Work in progress: not for reference or citation

HESG paper prepared for the meeting to be held on 3rd to 5th July 2002

Two ways to skin a cat: a comparison of two variants of Standard Gamble

John Brazier and Paul Dolan
Sheffield Health Economics Group
School of Health and Related Research
The University of Sheffield

Abstract

This study compares the 'ping pong' version (PPV) of standard gamble (SG) developed by Torrance and others at McMaster with the 'titration' version (TV) developed by Jones-Lee and colleagues. A sample of 58 respondents was randomly divided into two groups. All respondents were asked to rank seven SF-6D health states, and then to rate them using the visual analogue scale (VAS). Respondents in Group 1 then valued four states by PPV followed by three by TV whilst those in Group 2 valued the first four by TV and then three by PPV. All respondents were then asked to comment on the ease of completion and understanding of the two variants, and to explain any logical inconsistencies in their values. The two groups were comparable in terms of gender, age and health status. Their VAS ratings of the seven states were similar. Overall, TV values were higher than those of PPV and these differences exceeded 0.1 for three health states (and were significant at the 5% level for two). Furthermore, a higher proportion of respondents assigned a value of 1.00 to health states using TV as compared to PPV. There was also evidence of an ordering effect that appeared to offset this difference for some health states. This paper discusses some of the likely explanations for the differences found and the implication of these results for using these techniques. These findings support the view that health state preferences are partly constructed by the way a question is asked and are influenced by the previous questions respondents have answered at the same administration.

Acknowledgements

This paper is very much work in progress but has already benefited from comments provided by Aki Tsuchiya.

1. BACKGROUND

The increasing use of Quality Adjusted Life Years (QALYs) in health service decision-making raises questions about the methods used to put the 'q' into the QALY. There are a range of valuation techniques commonly used to do this, including standard gamble (SG), time trade-off (TTO) and visual analogue scales (VAS). It is well established in the literature that different valuation techniques generate different health state values (Dolan, 2000; Green et al, 2000). However, over the years a number of different variants of these techniques have been developed and there is less evidence as to whether different variants also generate different values.

Expected utility theory does not suggest one variant of SG should be preferred to another. Different variants have been developed in order to aid administration and to overcome possible framing effects. This paper is concerned with two of the more widely used variants of the SG technique: 1) the 'ping pong' method developed by Torrance and others originally at McMaster University in Canada; and 2) the 'titration' method first used to value health states by Jones-Lee and colleagues at the University of Newcastle in the UK.

The paper tests the hypothesis that these methods generate the same values for a given set of health states. Where there are differences, the study also to attempts to address the reasons for them from comments given by respondents. The paper begins with a description of the two variants, followed by sections describing the methods and results of the study. There is then a preliminary discussion of the results and their implications for consideration at HESG.

2. METHODS

Basically, the SG valuation technique asks the respondent to choose between the certainty of an intermediate health state and the uncertainty of a treatment with two possible outcomes, one of which is better than the certain outcome and one of which is worse. The objective is to find the probability, p, at which the respondent is indifferent between accepting the intermediate health state and accepting the uncertain treatment. The value for the intermediate state (relative to the gamble outcomes) is then simply given by p. The SG technique is directly derived from expected utility theory (EUT) and is regarded by many health economists as the 'gold standard' amongst valuation techniques in health care (Gafni and Birch, 1993; Torrance, 1986 - but see Richardson 1994 and Dolan 2000 for alternative views). There are a number of different ways of eliciting SG values, but EUT does not suggest one variant of SG should be preferred to another.

2.1 The Variants

The *'ping pong' variant* (PPV) of the SG was developed by a team at McMaster (Furlong et al, 1990). It is interview administered and employs a prop where the probabilities are displayed on a chance board, both numerically and in the form of a pie chart. The following questions are asked for each health state. The respondent is first asked which alternative they would choose if the probability of the best outcome occurring were 1.0. This test question is to ensure the respondent understands the health

state classification. Should they choose the certain intermediate state over the certain best state, they are asked the reason for this, but such a choice is rarely made.

