VALUE IN HEALTH 22 (2019) 446–452

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Methodology

Eliciting Health State Utilities Using Paired-Gamble Methods: The Role of the Starting Point

Eva Rodríguez-Míguez, PhD [1,2,*], *José Luis Pinto-Prades, PhD* [3,4], *Jacinto Mosquera-Nogueira, PhD* [2,5]

[1]*Department of Applied Economics, Group of Research in Empirical Economics (GRiEE), ECOBAS, University of Vigo, Vigo, Spain;* [2]*Fundación Biomédica Galicia Sur, Vigo, Spain;* [3]*University of Navarra, Pamplona, Spain;* [4]*Glasgow Caledonian University, Glasgow, UK;* [5]*Galician Health Service, Vigo, Spain*

ABSTRACT

Background: Paired-gamble methods have been proposed to avoid the "certainty effect" associated with standard gamble methods. **Objective:** This study examines the role of starting-point effects in paired-gamble methods. In particular, it examines how the utilities so derived vary as a function of the probabilities of the stimulus lottery. **Methods:** A sample of 455 members of the Spanish general population valued 9 health states via face-to-face interviews. Subjects were randomly placed into 3 subgroups, which differed in terms of the stimulus gamble's probability. Nonparametric tests and an interval regression model were used to test if utilities change when the probability distribution is modified. **Results:** Nonparametric tests showed that the probability of a health state being considered worse than death did not differ among subgroups. Nevertheless, changes in the stimulus gamble did produce significant differences in the distribution of utilities: the higher the probability of full health in the stimulus, the higher the utility elicited. Regression estimates support the existence of starting-point effects when the utilities are obtained under expected utility. According to the prospect theory, the conclusions depend on the reference point considered. When the reference points used are death or the health state evaluated, we observe differences among these groups. Nevertheless, when full health is used, these differences disappear. **Conclusion:** This research suggests that paired-gamble methods may also be susceptible to starting-point effects. Yet the differences are small, and they disappear when the data are analyzed using prospect theory with full health as the reference point.

Introduction

Farquhar[1] distinguishes 2 categories of preference elicitation methods for obtaining utility functions under risk: standard gamble methods and paired-gamble methods. Both approaches seek to elicit risk attitudes by establishing equivalences between the 2 options. In standard gamble methods, one option is a lottery and the other is a certain outcome—in the health domain, the certainty is usually a chronic health state and the lottery's outcomes are death or full health. Using the *probability equivalent*, which is the procedure most often used to estimate the utilities of health states, the subject is asked to estimate the probability P such that she is indifferent between a chronic state S and $[P,X;(1-P),Y]$; here X and Y are health states such that $X \succ S \succ Y$ ("$\succ$" denotes "is preferred to"). In paired-gamble methods, the 2 options are lotteries with at least 2 outcomes that have positive probabilities. Here all parameters (probabilities and outcomes) are predefined by the researcher in one of the lotteries, whereas the other lottery has some free parameters that the subject adjusts until indifference is reached. Under the *probability lottery equivalent*, the indifference between 2 gambles is reached by varying the probabilities in one lottery while keeping fixed the probabilities of the other lottery (the "stimulus"); that is, the subject must estimate P such that she is indifferent between (i) $[P,X;(1-P),Y]$ and (ii) either $[q,X;(1-q),S]$ or $[q,S;(1-q),Y]$, where q is predefined by the researcher.

The standard gamble is a method widely used to obtain health state utilities in risky decision contexts.[2,3] With only a few exceptions,[4–6] however, the paired-gamble method has seldom been employed to estimate health state utilities. Yet the increasingly popular "multiple price list" method[7] has led to the paired-gamble approach being used also to analyze risk attitudes in the health domain.[8–10] Nevertheless, this is a different use of paired gambles.

* *Address correspondence to*: Eva Rodríguez-Míguez, PhD, Departamento de Economía Aplicada, Facultad de Ciencias Económicas y Empresariales, Universidad de Vigo, As Lagoas-Marcosende s/n. 36310 Vigo, Spain.

E-mail: emiguez@uvigo.es

1098-3015/$36.00 - see front matter

https://doi.org/10.1016/j.jval.2019.01.007

Notwithstanding the standard gamble's widespread use, there are both theoretical reasons[11] and empirical evidence,[12–14] suggesting that standard gamble methods yield utilities that are too high (upward biased). One reason is what Kahneman and Tversky[15] call the *certainty effect*. When combined with loss aversion, this effect may generate upward-biased utilities. At the same time, there is some evidence that paired-gamble methods do not exhibit the certainty affect and hence result in less biased utilities.[12–14] For example, Abellán-Perpiñán et al.[4] use paired-gamble methods to avoid the "floor effect" exhibited under standard gambles[16] by the SF-6D index; also, Bleichrodt et al.[5] show that paired-gamble methods produce more consistent utilities than do their standard-gamble counterparts. Despite these promising results, paired-gamble methods can also be subject to biases and inconsistencies—a topic that the literature has not examined in depth.

