequently valued through the WTP exercise, these estimates will be used henceforth.

Table 3 here

The results of tests for sensitivity to scale in terms of quality of life (*Hypothesis I*) are presented in Table 3 as well. With respect to within-sample analysis, the parametric and non-parametric tests revealed no statistical difference between mean WTP estimates for gains of comparable size in scenarios 1 and 2 in all blocks. Although not statistically significantly different, the gains situated lower on the scale systematically received a higher WTP relative to similarly-sized gains positioned higher on the scale. This may signal that a similar health gain is considered more valuable when attained low on the QALY scale.

When testing sensitivity to scale by comparing the valuations in scenarios 1 and 2 with those obtained in scenario 3 (representing a considerably higher gain), a (marginal) statistically significant difference between WTP for smaller and larger gains was only observed in blocks 8-10 (p=0.1 for scenario 1 vs. 3). In terms of practical meaningfulness (or 'theoretical plausibility'), however, the increase in WTP appears implausibly low when compared to the increase in the health gain on offer. Indeed, a gain increase of 50% (i.e., from 0.310 in scenario 1 to 0.442 in scenario 3) resulted in an increase in WTP of no more than 7 Euro per month (+3.6%). In other blocks, no significant differences between the values obtained in scenarios 1 and 2 and those in scenario 3 were detected. Non-parametric tests did not detect any statistical difference between scenarios 1 and 2 compared to scenario 3 across blocks.

As indicated, in the context of *Hypothesis I*, it was explored whether on subgroup level the differences between the valuations obtained in the different scenarios would be more pronounced. The results indicate that respondents with a higher level of certainty regarding their answers exhibited only marginally higher level of sensitivity to scale: statistically significant difference was observed between scenario 1 and scenarios 2 and 3 in blocks 8-10 (p=0.02 and 0.01, respectively) and scenario 1 and 3 in blocks 5-7 (p=0.05). These respondents were on average younger respondents (p=0.00), in better health (p=0.00), more often employed (p=0.00) and devoted more time to solving the questionnaire (p=0.00). No difference in sensitivity to scale was detected between respondents belonging to different income groups (this issue is further explored below).

Between-sample tests revealed no statistical differences in WTP for similar gains in scenarios 1 and 2, across blocks (Table 3). For a range of gains between 0.268 and 0.348, the WTP ranges from 150 to 167 Euro a month. One may interpret this result, also in the relation to the previous results, in different ways. First, it might be encouraging that different samples, when presented with gains of similar size, elicit similar WTP estimates.

On the other hand, one may feel that a fluctuation of €17 per month between groups as compared to a fluctuation in health gain of 0.08 QALY indicates an insensitivity to scale. (The highest versus the lowest health gains amounted to a difference of 29.9%, the corresponding increase in WTP was 11.3%). The between-sample tests, however, revealed a significant difference between WTP estimates for health gains in scenarios 3 across blocks, although these were also comparable in size. This result prompted additional investigation, since it might be driven by income constraints, given that the gains in scenarios 3 were relatively large. However, this was not the case. First, the majority of WTP-to-income ratios was fairly low (mean was 7.4%, median was 3.6%). Second, those respondents who elicited bids corresponding to an above average WTP-to-income ratio, and therefore approached the income constraint more closely, did not exhibit lower variability in WTP estimates than other respondents (Smith, 2005). Third, the ratios between WTP for larger gains in scenario 3 and the smaller gains in scenario 1 in all blocks, both for respondents with an income below the mean and median and those above, were comparable in size (1.10 for low income respondents and 1.28 for high income respondents). If budget constraints would have played a substantial role, the ratio should be considerably larger for higher incomes, because their ability to express a 'true' WTP is less constrained (Pinto Prades et al., 2009). A modest difference in ratios (especially considering that it was not corrected for other variables such as education or age) gives little reason to expect that budget constraints were a main driver of our results.