Next, the probability of the best outcome would be set to 0.9. Should the respondent choose the certain prospect, he/she would be asked to explicitly consider the probability of the best outcome being 0.95[1]. Where the respondent continues to choose the certain prospect, then the interpolated value for indifference is 0.975 and where the respondent chooses the uncertain prospect, the value is 0.925. A respondent who chooses the uncertain prospect at a probability of 0.9 would be asked to consider the choice with a 0.1 chance of success. Continuing to choose the uncertain prospect would lead to a value of indifference of 0.05. Choosing the certain prospect would then result in the respondent being asked to consider a probability of 0.8. This procedure continues in a 'ping pong' fashion until the respondent's point of indifference is revealed.

This version of SG is used for its ease of use by interviewers. The chance board is designed to make the interview as straightforward as possible, by leading the interviewer through a set of questions depending on the respondents answer to the previous question, and minimise the risk of interviewer variation. The developers have tried and tested the procedure and its associated prop over many years and it has become widely used in health economics, including the valuation of the Health Utility Index-II&III and the SF-6D (Feeney et al, 2002; Brazier et al, 2002). The McMaster team provided training in this variant to the authors of this paper and the chance boards were manufactured to their recommended specification at McMaster University.

The titration variant (TV) of SG was originally developed by Jones-Lee and colleagues (1993). This version lists values for the chances of success from 0 in 100 to 100 in 100. Respondents are asked to place a tick against all the probabilities of success at which they are confident they would choose the treatment and a cross against all the values where they would reject the treatment (see the appendix). They are then asked to indicate all chances of success at which they would find it most difficult to choose. Where this region covers more than one probability value, then a mid-point is taken to be the indifference value. This version of SG has been found to produce more consistent and reliable data than an interview based variant using props (Dolan et al, 1996).

2.2 Design of the survey

A convenience sample of staff and students at the University of Sheffield were invited to participate in the survey. Once recruited respondents were randomised between the two groups. Each group completed a different questionnaire. In both groups, respondents were asked to complete the SF-6D, a health state classification developed from the SF-36 (see Figure 1). The SF-6D describes health across six multi-level dimensions: physical functioning, role limitations, social functioning, pain, mental health and vitality (Brazier et al, 2002). Health states are defined by the SF-6D by taking one level from each dimension.

[1] In the original version developed by Torrance and collegues, a respondent who chooses the certain prospect would not be asked any more questions about that health state but moved onto the next one. The point of indiference would be extrapolated to 0.95 and hence the assumed value for the intermediate value.

Respondents in both groups were asked to rank and rate the same seven health states defined by the SF-6D, where one was the best state and another was the worst defined by this classification. The remaining six health states are presented on Table 1. These health states were selected to represent a range of states from very mild through to the most severe (the 'pits') state defined by the SF-6D. These six states have been selected in pairs, where each pair of states differed in only a one level of one dimension. This closeness of the three pairs has been designed in order to examine the logical consistency across methods.

Respondents were then asked to value each of these states by SG (see Table 1). Each group valued these states in the same order. Group 1 valued the first four states using PPV and then valued the remaining three states using TV. Group two valued the first four states using PPV and the other three states using PPV. All respondents were interviewed by the same person had been trained in the two methods by the authors. It should be noted that the health states were valued using full health and pits as the reference states. Pits itself has been valued against full health and death.

At the end of the interview, respondents were asked their gender, age, employment and educational status. They were then asked to comment on the questionnaire. Respondents were asked to comment on the reason for any inconsistencies in their valuations of states between the three pairs of states: Z and X, Q and Y and T and R. Respondents were then asked how difficult they found each variant and how well they understood the tasks using five point likert scales.

2.3 Analysis

The results reported in this paper focus on mean health state values (that is, raw scores that have not been adjusted by the value of 'pits' to lie on the 0-1 scale) by variant and respondent group. The significance ($p<0.01$) of any differences found is tested using the t-test. An important check of the extent to which respondents were able to understand the tasks, and of the importance of differences between the variants, is to examine the logical consistency of responses. For some pairs of states, one state will be strictly better than another when it is better on one dimension and no worse on other dimensions. The health states were selected in adjacent pairs (Z>X, Q>Y and T>R) in order to scrutinise the consistency of respondents' valuations that differ by one level in one dimension.

3. RESULTS

A total of 58 respondents were recruited from staff in the Medical Faculty at the University of Sheffield. There were 28 respondents in Group 1 and 30 in Group 2. The groups were comparable in terms of gender (9 men in Group 1 versus 10 men in Group 2), mean age (38 vs. 36) and own health status according to the mean VAS rating (0.88 vs. 0.88). The two groups ranked the states in a very similar order and their VAS ratings of the states were virtually identical, with no statistically significant differences (see Table 2).

