

Ordering disparate health programmes through explicit ranking versus willingness to pay: 'sticky wicket' or 'sticking with it'?

Cam Donaldson^{1,2}, Phil Shackley¹ and Gillian Currie^{3,4}

1. Centre for Health Services Research, School of Population and health Sciences, University of Newcastle upon Tyne
2. Business School (Economics), University of Newcastle upon Tyne
3. Department of Community Health Sciences, University of Calgary
4. Department of Paediatrics, University of Calgary

Paper presented to Health Economists' Study Group and Collège des Économistes de la Santé, City University, London, 4-6th January 2006

Introduction

This paper continues the debate on whether willingness to pay (WTP) can be used to compare the relative value of alternative health programmes when these programmes are competing for the same resources within a fixed budget. Recent important results from the EuroWill project have demonstrated a lack of convergence between programme orderings based on explicit ranking and orderings (or implicit rankings) derived from respondents' WTP-values for the competing programmes (Olsen et al., 2005). To use a cricketing analogy, it would seem that continued use of WTP in such contexts would mean that one is batting on a 'sticky wicket'!

However, 'sticking with it', this paper reports results from two projects conducted in Canada and England, which followed EuroWill. The reasons these projects were undertaken are threefold. First, it can be seen from Table 1 of Olsen et al. (2005) that results on convergent validity improve in line with the chronological order of the six EuroWill studies – this order being as listed in that table. (Although the intention is for this paper to be free-standing, for purposes of the conference, we have included an Appendix in which the EuroWill project is summarised and Table 1 from the Olsen et al. paper is reproduced. This provides the reader with more background, but also highlights the significant amount of EuroWill work that was based on France-UK collaborations.) Second, because of language differences, detailed exploration of what lay behind respondents' valuations, through qualitative analysis, was not possible in EuroWill. The opportunity to explore this issue arose in Canada, although these results are not reported here. Third, as reported by Olsen et al. (2005), as EuroWill progressed, and the lack of convergent validity became apparent, a marginal approach was developed as a possible way forward, and tested in the final EuroWill country – Ireland (Shackley and Donaldson, 2002). As with the main EuroWill project, the essence of the approach is to elicit values for a number of proposed expansions (or contractions) to existing health care programmes. However, rather than asking for separate values for each of the competing programmes, the marginal approach involves taking the lowest ranked programme and eliciting a partial WTP value for this first. The next worst ranked programme is then presented to respondents who are asked to state how much *more* they would be willing to pay for this programme compared to the lowest ranked programme. A similar question is asked for successively higher-ranked programmes. Questions can also be devised for programmes which are tied in the initial simple ranking. Although the marginal approach did not enhance convergent validity to any great degree in Ireland, adjustments to this approach were made which required subsequent testing.

Thus, what we report here are results from Canada, using the standard EuroWill approach, as reported in Olsen et al (2005), and from a comparator sample using the marginal approach. In England, the marginal approach alone was used to compare four disparate health programmes (two expansions and two contractions) instead of the usual three (all expansions) as in EuroWill. In the following section, we speculate on why the EuroWill results as reported by Olsen et al. (2005) may have arisen and how they could be improved upon. Important differences between EuroWill and subsequent studies in Canada and England are outlined before reporting on results from these countries, which demonstrate significant improvements in convergent validity on earlier EuroWill results.

Sticking with it: improving on EuroWill?

It could be postulated that the EuroWill results were not as bad as might appear to be the case. One suggestion was that partially consistent responses in Shackley and Donaldson (2002) could be due to respondents having a relatively weak strength of preference between two or more programmes, thus leading them to assign the same WTP value to each. For example, even the results from Olsen et al. (2005) show that, despite not being a perfect match in terms of implied and explicit rankings, 316 respondents (25%) across the six countries who explicitly ranked between all three programmes, still used WTP to discriminate between two of these programmes (giving the other two equal WTP). This would seem to display some potential for further development of the method.