Robinson et al.[17] argue that a potential drawback of the paired-gamble method is its greater susceptibility (as compared with other methods, such as discrete choice) to starting-point biases. This may well be true, but only scant evidence has been marshaled to date in favor of that claim. Researchers who have used paired-gamble methods acknowledge the necessity of checking for whether "utilities would remain approximately the same for different baseline probabilities."[4] Oliver[6] points out that "larger studies, which vary […] the value of probability […] and which use severity rather than longevity as the basis for the health outcomes, need to be undertaken." This is precisely the objective of our article. We examine how much the utilities elicited vary when the stimulus gamble's probabilities $[q;(1-q)]$ change—that is, whether those utilities are affected by a starting-point bias. According to the expected utility theory, utilities should be *independent* of the baseline probabilities employed. Nevertheless, there is evidence in the literature that utilities elicited with the certainty-equivalent standard gamble technique change with the probabilities of the lottery used as stimulus.[18,19] In this article we explore whether that problem is present also in paired-gamble methods.

Empirical studies that used paired-gamble methods in the health domain have kept the baseline probabilities fixed, which makes it impossible to test the sensitivity of utilities to changes in the distribution of those probabilities.[4–6,13] We are aware of only one article in which this issue has been addressed.[20] The authors calculate the utility of 3 health states by asking 41 students to estimate the probability P that makes them indifferent between the gamble $[P,FH;(1-P),D]$ and the gamble $[q,S;(1-q),D]$ for 2 different values (50% and 75%) of q; here FH stands for full health and D for death. That study finds a clear pattern: utilities elicited with $q = 0.5$ were higher than those elicited with $q = 0.75$. Nevertheless, those differences were not statistically significant. In short, there is little evidence that addresses how the probabilities used in the stimulus lottery affect the elicited utilities for health states in paired-gamble methods. The study presented here offers new evidence on this topic.

Methods

The Paired-Gamble Method

The paired method was operationalized by asking each subject to establish indifference between 2 lotteries: $[P,FH;(1-P),D]$ and $[q,FH;(1-q),S]$. Here P and q are the probabilities of the *best* outcome in each lottery, and S is an intermediate health state between full health and death. In this study, the value of q was fixed and the subject was asked to state the probability P^* such that $[P^*,FH;(1-P^*),D]\sim[q,FH;(1-q),S]$ (where "~" signifies indifference).

If one assumes that preferences obey expected utility theory and that the linear QALY model is valid, then the utility U of state S can be estimated as

$$U(S) = \frac{P^* - q}{1 - q}. \quad (1)$$

We also analyzed the data under prospect theory to include some the major deviations from expected utility—namely, probability transformation and loss aversion.[15] Following Bleichrodt et al.,[5] we estimated utility while considering 3 reference points: death D, full health FH, and the impaired health state S. Then $U(S)$ was estimated as follows:

1. $U(S) = \frac{w^+(P^*)-w^+(q)}{1-w^+(q)}$ when the reference point was D;
2. $U(S) = \frac{w^-(1-q)-w^-(1-P^*)}{w^-(1-q)}$ when the reference point was FH; and
3. $U(S) = \frac{w^+(P^*)-w^+(q)}{w^+(P^*)-w^+(q)+\lambda w^-(1-P^*)}$ when the reference point was S.

To estimate $w(P)$, we used the probability weighting function proposed by Tversky and Kahneman[21]; thus,

$$w(P) = \frac{P^\gamma}{[P^\gamma + (1-P)^\gamma]^{1/\gamma}}.$$

We also used the parameters estimated by Tversky and Kahneman[21]; these were $\gamma = 0.61$ for w^+, $\gamma = 0.69$ for w^-, and $\lambda = 2.25$. There is evidence that these (or similar) parameters are good approximations in the case of health outcomes.[22,23]

The main hypothesis tested in this study is that $U(S)$ is independent of q. For that test, the 455-person sample was randomly split into 3 subsamples for which $q = 0.25$, $q = 0.50$, or $q = 0.75$. These subsamples are, respectively, labeled F25, F50, and F75 ("F" stands for "format").

Selection of States and Subjects

This study was part of a larger project aimed at identifying the quality-of-life dimensions most affected by alcohol abuse and to estimate their relative weight. The descriptive system employed consisted of 4 dimensions with 3 levels each, and it was generated by means of 2 focus groups conducted among patients and specialists (see Table 1). Nine of the 81 possible combinations were selected to yield an orthogonal experimental design. These 9 states (shown at the bottom of Table 1) were evaluated by all participants.