Sensitivity to scale was tested with respect to the duration of the effect (*Hypothesis II*). Table 4 shows that WTP values do increase when the duration increases but the increase is mostly statistically insignificant (p=0.09 or higher), both when using non-parametric or parametric tests and in both subgroups. Apparently, respondents were able to assign similar values to similarly sized gains in two different scenarios (even though they were not specifically informed about the equality of the health states) but they failed to assign significantly higher values to benefits when these last longer. For instance, in Block 1 respondents valued a health improvement of similar size in scenarios 2 and 4 (gains on the VAS of 0.224 and 0.220, respectively) and elicited almost equal mean WTP for both gains (194 and 196 Euro, respectively). However, the duration of the gain in scenario 4 is three times longer than in scenario 2 (Table 4) and, therefore, the total (undiscounted) gain in scenario 4 is three times higher than in scenario 2. An increase in WTP of 2 Euros seems, besides being statistically insignificant, practically meaningless (i.e. theoretically implausible) as well, adding to the negative evidence on sensitivity to scale of WTP per QALY estimates.

Table 4 here

The existence of a sub-additivity bias was confirmed (*Hypothesis IV*). Note that in all blocks scenario's 1 to 3 used three distinct health states. If we label these states, ordered from lowest to highest ranked health state A, B and C, scenario 1 valued the distance from A to B, scenario 2 from B to C and scenario 3 from A to C. Since we randomized the order of the different scenarios, the three health states were all valued two times (i.e. in each scenario in which they appeared). If the valuations of the health states would be equal in both valuations, the distance on the VAS from A to C in scenario 3 should equal the sum of the distance between A and B in scenario 1 and B and C in scenario 3. As can be seen in Table 3, this was not the case. In scenario 3, the distance between A and C was smaller than the sum of A to B and B to C (in scenarios 1 and 2). (For example, 0.348 + 0.268 is more than 0.467 in Block 1 – 4, Table 3). This difference was not caused by different VAS valuations of the highest (C) and lowest (A) health states (which were valued almost identically both times), but

especially caused by a difference in valuation of the middle health state (B) in scenarios 1 and 2 (on average getting a VAS score of 0.42 and 0.55). This may be caused by a combination of ceiling effects and the wish to clearly differentiate between health states on the VAS scale.

Analyzing the WTP for the two smaller gains makes clear that even considering the fact that the VAS gains valued in scenarios 1 and 2 may exceed the VAS gain valued in scenario 3 (even though the two extreme health states were (valued) identical in the different scenarios), there is clear evidence for a sub-additivity bias. For instance, considering Block 1 – 4 (Table 3), the WTP for the integral gain between the two extreme health states in scenario 3 (174 euro) is far exceeded by the valuations of the smaller gains in scenario 1 and 2 (which were 150 and 170 Euro, respectively). Here, the valuations of the parts add up to 320 euro, while the whole is valued at 174. On an individual level, over 90% of all respondents indicated a positive "scope" and only 28 respondents (2.5%) indicated neutral "scope" at one or more occasions., so that the valuations of the parts add up to the valuation of the whole.

Multivariate regression models explaining WTP estimates are presented in Table 5[11]. The signs of coefficients aligned with *a priori* expectations (Olsen et al., 2004; King et al., 2005). The size of coefficients, however, indicated a non-proportional relationship between WTP and the size of the health gain. Previous findings rarely showed income elasticity of WTP higher than one (Flores and Carson 1997), which is what we also find in the current study (0.05 and 0.06 in Models 1 and 2, Table 5), implying that a 10% increase in household income results, *ceteris paribus*, in less than one percent increase in WTP. This is largely in line with the other findings presented. The relationship between the main variables remained stable when introducing other variables in the regression. These results emphasize that while the relationship between WTP and size of the health gain (both per year and in terms of duration) is in the expected direction (i.e. 'theoretically valid'), the size of the coefficients raise important questions about the practically meaningfulness or 'theoretical plausibility' of the results.

Table 5 here

Median values were (predominately) independent of the size of the gain (Norinder et al., 2001) We tested for learning effects as well, but found no indication that respondents became more sensitive to changes in the size of the gain in consecutive exercises. Finally, we found no evidence of an ordering bias in our study.

4. DISCUSSION

Depending on the normative framework and decision rules adopted (Claxton et al., 2010), the monetary value of a QALY can be seen as representing the appropriate cost-effectiveness threshold. If not directly informative in that context, estimates of the monetary value of QALYs may at least provide an opportunity for a public discourse on health care limits and decisions (Gyrd-Hansen, 2005; Weinstein, 2008). However, before available estimates can be considered useful for decision making or even public debate, it is necessary to provide convincing evidence regarding their validity. Here, we have empirically explored the construct validity of

[11] All variables were initially tested for normality of the distribution and log-transformed if the distributions were found to be non-normal.

individual WTP per QALY estimates using a large scale study (Bobinac et al., 2010), designed to obtain monetary values for health gains and to explicitly test several aspects of validity.

