MONASH University
Business and Economics

Centre for Health Economics

Research Paper 2010 (53)

The Relative Social Willingness to Pay (RS-WTP) Instrument: Updated Results

Professor Jeff Richardson

Foundation Director, Centre for Health Economics
Monash University

Angelo Iezzi

Research Fellow, Centre for Health Economics
Monash University

Dr Kompal Sinha

Research Fellow, Centre for Health Economics
Monash University

Dr John McKie

Senior Research Fellow, Centre for Health Economics
Monash University

June 2010

Centre for Health Economics
ISSN 1833-1173
ISBN 1 921187 52 2

Correspondence:

Professor Jeff Richardson
Centre for Health Economics
Faculty of Business and Economics
Monash University Vic 3800
Australia

Ph: +61 3 9905 0754 Fax: +61 3 9905 8344
Email: Jeff.richardson@buseco.monash.edu.au

ABSTRACT

This paper describes a new Quality of Life (QoL) scaling instrument for measuring the value of health states, the Relative Social-Willingness to Pay (RS-WTP). Like the person trade-off (PTO), values are placed upon programs which move people from one health state to another. Also like the PTO it adopts a social perspective and asks respondents to evaluate programs on behalf of society. Unlike the PTO it uses the dollar as a measurement metric, but unlike the conventional willingness to pay (WTP) technique, the opportunity cost of funds spent on one program requires an offsetting reduction in funds for a second program. The amount spent on each program therefore indicates relative, not absolute, value. The dollar amount chosen for division, however, is not arbitrary, but can reflect, for example, the maximum that a NHS would be willing to pay for a life-saving treatment. The wording of the RS-WTP questionnaire is flexible and can incorporate some variations in social objectives. Results from the application of the instrument to a sample of the Australian population indicate that it rates well in terms of reliability and provides values similar to the PTO and time trade-off (TTO) techniques.

TABLE OF CONTENTS

List of Tables

List of Figures

The Relative social Willingness to Pay (RS-WTP) Instrument: Updated Results

1 Introduction

In his seminal article Torrance (1986) discusses 5 options for measuring the value of health improvement: namely, the standard gamble (SG), time trade-off (TTO), person trade-off (PTO), magnitude estimation (ME) and the rating scale (RS). Of these, the three instruments using trade-off techniques have been most widely used and discussed by health economists: although recently a case has been made for the use of the RS (Parkin and Devlin 2006). The techniques differ very significantly in their framing, visual props, perspective, and even their conceptual basis. The rating scale, which is used widely by psychologists in studies of subjective well-being, is claimed to "tap into people's mood affect" - their underlying feeling of well-being (Cummins, Lau et al. 2009). The trade-off instruments draw upon cognition: people must appreciate and manipulate probabilities or the number of years or people in different health states. The willingness to pay (WTP) measures reveal (or are stated to reveal) preferences and these may tap into affect, cognition, or a mix of these. The PTO adopts a "social perspective" (what a person would select on behalf of the society) while the SG, TTO and WTP usually incorporate a "personal perspective" (what a person would select for themself).

To date, none of the techniques has been shown to be "correct" in some technical sense (Smith, Brown et al. 2008) and it is unlikely that this will ever occur (Richardson and McKie 2009). It is not known whether instruments tap into the same concept of preferences , or whether potential sources of bias are quantitatively important. As a consequence, it is desirable that there is further enquiry into the instrument properties and that there is experimentation with new instruments. The Relative Social Willingness to Pay (RS-WTP) is one such example and might be seen as a response to Smith, Brown et al.'s recent call for expanded research into developing new tools for quantifying health-related quality of life that are more valid, more sensitive to changes in health status, and less biased (Smith, Brown et al. 2008 p 85).

The instrument is described in section 4, below. Before that, there is a brief discussion of the existing trade-off instruments in section 2 along with some of the conceptual ambiguities associated with them. In section 3 we consider the neglected issue of the criteria with which to evaluate instruments. The RS-WTP evolved as an attempt to meet these criteria. Section 5 describes the empirical methods. Section 6 reports the results of a survey designed to test the validity and reliability of the instrument using the TTO and PTO as the measures of convergent validity.