The mean SG health state values are shown on Table 3 by variant and group. Mean health state values generated by TV are significantly higher than those produced by the PPV variant for the first set of states valued (i.e. Y, Z, R and pits). The differences are

.107, .038, .130 and .138 respectively and are all significant except Z. For the second set of states X, Q and T, that are valued by each group using the other SG variant, there is little or no difference between the variants.

In terms of consistency with the logical ordering across the three pairs of states (i.e. Z>X, Q>Y, T>R), the pattern is mixed (see Table 4). VAS mean values are consistent across all three pairs. However, for the mean SG values there was inconsistency within variant and between variant. Strict inconsistencies arise between Z and X within the PPV variant and between Q and Y for TV. Between variant inconsistency was measured within group, since each group valued the same states but used different variants. Therefore, Group 1 was inconsistent between Z (PPV) and X (TV) and Group 2 between Q (TV) and Y (PPV). In terms of difficulty, most respondents found the two variants quite difficult through to fairly easy (see Table 5). Only one person found either technique very difficult. However, the responses indicate a significant difference in favour of the PPV. Whilst most respondents said they fully understood the tasks (see table 6), there was again a small and significant difference in favour of PPV.

Respondents who were found to have produced inconsistent responses were asked to comment on why they thought the difference arose, and 24 people responded to this question. The most common reason given by respondents was ‘confusion’, finding it ‘hard to remember’ and ‘forgot’. A number of respondents thought the tasks might have been responsible for the difference: "Board leads to more risk taking – more attractive" and PPV "‘yo’s, yo’s you around", whereas TV "makes you more cautious". Some respondents also pointed out that the two variants had different scales and some apparent inconsistencies. Finally, respondents were asked whether there were any other differences between the variants. The most common comment (n=16) was that they found PPV easier and many liked the visual aid provided (n=7). A smaller number of respondents (n=4) commented that they found the PPV props confusing and offering only a limited number of choices (n=2).

DISCUSSION

The evidence from this study does not support the hypothesis that these two variants of SG produce the same answers. Overall, it would seem that TV tends to generate higher health state values than PPV. For three states this exceeded 0.1, which could have a potentially important impact on the final incremental cost per QALY of an intervention, especially if the differences across other states were not the same. (If they were the same, the incremental cost per QALY would not be different for movements between dysfunctional health states but would still be different for those interventions that prevent death or return people to full health).

Different valuations across variants were only found for four out of the seven states - for three states the differences were very small. These mixed results suggest that there may be more than one reason for the differences found. An obvious reason for the difference is that the four states where TV exceeded PPV were valued first by both groups. This would suggest that the valuation of the second set of states may have been affected by the valuation of the first four states. The first four states therefore represent a pure comparison of variants whereas the other three states have been influenced by the task that went before.

The possibility of a *reference point effect* (see Dolan and Robinson, 2001) would suggest that the TV valuations in Group 1 of the last three states were lower than would otherwise be the case since they were 'dragged down' by the respondents memory of the valuation of the first four states by PPV. For Group 2, the PPV valuation of the last three states are 'dragged up' by the TV valuation of the first four states. The net effect of this reference point effect is then to offset the variant effect.

The finding that TV values exceed PPV values could be due to an *anchoring effect* (see Dolan et al, 1996), whereby TV starts eliciting preferences at the upper end of the scale. PPV, on the other hand, iterates respondents between the upper and lower ends and thereby avoids an anchor point bias. An alternative view would be that the ping pong procedure actually confuses people by asking them to 'jump around' between extreme values (imagine thinking that a state is only slightly worse than full health and then being taken from a 100% chance of success to only a 10% chance of success).

So, which variant should be used? PPV seems to be easier for respondents, many of whom like the visual aids and the iterative procedure. However, others found the ping pong procedure confusing, and it could be argued that TV allows people to think through their answer in their own time and to simultaneously consider the range of possible risks. The opposing argument would be that the TV method is prone to an anchoring effect. In relation to the choice of variant, it is worth noting that it is possible to use the titration method with the McMaster board (in much the same way as York MVH study used a form of titration with the TTO board). It might also be possible to ping pong between values in a column of responses, although this might be somewhat confusing in the context of a self-completion exercise.