Another suggestion put forward in relation to the Ireland survey reported by Shackley and Donaldson (2002) was that the wording of the questionnaire may have led respondents to believe they were being asked to pay for all the programmes, thus leading them to come up against a perceived budget constraint when asked to value their highest ranked programme. In the Ireland survey, the marginal WTP questions were phrased as follows:

“How much more would you be willing to contribute each year to expand the programme compared to the programme?”

It was speculated that the term “compared to” may have led respondents to believe they were being asked to pay for all three programmes from their budget rather than any one programme. The example used to illustrate this was of a respondent who has a nominal budget of £50 which they are willing to allocate to health care. If after valuing each programme s/he is effectively given this budget back (which is what was intended in the interview), then s/he has the potential to value each programme up to a maximum of £50. However, if s/he is under the impression that after valuing the first programme, his/her budget for the subsequent programmes has been reduced by the amount they said they were willing to pay for that first programme, then there is the potential for WTP values for the second and/or third programmes to be affected. Specifically these values may be lower than would otherwise have been the case.

To counter this possible effect, the term “compared to” was removed from the WTP questions in the studies reported here. In addition, attempts were made to make it more explicit to respondents that for each programme evaluation their ‘budget’ had not been diminished by any WTP values they may have stated for previous programmes. Specifically, the wordings for Canada and England were respectively as follows (the first example relates to a programme ranked second in a 123 ranking in the Canada survey and the second to a programme ranked third in a 1234 ranking from the survey in England):

“I would now like you to consider the health care program you ranked second – the programme. Please ignore the other two programs for now; the only thing I want you to consider is this program.”

“I would now like you to consider the programme you ranked third – the programme. In considering this programme, I would like you to imagine that you

have been given back the amount you said you were willing to pay for the previous programme. So forget about the previous programme and assume we are back to 'square one'. We're only concerned with the programme in this section."

By contrast, the wording used in the Ireland survey was much less specific, merely asking respondents to assume the other two programmes were not available.

The purchase of moral satisfaction or warm glow was suggested as another possible reason to explain the persistence of the inconsistencies in the Ireland (marginal) survey (Kahneman and Knetsch, 1992). Moral satisfaction is purchased by stating a positive WTP amount for what the respondent regards as a good cause. It accrues from the signalling of approval through a positive contribution, with the size of the contribution being of secondary importance. In some respects, the marginal approach is designed to avoid this, and it was hoped that the improved wording outlined above would help in this regard. One other way in which the surveys described in the paper attempted to do this was to have successive programmes valued over and above the one immediately below it in the respondent's ranking; in Ireland, every other programme was valued over and above the 'baseline' value given for the lowest ranked programme. The rationale for doing this in Ireland was so as not to totally force rationality, giving respondents an opportunity to still be inconsistent; thus a potential criticism of the surveys reported on here is that such rationality was forced.

Of course, a more fundamental reason why some respondents give inconsistent responses is that they perceived the ranking exercise and the valuation of the programmes through WTP to be different from one another. It was speculated in the Ireland marginal survey that different wording in the ranking and WTP questions might have provoked this interpretation among respondents. In that survey, the ranking exercise asked individuals to indicate how important they thought the programmes were (potentially causing respondents to think from the perspective of society) while the WTP questions focused on value (perhaps implying a more individualistic notion of value). This wording was also used in the survey in Canada (the actual wording being to "compare these programmes and order them in terms of their importance"), whereas in England in the surveys reported here, the wording was amended with the intention of conveying the notion of individual value in both contexts (so, for the ranking exercise, the wording asked the respondent to "place these programmes in order of how highly you value them starting with the one you like most. When doing this, concentrate on how much you value the proposed expansions and how you value preventing the proposed reductions from going ahead").

Results

As argued in Olsen et al. (2005), in accordance with microeconomic theory, we would expect consistency between the ranking implied from WTP and the previously stated *explicit rankings* of respondents. If a respondent ranks A before B, the expectation is that s/he would signal this preference by indicating a higher WTP for A over that for B. Thus, we would hypothesise that WTP shows convergent validity with explicit ranking.