2 The Trade-off Instruments

Each of the three trade-off methods has its claimed advantages and disadvantages. In his 1986 article and later Torrance endorsed the orthodox view that the standard gamble represents the welfare theoretic gold standard as it is based upon the axioms of von Neumann and Morgenstern; that is, the axioms of Expected Utility Theory (EUT). Evidence is growing, however, that these axioms are empirically flawed (Allais and Hagen 1979; Kahneman and Tversky 1979; Schoemaker 1982; Luce 2000). Responding to this evidence, some authors have argued that the axioms have normative status as they embody consistent (rational) behaviour which justifies its status as the preferred measurement instrument in normative economics (Marschak 1950; Savage 1954; Torrance and Feeney 1989). Supporters of the TTO have argued that it avoids the chief theoretical problem of the standard gamble and that it is the instrument that is most transparent in trading off life and quality of life for the individual, which is the defining (though little discussed) property of the QALY (Richardson 1994). Defenders of the PTO have argued that it embodies the appropriate social perspective for a collectively financed national health scheme as subjects are asked to imagine they are making a judgement on behalf of others. Additionally, the PTO measures the value of a movement from one health state to another as distinct from the value of the health state itself, and this is an advantage if that value is not simply equal to the difference between the values of the two health states (Nord 1995).

A problem with the three trade-off instruments is that each includes an instrument variable which is used to achieve the trade-off, but which may independently influence choice and therefore alter the value of the measurement by something other than the quality of life. Bleichrodt (2002) examines the affect of this in the case of the TTO and SG, and bias in the TTO is examined by Schwarzinger et al. (2004), Abellan-Perpinan et al. (2007), and Doctor et al. (2009).

The standard gamble employs probability as the core of the instrument. It equates the value of a certain health state (i.e. probability $p = 1$) with the value of a gamble where there is a probability, $p<1$, of full health ($U = 1.00$) and 1-p of death ($U = 0.00$), where the probability, p, is selected by interview subjects. If people behave according to the axioms of expected utility theory then the probability, p, will be an index of the utility of the health state. (If $1.00_* (U) = p*(1.00) + (1-p)*(0.00)$ then $p=U$.) The theoretical limitation of EUT was clearly acknowledged by Von Neumann and Morgenstern, and clearly articulated by Morgenstern (1979) but subsequently ignored. It has recently and eloquently been formalised and generalised by Pope et al. (2007).

The theoretical problem arises from the (dis)utility of risk per se as distinct from the objective probability and utility of the outcomes. Using Pope's terminology, this (dis)utility occurs in the 'pre-outcome' period and will be determined by a person's temperament and background, which may be unrelated to the objective circumstances that follow the realisation of the outcome. People are likely to experience quite different emotions at the prospect of immediate death compared with the prospect of some mild inconvenience caused by ill health, and as explicitly stated by Morgenstern (1974), this element disqualifies the NM axioms from being a general theory of risk. There are some health contexts which co-incidentally approximate the utility of a health state without treatment - where the outcome of a treatment may involve a risk of death - but this is an exception, and 'utilities' inferred from this gamble could not be compared with utilities from gambles where this co-incidence does not occur. Formally, if 'g' is the disutility of risk per se then the SG calculation becomes $1.00_* (U) = g + p_* (1.00) + (1-p)_* (0.00)$ from which measured $p = U-g$. If g is negative (people dislike the risk of death), the value of U is exaggerated by observed p. More pragmatically, the SG relies upon a person's capacity to appreciate and manipulate 'risk', and this ability is notoriously poor (Schoemaker 1982).