Overall, our results lend relatively little support to the validity of WTP per QALY estimates. The relationship between WTP and QALYs was clearly not 'nearly proportional' (NOAA, 1993; Hammitt and Graham, 1999). In fact, only sporadic statistically significant differences between WTP estimates for smaller and larger health gains were found. Our results largely confirm the existence of a sub-additivity bias. This result was not unexpected (although relatively new in this specific context), but the magnitude of the differences between the sum of two smaller gains and the integral gain is indeed considerable and raises questions regarding how to infer an accurate value for a health gain. This is compounded by the fact that our results are in line with previous studies in this area, also reporting a non-linear and often insensitive relationship between WTP and QALY gains (even for changes that might be considered marginal, Smith, 2005; Olsen et al., 2004; Pinto Prades et al., 2009).Budget constraints do not appear to have driven the construct invalidity (even though the health gains on offer cannot be considered marginal).

Larger scale sensitivity, in terms of the statistical difference between WTP estimates for smaller and larger gains of *equal* duration was found in the subgroup of respondents who indicated a higher level of certainty in their WTP values. This reveals the importance of sub-group analyses as sample-average sensitivity tests might lead to over-restrictive conclusions regarding validity (e.g. Heberlein et al., 2005). However, the extent of the scale sensitivity improved only marginally in this subgroup. If, alternatively, much larger sensitivity was found among more certain respondents (given also that their estimates were found to approach real WTP more closely, e.g. Johannesson et al., 1999; Blumenschein et al., 2001), it would signal a larger practical usefulness of these estimates.

The lack of validity observed in this study could stem from several sources, among others from design-related problems (in spite of careful consideration of various aspects of study design and pilot-testing) and from problems in the properties of and the relationship between WTP and QALY.

With respect to the former, the current study employed a web-based questionnaire and it may be hypothesized that such a format engages respondents less than face-to-face interviews or even postal questionnaires, resulting in increased use of heuristics and problems with "believing" the questionnaire. In the setting, the EuroQoL-5D description system might not be enough for respondents to fully appreciate the severity of health states within a web-based questionnaire. Since this study was not repeated, the quality (or reliability) of the contingent market could not be tested further, although it efforts were made to reduce the problem of "non-realism" though *ex post* and *ex ante* mitigation. A partial solution to invalidity caused by such procedural problems would be to (even more) strongly emphasize the differences in outcomes (Arrow, 1993; Corso et al., 2001). Providing respondents with an opportunity to form more stable preferences by restating these in repeated interviews is an alternative option. Heberlein et al. (2001), Macmillan et al. (2006) and Tisdell et al. (2008) are among many who found that WTP changed significantly when respondents were given additional information and time to think about an unfamiliar environmental good thus better forming their preferences. In fact, if knowing and thinking more about the good in question leads to more valid results then researchers might also consider using a more health literate population in future WTP studies. This issue raises an important, more general dilemma regarding the validity of WTP estimates. It may well be that the validity of WTP

responses may be better in some subgroups (such as a group of respondents more certain in their WTP estimates) than in a representative sample of the general population. Would it then be appropriate to focus on such sub-groups (loosing representativeness) or on the representative sample (loosing validity)?