In this paper, the same classification procedure as that used by Olsen et al. (2005) is employed. Using the example of three alternatives, if three different WTP values were given, it means *strong ranking* (P P), i.e. one programme was strictly preferred to a second that was strictly preferred to the third. If two equal values and one different, there is a *weak ranking* (P I or I P), i.e. respondents preferred one programme to the other two, between which they were indifferent, or, they were indifferent to the first two, both of which were preferred to the third. *No rankings* (I I) are of two kinds; either three identical positive WTP-values or three zeros. Again, as with EuroWill, the data arise from interviews with randomly selected members of the population.

Table 1 shows the numbers from each country. The first column shows the implied rankings based on the WTP-answers. The subsequent columns show which type of ranking the respondents within each of these groups had stated on the *explicit ranking* (ER) question. Thus, each number in the WTP-column represents the corresponding row sum under ER. As with Olsen et al. (2005), complete convergent validity would be revealed if numbers were seen in the diagonal only, i.e. combination P P in WTP and P P in ER; combination P I or I P in WTP and P I or I P in ER, and; combination I I in WTP and I I in ER. It can be seen that, much more so than was the case with Olsen et al. (2005), this is the case with these data, especially when focussing on results from the two surveys based on the marginal approach.

Again, as with Olsen et al. (2005), Table 2 summarises the results across the six countries. First, the (unweighted) average across all three surveys shows that 46% stated strong rankings (PP) in their WTP-answers and that 55% had done so in their explicit ranking – unlike the respective figures of 23% and 73% from Olsen et al. (2005). Second, in Olsen et al. (2005) the reluctance to rank the programmes through WTP was shown through 41% of respondents indicating indifference between the three programmes according to this method of valuation; the corresponding figure across the three surveys reported here being 10%. Third, an important type of inconsistency reported by Olsen et al. (2005) is the high proportion of respondents who stated indifference through WTP, but who in the previous preference elicitation exercise had stated strong ranking. This figure is much lower in the three Canada/England surveys; 21% as opposed to the earlier figure of 64%. Lastly, the column ‘Convergent validity’ shows the percentages of respondents in the diagonals of Table 1; 55% in Canada/England as opposed to only 34% in EuroWill, although, without including the results from the standard survey in Canada, the average (i.e. for the two surveys using the marginal approach) is 69%.

Discussion

The work presented here demonstrates the positive outcome of continually improving a method over a period in excess of ten years, starting with the original Northern Norway survey undertaken by Olsen and Donaldson (1998). The marginal approach has been shown to work well in the narrower context of patients valuing alternative forms of close substitutes (Donaldson, 2001) and now with more disparate alternatives.

Despite these promising results, however, the method will still have to stand up to the scrutiny of the qualitative research which also took place alongside the experiments reported above. In addition, experiments involving scope tests (currently underway in

Calgary) also need to be reported. However, it might be that we can be more optimistic about the potential for the application of WTP methodology in health care than many have been up to now.

References

Donaldson C (2001) Eliciting patients' values by use of 'willingness to pay': Letting the theory drive the method. *Health Expectations*, 4: 180-188.

Kahneman DJ and Knetsch J (1992) Valuing public goods: the purchase of moral satisfaction. *Journal of Environmental Economics and Management*, 22, 57-70.

Olsen JA and Donaldson C (1998) Helicopters, hearts and hips: Using willingness to pay to set priorities for public sector health care programmes. *Social Science and Medicine*, 46:1-12.

Olsen JA, Donaldson C, Shackley P and the EuroWill Group (2005) Implicit versus explicit ranking: on inferring ordinal preferences for health care programmes based on differences in willingness-to-pay. *Journal of Health Economics*, 24: 990-996.