The TTO technique uses as its instrumental variable the length of time in full health which is adjusted until it is deemed by respondents to have the same value as a stated number of years in a poorer health state. From this, an index of utility is calculated as the ratio of years in full health to years in poor health. If the calculation uses a 10 year time frame for poor health and n is the number of equivalent healthy years, $10_* (U) = n_* (1.00)$ from which U = n/10. However, as the time period in good health is adjusted it does not correspond to the time period of the poor health state and, consequently, there is a distortion arising from the rate of time preference. Coincidently, the reduced length of time in good health may, on occasions, correspond with a real world option where a treatment involves a reduced life expectancy. However, this would not be a general condition. Formally, if the time distortion reduces the present value of the longer period by a positive amount, t, then $10_* (U)-t = n_*1.00$ from which measured (n/10) = U – (t/10) and the value of U is understated by the calculated value, n/10.

Similarly, with the PTO there is a trade-off that alters the character of the options. In the version used in this study, a standard number of people may be moved from imminent death to a better state with social value or social utility, U*. Respondents select the number to be shifted, also from imminent death but to full health, which will make the two options of equivalent social value. If the standard number is 100 and the equivalent (smaller) number n, then $100_*(U-0.00) = n_* (1.00 – 0.00)$ from which U* = n/100. However, as people are sensitive to issues of equity, and the number obtaining a health service in the two options differ, the social value of the fairness of the options will influence the equilibrating numbers. If ‘e’ is the extra value from the fairness of more people receiving a (life saving) treatment, $100_* (U^*-0.00) + e = n_* (1.00-0.00)$ from which measured value, n/100 = U* + e/100. ‘Social utility’/‘social preference’ for health alone is overestimated with this version of the PTO. Note that life saving per se is ignored in this calculation. If it has special status (value) then this further increases the value of the second option where more people’s lives are saved. This effect is illustrated by increasing the numerical value of e; that is, any special value of life per se further exaggerates apparent utility.

With a second version of the PTO, the second program improves people’s health from U* to full health and utility is calculated as U*=1-n/100. Equity again favours the second option where benefits are received by the larger number, 100. But this option involves no life saving. Hence, any special life effect, f, increases the value of option 1. Consequently n(1-0.00)+f = 100 (1-U)+e from which measured value, (1-n/100) = U-(e-f)/100. The bias is ambiguous.

Just as the risk of immediate death has been claimed to be an advantage in the SG, so the distributive element of the PTO has been claimed to be an advantage of that technique, as it captures an element of fairness. However, as with the SG and TTO it is fortuitous whether or not the number emerging as a result of the disutility of the poor health state calculation corresponds with the number of people affected by the program being evaluated. In general, as in the previous two cases, the final score obtained from the instrument is not a pure quality of life score but a composite of the quality of life score and the influence of the additional factors associated with the technique.

The problems discussed above have a common origin. This is that each instrument has an instrumental variable – risk, time or equity – which has an independent effect upon people’s valuations. With respect to this problem the TTO is, arguably, the least compromised instrument, as it is possible to estimate the rate of time preference for each individual subject by altering the time horizon of the scenarios. Even the use of this correction factor requires the assumption that the rate is constant over other time periods. However, the TTO suffers from a second, intrinsic, problem (Bleichrodt 2002) . Typically, subjects are asked to sacrifice an amount of life in exchange for an improved quality of life and the amount sacrificed is “flip flopped” backwards and forwards until the options of reduced life expectancy in full health and greater life expectancy in poor health are equally attractive. However, people do not respond equally to a loss (length of life) and gain (quality of life). Independent testing has found that, in many contexts, people place a greater weight upon a loss, with a relative weight of about 2 being attached to the loss relative to the gain. This suggests that the TTO may have a significant upward bias in scores – people may resist loss of life and nominate higher equivalent numbers for healthy years. Consistent with this, Bleichrodt et al. (2003) found 'short gauge' duration TTO scores exceeded SG values for the same health state. More generally, if the bias caused by the asymmetric treatment of losses and gains is of the magnitude observed elsewhere it might more than offset the downward bias caused by time preference.