With respect to the problems in the properties and the relationship between WTP and QALY, there are three noteworthy issues to rise. First, this study used the VAS as a means of health state valuation. Not only has the relevance of VAS valuations sometimes been disputed (e.g. Parkin and Devlin, 2006), in this study it may have resulted in some noteworthy response patterns. Due to randomization of the scenarios (causing some respondents to start with scenario 3, rather than 1 or 2), all respondents had to value the three health states twice. As mentioned, the VAS valuation of the two extreme health states (the lowest one in scenarios 1 and 3 and the highest one in scenarios 2 and 3) were nearly identical at the two valuation moments. However, the middle health state was valued clearly differently between the two moments. Perhaps respondents attempted to clearly differentiate between the health states on the VAS while simultaneously being influenced by ceiling effects of the VAS. The latter implies that shifting the extreme health states further downward or upward was not an option, so that that placement of the middle state had to be shifted between scenarios 1 and 2. This implies that our testing of the sub-additivity was imperfect in terms of indicated utility distance (although clearly correct in terms of the described health gain). Although important to note, in the context of the main aim of this study, this result does not seem to have caused the observed insensitivity. It may, however, have contributed to the observed sub-additivity bias, as also the largest VAS gain (in scenario 3) was lower than the sum of the two smaller VAS gains (in scenarios 1 and 2). It may also have influenced the relative valuations of QALY gains high and low on the scale, to some extent. Second, WTP may measure broader outcomes than the QALY does, and therefore, a (nearly) proportional relationship between the two quantities need not exist. Given that there may be more associated with improved health than health-related utility, e.g. income changes, the relationship between WTP and QALY gains may behave seemingly unexpectedly. It seems unlikely, however, that such considerations could explain the results presented here in a convincing manner. It is hard to see for instance how this fact would explain that an additional 0.2 QALY gain would yield a 20 Euro increase in monthly WTP in a sample of the general public in the Netherlands. In fact, it would rather be expected to induce an increased sensitivity of WTP per QALY estimates for (some ranges of) health gains. Third, there is a difference between health state valuations focusing on valuing health *states* (after which the size of a health gain is calculated by comparing the value of the health states involved) and WTP valuations, which directly value a health *change*. In contrast to a health state valuation exercise, in a WTP exercise respondents thus have the opportunity to consider both the origin and destination of the gain, potentially influencing their valuation (Weinstein et al., 2009). It seems that using different procedures to value changes thus may hamper the comparison of methods and influence the observed validity of WTP per QALY estimates. This issue may be related to recent debates regarding the optimal way of valuing health improvements (Nord et al., 2008; Drummond et al., 2009). Moreover, differences between valuation methods could potentially arise if the properties underlying the conventional QALY model do not hold. It has been suggested, therefore, to consider non-linear specifications of the QALY model in future research, in order to explore whether this adds to validity of findings (Pinto Prades et al., 2009). Although important, the results presented here seem not easily explained solely by violations of the QALY model.

This study has emphasized that caution is required in considering outcomes of WTP per QALY estimates. While we do not wish to imply that practically meaningful WTP estimates cannot be found, our findings provide an indication of the type and the extent of problems that similar studies may face and the depth of inquiry required in order to make claims about the validity of results. Theoretical validity is a relatively undemanding requirement, but insufficient to demonstrate construct validity. If sound estimates of WTP for health gains are sought, it is pivotal to attempt to understand the insensitivity of these estimates as reported here, and, if possible, to unravel (and ideally counter) its causes. Whether this is possible has, in fact, been doubted (Kahneman et al., 1999, p.217)[12]. Further methodological analysis and testing thus seems to be necessary to investigate whether and how the CV-method can meaningfully inform health care decision making (Klose, 1999). Possibly, validity testing should become an integral part of piloting a CV study since at that stage there is still room for improvement.

For now, the theoretical validity and, especially, practical meaningfulness or theoretical plausibility of WTP estimates for health gains in general appear to be insufficiently demonstrated in order to consider current estimates useful for informing policy making or public debate. Future studies need to convince readers not only of the theoretical validity of the estimates of the value per QALY, but also of their construct validity, that is their theoretical plausibility. In that sense, it appears pivotal to pay more attention to these aspects in future studies in order to get more valid results: pay more, get more, therefore.

Acknowledgement: This study is part of a larger project investigating the broader societal benefits of health care, which was financially supported by Astra-Zeneca, GlaxoSmithKline, Janssen-Cilag, Merck and Pfizer BV. The researchers were free in study design; collection, analysis and interpretation of data as well as in writing and submitting the manuscript for publication. The views expressed in this paper are those of the authors.

REFERENCES

1. Arrow K, Solow R, Leamer E, Radner R, Schuman H. Report of the NOAA Panel on contingent valuation, Federal Register 1993; 58: 4601–4614.
2. Bateman IJ, Brouwer R. Consistency and construction in stated WTP for health risk reductions. A novel scope sensitivity test. Resource and Energy Economics 2006; 28: 199-214.
3. Bateman IJ, Jones AP. Contrasting conventional with multi-level modeling approaches to meta-analysis: Expectation consistency in U.K. woodland recreation values. Land Economics 2003; 79: 235-258.
4. Baker R, Currie GR, Donaldson C. What needs to be done in contingent valuation: have Smith and Sach missed the boat? Health Economics, Policy and Law 2010; 5:113-121
5. Bhatia MR, Fox-Rushby JA. Validity of willingness to pay: hypothetical versus actual payment. Applied Economics Letters 2003;10:737-740.

