HEALTH ECONOMICS
*Health Econ.* (2011)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.1792

# LOWERING THE ‘FLOOR’ OF THE SF-6D SCORING ALGORITHM USING A LOTTERY EQUIVALENT METHOD

JOSÉ MARÍA ABELLÁN PERPIÑÁN*, FERNANDO IGNACIO SÁNCHEZ MARTÍNEZ,
JORGE EDUARDO MARTÍNEZ PÉREZ and ILDEFONSO MÉNDEZ

*Applied Economics Department, School of Economics, University of Murcia, Murcia, Spain*

## SUMMARY

This paper presents a new scoring algorithm for the SF-6D, one of the most popular preference-based health status measures. Previous SF-6D value sets have a minimum (a floor), which is substantially higher than the lowest value generated by the EQ-5D model. Our algorithm expands the range of SF-6D utility scores in such a way that the floor is significantly lowered. We obtain the wider range because of the use of a lottery equivalent method through which preferences from a representative sample of Spanish general population are elicited. Copyright © 2011 John Wiley & Sons, Ltd.

Received 27 May 2010; Revised 2 July 2011; Accepted 19 July 2011

KEY WORDS: SF-36; SF-6D; floor effect; standard gamble; lottery equivalent methods

## 1. INTRODUCTION

The Short Form 36 (SF-36) is one of the most widely used generic health-related quality-of-life measures, although it cannot be directly used in cost–utility analysis because it does not produce utilities. Bridging the gap between descriptive information of the SF-36 and population preferences is achieved by algorithms, which convert item responses into utility scores. Pickard *et al.* (2005) compared various algorithms, concluding that the ‘Brazier’s algorithms for the SF-12 and SF-36 appear to be most favourable’ (p. 8). To estimate these algorithms, Brazier and colleagues used a subset of SF-36 items, which were grouped in a six-dimensional measure called the SF-6D (Brazier *et al*., 2002; Brazier and Roberts, 2004). One of the reasons why Pickard *et al.* claimed the superiority of the SF-6D over other algorithms is because preference measurements were performed using a particular elicitation method—the standard gamble (SG), which has usually been regarded as the ‘gold standard’, for example, Torrance *et al.* (2001).

Despite the presumed superiority of the SF-6D, it is commonly held that the SF-6D suffers from a problem known as the ‘floor’ effect, that is, ‘that the instrument does not appear to describe health states at the lower end of the scale’ (Longworth and Bryan, 2003: p. 1066). This weakness of the descriptive system emerges in some studies, which report large numbers of patients at the bottom level of certain dimensions, particularly ‘role limitations’ (Longworth and Bryan, 2003; Brazier *et al*., 2004; Ferreira *et al*., 2008). Notwithstanding, the evidence is far from being generalizable to all studies, and some researchers have not observed a clear floor effect (Szende *et al*., 2004; Bryan and Longworth, 2005; Bharmal and Thomas, 2006; Fryback *et al*., 2010).

Although the term ‘floor effect’ is mostly reported as a drawback of the descriptive component of the SF-6D, it has sometimes been used with a somewhat different meaning to denote that the range of utilities generated by the SF-6D algorithm, that is, the SF-6D ‘value set’ or ‘tariff’, has a minimum (a floor), which is considerably

*Correspondence to: Departamento de Economía Aplicada, Facultad de Economía y Empresa, Campus de Espinardo, 30100 Murcia, Spain. E-mail: dionisos@um.es

Copyright © 2011 John Wiley & Sons, Ltd.

higher than the lowest value generated by the EQ-5D algorithm.[1] This narrower range of values causes SF-6D utilities to be higher than EQ-5D utilities for less healthy individuals (Longworth and Bryan, 2003; Brazier *et al*., 2004; Bryan and Longworth, 2005; Petrou and Hockley, 2005; Lamers *et al*., 2006; Buxton *et al*., 2007; Barton *et al*., 2008), meaning that the utility benefits of improving health conditions tend to be higher according to the EQ-5D than the SF-6D.

In the extent that the 'floor effect' is a shortcoming of the descriptive system, nothing can be performed except modify the instrument by rewording the level descriptions or adding lower levels to some of its dimensions (Brazier *et al*., 2004). Notwithstanding, evidence by Fryback *et al*. (2010) suggests that the 'floor' observed in the range of utilities may have to do more with the algorithm than with the descriptive deficiencies of the classification system itself and, thus, that this effect may be mainly a product of the valuation method. Tsuchiya *et al*. (2006) provide empirical support to the hypothesis that the valuation method explains part of the discrepancy between the EQ-5D and the SF-6D algorithms because EQ-5D utilities were elicited using the time trade-off (TTO) method instead of the SG.

In this paper, we argue that, at least in part, the 'floor' of the SF-6D range of values is caused by the type of valuation method on which the SF-6D algorithm is based. Our claim is based on the fact that the SG usually gives higher utility scores than other methods do, suggesting a degree of risk aversion that is so strong that they cannot be properly described by expected utility. Evidence of the extreme risk aversion raised by the SG includes studies performed with both monetary outcomes (Hershey and Schoemaker, 1985; Johnson and Schkade, 1989; Delquié, 1993) and health outcomes (Wakker and Deneffe, 1996; Bleichrodt *et al*., 2001, 2007).

Empirical evidence (Rutten-van Mölken *et al*., 1995; Bleichrodt, 2001; Oliver 2003) suggests that the natural tendency of the SG to inflate valuations is reinforced if the method is applied in a chained way, that is, by replacing the death of a first SG by another health state and next by valuing that state with respect to death by a second SG. This was the case of the specific SG implemented by Brazier *et al*. (2002). Apparently, this drawback is shared by other valuation methods, for example, TTO, whose chained versions also lead to higher values than their unchained applications (Oliver, 2005; Pinto and Abellán-Perpiñán, 2005).

This paper reports the results of the first study conducted in Spain to estimate a SF-6D algorithm for the SF-36. The main novelty of such an algorithm is that it is not based on the SG. Instead, we use a variant of the so-called lottery equivalent (LE) procedures (McCord and de Neufville, 1986). The specific method we use is unchained, so it is not biased by the linking of two consecutive elicitations. Although these techniques have been used to elicit health state utilities (Pinto and Abellán-Perpiñán, 2005; Bleichrodt *et al*., 2007) they have never been applied before to estimate a scoring algorithm.

These procedures are based on the comparison of two gambles and were developed to avoid the dislike for gambling typical of the SG. People facing a SG question tend to overvalue the riskless outcome in comparison to the gamble, a phenomenon known as the certainty effect (Kahneman and Tversky, 1979). Such overweighting of the certainty is 'drastically reduced' (Cohen and Jaffray, 1988) when assessments are made by LE methods in which no sure outcome is involved (McCord and de Neufville, 1986; Wakker and Deneffe, 1996; Pinto and Abellán-Perpiñán, 2005). This seems to be the main reason why violations of expected utility are less pronounced when both alternatives are risky (Camerer, 1992). In our study, the upward bias of the SG is tested by comparing the utilities elicited by this method to those measured through a LE procedure.

## 2. THE VALUATION STUDY

### 2.1. General design

We designed two valuation surveys. Survey 1 was addressed at estimating the SF-6D algorithm and included the LE questions, which were asked without chaining. Our choice was based on the evidence that the chained

[1] Lam *et al*. (2008) state that the range of the Hong Kong SF-6D value set is greater, 'suggesting that it might have less floor effect than the UK algorithm'. (p. 302).