Shackley P and Donaldson C (2002) Should we use willingness to pay to elicit community preferences for health care? New evidence from using a 'marginal' approach. *Journal of Health Economics*, 21: 971-991.

Table 1: Comparing implicit rankings from WTP with explicit rankings (ER)

<i>Canada - standard version</i>	WTP	ER			
		PP	PI or IP	II	missing
PP	64	47	15	2	0
PI or IP	74	37	33	4	0
II – posWTP	14	5	4	5	0
II – zeros	8	3	3	2	0
Total	160	92	55	13	0

<i>Canada – Marginal version</i>	WTP	ER			
		PP	PI or IP	II	missing
PP	102	88	13	1	0
PI or IP	45	11	33	1	0
II – posWTP	6	0	6	0	0
II – zeros	7	2	1	4	0
Total	160	101	53	6	0

<i>England – Marginal</i>	WTP	ER			missing
		PPP	PPI/IPP/ IIP/PII	III	
PPP	35	33	2	0	0
PPI/IPP/IIP/PII	51	8	39	4	0
III – posWTP	11	0	1	10	0
III – zeros	3	3	0	0	0
Total	100	44	42	14	0

Table 2: Survey comparisons of WTP-ranking with explicit rankings (ER) (missing included) and comparisons with EuroWill

	PP(P) in WTP	PP(P) in ER	II(I) in WTP	II(I) in ER	% of II (I) in WTP giving PP(P) in ER	Convergent validity
Canada – standard	40	58	9	8	36	29
Canada – Marginal	64	63	8	4	0	55
England	35	44	14	14	27*	82
Survey average	48	55	10	9	21	57
EuroWill Country average	23 %	73 %	41 %	11 %	64 %	34%

** This 27% is made up people who were all protesters – i.e. were not willing to pay for any of the programmes on the grounds they felt they paid enough already through taxation. This is comparable with Olsen et al. (2005). In the equivalent groups in the Canadian versions, there were no protesters.*

Appendix: the EuroWill project

Olsen and Donaldson (1998) tested the feasibility of using WTP to value disparate alternatives which could be funded from the health care budget. The various problems and issues arising from that study, and in the broader literature through the 1980s and 1990s, led to a project which gathered together a wider network of researchers around Europe who were interested in improving the method of WTP in the context of eliciting community values for priority setting. 'EuroWill', a project funded by the European Commission, involved surveys in six European countries (Donaldson, 1999b).

From Table A1, it can be seen that the project has addressed several important issues and has shown: the importance the population puts on community *vis-à-vis* acute services (in Ireland); the existence of ordering effects, reasons for them and, thus, potential solutions; the difficulty of detecting scope effects (whether through giving respondents different sizes of effect, in Portugal, and different numbers of people benefiting from the programme, in Norway); that the closed-ended approach yields significantly higher WTP values than the payment scale, adding to the small body of literature which contradicts the recommendation of the NOAA Panel to use the former on the basis that it leads to more 'conservative results'; the different values that can be obtained by providing respondents with different amounts of information about the same option; and that improved econometric techniques for estimating factors associated with WTP are required for surveys in which three multiple alternatives are valued by each respondent, and are possible to achieve (O'Shea et al., 2002; Stewart et al., 2002; Olsen et al., 2004; Ryan et al., 2004; Protière et al., 2004; Luchini et al., 2002).

An overarching issue, across the six countries was the extent to which orderings based on a simple ranking procedure matched those derived from WTP questions. These are the results which were summarised in Olsen et al. (2005), reproduced here in Tables A2 and A3, and are what this particular paper is about.

For those who wish to follow up on any EuroWill papers on issues introduced above, the references are as follows:

Donaldson C (1999b) Developing the method of 'willingness to pay' for assessment of community preferences for health care. Final report to Biomed 2 Programme (PL950832) of the European Commission. Health Economics Research Unit, University of Aberdeen and Departments of Economics and Community Health Sciences, University of Calgary. (E-mail cam.donaldson@ncl.ac.uk for a copy.)