12 As Kahneman et al. (1999) state: *'Insensitivity to scope is the inevitable result of general rules that govern human judgment. It is naive to expect broad psychological laws to be overcome by minor methodological adjustments'* (p. 217).

6. Blomquist G, Blumenschein K, Johannesson M. Eliciting willingness to pay without bias using follow-up certainty statements: comparisons between probably/definitely and a 10-point certainty scale. Environmental and Resource Economics 2009; 43: 473-502.
7. Blumenschein K, Johannesson M, Yokoyama KK., Freeman PR. Hypothetical versus real willingness to pay in the health care sector: results from a field experiment. Journal of Health Economics 2001;20:441-457.
8. Bobinac A., van Exel NJA, Brouwer WBF, Rutten F. Willingness to pay for a QALY: the individual perspective. Value in Health 2010, In press
9. Bradford DF. Benefit–cost analysis and demand curves for public goods. Kyklos 1972; 23: 775–791.
10. Cameron AT, Quiggin J. Estimation Using Contingent Valuation Data from a "Dichotomous Choice with Follow-Up" Questionnaire. Journal of Environmental Economics and Management 1998; 35: 195-199
11. Carson R, Mitchell R. Sequencing and Nesting in Contingent Valuation Surveys. Journal of Environmental Economics and Management 1995; 28: 155-173.
12. CBS, http://www.cbs.nl/en-GB/menu/cijfers/kerncijfers/default.htm (accessed 21th of March 2009).
13. Corso P, Hammitt J, Graham J. Valuing Mortality-Risk Reduction: Using Visual Aids to Improve the Validity of Contingent Valuation. Journal of Risk and Uncertainty 2001;23, 165-184.
14. Claxton K, Paulden M, Gravelle H, Brouwer WBF, Culyer AJ. Discounting and decision making in the economic evaluation of health care technologies. Health Economics, forthcoming
15. Desvousges WH, Johnson FR, Dunford RW, Boyle KJ, Hudson SP, Wilson KN. Measuring natural resource damages with contingent valuation: tests of validity and reliability. In: J.A. Hausman, Editor, Contingent Valuation: A Critical Assessment, North-Holland, Amsterdam (1993), pp. 91–159.
16. Donaldson C, Shackley P, Abdalla M. Using willingness to pay to value close substitutes: carrier screening for cystic fibrosis revisited. Health Economics 1997;6:145–159.
17. Donaldson C, Shackley P, Abdalla M, Miedzybrodzka Z. Willingness to pay for antenatal carrier screening for cystic fibrosis. Health Economics 1995;4:439–452.
18. Drummond M, Brixner D, Gold M, Kind P, McGuire A, Nord E and Consensus Development Group (2009), Toward a Consensus on the QALY. Value in Health 2009;12:31–35.
19. Dubourg WR, Jones-Lee MW, Loomes G. Imprecise preferences and survey design in contingent valuation. Economica 1997;64:681-702.
20. Fisher AC, The conceptual underpinnings of the contingent valuation method. In: D.J. Bjornstad and J.R. Kahn, Editors: The Contingent Valuation of Environmental Resources: Methodological Issues and Research Needs, Edward Elgar, Cheltenham (1996), pp. 19–37.
21. Flores, N. and Carson, R. The relationship between the income elasticities of demand and willingness to pay. Journal of Environmental Economics and Management 1997 33; 287-295
22. Frederick S, Fischhoff B. Scope insensitivity in elicited valuations. Risk Decision and Policy 1998; 3: 109-123.
23. Gyrd-Hansen D. Willingness to pay for a QALY. Health Economics 2003;12:1049-1060.
24. Gyrd-Hansen D. Willingness to pay for a QALY: theoretical and methodological issues. Pharmacoeconomics 2005; 23: 423-432.
25. Hammitt JK, Graham JD. Willingness to Pay for Health Protection: Inadequate sensitivity to probability? Journal of Risk and Uncertainty 1999; 18: 32-62
26. Heberlein TA, Wilson MA, Bishop RC, Schaeffer NC. Rethinking the scope test as a criterion for validity in contingent valuation. Journal of Environmental Economics and Management 2005; 50: 1-22