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

SG method tends to generate higher valuations and, additionally, that linking two successive stages may amplify errors that respondents make in reporting their responses (Wakker and Deneffe, 1996). Through survey 2, we elicited preferences from an independent sample to test whether the typical (unchained) SG indeed yielded higher utility scores than our (also unchained) LE method. We posed this comparison to discard the idea that differences between our valuation method and that of Brazier *et al*. (2002) are exclusively due to chaining.

### 2.2. The sample

The main sample (survey 1) consisted of 1020 subjects grouped into 17 subsamples ($n = 60$ each). The same sample size ($n = 60$) was used in survey 2.

Respondents were selected by using stratified random sampling. First, the population was divided into groups according to age and sex quotas representative of the overall Spanish population. Then, the participants were drawn at random from each group. Given this design, we expected that, although the two samples involved in surveys 1 and 2 were independent, their preferences would be similar as long as a common elicitation procedure was applied. Such an *ex ante* homogeneity condition was tested by including a visual analogue scale (VAS) in the questionnaires.

Both surveys took place in the region of Murcia over a period of 2 months. All the interviews were face to face and run on laptops. The average time per interview was around 20 minutes.

### 2.3. The health states

A total of 78 SF-6D health states (see Table II) were used in survey 1. Of these, 49 were obtained by running the orthoplan module of SPSS version 17, which yields the minimum subset of states, which allows the estimation of an additive model. The remaining states (including the worst possible SF-6D state, the so-called ‘pits’ state) up to 78 were included to estimate potential interaction effects between attributes. A similar approach has been followed by most SF-6D studies (Brazier *et al*., 2002, 2009; Ferreira *et al*., 2010). Each of the 17 groups of respondents included in survey 1 valued a different subset of five health states, although 7 of the 78 states were included in two subsets and then valued by two different groups.[2] The only group involved in survey 2 valued the five health states shown in Table III.

### 2.4. The questionnaire

The questionnaire was organized as follows. First, the SF-6D classification system was explained to the respondents. Second, they were asked to rate five SF-6D health states by means of a VAS similar to the ‘thermometer’ used by the EuroQol Group (1990). Next, the main task was administered (the LE method in survey 1 and the SG in survey 2). Finally, respondents answered some sociodemographic questions (sex, age, etc.) and described their health status using the EQ-5D and the SF-36.

### 2.5. Elicitation procedures

*2.5.1. The probability lottery equivalent method.* We call our LE procedure a probability LE (PLE) method because the equivalence between the two gambles is reached by varying the probability of one of them. The method asks for the probability $p$ that makes the respondents indifferent between the gamble denoted by (full health, $p$; death), yielding full health with probability $p$ and death with probability $1–p$, and the gamble denoted by (full health, 0.5; $h$), yielding full health and the health state $h$ with the same probability. Appendix A shows how this question was displayed to participants for the first time, when probability $p$ was fixed at 0.5.

This framing allowed us to elicit preferences for both better-than-death and worse-than-death states. If the respondent preferred the second gamble to the first in the opening question, that is, for $p = 0.5$, it meant that

[2] We needed a higher sample size for those health states to address a different investigation with the SF-6D, which will be reported elsewhere.

Copyright © 2011 John Wiley & Sons, Ltd.

$h$ was regarded as better than death. In consequence, the final probability of indifference $p^*$ was elicited between 0.5 and 1. On the contrary, if the first gamble was preferred to the second for $p = 0.5$, then $h$ was considered as worse than death, and $p^*$ was elicited between 0 and 0.5. Under expected utility, assuming the convention that the utility of full health is 1 and the utility of death is 0, the utility of the health state $h$ is calculated as $U(h) = 2p^* - 1$.

*2.5.2. The standard gamble method.* The SG method we used in survey 2 asks the respondents for the probability $r$ that makes them indifferent between health state $h$ for sure and a gamble, denoted by (full health, $r$; death). Under the same assumptions as before, the utility $U$ of the health state $h$ equals $r^*$.

### 2.6. Search procedure

According to Lenert *et al*. (1998), a search procedure is the method used to find the point at which the respondent is indifferent between the offered alternatives. In our study, we used the parameter estimation by sequential testing procedure suggested by Luce (2000). This is a specific choice-based method that appears to be less prone to inconsistencies than other usual search procedures, for example, ping-pong (Fischer *et al*., 1999).

### 2.7. The modeling

Our initial specification is a model without interactions between variables, that is, a main effects model, whose constant term was forced to unity to ensure that the utility of full health equals one. The model was estimated using both ordinary least squares (OLS) and random effects (RE) estimators, as in models (5) and (6) shown in Table V of Brazier *et al*. (2002).

We also estimated extended versions of the model including variables denoting the presence in the state of the highest (worst) level in, at least, one of the dimensions, as well as interactions between variables in the main effects model, for example, the so-called MOST term (a dummy variable which takes a value of 1 if any dimension in the health state is at the most severe level, and 0 otherwise). The optimal specification was chosen according to the usual criteria of consistency, goodness of fit, and parsimony.

## 3. RESULTS

### 3.1. The data set

A number of 15 individuals were left out of the analysis because of inconsistencies (in an ordinal sense) in their valuations of health states either by the VAS or by the PLE (survey 1). In general, an ordinal inconsistency between two health states arises if the less severe state is valued lower than the more severe state. Because SF-6D health states are coded by assigning a higher level as the severity of the dimension becomes worse, one health sate will be logically better than another one when the levels in all dimensions of the former are equal or lower than those of the latter, for example, state 41111 is logically better than state 411142. Sets of health states include one or more cases of this type of dominance for 94.1% of respondents (960/1020).

Another seven respondents were excluded because they never chose the gamble (full health, $p$; death) for any $p < 1$ in at least three of the five health states valued. This means that they were not willing to accept any risk of death to improve their health. No exclusion was performed in the sample belonging to survey 2.

After exclusions, the final sample used to estimate the SF-6D algorithm consisted of 998 individuals, whose sociodemographic characteristics are shown on Table I. Compared with the Spanish population, our sample has a larger proportion of people in the highest and lowest educational levels and more people with higher earnings. As noted in Section 2.2, the sample was representative of the Spanish adult general population in terms of age and sex.

Copyright © 2011 John Wiley & Sons, Ltd.

Table I. Sociodemographic characteristics of subjects

| | Sample ($n=998$) | Spanish population |
|---|---|---|
| *Male/Female (%)* | 50/50 | 51/49 |
| *Mean (SD) age in years* | 43.6 (16.64) | 42.7 (16.9) |
| *Marital status* | | |
| Single | 33.7 | 32.4 |
| Married/cohabiting | 59.8 | 63.1 |
| Separated/divorced/widow | 6.5 | 4.5 |
| *Education level* | | |
| Illiterate/primary studies | 34.5 | 30.1 |
| Secondary studies | 34.4 | 45.1 |
| University studies | 31.1 | 24.7 |
| *Income level* | | |
| Up to €2000 | 51.3 | 69.5 |
| €2001–3000 | 29.8 | 19.5 |
| More than €3000 | 18.9 | 11.0 |
| *Smoker (%)* | 27.0 | 29.8 |
| *Self-assessed health state (EQ-5D)* | | |
| 11111 | 60.8 | |
| 11121 | 15.8 | |
| 11112 | 4.3 | |
| Other | 19.1 | |
| *Self-assessed health state (SF-6D/SF-36)* | | |
| 111122 | 6.0 | |
| 111112 | 4.3 | |
| 111222 | 3.1 | |
| 111111 | 2.9 | |
| Other | 83.7 | |

## 3.2. Direct health state valuations

*3.2.1. Probability lottery equivalent utilities.* Basic descriptive statistics for the 78 health states directly valued are shown in Table II. Each of the states was valued by 64 individuals on average, ranging from a minimum of 56 subjects to a maximum of 119.[3] Mean values range from $-0.515$ to 0.988. Contrary to Brazier *et al*. (2002), we obtain negative mean values in 2 of the 78 states. The proportion of utilities below zero is lower than that in Brazier *et al*. (2002) (4.8% versus 7%), but our negative valuations are generally higher in absolute value.