Luchini S, Protière C and Moatti J P (2003) Evaluating several willingness to pay in a single contingent evaluation: application to health care. *Health Economics* 12: 51-64.

Olsen JA, Donaldson C and Periera J. The insensitivity of 'willingness to pay' to the size of the good: new evidence for health care. *Journal of Economic Psychology* 2004, 25: 445-460.

Olsen JA, Donaldson C, Shackley P and the EuroWill Group (2005) Implicit versus explicit ranking: on inferring ordinal preferences for health care programmes based on differences in willingness-to-pay. *Journal of Health Economics*, 24: 990-996.

Olsen JA, Kidholm K, Donaldson C and Shackley P. Willingness to pay for public health care: a comparison of two approaches. *Health Policy* 2004; 70: 217-228.

O'Shea E, Stewart J, Donaldson C and Shackley P (2002) Eliciting preferences for resource allocation for health care. *Economic and Social Review* 32: 217-238.

Protière C, Donaldson C, Luchini Stéphane, Moatti JP and Shackley P. The impact of information on non-health attributes on willingness to pay for multiple health care programmes. *Social Science and Medicine* 2004, 58: 1257-1269.

Ryan M , Scott D and Donaldson C (2004) Valuing health care using willingness to pay: a comparison of the payment card and dichotomous choice methods. *Journal of Health Economics* 23: 237-258.

Shackley P and Donaldson C (2002) Should we use willingness to pay to elicit community preferences for health care? New evidence from using a 'marginal' approach. *Journal of Health Economics*, 21: 971-991.

Stewart J, O'Shea E, Donaldson C and Shackley P (2002) Do ordering effects matter in willingness to pay studies of health care? *Journal of Health Economics* 21: 585-599.

The EuroWill team(with current positions) comprised:

Professor Cam Donaldson, University of Newcastle upon Tyne UK [EuroWill Coordinator]

Dr Kristian Kidholm, MUUSMANN Research & Consulting, Denmark

Professor Andrew Jones, University of York, UK

Professor Michael Jones-Lee, University of Newcastle upon Tyne, UK

Professor Graham Loomes, University of East Anglia, UK

Professor Jean-Paul Moatti, Inserm Research Unit 379 (Marseilles), France

Professor Jan Abel Olsen, University of Tromsø, Norway (joint co-ordinator)

Dr Eamon O'Shea, National University of Ireland Galway, Republic of Ireland

Dr Joao Pereira, Universidade Nova de Lisboa, Portugal

Dr Christel Protière, Inserm Research Unit 379 (Marseilles), France

Professor Mandy Ryan, University of Aberdeen, UK

Dr David Scott, Fourth Hurdle Consulting, London, UK

Dr Phil Shackley, University of Newcastle upon Tyne, UK

Dr Jennifer Stewart, McMaster University (Ontario), Canada

Table A1: EuroWill: issues addressed and numbers of responses in each country

Country (dates of survey)	Areas of care	Issues addressed	Numbers receiving different versions of the questionnaire	Total in each country
Norway (March 1997)	More heart operations More cancer treatments Helicopter ambulance	Insurance versus community-based questions Size of effects	Community-based = 80 Insurance-based = 83 Community-based (2 cancer progs) = 79 Community-based (all progs for less people) = 81	323
Portugal (Oct-Nov 1997)	More heart operations More cancer treatments Improved car ambulance	Size of risk	Cancer, hearts, ambulance = 104 Cancer, hearts, improved hearts = 103 Cancer, improved hearts, ambulance = 103	310
Denmark (April 1998)	More heart operations More hip operations More cataract operations	Insurance-based versus community- based questions Test-retest	Insurance = 168 Community = 168 Test-retest = 50	386
UK (May-June 1998)	More heart operations More cancer treatments Helicopter ambulance	Payment scale versus closed- ended	Payment scale = 236 Closed-ended = 342	578
France (Oct-Nov 1998)	More heart operations More cancer treatments Helicopter ambulance Reduction in pollution	Process utility Cognitive capacity Test-retest	No process = 100 Neutral process = 104 Positive process = 99 Pollution/no process = 154 Pollution/process = 149 Test-retest = 50	353
Ireland (April 1999)	More heart operations More cancer treatments More long-term care	Marginal approach Ordering effects	Basic approach = 113 Marginal approach = 121 Different ordering = 101	335
Grand total				2285