27. Jakobsson K, Dragun J. Contingent Valuation and Endangered Species, Edward Elgar, Cheltenham, UK and Brookfield, US (1996)
28. Johannesson M, Blomquist GC, Blumenschein K, Johansson P, Liljas B, O'Conor RM. Calibrating Hypothetical Willingness to Pay Responses Journal of Risk and Uncertainty 1999; 18:21-32.
29. Johnson R, Banzhaf MR, Desvousges WH. Willingness to pay for improved respiratory and cardiovascular health: a multiple-format, stated-preference approach. Health Economics 2000;9:295-317.
30. Jones-Lee M, Loomes G, Philips P. Valuing the Prevention of Non-Fatal Road Injuries: Contingent Valuation vs. Standard Gambles. Oxford Economic Papers 1995; 47: 676-695.
31. Jorgensen BS, Syme GJ, Smith LM, Bishop BJ. Random error in willingness to pay measurement: a multiple indicators, latent variable approach to the reliability of contingent values, Journal of Economic Psychology 2004; 25: 41–59.
32. Kahneman D, Ritov I, Schkade D. Economic Preferences or Attitude Expressions?: An Analysis of Dollar Responses to Public Issues. Journal of Risk and Uncertainty1999;19:203-235.
33. Kind P, Dolan P, Gudex C, Williams A. Variations in population health status: results from a United Kingdom national questionnaire survey. BMJ 1998;316:736-741.
34. King JT Jr., Tsevat J, Lave JR, Roberts MS. Willingness to pay for a Quality-Adjusted Life year: implications for societal health care resource allocation. Medical Decision Making 2005;25:667-677.
35. Klose, T. The contingent valuation method in health care. Health Policy 1999; 47:97-123
36. Lamers LM, McDonnell J, Stalmeier PF. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Economics 2006;15:1121-1132.
37. Lienhoop N, MacMillan D. Valuing wilderness in Iceland: Estimation of WTA and WTP using the market stall approach to contingent valuation. Land Use Policy 2007; 24: 289-295
38. MacMillan D, Hanley N, Lienhoop N, Contingent valuation: environmental polling or preference engine, Ecological Economics 2006; 60: 299–307
39. McFadden D, Leonard G. Issues in the Contingent Valuation of Environmental Goods: Methodologies for Data Collection and Analysis. In: J.A. Hausman, Editor, Contingent Valuation: A Critical Assessment, North-Holland, Amsterdam (1993), pp. 91–159.
40. NOAA Panel, (http://www.cbe.csueastbay.edu/~alima/courses/4306/articles/NOAA%%20on%%20contingent%%20valuation%%201993.pdf.) (accessed 14th of April 2010).
41. Norinder A, Krister H, Ulf P. Scope and Scale Insensitivities in a Contingent Valuation Study of Risk Reductions. Health Policy 2001; 57: 141-153.
42. Nord E, Daniels N, Kamlet M. QALYs: some challenges. Value Health 2008;12 (Suppl. 1):10–5.
43. Olsen JA, Donaldson C, Pereira J. The insensitivity of 'willingness-to-pay' to the size of the good: New evidence for health care. Journal of Economic Psychology 2004; 25: 445-460
44. Olsen JA, Donaldson C. Helicopters, hearts and hips: using willingness to pay to set priorities for public sector health care programmes. Social Science and Medicine 1998;46:1–12.
45. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? Health Economics 2006;15:653-664.
46. Pinto Prades JL, Loomes G, Brey R. Trying to estimate a monetary value for the QALY. Journal of Health Economics 2009; 28: 553-562.

47. Ryan, M. A comparison of stated preference methods for estimating monetary values. Health Economics 2004; 13: 291-296
48. Smith RD. Sensitivity to scale in contingent valuation: the importance of the budget constraint, Journal of Health Economics 2005; 24: 519–529.
49. Smith RD. The relative sensitivity of willingness-to-pay and time-trade-off to changes in health status: an empirical investigation, Health Economics 2001; 10: 487–497.
50. Smith RD. It's not just what you do, it's the way that you do it: the effect of different payment card formats and survey administration on willingness to pay for health gain. Health Economics 2006;15: 281-293.
51. Smith RD, Sach TH. Contingent valuation: what needs to be done? Health Economics, Policy and Law 2010; 5: 91-111
52. The EuroQol Group. EuroQol – a new facility for the measurement of health related quality of life. Health Policy 1999; 16: 199–208.
53. Tisdell C, Wilson C, Nantha HS. Contingent Valuation as a Dynamic Process. Journal of Socio-Economics 2008; 37: 1443-1458
54. Van Exel NJA, Brouwer WBF, van den Berg B, Koopmanschap MA. With a little help from an anchor: Discussion and evidence of anchoring effects in contingent valuation. Journal of Socio-Economics 2006; 35: 836-853
55. Van Houtven G, Powers J, Jessup A, Yang J. Valuing avoided morbidity using meta-regression analysis: what can health status measures and QALYs tell us about WTP? Health Economics 2006; 15: 775-795.
56. Weinstein MC. How much are Americans willing to pay for a Quality-Adjusted Life Year? Medical Care 2008; 46: 343-345.
57. Yeung R, Smith RD, McGhee SM, Willingness to pay and size of health benefit: an integrated model to test for 'sensitivity to scale', Health Economics Letters 2003; 12: 791–796.

