Is it always irrational to provide cardinally intransitive person trade-off responses?

Richard Edlin
Sheffield Health Economics Group, ScHARR
University of Sheffield

1.1. Introduction

Within health economics, the person trade-off (PTO) is used to assess the relative merits of treating two groups. In this method, a point of indifference is found between two interventions by holding the numbers benefiting in one group constant and varying the numbers in the other. This paper attempts to show how certain types of intuitive reasoning in forming person trade-off responses can lead to outcomes that violate the assumptions underlying this method.

A PTO response is assumed to indicate an individual judgement of the unique societal trade-off between people in two groups. Overall, such responses are assumed to be consistent with a wider social welfare function that could be used as a basis for policy. If these statements are true, then the ratio of PTO responses for groups {X, Y} and {X, Z} will allow us to infer the societal trade-off between groups Y and Z. This assumption of cardinal transitivity is essential if this method can be directly applied as a basis for policy, as recommended by Nord et al (1999). This paper considers a potential problem whereby well-reasoned, and hence rational, responses do not have this property and hence may be cardinally intransitive.

Of course, as judgements of “rationality” hinge on what is intended by the term, I begin by briefly discussing its definition under two different models of preference. Following this, I outline a potential cause of intransitivity and present some preliminary empirical evidence as to its significance in elicited PTO responses.

1.2. Rationality under articulated and constructed preferences

Welfare economics typically assumes that preferences over an individual's goals can be approximated by a single utility function. Welfare economics does not necessarily argue, however, that we use such a utility function when making decisions. For instance, Lucas (1987) argues that we make decisions using adaptive rules that have been found to be useful over a range of situations and which are no longer revised. In this case, our preferences will be stable and hence can be modelled by a utility function that can reliably predict preference over outcomes without necessarily reproducing these adaptive rules.

In the case where such a function exists, "rational" can refer to the properties of the approximating utility function. For instance, expected utility theory expects and requires that if one is indifferent between any two lotteries X and Y, then one will also be indifferent between lottery X and a mixture of lotteries, $b\mathrm{X} + (1-b)\,\mathrm{Y}$ for some b on [0,1]. However, this "betweenness" assumption is not borne out by empirical evidence (Chew and Walker, 1986; Camerer, 1989; Camerer and Ho, 1994). If we continue to assume that a set of stable decision rules exist, then this descriptive inaccuracy suggests that a different functional form incorporating a weaker axiom may provide a better fit. Experimental economics has expended considerable effort in this direction, with only moderate success in resolving anomalies in individual decision-making under risk (see Starmer (2000)).

An alternative conclusion, which leads us closer to a psychologist's view of preference, suggests that the decision rules we use to make a decision are not fixed in a single context, let alone across a range of situations. Within this view, preferences are constructed within a given situation and construction allows considerable leeway in the preferences an individual exhibits.[1]

Loomes and Sugden (1995) provide a useful welfare economic model of behaviour that is generally consistent with constructed preferences. In this random preference model, the preference function is randomly selected from a family of utility functions that differ in their coefficients. Whilst an individual always behaves consistently with the core theory, they will not necessarily answer consistently when asked a question on different occasions. Note, however, that this is not a complete explanation of preference, as it does not predict where, why, or how such indeterminacy occurs.

More generally, "rationality" does not require that the same question be answered in precisely the same way over time or across contexts. If individuals behave consequentially, the degree of rationality exhibited refers to whether preferences are consistent with the ends

[1] For a discussion of constructed preferences, see Fischhoff (1991) and Slovic (1995) within psychology and McFadden (1999) within economics.

considered relevant at that point in time. In this sense, rationality is a much weaker criterion under constructed preferences. In practice, however, the use of preferences at a policy level must also require that the preferences are relatively stable.

The view taken of preferences therefore determines "rationality". Where preferences are assumed to reflect stable decision rules modelled by a utility function, "rationality" refers to a person's adherence to a set of axioms that not only guarantee consistency with individual goals but which also give structure to the approximating utility function. Under this view, one can be perfectly rational in the sense of using a consistent and well-reasoned argument but irrational in the sense that the utility function used does not encompass your particular method of decision-making. Conversely, under the "constructed" version of rationality, it is difficult to ever substantively disqualify preferences as irrational. What can be done under this weaker definition of rationality, however, is to show how a well-reasoned set of preferences can be acceptable (rational) under constructed preferences but violate the assumptions used to structure the *interpretation* of preferences.