A histogram and descriptive statistics for the 4999 individual valuations are shown in Figure 1. Compared with those in Brazier *et al*. (2002), mean and median values are lower in our study (0.499 and 0.50 versus 0.542 and 0.65, respectively), and the degree of negative skewness is higher ($-1.23$ versus $-0.78$, Fisher's skewness coefficient). Moreover, most respondents in their study judged the pits state as better than death (73%), whereas 77.5% of individuals in our study who valued that state assigned it utilities under $-0.30$, and one-third of health states (26/78) were considered worse than death by, at least, one of the respondents.

To check to what extent the mean health state values were logically consistent in the sense explained in Section 3.1, we examined all the possible ordinal pairwise comparisons for the 78 health states. When mean values for these states are confronted, logical inconsistencies only emerge for 2.51% of all possible comparisons (14/558).

*3.2.2. Comparison between probability lottery equivalent and standard gamble utilities.* Table III shows that VAS scores for the five health states valued in both surveys were very similar ($p > 0.05$). Consequently, the comparison between PLE and SG utilities for the same states, shown in Table IV, could be considered as meaningful, although they came from two independent samples. Both mean and median utilities measured by the SG were significantly higher than those assessed through the PLE, corroborating our initial expectation.

[3] See footnote 2.

Copyright © 2011 John Wiley & Sons, Ltd.

Table II. Statistics for the SF-6D health state valuations

| State | *n* | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|---|
| 111115 | 114 | −0.960 | 1.000 | 0.855 | 0.900 | 0.197 |
| 111131 | 120 | 0.500 | 1.000 | 0.878 | 0.900 | 0.110 |
| 111411 | 119 | 0.540 | 1.000 | 0.803 | 0.780 | 0.105 |
| 112451 | 60 | 0.200 | 0.800 | 0.515 | 0.500 | 0.137 |
| 113131 | 58 | 0.800 | 1.000 | 0.988 | 1.000 | 0.036 |
| 115111 | 118 | 0.300 | 0.800 | 0.649 | 0.660 | 0.122 |
| 115533 | 57 | −0.900 | 0.960 | 0.383 | 0.400 | 0.298 |
| 121525 | 60 | 0.160 | 0.900 | 0.569 | 0.600 | 0.145 |
| 121622 | 60 | 0.180 | 0.700 | 0.451 | 0.460 | 0.103 |
| 122255 | 59 | 0.200 | 0.900 | 0.469 | 0.460 | 0.167 |
| 124123 | 60 | 0.200 | 1.000 | 0.760 | 0.800 | 0.172 |
| 132144 | 59 | 0.320 | 0.900 | 0.605 | 0.600 | 0.136 |
| 132612 | 59 | 0.300 | 1.000 | 0.710 | 0.700 | 0.174 |
| 133322 | 59 | 0.360 | 1.000 | 0.671 | 0.660 | 0.136 |
| 135242 | 56 | 0.100 | 0.900 | 0.595 | 0.600 | 0.170 |
| 141314 | 56 | 0.240 | 1.000 | 0.755 | 0.800 | 0.176 |
| 144411 | 59 | −0.200 | 1.000 | 0.675 | 0.700 | 0.234 |
| 211213 | 59 | 0.200 | 1.000 | 0.903 | 0.960 | 0.146 |
| 213615 | 58 | −0.300 | 0.900 | 0.449 | 0.400 | 0.254 |
| 222222 | 60 | 0.520 | 1.000 | 0.891 | 0.930 | 0.130 |
| 222332 | 58 | 0.400 | 1.000 | 0.826 | 0.820 | 0.136 |
| 223534 | 56 | 0.060 | 0.820 | 0.474 | 0.420 | 0.190 |
| 224152 | 60 | 0.100 | 0.940 | 0.411 | 0.400 | 0.158 |
| 224635 | 60 | 0.060 | 0.700 | 0.297 | 0.260 | 0.159 |
| 231424 | 60 | 0.160 | 0.680 | 0.340 | 0.300 | 0.102 |
| 234243 | 59 | 0.180 | 0.620 | 0.381 | 0.360 | 0.114 |
| 235121 | 59 | 0.400 | 1.000 | 0.610 | 0.600 | 0.134 |
| 242541 | 59 | 0.300 | 1.000 | 0.613 | 0.560 | 0.193 |
| 243543 | 59 | 0.160 | 0.740 | 0.400 | 0.380 | 0.125 |
| 245354 | 56 | 0.060 | 0.700 | 0.375 | 0.400 | 0.161 |
| 311112 | 57 | 0.580 | 1.000 | 0.908 | 0.940 | 0.089 |
| 314345 | 59 | 0.200 | 0.800 | 0.395 | 0.400 | 0.122 |
| 322134 | 60 | 0.000 | 1.000 | 0.598 | 0.600 | 0.211 |
| 325412 | 58 | 0.000 | 0.900 | 0,552 | 0.510 | 0.182 |
| 325554 | 58 | −0.900 | 0.700 | 0.191 | 0.200 | 0.323 |
| 331551 | 59 | −0.800 | 0.960 | 0.522 | 0.560 | 0.358 |
| 333221 | 58 | −0.940 | 1.000 | 0.865 | 0.980 | 0.296 |
| 333433 | 60 | 0.100 | 0.900 | 0.490 | 0.500 | 0.174 |
| 333633 | 59 | 0.200 | 0.900 | 0.514 | 0.480 | 0.184 |
| 335244 | 59 | −0.600 | 0.980 | 0.557 | 0.640 | 0.329 |
| 342623 | 58 | −0.980 | 1.000 | 0.227 | 0.200 | 0.516 |
| 343333 | 60 | 0.200 | 0.800 | 0.443 | 0.420 | 0.151 |
| 344425 | 60 | 0.060 | 0.580 | 0.160 | 0.160 | 0.085 |
| 411111 | 117 | 0.600 | 1.000 | 0.895 | 0.900 | 0.100 |
| 411142 | 56 | 0.180 | 1.000 | 0.890 | 0.900 | 0.127 |
| 412422 | 60 | 0.300 | 1.000 | 0.599 | 0.600 | 0.178 |
| 422211 | 60 | 0.300 | 0.860 | 0.757 | 0.760 | 0.099 |
| 423433 | 60 | 0.340 | 0.680 | 0.458 | 0.460 | 0.077 |
| 423514 | 59 | −0.200 | 0.600 | 0.138 | 0.100 | 0.149 |
| 431353 | 56 | 0.200 | 0.900 | 0.584 | 0.600 | 0.156 |
| 434545 | 59 | 0.060 | 0.520 | 0.241 | 0.220 | 0.102 |
| 434631 | 60 | −0.880 | 0.780 | 0.153 | 0.200 | 0.304 |
| 444245 | 58 | −0.400 | 0.900 | 0.260 | 0.200 | 0.251 |
| 444544 | 59 | 0.080 | 0.940 | 0.416 | 0.400 | 0.206 |
| 445125 | 60 | 0.100 | 0.800 | 0.369 | 0.330 | 0.140 |
| 445354 | 59 | −0.980 | 0.920 | 0.265 | 0.400 | 0.466 |
| 512522 | 59 | 0.200 | 0.800 | 0.476 | 0.500 | 0.130 |
| 514224 | 60 | −0.760 | 0.920 | 0.446 | 0.420 | 0.270 |
| 521641 | 58 | −0.200 | 0.900 | 0.384 | 0.400 | 0.229 |