Subjects were chosen from the general population of Galicia (a region in northwest Spain) by way of 4-stage random sampling using a stratified cluster design. Trained interviewers conducted a total of 455 face-to-face interviews. Each participant evaluated 9 health states, after which they reported their socioeconomic characteristics (age, sex, education, income, labor status, and type of cohabitation), their alcohol consumption, and whether they had friends or relatives with alcohol problems. All participants gave their informed consent before being included in the study.

Tasks by Group

Each subject was asked to establish indifference between lotteries, such as $[P,FH;(1-P),D]$ and $[q,FH;(1-q),S]$, for 9 health states that were presented randomly. Subjects were randomly assigned to 3 subgroups identified as F25 ($n = 147$), F50 ($n = 155$), and F75 ($n = 153$).

Indifference was obtained through a sequence of binary choices. Empirical evidence suggests that a choice-based procedure tends to result in fewer inconsistencies than do matching methods.[24–26] Figure 1 shows the sequence of potential questions in the 3 formats. Visual aids were used in all questions to help subjects understand probabilities; see Figure 2 for the visual aid

Table 1 – Dimensions, levels, and states evaluated

Dimensions and levels
Family consequences
1. No or almost no family problems
2. Moderate family problems, such as a frequent arguments, distrust, verbal abuse, or cohabitation problems
3. Serious family problems, such as traumatic separation of the couple, physical abuse within the family, or no relationship with the family
Physical health consequences
1. No or almost no effects on physical health
2. Moderate health problems, such as falls or liver inflammation
3. Serious health problems, such as cirrhosis or serious fractures
Psychological consequences
1. No or almost no psychological problems
2. Moderate psychological problems, such as guilt or shame, low self-esteem, minor depression, or memory problems
3. Serious psychological problems, such as severe depression or inconsistent behaviour
Social consequences
1. No or almost no social problems
2. Moderate social problems, such as difficulty relating to other persons or loss of interest in hobbies
3. Serious social problems, such as absence of social relationships or inappropriate social behaviour
States evaluated (numbers indicate the level of each dimension)
State 1:3132; State 2:3321; State 3:1231; State 4:1312; State 5:1123; State 6:2333; State 7:2222; State 8:3213; State 9:2111

presented with the first question. That first question was always a choice between the gambles [P,FH;(1−P),D] and [q,FH;(1−q),S], with $P = q = 0.25$ in F25, $P = q = 0.50$ in F50, and $P = q = 0.75$ in F75. The first choice was meant to determine whether the respondent considered the state S to be better or worse than death ($U(S) \gtrless 0$). The rest of the questions in the sequence of binary choices were such that the implicit utility was the same in each question that had the same position in the sequence.

Because the stimulus gambles were different for each subsample, probabilities in the matching gamble were adjusted accordingly. Suppose, for example, that the subject viewed state S as being preferable to death, then he would be presented with a second binary choice, depending on his subgroup (format) membership, as follows:

- F25: [0.7,FH;0.3,D] versus [0.25,FH;0.75,S];
- F50: [0.8,FH;0.2,D] versus [0.50,FH;0.50,S];
- F75: [0.9,FH;0.1,D] versus [0.75,FH;0.25,S].

In each of these 3 formats, the subject's choice revealed whether $U(S) \gtrless 0.6$ in Equation (1). If the second choice revealed that $U(S) > 0.6$, then the third question was used to elicit whether $U(S) \gtrless 0.8$; but if the second choice revealed that $U(S) < 0.6$, then the third one elicited whether $U(S) \gtrless 0.2$. If the subject's response to the third question revealed that $U(S) > 0.2$, then a final question was asked to elicit whether $U(S) \gtrless 0.4$. Thus each subject's final utility was estimated as an interval (0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, or 0.8-1.0).

Pairing questions among formats in this manner is not feasible when a health state is considered to be worse than death because the lowest potential utility differs depending on the format. As P becomes infinitely small, the lowest utility for F25 is −0.33, for F50 is −1, and for F75 is −3. Hence it is not possible to generate common intervals for the 3 formats below a certain point—namely, the lowest potential utility of F25. For this reason, only those answers revealing $U(S) > 0$ were considered for hypothesis testing.

Analysis

Consistency

Dominance tests were used to analyze the internal consistency of responses. One state is said to "dominate" another when the former is better in at least one dimension and it is not worse in any other dimension. In this experiment, state 9 dominated states 1, 2, 6, 7, and 8, whereas state 6 was dominated by states 3, 4, 5, 7, and 9. A dominance criterion is considered to be violated if the dominated state is assigned a greater utility interval than is the better state. Hence there were 9 possible violations of dominance to be checked for each participant.

Hypotheses

Because we accounted only for those responses that implied the focal health state was preferable to death, the first hypothesis was that the probability of a health state being viewed as better than death would not change between formats. A total of 27 proportion tests (9 states × 3 frames) were run to test this hypothesis. The second hypothesis is that utilities do not change systematically between formats; it is evaluated using both nonparametric and parametric tests.