I argue here that the PTO method may suggest some intuitive arguments that, if given weight in forming preferences, will lead to cardinally intransitive responses. That is, preferences that are rational in the sense that they are well reasoned and likely to be robust to repeated enquiry could violate the crucial assumption of cardinal transitivity. By concentrating on the weaker rationality assumption used under constructed preferences, it is thus possible to argue that the PTO may be ill-suited for use as a method of eliciting societal preferences over health care.

1.3. <u>The PTO method</u>

When Nord (1992) proposed the PTO, he also viewed it primarily as a method for the valuation of health states. This has remained the major empirical interpretation of the PTO method, as evidenced by Green (2001), in which all eight empirical papers used the PTO as a valuation technique. A search of MEDLINE in July 2002 for more recent papers revealed three new articles (Sanderson and Andrews, 2001; Jelsma et al, 2002; Baron and Ubel, 2001), all of which used the PTO as a valuation tool. Whilst potentially intuitive, the PTO has been argued to lack a theoretical underpinning and only limited evidence exists as to its acceptability, feasibility and empirical validity (Brazier et al, 1999; Green, 2001). As such, the intuitive appeal of the PTO as "asking the right question" is the main basis for claims as to its acceptability as a basis for policy.

Dolan (1998) argued that PTO "responses … contain the relative weights a respondent attaches to at least four things: (1) the severity of the pre-intervention health state; (2) the severity of the post-intervention health state; (3) the health gain as a result of intervening; and (4) the number of persons treated." As PTO responses may include information other than the health states concerned, it represents a general method to find distributive preferences across changes in health rather than a method of valuing health states. As such, it may be necessary to decompose PTO weights to show how much each factor contributes towards the final weighting given to one group over another. If societal values are elicited and are not decomposed, for instance by using the value of cures of severe vs. milder illnesses from the PTO-3, then we would know the societal values placed on cures but not necessarily the weight placed on treatments that do not return a person to full health.

Nord et al (1999) "decompose" societal value into three parts. Here, the first part of their equation is the standard HRQoL gain from treatment, the second a "severity weight" ($S(\bullet)$) based on the pre-treatment level of health, and the final part a measurement of preferences for the realisation of potential health improvements ($P(\bullet)$). This weighting function can be summarised for a treatment taking health from h_i to h_i' as follows.

$$v(h_i, h_i') = (h_i' - h_i) \times S(h_i) \times P\left(\tfrac{h_i' - h_i}{1 - h_i}\right) \qquad (1)$$

where $h_i > h_i' \geq 0, S:[0,1] \rightarrow [0,1]$, and $P:[0,1] \rightarrow \Re^+$

Whilst Nord et al (1999) argue that the severity and potential weights are "well documented" and "highly relevant" concerns that should be included in all economic evaluations, no published study at present uses the PTO explicitly as a tool to show the societal weightings of two benefits enumerated in health status terms. Therefore, the ability of the PTO to produce reliable, soundly constructed preferences of this type is particularly uncertain.

One potential weakness of treating each group's society value separately (as in (1) above) is that this ignores a class of arguments that individuals may use in answering PTO questions. That is, individuals may choose to make direct comparisons across the two groups highlighted in the PTO. If such comparisons are made then it is possible that differences between the two groups highlighted in the PTO may lead to cardinal intransitivity.

1.3.1 Cardinal transitivity and the PTO

When answering a PTO question, individuals may construct a response by considering different arguments that fit the context of a choice between two groups. These arguments will not necessarily be appropriate for choices in a wider setting that considers all groups in society. If arguments are used that rely on the 2-group comparison but do not hold across an *n*-group society, then PTO responses may prove incompatible with societal value judgments. That is, whilst the PTO has an intuitive appeal as a method of finding societal value, not all intuitive arguments are necessarily suitable for use in policy considerations.

If person trade-offs are to be useful in setting policy, then we must assume internal consistency (approximately) holds in the sense of ordinal and cardinal transitivity. Ordinal transitivity requires that if greater weight is put on group *X* than group *Y*, and more in group *Y* than group *Z*, we must also have a greater weighting on group *X* than group *Z*. Cardinal transitivity is more stringent and requires that the ratios of benefit for those in *X*, *Y* and *Z* must satisfy the restriction that they can be represented $v(X):v(Y):v(Z)$.

The experimental results on this stronger assumption are mixed. Ubel et al (1996) and Baron et al (1999) both find that this later assumption is violated in the majority of cases considered. In a study by Dolan and Tsuchiya (2002), they too found that neither ordinal nor cardinal transitivity hold in a sample asked to provide PTO values for age weights. However, when those who violate ordinal transitivity were removed from the sample, the evidence for cardinal intransitivity became insignificant.