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

Table II. *Continued*

| State | *n* | Min | Max | Mean | Median | SD |
|---|---|---|---|---|---|---|
| 524345 | 59 | −0.400 | 0.700 | 0.285 | 0.300 | 0.142 |
| 525311 | 58 | −0.200 | 1.000 | 0.729 | 0.800 | 0.240 |
| 531435 | 60 | 0.020 | 0.900 | 0.368 | 0.340 | 0.226 |
| 532113 | 60 | 0.040 | 0.900 | 0.507 | 0.500 | 0.218 |
| 532454 | 58 | −0.980 | 0.860 | 0.161 | 0.200 | 0.436 |
| 543152 | 57 | −0.780 | 0.960 | 0.509 | 0.500 | 0.271 |
| 543233 | 58 | 0.100 | 0.900 | 0.610 | 0.600 | 0.197 |
| 545654 | 60 | 0.020 | 0.400 | 0.168 | 0.110 | 0.090 |
| 612321 | 56 | 0.160 | 1.000 | 0.677 | 0.700 | 0.156 |
| 615654 | 60 | −0.960 | 0.620 | −0.263 | −0.310 | 0.346 |
| 621121 | 60 | 0.240 | 0.980 | 0.657 | 0.700 | 0.176 |
| 623443 | 59 | −0.300 | 0.800 | 0.242 | 0.200 | 0.168 |
| 632115 | 60 | 0.100 | 0.800 | 0.580 | 0.600 | 0.169 |
| 634512 | 56 | −0.200 | 0.600 | 0.158 | 0.100 | 0.172 |
| 641111 | 115 | −0.100 | 1.000 | 0.672 | 0.700 | 0.264 |
| 641232 | 60 | −0.900 | 0.800 | 0.329 | 0.380 | 0.286 |
| 643233 | 60 | 0.060 | 0.700 | 0.315 | 0.300 | 0.178 |
| 644342 | 57 | −0.980 | 0.660 | 0.004 | 0.060 | 0.366 |
| 645655 | 116 | −0.980 | 0.500 | −0.515 | −0.600 | 0.426 |

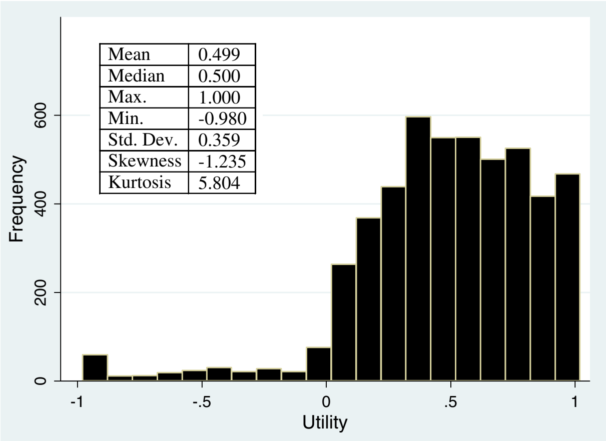


Figure 1. Histogram and descriptive statistics for health state valuations (probability lottery equivalent)

Table III. Comparison of visual analogue scale valuations of Survey 1 (probability lottery equivalent) and Survey 2 (standard gamble)

| | Mean valuations | | | Median valuations | | |
|---|---|---|---|---|---|---|
| Health states | Survey 1 | Survey 2 | *t*-test (*p*-value) | Survey 1 | Survey 2 | Wilcoxon (*p*-value) |
| 132612 | 57.083 | 54.067 | 0.207 | 55.500 | 60.000 | 0.614 |
| 141314 | 67.700 | 65.167 | 0.243 | 70.000 | 66.000 | 0.328 |
| 222332 | 66.517 | 64.167 | 0.441 | 70.000 | 65.500 | 0.130 |
| 311112 | 88.450 | 87.017 | 0.377 | 90.000 | 88.500 | 0.192 |
| 412422 | 50.050 | 47.267 | 0.224 | 50.000 | 53.000 | 0.678 |

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

Table IV. Probability lottery equivalent versus standard gamble valuations

| | Mean valuations | | | Median valuations | | |
|---|---|---|---|---|---|---|
| Health states | PLE | SG | *t*-test (*p*-value) | PLE | SG | Wilcoxon (*p*-value) |
| 132612 | 0.711 | 0.815 | 0.000 | 0.700 | 0.800 | 0.001 |
| 141314 | 0.754 | 0.846 | 0.000 | 0.780 | 0.850 | 0.002 |
| 222332 | 0.832 | 0.905 | 0.000 | 0.820 | 0.900 | 0.000 |
| 311112 | 0.880 | 0.955 | 0.025 | 0.940 | 0.950 | 0.025 |
| 412422 | 0.599 | 0.780 | 0.000 | 0.600 | 0.800 | 0.000 |

PLE, probability lottery equivalent; SG, standard gamble.