Nonparametric test of the equality of distributions. The Kruskal–Wallis H test was conducted to determine whether there were significant differences between formats. When statistical differences were found, Mann–Whitney tests were applied to determine which of these formats differed from each other. The P-values were Bonferroni corrected for multiple comparisons.

Parametric test. A regression was performed to analyze the effect of the different formats on average utility. More specifically, the following model was estimated:

$$U_{ij} = \alpha + \sum \beta_j s_j + \gamma_t(Format_t) + \varepsilon_{ij}.$$

Here U_{ij} is the utility interval assigned by respondent i to health state j ($j = 1,\ldots,9$); s_j is a dummy variable that identified the state being evaluated; $Format_t$ ($t = 1, 2, 3$) is a dummy variable that identified the format used; ε_{ij} is an error term; and α , β_j , and γ_t are the parameters to be estimated. It is assumed that $\varepsilon_{ij} = u_j + e_{ij}$, where e_{ij} and u_i are the unobservable error terms owing to differences among (respectively) observations and respondents. Because the dependent variable is an interval variable, interval regression analysis was used to estimate the parameters. The main hypothesis was tested by analyzing the significance of γ_t.

The Stata software package was used for all statistical analyses.

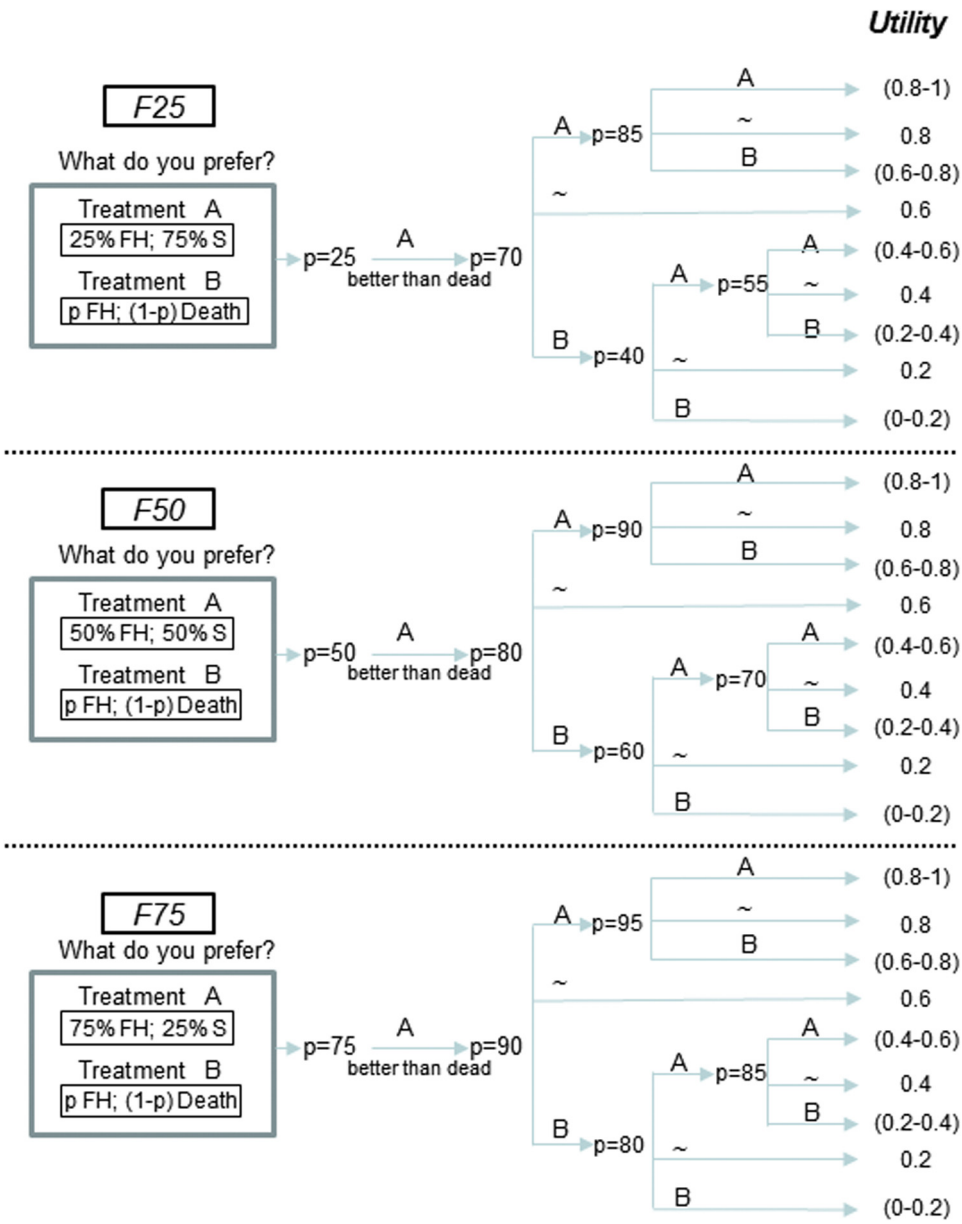

Fig. 1 – Protocol used in the questionnaire according to format.