As cardinal transitivity is a desirable property, violations may be interpreted as errors or a failure of individual rationality in the sense that individuals do not appear to naturally produce results that are internally consistent. In this case, there is little that can be done beyond revising methods to encourage the respondent to double-check their results. A second alternative is possible, however, if transitivity fails not because of a failure in individual rationality but in incorrect modelling assumptions. Multiplicative transitivity holds under a societal value formulation such as that in (1) because the value assigned to treating a member of each group is assessed separately. That is, for a PTO between two groups X and Y, the PTO response gives the ratio of societal values:

$$\frac{N_X}{N_Y} = \frac{v(h_{iY}, h'_{iY})}{v(h_{iX}, h'_{iX})}$$

After adjusting for any small cognitive errors, the observed ratio between two groups should equal the chained value.

$$\left.\frac{N_Z}{N_Y}\right/\frac{N_Z}{N_X} = \left.\frac{v(h_{iY}, h'_{iY})}{v(h_{iZ}, h'_{iZ})}\right/\frac{v(h_{iX}, h'_{iX})}{v(h_{iZ}, h'_{iZ})} = \frac{v(h_{iY}, h'_{iY})}{v(h_{iX}, h'_{iX})} = \frac{N_X}{N_Y}$$

Conversely, if the societal value to each group is not evaluated separately, then persistent biases will exist in the method. Specifically, suppose people use arguments that make judgements based on the relative levels of the two groups in a PTO. In this case, the PTO-based weight given to a group depends on which comparison (chained or direct) is used to obtain its relative weight. Further, note that if an argument holds in some comparisons but not others, multiplicative consistency will not generally hold.

Now PTO values are formed as a distributive judgement between *two* groups of people. Consider a very simple 3-group society in which some in the first group are in health 0.1 but with available funds the health of some could be improved to 0.2, those in the second are in health 0.2 but some could be at 0.4 and those in group three are in 0.3 but some could be in 0.7. Suppose that the PTO was to be used to determine which group received better health.

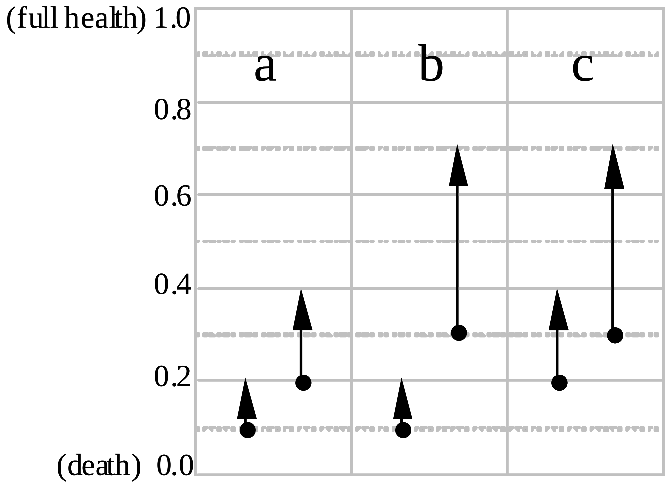


Figure 1: Three-link PTO chain

In Figure 1, the PTO values for two of these values will allow us to find the PTO value for the third if the PTO responses satisfy cardinal transitivity. In (a) and (c), those in worse-off group can achieve the same health as the better-off group when treated. In PTO (b), however, the worse-off group will lag behind the better-off group even when treated. If given positive weight, this argument suggests that the observed values in (b) will give greater weight to benefiting the worse-off group than is suggested by the chaining of (a) and (c). Here, cardinal transitivity fails and there is no clear evidence of irrationality. Worse, without transitivity, it is no longer possible to characterise PTO responses as uniquely suggesting the societal value of each programme. At this point, a rational individual (in the constructed sense) may answer in a way that suggests the PTO in inconsistent with a wider social welfare function. Here, the SWF's assumption that implicitly assumes that each group is evaluated separately is descriptively false and should not be viewed as an error. Instead, a revised method to estimate societal preferences would be required.

1.4. A study of factors affecting choice in PTO data

The experiment here combines data from two small studies to consider a range of potential factors that could be used in making decisions. The first study considered the effect of perspective on PTO responses, whilst the second considered which factors were likely to be important in PTO responses.

Any tests for which factors are significant in PTO judgements will require that we isolate a group of factors.[2] Further, given the limited number of questions available in a single study, it was considered important to focus on a small number of factors that may enter into reasoning but which would not require a large amount of data. As all published studies are valuation-based, most attention is focused on the effect of the pre-treatment health state on the PTO. Such a focus is not required here and as such, I focus on four additional factors that may be considered important in constructing a PTO response.