A further important class of problems should be mentioned, which affects all of the techniques to varying degrees. This is the problem of the cognitive complexity or unreality of the instrumental variable. People do not commonly, if ever, manipulate the variables used by the trade-off instruments. Few have had experience with decisions involving the possibility of instant death, explicit trade-offs between quality of life and reduced life expectancy, or the manipulation of numbers of people. The unfamiliarity of these tasks introduces further unknown bias including the framing affects and start point bias. (For example, should the PTO use 10 or 100 people as its reference case? See Schwarzinger et al. (2004).)

In contrast with the trade-off techniques, the use of Willingness to Pay employs a common and easily understood form of decision-making, namely the sacrifice of money. Although this might also be considered a trade-off instrument, money is a direct measure of (one notion) of value and the use of money per se as a metric does not introduce extraneous factors. However, personal willingness to pay encounters problems associated with its dependence upon an individual's income and wealth. This is a serious problem in the context of a national health scheme which is usually created to ensure the separation of allocative decisions from considerations of socio-economic status and, however massaged, these personal willingness to pay estimates are derived from wealth-based preferences and cannot be easily corrected. (For a discussion see Smith and Richardson (2005) and Richardson and Smith (2004)).

The discussion here does not purport to exhaust the possible problems with the main techniques. The purpose is to argue that, as each has problems of unknown magnitude, there is a case for experimentation. Validation commonly occurs by increasing the number of comparisons between an instrument and plausible alternatives, and the same logic applies here. One approach is to modify existing instruments (for example, see Abellan-Perpinan et al.(2006); Attema and Brouwer (2009)). The present paper outlines an alternative, more ambitious approach – namely, the development of a new instrument.

3 Criteria for Evaluating a Scaling Instrument

There has been relatively little discussion in the health economics literature of how we should select a technique for measuring QoL – what criteria should be used for their evaluation? Historically the theoretical literature has been dominated by the assumption of EUT which implies the use of the SG, and by the less restrictive assumption that we should measure "utility" or individual preferences, which implies, inter alia, that an individual perspective should be adopted in measurement. This latter assumption has been challenged theoretically (e.g. by Nord (1999)) and evidence indicates that different values will be obtained by adopting a social perspective (Mann, Brazier et al. 2009). More recently, the assumption that either social or individual preferences should be the object of measurement has been challenged. Following the experimental work of Kahneman et al.(1997), Dolan has argued that subjective wellbeing – as experienced - should replace preferences (Dolan 2008a; Dolan 2008b; Dolan and Kahneman 2008). Hausman (2008), in reply, has argued for the measurement of functioning and capabilities, a suggestion derived in recent times from Sen (1993; 2009) and operationalised in the recently developed ICECAP instrument (Grewal, Lewis et al. 2006; Coast, Flynn et al. 2008).

The present study reports an attempt to devise an instrument which, as far as possible, meets the general criteria discussed by Richardson (1994; 2002) and to further develop a version of the instrument which was found to be valid and reliable in postal version. These criteria emphasise technical aspects of scaling and are broadly compatible with the underlying values in all four options above (individual utility, social utility, subjective well-being, and capabilities, which differ primarily in their normative content). The criteria relevant to the present task are:

1. Validity:
   - The concept embodied in the instrument should be the one which satisfies (normative) social objectives.
   - The numerical value obtained by the instrument should only depend on the value of the desired object of measurement (avoidance of ‘instrument bias’).

2. Scale Properties:
   - The unit should have a clear interval property - i.e. an equal numerical increase (e.g. 0.2) at any point along the scale should have the same meaning.
   - A consideration of the opportunity cost of decisions should be unavoidable; that is, the negative consequences of one choice compared with another should be apparent.

3. Cognitive Simplicity (Reliable Measurement, correct usage):
   - The question asked or task undertaken should be simple, to avoid, as far as possible, ambiguities of language or concepts and to avoid framing effects.

It should be emphasised that these are not arbitrary criteria, but derived from the purpose of the instrument. Criterion 1 requires an explicit decision with respect to what is to be measured. There are numerous options; hedonic utility (subjective well-being), preferences (decision or ex-ante) utility, social value (as assessed by citizens or others); a metric which takes account of equity, severity, health potential, etc.