Results

Table 2 presents the characteristics of each subsample. There were no statistically significant differences (χ^2 test at the 5% level) among the 3 subsamples except with respect to personal alcohol consumption, where the F50 subsample differed significantly from the other 2.

As for internal consistency, 314 participants satisfied all the dominance tests and another 64 participants failed only one of those 9 tests. Table 2 reports the internal consistency results by format. There were no statistically significant differences (χ^2 test at the 5% level) among the 3 subsamples. This level of consistency is similar to that observed in the literature.[27]

Table 3 shows the utility distribution for each state in each format. On average, the states were considered to be better than death by 86% of participants. Proportion tests showed that the probability of a health state being considered better than death was not significantly different among the formats for all states ($P > 0.1$ in all proportion tests).

The Kruskal–Wallis H test revealed that there was a statistically significant difference in utility between the 3 groups ($\chi^2(2) = 32.36$, $P < 0.001$). At the aggregate level, Mann–Whitney tests further showed that (1) F25 yielded lower utilities than did both F50 (Bonferroni-corrected $P = 0.009$) and F75 (Bonferroni-corrected $P < 0.001$), and (2) F50 yielded lower utilities than did F75 (Bonferroni-corrected $P = 0.021$). Similar conclusions are obtained if instead Dunn's test is employed. In terms of states, the last column of Table 3 also suggests that the higher the probability of *FH* in the stimulus, the higher the utility; for example, in all states the percentage of participants considering $U(S) \geq 0.8$ was higher for $q = 75$ than for $q = 25$, and in 8 states, this percentage increases with q. Nevertheless, the separate Kruskal–Wallis H test conducted for each state showed that only for state 1 were the differences large enough to be significant.

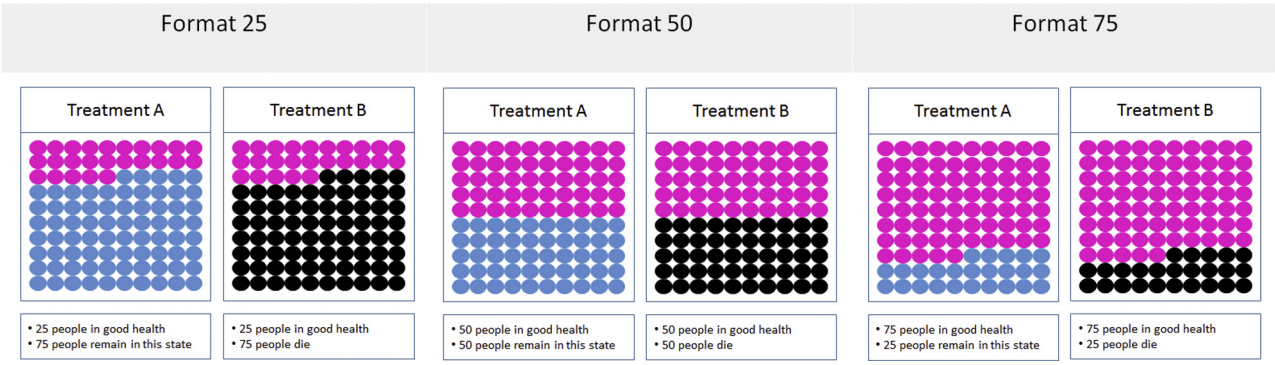

Fig. 2 – Initial visual aids by format.

The main results of our study are shown in Table 4. Under expected utility, the parameters of the *format* variable tend to increase with higher probabilities of full health in the stimulus, although significant differences (at the 5% level) are observed only between F75 and F25 (an additional test showed that there were no significant differences between F50 and F75). Under prospect theory, the results depend on the reference point. When *D* is the reference point, results are the same as with expected utility. When *S* is the reference point, there are statistically significant differences between F75 and the rest of formats (differences between F50 and F75 were checked with an additional test), and the differences between F75 and F25 are even greater than under expected utility (0.080 vs 0.046). Yet when *FH* is the reference point, we find no differences among the 3 formats. Interactions were used to see whether format-related differences changed with the health state's severity; none of the interaction terms was statistically significant.