Table V. SF-6D (SF-36) health state models

| | Random effects models | | | OLS mean model | |
|---|---|---|---|---|---|
| | 'Raw' | | Efficient | | Mean |
| | (1) | | (2) | | (3) |
| Cons | 1 | Cons | 1 | Cons | 1 |
| PF2 | −0.025 | PF2 | −0.022 | PF2 | −0.015 |
| PF3 | −0.056 | PF3 | −0.062 | PF3 | −0.034 |
| PF4 | −0.120 | PF4 | −0.122 | PF4 | −0.090 |
| PF5 | −0.107 | PF5 | −0.109 | PF5 | −0.111 |
| PF6 | −0.335 | PF6 | −0.340 | PF6 | −0.338 |
| RL2 | 0.007 | | | RL2 | −0.014 |
| RL3 | −0.045 | RL23 | −0.018 | RL3 | −0.038 |
| RL4 | −0.089 | RL4 | −0.085 | RL4 | −0.070 |
| SF2 | −0.071 | SF2 | −0.069 | SF2 | −0.037 |
| SF3 | −0.078 | SF3 | −0.079 | SF3 | −0.060 |
| SF4 | −0.194 | SF4 | −0.194 | SF4 | −0.203 |
| SF5 | −0.239 | SF5 | −0.234 | SF5 | −0.208 |
| PAIN2 | −0.044 | | | PAIN2 | −0.018 |
| PAIN3 | −0.047 | PAIN23 | −0.044 | PAIN3 | −0.034 |
| PAIN4 | −0.172 | PAIN4 | −0.178 | PAIN4 | −0.198 |
| PAIN5 | −0.230 | PAIN5 | −0.225 | PAIN5 | −0.202 |
| PAIN6 | −0.343 | PAIN6 | −0.345 | PAIN6 | −0.318 |
| MH2 | −0.026 | MH2 | −0.029 | MH2 | −0.066 |
| MH3 | −0.050 | MH3 | −0.053 | MH3 | −0.078 |
| MH4 | −0.072 | MH4 | −0.075 | MH4 | −0.096 |
| MH5 | −0.196 | MH5 | −0.199 | MH5 | −0.224 |
| VIT2 | −0.043 | VIT2 | −0.042 | VIT2 | −0.058 |
| VIT3 | −0.093 | VIT3 | −0.091 | VIT3 | −0.121 |
| VIT4 | −0.158 | VIT4 | −0.156 | VIT4 | −0.157 |
| VIT5 | −0.181 | VIT5 | −0.179 | VIT5 | −0.199 |
| N | 4,990 | N | 4,990 | N | 78 |
| **Predictive ability** | | | | | |
| MAE | 0.0871 | | 0.0872 | | 0.0812 |
| $\lvert$pred. error$\rvert < k$ | | | | | |
| k = 0.01 | 8.13 | | 4.72 | | 11.72 |
| k = 0.05 | 36.41 | | 35.25 | | 36.49 |
| k = 0.10 | 63.50 | | 62.24 | | 70.50 |

All coefficients are significant at a 99% confidence level except for PF2 in models (1) and (2); RL23 in model (2), which are significant at the 95% level; and RL2 in model (1), which is statistically non-significant.

The estimation of the mean model incorporates corrective weights proportional to the number of individuals valuing each of the health states.

MAE, mean absolute error.

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

*3.3. SF-6D algorithms.* Estimated coefficients are shown in Table V for the three models, which led to the best results in terms of goodness of fit and parsimony. Two are RE models, whereas the third is an OLS model using mean values. The RE model labeled as the 'raw model' is a model without the removal of non-significant variables. The RE model labeled as the 'efficient model' was constructed by eliminating non-significant regressors from the 'raw model' and by grouping the variables of any two consecutive levels when their coefficients are not significantly different from each other according to the value of the Wald statistic. We did not find any significant interaction term, so all our algorithms only reflect main effects.

All the coefficients have the expected sign and are highly significant, except for the level 2 of the 'role limitation' dimension in the 'raw' RE model.[4] Mean absolute error (MAE) attached to any of our models is only slightly higher than those reported by Brazier *et al*., who used much more health states, which underlines the quality of fit we obtained.

The coefficients for the two RE models suggest that the greatest utility losses associated to the maximum level of severity in a dimension occur for 'pain', 'physical functioning' and 'social functioning' attributes, in that order. However, the conclusion differs slightly for the OLS model at mean level because, in this case, 'physical functioning' is the dimension that produces the larger disutility. The lower weight of 'mental health' in comparison to the UK algorithm should be noted.

As in Brazier *et al*. (2002) and in Brazier and Roberts (2004), who reestimate the SF-6D (SF-36) after removing ordinal inconsistencies (see Equation 2 in their paper), we also recommend the OLS mean model for cost–utility analysis because it yields the lowest MAE and the highest predictive precision of our estimated models.

Figure 2 shows the distribution of predicted utilities by both our mean OLS model and the mean consistent model in Brazier and Roberts (2004). It is apparent that our model lowers the minimum threshold of Brazier and Roberts's algorithm, expanding the left tail of the distribution below zero. The minimum score predicted by our mean model is −0.357, a value very far from 0.354, the floor predicted by the UK algorithm.

Because the percentage of negative valuations in our study was lower than those found by Brazier and colleagues, one possible objection to our estimates could be that they are not consistent once a minimum fraction of observations are removed from the data set. To explore this possibility, we repeated the OLS estimation by excluding from the data successively the lowest 5%, 10%, and 20% of valuations. The predicted utilities derived from the most demanding case also are shown in Figure 2. As can be observed, even when 20% of the observations are excluded (which implies that most of the worst state's valuations are left aside), the floor of our SF-6D algorithm is still lower than those in previous studies.

## 4. DISCUSSION

This paper reports the estimation of a SF-6D algorithm using a LE method termed PLE. An implication of using this procedure is that the 'tariff' that predicts our algorithm is shifted to the left compared with that predicted by the consistent model by Brazier and Roberts (2004). We have a significant part of the distribution (around one-fourth) below 0.354, which is the minimum threshold of the algorithm by Brazier and Roberts. In fact, the value predicted by our algorithm for the worst SF-6D health state is far below zero, –0.357. This is the 'lowering' of the SF-6D floor, which figures in the title of this article.

We claim the relevance of the PLE to explain the differences between our algorithm and the previous ones for three reasons. First, we applied the same econometric models as most previous studies. It is true that none of our models included the interaction term MOST, but if we compare the range of the SF-6D values predicted by our mean model with that predicted by the main effects model (6) by Brazier *et al*. (2002), our range continues to be larger. The same occurs if the comparison is performed with respect to the main effects models derived

[4] As Brazier *et al*. (2002) remarked, there is no clear ordinal relationship between levels 4 ('Your health limits you a lot in moderate activities') and 5 ('Your health limits you a little in bathing and dressing') of the physical functioning dimension, so the apparent inconsistency between coefficients PF4 and PF5 is not real.

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

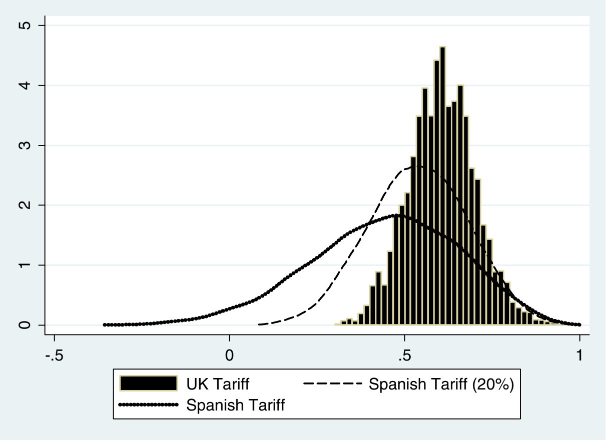


Figure 2. A comparison of the Spanish and UK tariffs' predicted values.
Note: The Spanish tariff corresponds to our OLS mean model in Table V. The UK tariff is the SF-6D (SF-36) 'consistent' model at mean level (column 2 of Table 4 in Brazier and Roberts, 2004). Spanish tariff (20%) shows the result of reestimating the algorithm after excluding the lowest 20% of observations

either for Hong Kong (Lam *et al*., 2008), Portugal (Ferreira *et al*., 2010), or Japan (Brazier *et al*., 2009). Hence, it seems necessary to look for other explanations to our findings.