The first three arguments are generally compatible with the PTO methodology, since there is no difficulty in including each within a societal value function similar to that in (1). These are the gain from treatment, whether death follows non-treatment and whether a group can be returned to full health. The fourth argument suggests that the response from a PTO weight will favour a group more when they appear to lag behind another group, even when treated.

If PTO responses are used to form a generalised (societal) CEA, then the first three factors can be operationalised as a weighting on individual QALYs. That is, the PTO is used to form a rule that determines what the appropriate cost-per-QALY threshold will be for a particular intervention. For instance, the QALY threshold for life-saving treatments will be lower than the QALY threshold for non life-saving treatments if societal preferences indicate a preference towards saving lives. The fourth factor, however, cannot be incorporated within a generalised CEA. Worse still, if such a factor exists, then policy based on the PTO (or a similar tool) may use values that are biased by which of the (n-1) possible trade-offs is used to obtain the societal value for that particular group.

1.4.1 DATA

The PTO question used here differs from the approach suggested by Nord et al (1999). In particular, the questions do not give descriptions of particular health states but instead give changes in terms of HRQoL values. In these questions, interventions were characterised as an arrow between the *pre- and post-treatment health state values.*

This chapter considers PTO data from two sources. In each study, 100 people in X were compared with an unknown number of people in Y, where those in X had worse pre-treatment health and those in Y had a greater capacity to benefit. The first study presented six questions in the format of Figure 2. The subjects were led through a sample question and presented with a suggested method focussing on a “scales” analogy where “balance” represented indifference. All non-negative values of the PTO were accepted and the interpretation of both zero and infinite responses was given verbally.

[2] For instance, the linear dependence between no-treatment health, post-treatment health and benefit requires that at least be dropped.

Sample Question:

If programme X is chosen then 100 people benefit by 4 units of health, increasing from 2 to 6.

If programme Y is chosen then an unknown number of people benefit by 6 units of health, increasing from 3 to 9.

How many people would have to be helped by programme Y for you to find it difficult to choose between the groups?

Answer: ________

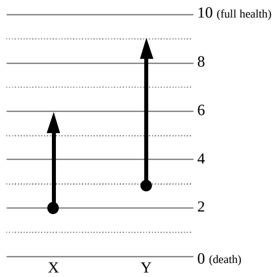


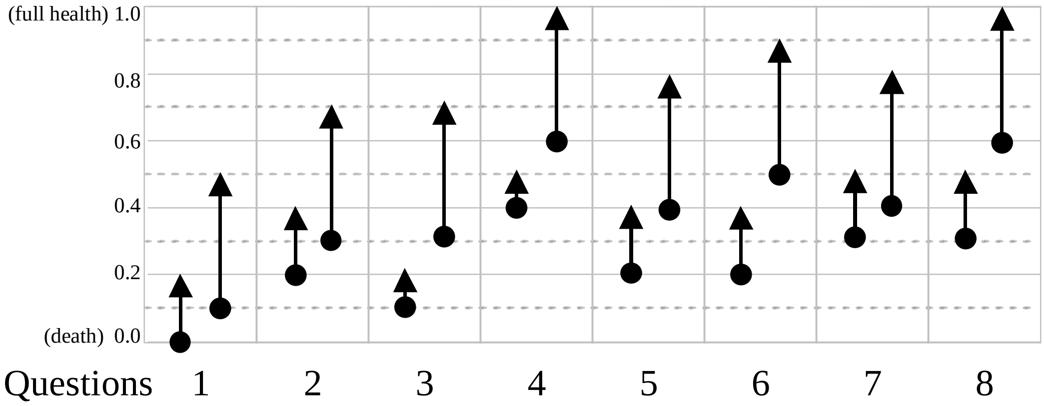
Figure 2: Sample question

The data from this source was composed of 47 students from a hall of residence in Sheffield. Each person answered questions from two societal perspectives. In the first (unconnected) perspective, a question was presented in which the answers provided were intended to apply as "policy advice" from the Department of Health to another country. In the second (connected) perspective, the answers were intended to apply to a UK context.

In the second study 41 economics students from the University of Sheffield answered questions at the end of a lecture. The students were asked to choose between two groups of people from a perspective where they were not immediately affected by their choice, although no guidance was given as to whether they were ultimately affected. Those answering in this study answered six questions that provided pre-treatment and post-treatment health state values under programmes X and Y in a table.

1.4.2 Methods

Questions 1 and 2 are common to all three perspectives, with the unconnected societal perspective also asking Question 3. The connected societal perspective presents Question 4, whilst the remaining questions (5 through 8) are specific to the societal perspective of the second dataset. Figure 3 provides a breakdown of question incidence by dataset. In non-parametric testing below the effects of different perspectives are ignored, although dummies are used in the regression analysis to capture differences in average weight.