Table 2 – Description of the samples (%)

	F50 ($N = 155$)	F75 ($N = 153$)	F25 ($N = 147$)
Sex (males)	51.0	45.1	44.2
Age (mean), years	52.1	50.0	52.2
Age distribution			
From 18 to 29 years	12.3	14.4	10.2
From 30 to 44 years	27.7	30.1	30.6
From 45 to 59 years	25.2	24.2	25.9
From 60 to 74 years	18.7	20.3	19.1
75 years and older	16.1	11.1	14.3
Level of education			
Less than primary	14.2	11.1	8.3
Primary education	37.4	38.6	44.1
Intermediate education	32.3	34.7	28.3
Higher education	16.1	15.7	19.3
Employment			
Working	41.3	41.2	44.9
Unemployed	10.3	15.7	11.6
Inactive population	48.4	43.1	43.5
Family income distribution (€ per month)			
Less than 1000	23.23	18.95	21.77
1000-1499	31.61	26.8	25.17
1500-1999	20.00	16.99	18.37
2000-2999	10.97	16.99	19.05
3000 or more	4.52	6.53	4.76
Do not know/did not answer	9.68	13.73	10.88
Lives alone	7.70	6.50	8.10
Lives with partner	72.9	71.2	71.4
Lives with children	41.3	42.5	44.2
Friend or relative with alcohol problems	22.6	23.5	24.5
Personal consumption			
Does not drink, or drinks occasionally	52.3	51.6	55.8
Drinks weekly	40.7	32.0	30.0
Drinks daily or has drunk excessively	7.1	16.3	14.3
Internal consistencies			
0 inconsistencies	64.5	73.2	69.4
1 inconsistencies	16.8	13.1	12.2
>1 inconsistencies	18.7	13.7	18.4

Discussion

This study used the *probability lottery equivalent* technique to estimate health state utilities using 3 different starting points. We found that the probability of a health state being considered worse than death does not vary with the probabilities used in the stimulus gamble. Nevertheless, under expected utility, changes in the probabilities of the stimulus gamble produced significant differences in utilities. This finding is important because expected utility is the theory that prevails in health economics when gambles are used to estimate health state utilities. In this section we deal with 2 issues: first, the explanation of the results, and second, the practical implications of our results to elicit preferences for health states.

The psychological effect of *similarity* can help explain why higher utilities are obtained when the stimulus lottery's probabilities are higher.[28] On this account, the worst outcome (death) receives greater attention when subjects choose between (0.75,*FH*;0.25,S) and [*p*,*FH*;(1−*p*),*D*] than when they choose between (0.25,*FH*;0.75,S) and [*p*,*FH*;(1−*p*),*D*] because, in the first comparison, the probabilities are more similar. It follows that, when the stimulus is the lottery (0.75,*FH*;0.25,S), the matching probabilities in the lottery [*p*,*FH*;(1−*p*),*D*] must be between 0.75 and 1.0 because the analysis is restricted to the answers revealing health states preferred to death. Because one of the outcomes (full health) is common, subjects may have focused on what is clearly different between the 2 lotteries—namely, *S* and *D*. In contrast, if the stimulus lottery is (0.25,*FH*;0.75,S) then the matching probabilities in the lottery [*p*,*FH*;(1−*p*),*D*] must lie between 0.25 and 1.0. The matching and stimulus probabilities are less similar in this case, and so subjects may have split their attention more evenly between probabilities and the 2 different outcomes (*S* and *D*).

The second explanation is that expected utility is not the best theory for analyzing these data. When the data are analyzed

Table 3 – Distribution of utilities by format and state (%)

State	Format	U > 0	0 ≤ U < 0.2	0.2 ≤ U < 0.4	0.4 ≤ U < 0.6	0.6 ≤ U < 0.8	0.8 ≤ U ≤ 1
State 1	F75	83.66	7.84	7.85	5.22	33.34	29.41
	F50	83.23	8.39	10.32	12.26	29.03	23.23
	F25	85.03	14.96	9.53	14.28	30.61	15.65
State 2	F75	80.39	16.99	9.81	1.96	24.83	26.80
	F50	78.71	9.03	12.26	12.26	23.22	21.94
	F25	80.95	13.60	8.85	10.20	32.65	15.65
State 3	F75	93.46	5.88	2.61	7.85	35.29	41.83
	F50	92.26	9.03	5.17	11.61	23.87	42.58
	F25	92.52	7.49	6.12	10.88	32.66	35.37
State 4	F75	89.54	12.42	7.19	5.22	29.42	35.29
	F50	91.61	7.74	6.45	14.84	31.61	30.97
	F25	89.12	8.85	9.52	9.53	31.97	29.25
State 5	F75	95.42	4.57	3.92	2.62	33.98	50.33
	F50	95.48	4.51	6.45	14.20	27.74	42.58
	F25	93.2	6.13	4.76	9.52	32.65	40.14
State 6	F75	60.78	16.99	3.92	3.92	11.11	24.84
	F50	59.35	12.25	7.75	9.67	14.20	15.48
	F25	62.59	14.29	6.12	12.93	18.37	10.88
State 7	F75	88.24	10.46	4.58	5.88	29.41	37.91
	F50	90.97	7.10	10.97	14.19	23.23	35.48
	F25	87.76	8.17	8.84	17.01	27.89	25.85
State 8	F75	84.97	13.73	9.15	5.23	26.14	30.72
	F50	84.52	5.81	9.68	13.55	25.80	29.68
	F25	84.35	8.84	13.61	12.24	27.89	21.77
State 9	F75	94.77	3.27	1.30	3.93	18.95	67.32
	F50	95.48	3.22	3.87	5.81	16.77	65.81
	F25	93.20	3.40	2.73	8.84	23.81	54.42
Mean	F75	85.69	10.24	5.59	4.65	26.94	38.27
	F50	85.73	7.45	8.10	12.04	23.95	34.19
	F25	85.41	9.52	7.79	11.71	28.73	27.66