Second, although part of the discrepancy may come from different preferences of the Spanish population in comparison to other countries, differences between ranges of utilities are so large as to be successfully explained merely on a country-specific basis. The comparison between Spanish EQ-5D tariff (Badia *et al*., 2001) and the UK one suggests that Spanish and UK preferences are indeed different. However, variation in preferences did not cause a change in the shape of the distribution of EQ-5D scores from one country to another as drastic as in our case with the algorithm for the SF-6D. Note that the absolute range of variation for the mean observed EQ-5D utilities in Spain and the UK was very close (1.5 for the Spanish utilities versus 1.4 for the UK utilities). In contrast, the range for UK SF-6D utilities was 0.9, whereas the range for our observed utilities was 1.5. Moreover, we obtain that Spanish SF-6D utilities are mostly lower than the UK ones, just the opposite pattern to that reported for the EQ-5D. Therefore, it seems that country-specific differences, although likely to affect results, cannot explain by themselves all the differences.

Third, although the design of our survey is different from that implemented by the study by Brazier *et al*. (2002), it is unclear that this fact can explain by itself all the differences found. Our respondents valued fewer health states, and the interview protocol also was different. In addition, Brazier *et al*. (2009), using a different design, report different results from those reached by Brazier *et al*. (2002). Hence, our setup may explain some of the differences. Nevertheless, the design followed by Lam *et al*. (2008) to estimate the SF-6D algorithm in Hong Kong was not the same either, yet despite this, their results were similar to those obtained for the UK algorithm, as they were for the Portuguese model (Ferreira *et al*., 2010).

Therefore, it seems that our findings cannot be fully explained unless we focus on the different valuation method used. We compared our PLE method and the SG for five different SF-6D health states confirming the starting hypothesis that the SG, even without chaining, yields higher values. This result is coherent with previous empirical evidence (Pinto and Abellán-Perpiñán, 2005) and also with the theoretical predictions of the context-dependent model by Bleichrodt and Schmidt (2002). Assuming this theoretical framework offers explanations for well-known violations of expected utility such as the referred certainty effect, so the model can be taken as a reasonable justification for the behavioral observation that LE utilities are lower than SG ones.

A possible alternative to the usage of the PLE would have been to use the TTO, the riskless procedure employed to estimate the EQ-5D algorithm. Surely, a TTO-based SF-6D algorithm would yield utilities more comparable to those generated by the EQ-5D instrument than the SG-based SF-6D model. However, our main

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

goal was not to produce an SF-6D algorithm more comparable to the EQ-5D *per se* but to estimate a better (a less biased) SF-6D algorithm than the previous one. We agree with the decision made by Brazier and colleagues to use a risky framing to value SF-6D health states. Our disagreement concerns the specific method they chose, the SG. There is recent evidence (Abellan-Perpiñan *et al*., 2009a, 2009b; Attema *et al*., 2010), which suggests that it is unclear whether the utilities obtained in a riskless environment can be freely transferred to a risky one (and vice versa). Given this evidence, we opted for a valuation method framed in terms of risk because we think that cost–utility analysis mainly deals with risky decision contexts.

We are aware that the PLE also might be affected by biases. For example, a first issue concerns its possibly higher cognitive burden. The PLE has more attributes than the SG, so it may be more difficult for respondents to reach indifference. This is a plausible objection, although, at least in our study, it does not seem to be disabling. Average time per interview in survey 1 (PLE) was only 2 minutes longer than that in survey 2 (SG).

As the empirical validity of the PLE has been insufficiently studied to date, future research should address this issue. For example, our PLE fixes the probability of one of the gambles at 0.5. This property means that utilities are bounded between −1 and +1, so avoiding the need to rescale the utilities. However, it is unknown if utilities would remain approximately the same for different baseline probabilities. If not, utilities could be biased in a similar way as the utility evaluation effect coined by Machina (1983), which predicts that the utility is more concave when higher probabilities are used in the elicitation. This issue surely deserves to be explored in future investigations.

Another issue deals with differences found between the number and magnitude of our negative utilities and those reported by Brazier *et al*. (2002). We think that these differences may be explained by three factors. First, the health states are not the same, so it is possible that our study contains (in relative terms) more severe health states than theirs. However, we cannot be sure about this because we can only compare our health states with the 30 states shown in Table 4 of Brazier *et al*. (2002: p. 280). Second, note that the two-stage valuation procedure used by Brazier *et al*. requires the pits state to be regarded as worse than death to produce a negative utility, but in fact, most of their respondents judged it as better than death, whereas most of ours believed the opposite. It might be thought that this discrepancy is caused by a different way of determining when a health state is better or worse than death. Brazier *et al*. (2002) considered that the pits state was worse than death if it was ranked below death in a ranking exercise with another six cards. In our case, any health state is taken as worse than death if the gamble (full health, 0.5; death) is preferred to (full health, 0.5; *h*). However, we also included a VAS in our questionnaire. A VAS, because it is a rating task, also requires rank ordering the different health states. More than 95% of the subjects who valued the pits state as worse than death in the PLE also ranked that state below death in the VAS. In consequence, it seems that the different perception of the severity of the pits state does not come from the way the health states were compared with the death.

A final distinction between the utilities obtained in the two studies lies in the fact that our utilities were not subject to any transformation, whereas those elicited by Brazier *et al*. (2002) were rescaled to be bounded between −1 and +1. One effect of rescaling is that final utilities will tend to be higher than raw utilities. Consider as an example the case of a respondent who is indifferent between death for certain and a gamble offering full health and the pits state for 50/50 probabilities. This leads to a raw utility of −1. After transformation, the same utility is now −0.5. It is apparent then that transforming negative utilities boosts valuations. Our PLE, on the contrary, yields utilities automatically bounded between −1 and +1, so no further transformation is required. Therefore, our procedure may be interpreted as being like the analogue under risk to the 'life profile' approach developed by Robinson and Spencer (2006) for decisions under certainty.

Further research is needed to explore in depth the validity of our new algorithm. Future investigations might develop new algorithms based on our data set by adopting a non-parametric approach. Moreover, different health states to those used in our study could be employed to compare whether differences between PLE and SG utilities persist. Comparisons with EQ-5D tariffs also should be made to obtain direct evidence as to what extent the two instruments, the SF-6D and the EQ-5D, are more comparable, after the 'floor' has been lowered. Implications for cost–utility ratios from that greater similarity between both instruments also should be