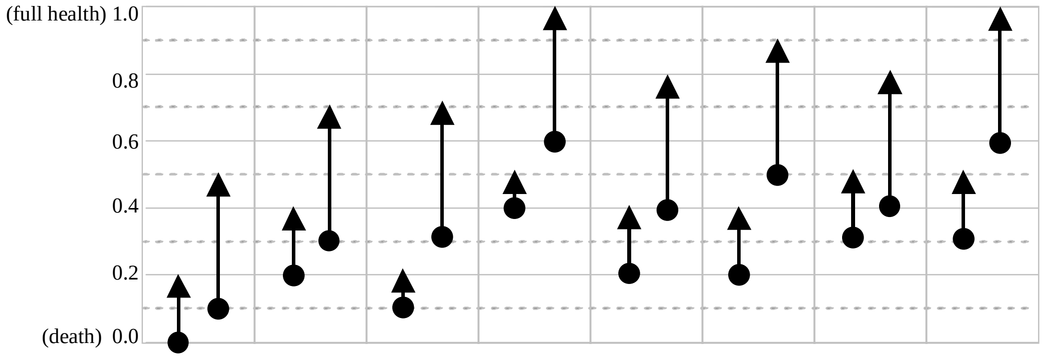
Connectedness & Dataset									
Unconnected	1	✓	✓	✓					
Connected	1	✓	✓		✓				
Uncertain	2	✓	✓			✓	✓	✓	✓

Figure 3: Questions 1-8 and appearance by their perspective

In all questions, those in Y benefit by 0.4 where treated. The LOWGAIN variable discriminates between those in which those in X benefit would benefit by 0.1 (LOWGAIN = 1) and where the benefit is 0.2 (LOWGAIN = 1). The relative weight placed on X vs. Y is expected to increase in LOWGAIN as it is hypothesised that will be greater weightings (per unit of benefit) on small improvements in health relative to large improvements.

The DEATH variable identifies those questions in which death is a potential outcome for those in X (DEATH = 1) and where it is not (DEATH = 0). Weightings are also expected to generally increase in DEATH due to a premium on life-saving treatments. The FULL variable signifies those questions in some in Y may reach full health when treated (FULL = 1, and 0 otherwise). FULL may increase the weighting on benefiting X against Y because it may considered unjust to return those in Y to full health where those in X are left in very poor health. Equally, if there is a premium placed on returning people to full health then the weighting on X vs. Y will fall.

The final factor is modelled using the variable GAP. GAP = 1 where those in the first group will have lower final health than the those in the second regardless of treatment decision, and zero otherwise. Figure 4 shows the values of LOWGAIN, DEATH, FULL and GAP for each of the eight questions.



Questions	1	2	3	4	5	6	7	8
LOWGAIN	0	0	1	1	0	0	0	0
DEATH	1	0	0	0	0	0	0	0
FULL	0	0	0	1	0	0	0	1
GAP	0	0	1	1	0	1	0	1

Figure 4: Questions 1-8 and associated explanatory variables

The study uses 2-tailed non-parametric statistics to test for differences in the distribution of responses to individual questions that differ in respect of only one of these factors. (For instance, a comparison between Questions 6 and 8 can test FULL.) As there is little control for the starting point of both groups in the PTO, a simple regression analysis is used to give some indication as to the strength of findings.

1.4.3 RESULTS

The responses obtained vary widely; in only Questions 5 through 7 is the maximum response is within a factor of 10 of the minimum. Infinite responses were received for Questions 1, 2 and 3.[3] This wide dispersal of responses required the use of non-parametric statistics. Note that in order to imply differences in the means or medians using non-parametric statistics we must assume that there are no other differences in the distributions.

[3] The "infinite" responses and one response of 100,000 were scaled back to 50,000. .

Table 1 presents the results of the non-parametric tests between questions that isolate an individual factor. In each comparison, the relevant questions appear in the table alongside the median weight on that question and the p-value for the non-parametric test (Wilcoxon matched-pairs signed-ranks and Mann-Whitney U test, as appropriate).