under prospect theory, we find that the results depend on the reference point. If the reference point is full health then differences between framings disappear, but if the reference point is the state evaluated then differences increase. The main issue here is whether we have evidence of the reference point that the participants themselves use when performing the probability

Table 4 – Effect of format (baseline probabilities) in the utility

	Expected utility theory		Prospect theory†					
			RP: D		RP:S		RP:FH	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Format *[Ref: F25]*								
F50	0.026	0.020	0.012	0.018	0.025	0.017	−0.005	0.019
F75	0.046*	0.022	0.047*	0.020	0.080‡	0.019	0.010	0.021
State *[Ref: S9]*								
S1	−0.176‡	0.014	−0.185‡	0.014	−0.184‡	0.014	−0.186‡	0.014
S2	−0.204‡	0.016	−0.205‡	0.015	−0.201‡	0.015	−0.209‡	0.015
S3	−0.091‡	0.012	−0.106‡	0.013	−0.108‡	0.013	−0.102‡	0.013
S4	−0.137‡	0.013	−0.150‡	0.013	−0.150‡	0.013	−0.148‡	0.013
S5	−0.066‡	0.011	−0.083‡	0.012	−0.085‡	0.013	−0.077‡	0.012
S6	−0.242‡	0.019	−0.228‡	0.016	−0.221‡	0.016	−0.238‡	0.018
S7	−0.135‡	0.013	−0.146‡	0.013	−0.146‡	0.013	−0.144‡	0.013
S8	−0.165‡	0.014	−0.173‡	0.014	−0.172‡	0.014	−0.173‡	0.014
Constant	0.763‡	0.016	0.554‡	0.016	0.474‡	0.016	0.670‡	0.016

N = 455; observations = 3506.
RP, Reference point; D, death; S, state evaluated; FH, full health; SE, standard error; Coef, coefficient.
* Denotes statistical significance at $\alpha = 0.05$.
† Prospect theory with parameters obtained by Tversky and Kahneman.[21]
‡ Denotes statistical significance at $\alpha = 0.001$.

equivalent task. We do not have direct evidence about that, and we found only one study that does. Van Osch and Stiggelbout[29] used the standard gamble and found that the certain outcome was used as the reference point by 52% of their subjects; another 45.5% of them used full health as the reference point, but almost nobody used death for that purpose.

Use of the certain outcome as the reference point in a standard gamble is easily explained by the certainty effect—that is, by participants focusing too much on the certain outcome. With regard to the use of full health or death as reference points, Van Osch and Stiggelbout explained their results by proposing that "respondents involved their status quo, which likely was most equal to the high outcome." Because most subjects were in good health, full health was substantially favored (as a reference point) over death. Those results suggest that full health may have been used frequently in our experiment because the certainty effect is absent in paired gambles.

We now discuss our findings' practical relevance for eliciting utilities. One could argue that these results show that, under expected utility, paired gambles are no less prone to bias than is the standard gamble. In that case, there would be no reason to prefer the former to the latter. Nevertheless, we believe that paired-gamble biases are less severe than are those observed in the case of the standard gamble. In comparing 3 different versions of the standard gamble for 2 health states, Bleichrodt et al.[5] observe considerable differences. The same group of subjects produced utilities that varied by about 0.4 for state A and 0.25 for state B. Of course, the manipulations used by Bleichrodt et al. to compare different versions of the standard gamble are quite different from our manipulations for paired gambles. One limitation of our study is the lack of a comparison group that faced the standard gamble, and so we have no way of knowing whether such large differences would have been found for our participants. Yet we argue that the differences observed in our study are much smaller than what other scholars have found in research on the standard gamble.

Finally, we remark that—in the monetary domain—it has also been observed that utilities elicited using the certainty equivalent technique are higher the higher the probability of the best outcome in the lottery used as stimulus.[18,19] Nevertheless, the differences reported in those studies were extremely large. Thus McCord and Neufville[19] found that the average differences in certainty equivalents ranged from 30% to 100% with changes in the probabilities of the lottery used as stimulus. In contrast, the same study found—in line with the results obtained by Law et al.[20] in the health domain—no significant differences between the utilities obtained under F50 and F75 under expected utility theory.