Table 1: Design of scenarios

	Scenarios 1-3							Scenario 4: Pairing and duration in Blocks									
	Scenario	**HS1**	**HS2**	**HS1 (tariff)**	**HS2 (tariff)**	**QALY gain**	**Duration in years**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Blocks 1-4	1	21312	12111	0.478	0.847	0.369	1			3	5**						
	2	22323	21312	0.109	0.478	0.369	1	3*	5								
	3	22323	12111	0.109	0.847	0.738	1										
Blocks 5-7	1	12311	11211	0.556	0.897	0.341	1							3			
	2	32311	12311	0.395	0.556	0.161	1					3	5				
	3	32311	11211	0.395	0.897	0.502	1										
Blocks 8-10	1	11312	11211	0.514	0.897	0.383	1								3		
	2	11332	11312	0.185	0.514	0.329	1									3	5
	3	11332	11211	0.185	0.897	0.712	1										

Note: * The "3" indicates that respondents in Block 1 (see column in table), in addition to the three scenarios listed in columns 2-4 of the table, evaluated a fourth scenario, which was identical to scenario 2 (see row in table) in terms of the health states presented but differed in terms of duration (i.e. 3 years instead of 1 year). ** The "5" indicates that respondents in Block 4 (see column in table), in addition to the three scenarios listed in columns 2-4 of the table, evaluated a fourth scenario, which was identical to scenario 1 (see row in table) in terms of the health states presented but differed in terms of duration (i.e. 5 years instead of 1 year).

Table 2: Summary statistics (n=1,091)

Variable	Mean	St. Dev	Min	Max
Age	42.1	12.1	18	65
Sex (% male)	0.47	0.50		
Marital status:				
- Married (% yes)	0.61	0.49		
- Divorced (% yes)	0.10	0.31		
- Single (% yes)	0.24	0.43		
- Widow (% yes)	0.03	0.16		
- Unknown (% yes)	0.02	0.14		
Children (% yes)	0.56	0.50		
- Number of children (n=3070)	2.23	10.1	1	10
Income (mean/median €)	2,564/2,499			
Income groups:				
- group 1 (% <1000€)	0.13	0.33		
- group 2 (% >999 & <2000€)	0.34	0.48		
- group 3 (% >1999 & <3500€)	0.40	0.49		
- group 4 (% >3499€)	0.12	0.33		
Number of people living on household income	2.44	10.4	1	20
University education (% yes)	0.36	0.48		
Employment status:				
- Employed (% yes)	0.62	0.48		
- Unemployed (% yes)	0.17	0.38		
- Student (% yes)	0.06	0.25		
- Housewife/husband or retired (% yes)	0.14	0.35		
Health status:				
- EQ-5D (Dutch tariff)	0.84	0.22	-0.26	1.00
- EQ-VAS	78.5	170.1	0	100
Suffering a chronic illness (% yes)	0.39	0.94		
Subjective life expectancy	81.9	11.2	30	120
Completion time of the questionnaire	18.8	60.13	9	61
Levels of certainty (%):				
- totally sure	14.4			
- pretty sure	41.7			
- neither sure nor unsure	32.9			
- not very sure	8.0			
- unsure	3.0			

Table 3: Results of the scale sensitivity test: scenarios with health gains of different size & equal duration