Argument	N	Questions Tested				p-value
		Q_1	Median Weights	Q_2	Median Weights	
LOWGAIN = 1 vs. LOWGAIN = 0	47,41	Q3	2.00	Q6	1.00	**0.045**
	47,41	Q4	2.00	Q8	1.60	0.879
DEATH = 1 vs. DEATH = 0	135	Q1	1.70	Q2	1.20	**0.000**
	41	Q1	1.70	Q7	1.40	**0.057**
FULL = 1 vs. FULL = 0	47,47	Q4	2.00	Q3	2.00	**0.093**
	41	Q8	1.60	Q6	1.40	0.640
GAP = 1 vs. GAP = 0	41	Q6	1.40	Q2	1.20	**0.042**
	41	Q6	1.40	Q5	1.30	0.382
	41	Q6	1.40	Q7	1.40	**0.052**

Table 1: Results of factor tests for matched pair data

The LOWGAIN factor appears to change the distribution of responses between Questions 3 and 6, causing the weighting on the benefit to group X to double relative to group Y. However, no significant difference was found in the comparison between Questions 4 and 8, even though the median weighting is greater on Q4 than Q8. DEATH appears highly significant on comparisons between Q1 and both Q2 and Q7. The FULL variable does not appear to change the median weighting between PTO groups, even though there is a significant difference in overall distribution between Questions 3 and 4.

The GAP variable appears to be significant in only two of the three tests. In comparisons of Question 6 against Questions 2 and 7, there appears to be a significant difference between the response distributions. In the comparison between Question 6 and 7, the median weightings are identical but there appears to be a greater skew to the left in the distribution of Question 6. This causes the UQ weighting on Q6 to be larger than the equivalent value for Q7.

All four factors are therefore significant in at least one case. It is not possible, however, to assess what the relationships between these factors are because of the limitations of non-parametric methods. An alternative here is to run a regression of relative weight against the factors LOWGAIN, DEATH, FULL and GAP, with two dummy variables added to remove some of the variability between the three datasets. This regression appears as (a) in Table 2 and explains very little of the variation in the relative weightings. Here, only the DEATH variable is significant within this regression and the parameter estimates found are very large. As this regression is likely to reflect only the largest responses, the data set was "pruned" to remove some of the excess variability. Whilst conclusions from an incomplete sample from a data set are likely to be biased, such conclusions could give a better indication of the relative strength of different factors.[4]

The model was modified by regressing only on those observations with a PTO response that had a relative weight for benefiting the worse-off group of less than 10. This removed only 4% of the observations and so is unlikely to greatly bias the regression. The resulting relationship in (b) gives far smaller estimates on effect size and significant coefficients on the LOWGAIN, DEATH and GAP variables in the expected directions. The effect from the perspective dummy variables is minimal. The model, as a whole, becomes significant (p = 0.0003). However, the model still captures very little of the underlying variation in the data (R^2-*adj* = 0.038).

In order to rectify this, a GLS random effects model was constructed with effects at the subject level. The coefficients on the models remain very similar in Regression (c) but the significance of both the model and the GAP variable improve. 20% of the underlying model is identified as due to the random effects.

That the constant is above one in (c) suggests that a benefit of 0.2 is weighted more heavily (per-unit of health gain) than a benefit of 0.4. Given that LOWGAIN is also positive, it further suggests that a benefit of 0.1 is considered to have a higher societal value (per-unit of health gain) than a benefit of 0.2. The sign and significance of DEATH suggests that a premium is placed on life-saving treatments. The FULL variable is again insignificant, whilst the significance of the GAP variable suggests that individuals may be comparing the two groups in the PTO in a way that does not allow the construction of a societal value function from PTO data.

Finally, regression (d) adds two variables that measure the starting health state of the two groups. Neither of these variables is significant, and they cause the magnitude of the remaining variables (except the constant and LOWGAIN) to fall appreciably. Here, the DEATH and GAP variables remain positive but are now insignificant.

1.5. Discussion

It appears unlikely that the data provided here is consistent with previous PTO valuation studies, since only 4% of responses provided a relative weight above 10. In contrast, Nord (1996) argues that a movement from “death” to “completely disabled” was judged to be 4,000 times as important as the curing of a slight illness, where both these changes were

[4] A repeat of this type of study would possibly include a pre-determined test that a person has understood the questions asked. Picking the “rational” people from a study may yield less variable results, as in Dolan and Tsuchiya (2002*).