In summary, the health state utilities obtained using paired-gamble methods are influenced by the probabilities appearing in the lottery used as stimulus. Yet there are 2 mitigating factors. On the one hand, the differences observed are lower than the differences reported in research on the standard gamble. On the other hand, our observed differences disappear when the data are analyzed using prospect theory with full health as the reference point. So even as we acknowledge the need to conduct more direct comparisons between paired-gambles and the standard gamble, we conclude that the paired-gamble approach seems a better candidate to elicit utilities for health states in situations that involve risk.

Acknowledgments

Funding for this study was provided by a grant from Spanish Ministry of Economy and Competitiveness (ECO2015-69334-R) and Regional Government of Galicia (10SEC300038PR). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

REFERENCES

1. Farquhar PH. State of the art—utility assessment methods. *Manage Sci.* 1984;30:1283–1300.
2. Brazier J, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Med Care.* 2004;42:851–859.
3. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care.* 2002;40:113–128.
4. Abellán-Perpiñán JM, Sánchez-Martínez FI, Martínez-Pérez JE, Méndez I. Lowering the 'floor' of the SF-6D scoring algorithm using a lottery equivalent method. *Health Econ.* 2012;21:1271–1285.
5. Bleichrodt H, Abellan-Perpiñan JM, Pinto-Prades JL, Mendez-Martinez I. Resolving inconsistencies in utility measurement under risk: tests of generalizations of expected utility. *Manage Sci.* 2007;53:469–482.
6. Oliver A. Testing the internal consistency of the lottery equivalents method using health outcomes. *Health Econ.* 2005;14:149–159.
7. Holt CA, Laury SK. Risk aversion and incentive effects. *Am Econ Rev.* 2002;92:1644–1655.
8. Arrieta A, García-Prado A, González P, Pinto-Prades JL. Risk attitudes in medical decisions for others: an experimental approach. *Health Econ.* 2017;26:97–113.
9. Galizzi MM, Miraldo M, Stavropoulou C, van der Pol M. Doctors-patients differences in risk and time preferences: a field experiment. *J Health Econ.* 2016;50:171–182.
10. Galizzi MM, Miraldo M, Stavropoulou C. In sickness but not in wealth: field evidence on patients' risk preferences in financial and health domains. *Med Decis Making.* 2016;36:503–517.
11. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Econ.* 2002;11:447–456.
12. McCord M, de Neufville R. Lottery equivalents: reduction of the certainty effect problem in utility assessment. *Manage Sci.* 1986;32:56–60.
13. Pinto-Prades JL, Abellán-Perpiñán JM. Measuring the health of populations: the veil of ignorance approach. *Health Econ.* 2005;14:69–82.
14. Wakker P, Deneffe D. Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Manage Sci.* 1996;42:1131–1150.
15. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica.* 1979:263–291.
16. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim JS. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Med Decis Making.* 2010;30:5–15.
17. Robinson A, Spencer A, Moffatt P. A framework for estimating health state utility values within a discrete choice experiment modeling risky choices. *Med Decis Making.* 2015;35:341–350.
18. Cohen M, Jaffray JY. Certainty effect versus probability distortion: an experimental analysis of decision making under risk. *J Exp Psychol.* 1988;14:554.
19. McCord M, Neufville R. Empirical demonstration that expected utility decision analysis is not operational. In: Stigum BP, Wenstop F, eds. *Foundations of Utility and Risk Theory with Applications.* Dordrecht: D. Reidel Publishing Company, Boston; 1983:181–199.
20. Law AV, Pathak DS, McCord MR. Health status utility assessment by standard gamble: A comparison of the probability equivalence and the lottery equivalence approaches. *Pharm Res.* 1998;15:105–109.
21. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertainty.* 1992;5:297–323.
22. Bleichrodt H, Pinto JL, Wakker PP. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Manage Sci.* 2001;47:1498–1514.
23. Bleichrodt H, Pinto JL. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Manage Sci.* 2000;46:1485–1496.
24. Attema AE, Brouwer WB. In search of a preferred preference elicitation method: a test of the internal consistency of choice and matching tasks. *J Econ Psychol.* 2013;39:126–140.
25. Bostic R, Herrnstein RJ, Luce RD. The effect on the preference-reversal phenomenon of using choice indifferences. *J Econ Behav Organ.* 1990;13:193–212.
26. Hey JD, Morone A, Schmidt U. Noise and bias in eliciting preferences. *J Risk Uncertainty.* 2009;39:213–235.
27. Krucien N, Watson V, Ryan M. Is best-worst scaling suitable for health state valuation? A comparison with discrete choice experiments. *Health Econ.* 2017;26:e1–e16.
28. Rubinstein A. Similarity and decision-making under risk. *J Econ Theory.* 1988;46:145–153.
29. Van Osch S, Stiggelbout AM. The construction of standard gamble utilities. *Health Econ.* 2008;17:31–40.