From the results presented here, it appears that SG values elicited using PPV after TV are the same as SG values elicited using TV after PPV. Since the reference point and anchoring effects seem to exactly cancel one another out, it could be argued that it doesn't matter which method you use - so long as you first elicit some 'warm-up' values with the other method. So, you can drag PPV values up by using TV first, or drag TV values down by using PPV first. But the fact that there is a very real variant effect to start off with, and a strong order effect thereafter, should warn against such a strategy. These results highlight still further the fact that health state valuations are partly constructed during the process of elicitation, and suggest that stated preferences in health must be elicited with care.

REFERENCES

Brazier J, Roberts J, Deverill M. The estimation a preference-based single index measure for health from the SF-36. *Journal of Health Economics 2002; 21(2):271-292*

Dolan P, The measurement of health-related quality of life for use in resource allocation decisions in health care, *Handbook of Health Economics*, North Holland, Culyer AJ and Newhouse J, 1723-1760, 2000.

Dolan P., Gudex C., Kind P, and Williams A. "Valuing health states: a comparison of methods", *Journal of Health Economics* 1996, 15, 209-231.

Dolan P and Robinson A, The measurement of preferences over the distribution of benefits: The importance of the reference point, *European Economic Review*, 45, 9, 1697-1709, 2001.

Feeney, D., Furlong, W., Torrance, G., Goldsmith, CH., Zenglong, Z., DePauw, S., Denton, D., Boyle, M. Multi-attribute and Single-Attribute Utility Functions for the Health Utility Index Mark 3 System. *Medical Care* 2002,40(2) :113-128.

Furlong, W., Feeny D., Torrance, G.W., Barr, R., Horsman J. (1990) Guide to design and development of health state utility instrumentation. Cntre for Health Economics and Policy Analysis Paper 90-9, McMaster University, Hamilton, Ontario.

Green C, Brazier J, Deverill M (2000) A Review of health state valuation techniques. *Pharmacoeconomics* 17(2):151

Jones-Lee M, Loomes G, O'Reilly D, Phillips, P. "The value of preventing non- fatal road injuries: findings of a willingness-to-pay national sample survey." Transport Research Laboratory, 1993.

Mehrez A., Gafni A. "Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress" Medical Decision Making 1993; 287-292.

Richardson, J. (1994) "Cost-utility analysis: what should be measured*?", Social Science and Medicine* 39(1), 7-21.

Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics, 5,* 1-30.

Figure 1: The Short Form 6D

Level	**Physical Functioning**
1	Your health does not limit you in vigorous activities
2	Your health limits you a little in vigorous activities
3	Your health limits you a little in moderate activities
4	Your health limits you a lot in moderate activities
5	Your health limits you a little in bathing and dressing
6	Your health limits you a lot in bathing and dressing

Level	**Pain**
1	You have no pain
2	You have pain but it does not interfere with your normal work (both outside the home and housework)
3	You have pain that interferes with your normal work (both outside the home and housework) a little bit
4	You have pain that interferes with your normal work (both outside the home and housework) moderately
5	You have pain that interferes with your normal work (both outside the home and housework) quite a bit
6	You have pain that interferes with your normal work (both outside the home and housework) extremely

	Role limitations
1	You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems
2	You are limited in the kind of work or other activities as a result of your physical health
3	You accomplish less than you would like as a result of emotional problems
4	You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems

	Mental health
1	You feel tense or downhearted and low none of the time
2	You feel tense or downhearted and low a little of the time
3	You feel tense or downhearted and low some of the time
4	You feel tense or downhearted and low most of the time
5	You feel tense or downhearted and low all of the time

	Social functioning
1	Your health limits your social activities none of the time
2	Your health limits your social activities a little of the time
3	Your health limits your social activities some of the time
4	Your health limits your social activities most of the time
5	Your health limits your social activities all of the time

	Vitality
1	You have a lot of energy all of the time
2	You have a lot of energy most of the time
3	You have a lot of energy some of the time
4	You have a lot of energy a little of the time
5	You have a lot of energy none of the time

Footnote: The SF-36 items used to construct the SF-6D are as follows: physical functioning items1, 2 and 10; role limitation due to physical problems item 3; role limitation due to emotional problems item 2; social functioning item 2; both bodily pain items; mental health items 1 (alternate version) and 4; and vitality item 2.