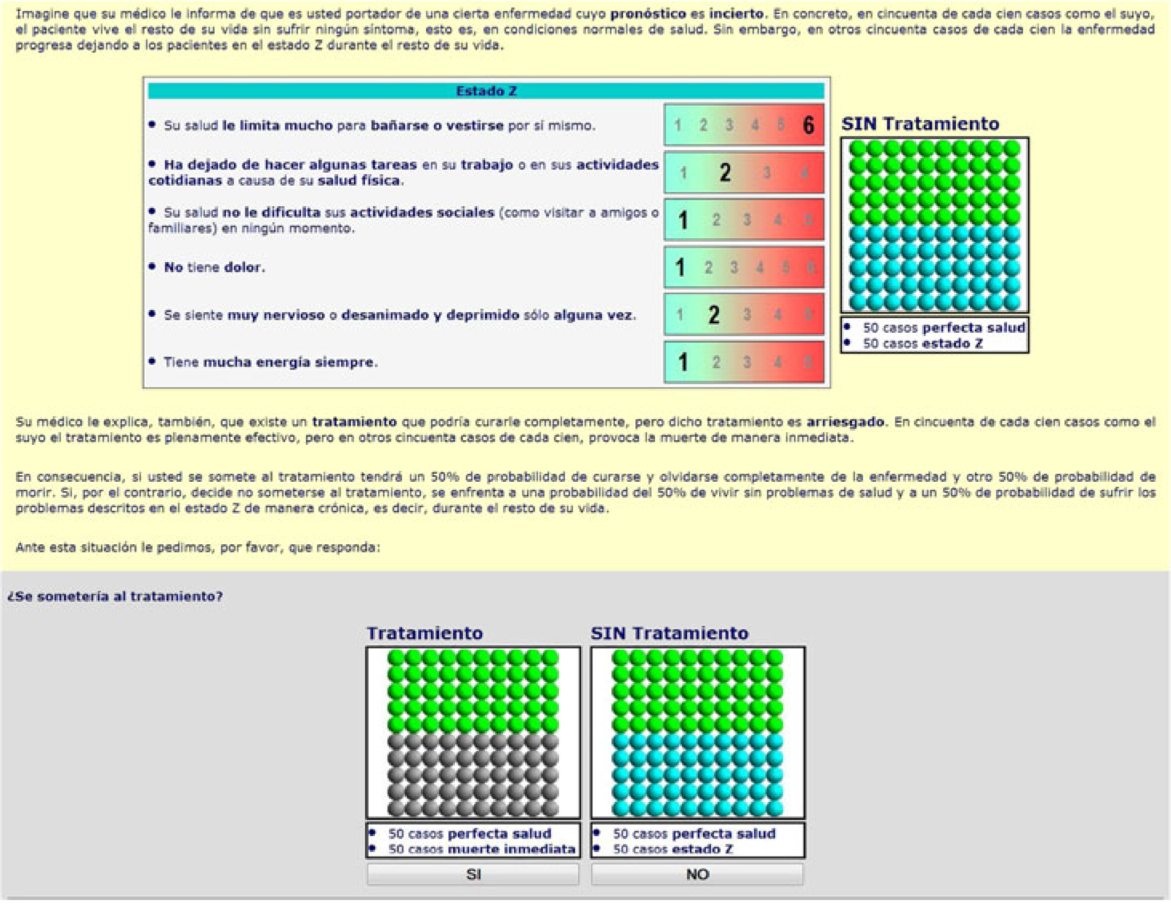
Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

analyzed. This could have an impact on applied economic evaluations and eventually on current recommendations of public agencies such as NICE, whose guidelines encourage the use of EQ-5D to measure changes in health-related quality of life over other preference-based instruments (NICE, 2008: p. 38).

This paper shows how using a different method to the SG, the PLE, to elicit health state utilities can lower the floor of the SF-6D value set. Neither cultural differences between the Spanish population and citizens of other countries nor the different interview design used seem able to explain the discrepancy in terms of utility range found in our study in comparison to previous works. Our results suggest that the PLE is a feasible technique, not much more cognitively demanding than the SG and yielding less biased values instead. We hope this finding encourages researchers to use non-conventional elicitation procedures, such as the PLE, as a basis for estimating preference-based algorithms.

## APPENDIX A: SCREENSHOT OF A PLE QUESTION



Imagine que su médico le informa de que es usted portador de una cierta enfermedad cuyo **pronóstico** es **incierto**. En concreto, en cincuenta de cada cien casos como el suyo, el paciente vive el resto de su vida sin sufrir ningún sintoma, esto es, en condiciones normales de salud. Sin embargo, en otros cincuenta casos de cada cien la enfermedad progresa dejando a los pacientes en el estado Z durante el resto de su vida.

**Estado Z**

- Su salud **le limita mucho** para **bañarse o vestirse** por sí mismo.
- **Ha dejado de hacer algunas tareas** en su **trabajo** o en sus **actividades cotidianas** a causa de su **salud física**.
- Su salud **no le dificulta** sus **actividades sociales** (como visitar a amigos o familiares) en ningún momento.
- **No** tiene **dolor**.
- Se siente **muy nervioso** o **desanimado y deprimido** sólo **alguna vez**.
- Tiene **mucha energía siempre**.

**SIN Tratamiento**

- 50 casos **perfecta salud**
- 50 casos **estado Z**

Su médico le explica, también, que existe un **tratamiento** que podría curarle completamente, pero dicho tratamiento es **arriesgado**. En cincuenta de cada cien casos como el suyo el tratamiento es plenamente efectivo, pero en otros cincuenta casos de cada cien, provoca la muerte de manera inmediata.

En consecuencia, si usted se somete al tratamiento tendrá un 50% de probabilidad de curarse y olvidarse completamente de la enfermedad y otro 50% de probabilidad de morir. Si, por el contrario, decide no someterse al tratamiento, se enfrenta a una probabilidad del 50% de vivir sin problemas de salud y a un 50% de probabilidad de sufrir los problemas descritos en el estado Z de manera crónica, es decir, durante el resto de su vida.

Ante esta situación le pedimos, por favor, que responda:

**¿Se sometería al tratamiento?**

**Tratamiento**

- 50 casos **perfecta salud**
- 50 casos **muerte inmediata**

**SIN Tratamiento**

- 50 casos **perfecta salud**
- 50 casos **estado Z**

Copyright © 2011 John Wiley & Sons, Ltd.

*English translation of the image shown in previous page:* Imagine that your doctor informs you that you have a disease of uncertain prognosis. Precisely, in fifty of a hundred cases like yours, the patient lives for the rest of his/her life without suffering from any symptoms, that is, in a good health condition. Nevertheless, in the other fifty cases of a hundred, the disease progresses, leaving patients in a health state as Z for the rest of their lives.

(The card describes the 621121 SF-6D health state, labeled as Z state.)

On the right side: WITHOUT Treatment. 50 cases full health; 50 cases state Z.

Your doctor also explains to you that a treatment could cure you but that the treatment is risky. In fifty of a hundred cases as yours, the treatment is fully effective, but in the other fifty cases of a hundred, it causes immediate death.

In consequence, if you undergo the treatment, you will have a 50% probability of recovering and completely forgetting the disease, and another 50% probability of dying. If, on the contrary, you decide not to undergo the treatment, you face a 50% probability of living without any health trouble and a 50% probability of suffering problems described in state Z chronically, that is, for the rest of your life.

In the face of this situation, we ask you, please, to respond:

Would you undergo the treatment?

Treatment: 50 cases full health; 50 cases immediate death. WITHOUT Treatment. 50 cases full health; 50 cases state Z.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the Regional Health Ministry of the Autonomous Community of Murcia and from Fundación Séneca. José María Abellán Perpiñán, Jorge Eduardo Martínez Pérez and Fernando Ignacio Sánchez Martínez also acknowledge financial support from Ministerio de Ciencia e Innnovación grant ECO2010-22041-C02-02. No conflict of interest is reported.

## REFERENCES

Abellan-Perpiñan JM, Bleichrodt H, Pinto-Prades JL. 2009a. The predictive validity of prospect theory versus expected utility in health utility measurement. *Journal of Health Economics* **28**: 1039–1047.

Abellan-Perpiñan JM, Martinez JE, Sanchez FI, Mendez I. 2009b. The QALY model which came in from a general population survey: roughly multiplicative, broadly nonlinear, and sometimes context-dependent. Documento de Trabajo CENTRA, E2009/04. Available from: http://www.centrodeestudiosandaluces.info/PDFS/E200904.pdf [Accessed 13 April 2010].