Table A2: Comparing the degree of explicit ranking (*strong, weak, no*) with the degree of ranking as implied from WTP.

<i>Norway</i>	<i>WTP-ranking</i>	<i>Explicit ranking</i>			
		<i>strong</i>	<i>weak</i>	<i>no</i>	<i>missing</i>
<i>Strong</i>	51	48			3
<i>Weak</i>	39	34			5
<i>no – posWTP</i>	47	39		1	7
<i>no – zeros</i>	25	21			4
Total	162	142		1	19

<i>Denmark</i>	<i>WTP-ranking</i>	<i>Explicit ranking</i>			
		<i>strong</i>	<i>weak</i>	<i>no</i>	<i>missing</i>
<i>Strong</i>	40	39			1
<i>Weak</i>	101	100			1
<i>no – posWTP</i>	83	78	1	2	2
<i>no – zeros</i>	87	75	1		11
Total	311	292	2	2	15

<i>France</i>	<i>WTP-ranking</i>	<i>Explicit ranking</i>			
		<i>strong</i>	<i>weak</i>	<i>no</i>	<i>missing</i>
<i>Strong</i>	80	79	1		
<i>Weak</i>	75	67	6	2	
<i>no – posWTP</i>	45	41	1	3	
<i>no – zeros</i>	101	86	11	4	
Total	301	271	19	9	

<i>Portugal</i>	<i>WTP-ranking</i>	<i>Explicit ranking</i>			
		<i>strong</i>	<i>weak</i>	<i>no</i>	<i>missing</i>
<i>Strong</i>	63	54	7	2	
<i>Weak</i>	92	71	16	5	
<i>no – posWTP</i>	19	13	3	3	
<i>no – zeros</i>	33	19	11	3	
Total	207	157	37	13	

<i>Scotland</i>	<i>WTP-ranking</i>	<i>Explicit ranking</i>			
		<i>strong</i>	<i>weak</i>	<i>no</i>	<i>missing</i>
<i>Strong</i>	38	25	10	3	
<i>Weak</i>	53	23	24	5	1
<i>no – posWTP</i>	52	14	16	22	
<i>no – zeros</i>	3	2	1		
Total	146	64	51	30	1

<i>Ireland</i>	<i>WTP-ranking</i>	<i>Explicit ranking</i>			
		<i>strong</i>	<i>weak</i>	<i>no</i>	<i>missing</i>
<i>Strong</i>	15	11	3	1	
<i>Weak</i>	34	21	6	7	
<i>no – posWTP</i>	46	16	5	25	
<i>no – zeros</i>	18	4	5	9	
Total	113	52	19	42	

Table A3: Country comparisons of WTP-ranking with explicit rankings (ER) (missing included)

	<i>Strong</i> in WTP	<i>Strong</i> in ER	<i>No</i> in WTP	<i>No</i> in ER	% of <i>no</i> in WTP giving <i>strong</i> in ER	Convergent validity
Norway	31 %	88 %	44 %	1 %	83 %	30 %
Denmark	13 %	94 %	55 %	1 %	90 %	13 %
France	27 %	90 %	49 %	3 %	87 %	31 %
Portugal	30 %	76 %	25 %	6 %	62 %	37 %
Scotland	26 %	44 %	38 %	21 %	29 %	49 %
Ireland	13 %	46 %	57 %	37 %	31 %	45 %
Country average	23 %	73 %	41 %	11 %	64 %	34%