Table 1: Health states valued by SG variant

SF-6D state[1]	State code	Group 1	Group 2
433333	Y	PPV[2]	TV
211211	Z	PPV	TV
535554	R	PPV	YV
645655	Pits	PPV	TV
333333	Q	TV	PPV
211212	X	TV	PPV
53554	T	TV	PPV

1. Each digit in the health state represents the level on each of the six dimension
2. PPV - ping pong variant; TV - titration variant

Table 2: Mean health state VAS rating

SF-6D state	State code		
		Group 1	Group 2
111111	H	.993	.994
433333	Y	.528	.556
211211	Z	.877	.899
535554	R	.306	.298
645655	Pits	.185	.159
333333	Q	.624	.667
211212	X	.833	.842
535544	T	.373	.380

Table 3: Mean health state SG values by group and variant

SF-6D state	State code	PPV		TV	
		Group 1	Group 2	Group 1	Group 2
433333	Y	.647			.754*
211211	Z	.875			.913
535554	R	.482			.612*
645655	Pits	.458			.596*
333333	Q		.748	.750	
211212	X		.883	.882	
535544	T		.626	.615	

* where the difference between variants reaches significance

Table 4: Consistency with logical ordering

Pair	VAS	PPV	TV	Group 1	Group 2	
Z>X	✓		✓		✓	
Q>Y	✓	✓		✓		
T>R	✓	✓	✓	✓	✓	

Table 5: Degree of difficulty

	PPV %	**TV %**
Very difficult	0.0	1.7
Quite difficult	24.1	41.4
Neither difficult nor easy	27.6	20.7
Fairly easy	32.8	27.6
Very easy	8.6	1.7

Wilcoxon rank test = 0.019

Table 6: Understanding of task

	PPV %	**TV %**
Fully understood	87.9	74.1
Partially understood	5.2	13.8
Did not really understand	0.0	5.2

Wilcoxon rank test = 0.005

Appendix

Please put a √ against all cases where you are CONFIDENT that you would CHOOSE the risky treatment (Choice B).
Please put an X against all cases where you are CONFIDENT that you would REJECT the treatment (Choice B) and accept the certain health state (Choice A).
Please put an = against all cases where you think it would be most difficult to choose between the treatment (Choice B) and accept the certain health state (Choice A).

Outcome of treatment

Chances of success		Chances of failure
100 in 100*		0 in 100*
95 in 100*		5 in 100
90 in 100		10 in 100
85 in 100		15 in 100
80 in 100		20 in 100
75 in 100		25 in 100
70 in 100		30 in 100
65 in 100		35 in 100
60 in 100		40 in 100
55 in 100		45 in 100
50 in 100		50 in 100
45 in 100		55 in 100
40 in 100		60 in 100
35 in 100		65 in 100
30 in 100		70 in 100
25 in 100		75 in 100
20 in 100		80 in 100
15 in 100		85 in 100
10 in 100		90 in 100
5 in 100		95 in 100
0 in 100		100 in 100

* You may be willing to accept the treatment but *only* if it has a chance of success of *higher* than 95 in 100 (*i.e.* a chance of failure which is less than 5 in 100). If so, at what level of success would you accept treatment?

Choice A

Q

Your **health** limits you a little in moderate activities

You **accomplish less** than you would like as a result of emotional problems

Your health limits your **social activities** some of the time

You have **pain** that interferes with your normal work (both outside the home and housework) a little bit

Your feel **tense or downhearted** and low some of the time

You have a lot of **energy** some of the time

Choice B

Success

H

Your **health** does not limit you in vigorous activities

You have no problems with your **work or other regular daily activities** as a result of your physical health or any emotional problems

Your health limits your **social activities** none of the time

You have no **pain**

You feel **tense or downhearted** and low none of the time

You have a lot of **energy** all of the time

OR

Failure

P

Your **health** limits you a lot in bathing and dressing

You are limited in the kind of **work or other activities** as a result of your physical health and **accomplish less** than you like as a result of emotional problems

Your health limits your **social activities** all of the time

You have **pain** that interferes with your normal work (both outside the home and housework) extremely

You feel **tense or downhearted** and low all of the time

You have a lot of **energy** none of the time