	Regression (a) ALL DATA Ordinary Least Squares	Regression (b) *w* < 10 Ordinary Least Squares	Regression (c) *w* < 10 Generalised Least Squares (Random Effects)	Regression (d) *w* < 10 Generalised Least Squares (Random Effects)
Total responses used	527	504	504	504
Constant	97.65	1.478 ***	1.474 ***	1.574 **
LOWGAIN	190.18	0.596 **	0.530 **	0.705 **
DEATH	219.48 **	0.754 ***	0.810 ***	0.529
FULL	- 4.514	- 0.401	- 0.356	- 0.200
GAP	- 92.46	0.408 *	0.452 **	0.043
StartX				- 3.382
StartY				2.098
zerorisk (dummy variable)	- 155.99	- 0.006	0.017	- 0.065
lowrisk (dummy variable)	- 182.88	- 0.017	0.014	0.011
	R^2 (adj.) = 0.003	R^2 (adj.) = 0.038	Overall R^2 = 0.049	Overall R^2 = 0.050
	F(6,521) = 1.24	F(6, 498) = 4.26 ***	$\chi^2(6)$ = 32.07 ***	$\chi^2(8)$ = 33.16 ***
ρ (variance from random effects)			0.200	0.196

*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level

Table 2: Regression results

argued to be approximately equal in benefit terms. Given that few responses were of a large magnitude, and the estimates of additional weight placed on life-saving treatments varied from 0.53-0.81 when these responses were removed, it appears that very different results were obtained here.

Only two of the three non-parametric tests of GAP found significant difference. This could be explained using a second factor that is also incompatible with a SWF. This second comparison-based factor would relate to whether or not the worse-off group can achieve better health than the other group when treated. Suppose people give lower weight to benefiting the worse off group in instances where they achieve better health than the group who are (initially) healthier. Here, the arrows on the PTO diagram overlap.

Consider Figure 5. Comparisons between Question 6 and either Question 2 or 7 finds significant results. The comparison between Question 6 and Question 5 find no such difference in distribution. Whilst there is no “gap” in Question 5, it is only just possible for the worse-off group to “catch-up” to the better off group, that is, there is no overlap. In Questions 2 and 7, however, the arrows do overlap and a significant finding exists. Whilst this second factor is potentially important, it cannot be included in the regression because any OVERLAP variable would be very nearly collinear to the combination of the constant and GAP variable.

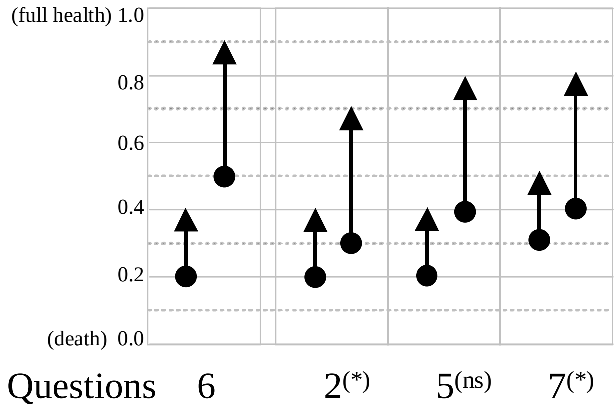


Figure 5: Tests for GAP – Question 6 versus 2, 5 and 7

Given that any use of a “gap” or “overlap” argument calls the existence of a PTO-derived societal value into question, the issue of its potential significance is important. In the non-parametric tests it appears that both may be important and the issue of whether or not such factors exist appears to merit further research. In the final regression above, the addition of starting health levels in regression (d) causes the GAP variable to become insignificant. One modification to the method (alongside direct tests of cardinal transitivity) would be to see whether comparison-based arguments are deemed important at both high and low levels of health.

1.6. Summary

The arguments identified here may lead to situations in which 2-group PTOs give a misleading impression of how sensitive we are as people to inequality. This would not only question the degree to which PTO responses can be incorporated within a SWF but would also have implications for any “valuation” PTO. In particular, consider a case where an “overlap” argument is used when constructing responses. Suppose a series of health states A, B, C, D, E and F exist, where A-E are ordered in terms of decreasing illness and F is full health. The PTO-3 measure of Pinto Prades (1997) starts by eliciting a PTO-1 on health state A. From here, PTOs are presented for movements from A to F versus B to F, B to F versus C to F, etc. In the presence of an “overlap” argument, these will produce different answers than successive evaluations of A to B versus B to C; B to C versus C to D; and so on. It is not clear which would be “correct” as such a judgment must be based on a normative and psychological examination of the argument.

Given the empirical findings here (and the possibility of other, as-yet-unobserved effects), further research should be taken into the way that PTO questions are answered and the justifications used when giving responses. In the absence of such work, any application of PTO data to policy issues is concerning. Whilst the PTO, suitably defined and analysed, can indicate the types of arguments people make to themselves when making decisions, not all such arguments are necessarily appropriate for an *n*-group society. In particular, some people may the PTO question in a context where *only those two groups are assumed to count*.

In a societal question, the comparison between two arbitrary groups is not sufficient to make decisions for society, as these decisions may also depend on other groups. Suppose two groups exist which have the property that the best conceivable health for X is slightly worse than the current health of Y. We could, presumably, give greater weight to benefiting X than Y due to equity concerns. If we then gave additional weight to X on the basis of a small "gap" in health, then we do nothing at all wrong within the intuition of the PTO exercise. Let this increase the relative weight on X vs. Y from 1.5 to 2.0.