Attema AE, Bleichrodt H, Wakker PP. 2010. Measuring discounting and QALYs more easily and reliably. *Working paper*. Available from: http://people.few.eur.nl/wakker/pdf/ulifedm.pdf [Accessed 13 April 2010].

Badia X, Roset R, Herdman M, Kind P. 2001. A comparison of GB and Spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making* **21**: 7–16.

Barton G, Sach T, Avery A, Jenkinson C, Doherty M, Whynes D, Muir K. 2008. A comparison of the performance of the EQ-5D and SF-6D for individuals aged $\geq$ 45 years. *Health Economics* **17**: 815–832.

Bharmal M, Thomas J. 2006 Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value in Health* **9**(4): 262–271.

Bleichrodt H. 2001. Probability weighting in choice under risk: an empirical test. *Journal of Risk and Uncertainty* **23**: 185–198.

Bleichrodt H, Schmidt U. 2002. A context-dependent model of the gambling effect. *Management Science* **48**: 802–812.

Bleichrodt H, Abellan-Perpiñan JM, Pinto JL, Mendez I. 2007. Resolving Inconsistencies in Utility Measurement under Risk: Tests of Generalizations of Expected Utility. *Management Science* **53**: 469–482.

Bleichrodt H, Pinto JL, Wakker P. 2001. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science* **47**: 1498–1514.

Brazier J, Roberts J. 2004. The estimation of a preference-based measure of health from the SF-12. *Medical Care* **42**: 851–59.

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

Brazier J, Fukuhara S, Roberts J, Kharroubi S, Yamamoto Y, Ikeda S, Doherty J, Kurokawa K. 2009. Estimating a preference-based index form the Japanese SF-36. *Journal of Clinical Epidemiology* **62**: 1323–1331.

Brazier J, Roberts J, Deverill M. 2002. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics* **21**: 271–92.

Brazier J, Roberts J, Tsuchiya A, Busschbach J. 2004. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics* **13**: 873–884.

Bryan S, Longworth L. 2005. Measuring health-related utility: Why the disparity between EQ-5D and SF-6D? *The European Journal of Health Economics* **50**: 253–260.

Buxton MJ, Lacey LA, Feagan BG, Niecko T, Miller DW, Townsend RJ. 2007. Mapping from disease-specific measures to utility: An analysis of the relationships between the Inflammatory Bowel Disease Questionnaire and Crohn's Disease Activity Index in Crohn's disease and measures of utility. *Value in Health* **10**(3): 214–220.

Camerer C. 1992. Recent tests of generalizations of expected utility theory. In Utility: Theories, Measurement and Applications, Edwards W (ed.), Kluwer Academic Publishers: Boston, MA; 207–251.

Cohen M, Jaffray J. 1988. Certainty effect versus probability distortion: an experimental analysis of decision making under risk. *Journal of Experimental Psychology* **14**: 554–560.

Delquié Ph. Inconsistent trade-offs between attributes: new evidence in preference assessment biases. *Management Science* 1993; **39**: 1382–1395.

Dolan P. 1997. Modeling valuations for EuroQol health states. *Medical Care* **35**: 1095–1108.

EuroQol Group. 1990. EQ-5D Health Questionnaire. English version for the UK (validated for Ireland). Available from: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Languages/Sample_UK__English__EQ-5D-3L.pdf [Accessed 13 April 2010].

Ferreira PL, Ferreira LN, Pereira LN. 2008. How consistent are health utility values? *Quality of Life Research* **17**: 1031–1042.

Ferreira LN, Ferreira PL, Pereira LN, Brazier J, Rowen D. 2010. A Portuguese Value Set for the SF-6D. *Value in Health* **13**(5): 624–630.

Fischer GW, Carmon Z, Ariely D, Zauberman G. 1999. Goal-based Construction of Preferences: Task Goals and the Prominence Effect. *Management Science* **45**: 1057–75.

Fryback DG, Palta M, Cherepanov D, Bolt D, Kim J-S. 2010. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Medical Decision Making* **30**: 5–15.

Hershey JC, Schoemaker PJ. 1985. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science* **31**: 1213–1231.

Johnson E, Schkade D. 1989. Bias in utility assessments: Further evidence and explanations. *Management Science* **35**: 406–424.

Kahneman D, Tversky A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263–291.

Lam CL, Brazier J, McGhee SM. 2008. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. *Value in Health* **11**: 295–303.

Lamers LM, Bouwmans CAM, van Straten A, Donker MCH, Hakkaart L. 2006. Comparison of EQ-5D and SF-6D utilities in mental health patients. *Health Economics* **15**: 1229–1236.

Lenert LA, Cher D, Goldstein M, Bergen MR, Garber A. 1998. The effect of search procedures on utility elicitations. *Medical Decision Making* **18**: 76–83.

Longworth L, Bryan S. 2003. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Economics* **12**: 1061–1067.

Luce RD. 2000. Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches. Lawrence Erlbaum Associates, Inc.: New Jersey.

Machina M. 1983. Generalized Expected Utility Analysis and the Nature of Observed Violations of the Independence Axiom. In Foundations of Utility and Risk Theory with Applications, Stigum BP, Wenstop F (eds.), D. Reidel: Dordrecht.

McCord M, de Neufville R. 1986. Lottery equivalents: Reduction of the certainty effect problem in utility assessment. *Management Science* **32**: 56–60.

National Institute for Health and Clinical Excellence. 2008. Guide to the methods of technology appraisal Issued: June 2008. Available from: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf [Accessed 13 April 2010].

Oliver A. 2003. The internal consistency of the standard gamble: tests after adjusting for prospect theory. *Journal of Health Economics* **22**: 659–674.

Oliver A. 2005. Testing the internal consistency of the lottery equivalents method using health outcomes. *Health Economics* **14**: 149–159.

Petrou S, Hockley C. 2005. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. *Health Economics* **14**: 1169–1189.

Pickard SA, Wang Z, Walton SM, Lee TA. 2005. Are decisions using cost-utility analyses robust to the choice of SF-36/SF-12 preference-based algorithm? *Health and Quality of Life Outcomes* **3**: 1–9.

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec

Pinto JL, Abellán-Perpiñán JM. 2005. Measuring the health of populations: the veil of ignorant approach. *Health Economics* **14**: 69–82.

Robinson A, Spencer A. 2006. Exploring challenges to TTO utilities: valuing states worse than dead. *Health Economics* **15**: 393–402.

Rutten-van Mölken MP, Bakker CH, van Doorslaer EKA, van der Linden S. 1995. Methodological issues of patient utility measurement. Experience from two clinical trials. *Medical Care* **33**: 922–937.

Szende A, Svensson K, Stähl E, Mészáros A, Berta GY. 2004. Psychometric and Utility-Based Measures of Health Status of Asthmatic Patients with Different Disease Control Level. *PharmacoEconomics* **22**(8): 537–547.

Torrance GW, Feeny D, Furlong W. 2001. Visual Analog Scales: do they have a role in the measurement of preferences for health states? *Medical Decision Making* **21**(4): 329–334.

Tsuchiya A, Brazier J, Roberts J. 2006. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. *Journal of Health Economics* **25**: 334–346.

Wakker P, Deneffe D. 1996. Eliciting von Neumann-Morgenstern Utilities when Probabilities are Distorted or Unknown. *Management Science* **42**: 1131–1150.

Copyright © 2011 John Wiley & Sons, Ltd.

DOI: 10.1002/hec