	Scenario	n	HS1 (tariff)	HS2 (tariff)	QALY gain (tariff)	Valuation using VAS (rescaled)				WTP (month)	
						HS1	HS2	QALY gain*	Ratio**	Mean (sd)	Median
Blocks 1-4	1	444	0.478	0.847	**0.369**	0.412	0.719	**0.348**	0.94	150 (319)	75
	2	444	0.109	0.478	**0.369**	0.305	0.530	**0.268**	0.73	170 (349)	75
	3	444	0.109	0.847	**0.738**	0.303	0.737	**0.476**	0.64	174 (358)	75
Blocks 5-7	1	329	0.556	0.897	**0.341**	0.402	0.728	**0.337**	0.99	167 (318)	100
	2	329	0.395	0.556	**0.161**	0.262	0.570	**0.340**	2.11	178 (320)	100
	3	329	0.395	0.897	**0.502**	0.256	0.743	**0.496**	0.99	197 (366)	100
Blocks 8-10	1	318	0.514	0.897	**0.383**	0.465	0.748	**0.310**	0.81	167 (353)	75
	2	318	0.185	0.514	**0.329**	0.323	0.570	**0.294**	0.89	196 (360)	100
	3	318	0.185	0.897	**0.712**	0.319	0.719	**0.442**	0.62	203 (385)	100

*Gains were calculated for each individual and then averaged; therefore the difference between two gains on average is not equal to the average gain presented in this column. ** Ratio of QALY gain estimated using VAS to QALY gain by tariff.

Table 4: Results of the scale sensitivity test: scenarios with health gains of equal size & different duration

	Scenario	n	HS1 (tariff)	HS2 (tariff)	QALY gain	Duration	Valuation using VAS (rescaled)			WTP (month)	
							HS1	HS2	QALY gain	Mean (sd)	Median
Block 1	2	113	0.109	0.478	**0.369**	1	0.29	0.514	**0.224**	194 (460)	80
	4					3	0.284	0.504	**0.22**	196 (447)	70
Block 2	2	110	0.109	0.478	**0.369**	1	0.3	0.54	**0.24**	127 (157)	75
	4					5	0.311	0.53	**0.219**	169 (301)	75
Block 3	1	110	0.478	0.847	**0.369**	1	0.41	0.751	**0.341**	168 (308)	80
	4					3	0.41	0.734	**0.324**	199 (368)	100
Block 4	1	111	0.478	0.847	**0.369**	1	0.424	0.71	**0.286**	92 (90)	75
	4					5	0.418	0.734	**0.316**	107 (116)	70
Block 5	2	109	0.395	0.556	**0.161**	1	0.28	0.55	**0.27**	159 (206)	100
	4					3	0.29	0.546	**0.256**	158 (169)	120
Block 6	2	110	0.395	0.556	**0.161**	1	0.24	0.59	**0.35**	182 (360)	100
	4					5	0.255	0.552	**0.297**	211 (435)	100
Block 7	1	110	0.556	0.897	**0.341**	1	0.393	0.78	**0.387**	176 (350)	100
	4					3	0.44	0.8	**0.36**	146 (260)	100
Block 8	1	108	0.514	0.897	**0.383**	1	0.47	0.75	**0.28**	121 (210)	58
	4					3	0.48	0.75	**0.27**	129 (181	75
Block 9	2	109	0.185	0.514	**0.329**	1	0.32	0.59	**0.27**	202 (374)	100
	4					3	0.32	0.56	**0.24**	216 (374	120
Block 10	2	101	0.185	0.514	**0.329**	1	0.315	0.55	**0.235**	223 (400)	105
	4					5	0.296	0.56	**0.264**	249 (432)	101

Table 5: Multivariate clustered regression analysis

	Model 1			Model 2		
DV: log(WTP)	**Coef**	**R st**	**p**	**Coef**	**R st**	**p**
Log(health gain)	0.05	0.03	0.07	0.06	0.03	0.02
Log(duration)	0.03	0.02	0.04	0.04	0.02	0.02
Income groups:						
- group 1 (≤999)				-	-	-
- group 2 (999 & <2000)				0.37	0.12	0.00
- group 3 (1999 & <3500)				0.72	0.12	0.00
- group 4 (≥3500)				1.34	0.16	0.00
Number of people living on household income				-0.07	0.03	0.01
Log(age)				-0.34	0.11	0.00
Higher education				0.26	0.07	0.00
Gender (1 = female)				0.14	0.07	0.04
Intercept	4.52	0.05	0.00	5.22	0.4	0.00
		n=4,018			**n=4,018**	
		R2=0.01			R2=0.12	

Note: DV=Dependant variable; “R st” stands for Robust standard error; “p” stands for P>|t|