In doing so, we make the implicit assumption that only X and Y count, therefore ignoring group Z, whose health is almost as bad as group X and who could reach the same health as group Y when treated. Group X, when treated, receives better health than Z's current state and so it receives no additional weight. Suppose we would choose to give a relative weight of 1.05 for benefiting X vs. Z. If our preferences (without the additional weight given to the gap) are cardinally transitive then our weight on Z vs. Y is around 1.43, compared to a chained value of 1.90. The implied weight on X vs. Y would be 1.5 compared to the direct comparison of 2.0.

If we accept the direct comparison of X and Y as a societal weight then by ignoring all other groups, then we answer as if society is composed of types X and Y only. PTO-derived weights may therefore by misleading as they suggest an improper context. Instead, the method used to find societal value – that is, the relative weight placed on benefiting different groups in society – should emphasise the full range of groups in society.

References

Baron J., Z. Wu, D.J. Brennan, C. Weeks, P. Ubel (1999), “Analog scale, magnitude estimation and person trade-off as measures of health utility: biases and their correction”, *Journal of Behavioral Decision Making* 14: 17–34

Baron, J., and P. Ubel (2001), “Revising a priority list based on cost-effectiveness: the role of the prominence effect and distorted utility judgements, *Medical Decision Making* 21:278-87

Brazier, J., M. Deverill, C. Green. R. Harper and A. Booth (1999), “A review of the use of health status measures in economic evaluation”, *NHS R&D HTA Programme*, Vol. 3 No. 9

Camerer, C.F. (1989), “An Experimental Test of Several Generalized Utility Theories”, *Journal of Risk and Uncertainty* 2:61-104

Camerer, C.F. and T. Ho (1994), “Violations of the Betweenness Axiom and Non-linearity in Probability”, *Journal of Risk and Uncertainty* 8:167-96

Chew, S.H. and W.S. Waller (1986), “Empirical Tests of Weighted Utility Theory”, *Journal of Mathematical Psychology* 30:55-72

Dolan, P. (1998), “The measurement of individual and social welfare”, *Journal of Health Economics* 17:39-52

Dolan, P. and A. Tsuchiya (2002*), “The person person trade-off and the transitivity principle: An example from preferences over age weighting”, Forthcoming in *Health Economics*

Fischhoff, B (1991), “Value Elicitation: Is There Anything There?”, *American Psychologist* 8:835-47

Green, C. (2001), “On the societal value of health care: what do we know about the person trade-off technique?”, *Health Economics* 10:233-43

Jelsma, J, D. Shumba, H. Kristian, W. De Weerdt, and P. De Cock (2002), “Preferences of urban Zimbabweans for health and life lived at different ages”, *Bulletin of the World Health Organization* 80:204-9

Loomes, G.C. and R. Sudgen (1995), “Incorporating a stochastic element into decision theories”, *European Economic Review* 39:641-8

Lucas, R. (1987), "Adaptive Behavior and Economic Theory", in R. Hogarth and M. Reder (eds) *Rational Choice: The Contrast Between Economics and Psychology*. (University of Chicago Press: Chicago)

McFadden, D. (1999), “Rationality for Economists?”, *Journal of Risk and Uncertainty* 19:73-105

Nord, E. (1992), “Methods for the quality adjustment of life years”, *Social Science and Medicine* 34:59-69

Nord, E. (1996), “Health status index models for use in resource allocation decisions”, *International Journal of Technology Assessment in Health Care* 12:31-44

Nord, E., J.L. Pinto, J. Richardson, P. Menzel and P. Ubel (1999), “Incorporating societal concerns for fairness in numerical valuations of health programmes”, *Health Economics* 8:25-39

Pinto Prades, J.L. (1997), “Is the person trade-off a valid method for allocating health care resources?”, *Health Economics* 6:71-81

Sanderson, K. and G. Andrews (2001), “Mental disorders and burden of disease: how was disability estimated and is it valid?”, *Australian and New Zealand Journal of Psychiatry,* 35:668-76

Slovic, P. (1995), “The Construction of Preference”, *American Psycologist* 50:364-71

Starmer, C. (2000), “Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice Under Risk”, *Journal of Economic Literature* 38:332-82

Ubel PA, G Loewenstein, D Scanlon and M Kanlet (1996) Individual utilities are inconsistent with rationing choices: a partial examination of why Oregon’s cost-effectiveness list failed. *Medical Decision Making* 16: 108-16