The need for scale properties arises from the fact that QALY gains must be similar when improvement commences from any starting point along the QoL scale. The emphasis upon cognitive simplicity is to facilitate correct judgement. The essence of QALYs is that they provide an exchange rate between the quality or value and length of life. But the exchange rate can only be appreciated if the unit is easily understood: survey respondents cannot make this pivotal judgement if they are uncertain about the metric and its meaning. Similarly, policy makers will hopefully wish to understand and fully appreciate the implications of the unit of benefit when they select between programs as they may wish to adjust measured benefits because of contextual or other program specific characteristics excluded from routine measurement.

These considerations do not cover one pivotal element, and the one which separates the options noted above. This is the method for determining social values. There is no objective way of settling this normative issue and the present proposal was designed to be adaptable in this respect as illustrated below.

4 The (Relative) Social-Willingness to Pay (RS-WTP) Instrument

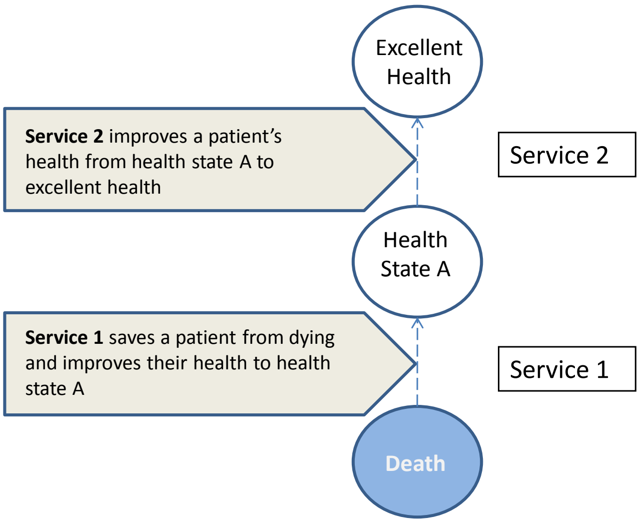
The RS-WTP was designed in an attempt to meet these criteria. Its chief characteristics are embodied in its title. It measures the *willingness to pay* from a fixed budget for one service (which takes a person from one health state to another) *relative* to another service, and does so from a *social* perspective (although it may be personalised). The essential features of the instrument are captured in Figure 1, which reproduces the original question used in surveys.

Figure 1 The Relative Social Willingness to Pay (RS-WTP) Instrument

Question

"How would you divide the $40,000 between Service 1 and Service 2 so that the amounts given for Service 1 and Service 2 indicate the value of the benefits to health from Service 1 compared with Service 2?"

Note: The health benefit from each service lasts for one year. Each following year will require further funding.


Excellent Health
Service 2 improves a patient's health from health state A to excellent health
Service 2
Health State A
Service 1 saves a patient from dying and improves their health to health state A
Service 1
Death

Respondents are asked to imagine that they are on a government committee which must decide how much Medicare (Australia's universal health insurance scheme) should pay for various medical services. They are told that the government will pay $40,000 for a service that will save a person's life and restore them to full health for one year (approximately the threshold observed by the Australian Pharmaceutical Benefits Advisory Committee (PBAC) (Henry, Hill et al. 2005)). Alternatively, it is the amount that will be spent on two services that together will achieve this - i.e. Service 1 saves life and improves health to state A for one year and Service 2 improves health from state A to full health. The services may be received by different people. Respondents are given the instruction '*to divide the budget between the two services' so that 'the amounts indicate the value of the benefits to health from service 1 and service 2*'.

The relative value of the two services is given by the relative amounts a respondent allocates to each service. Alternatively, on a 0-1 scale, indices of value may be obtained as the amount allocated to each service divided by $40,000.

The instrument is flexible. The key question may be varied to alter the constraints or perspective. In the first part of the empirical analysis reported below the question in Figure 1 was used. In the second and main application of the instrument, the key question was altered to read "*taking everything you believe to be important into account, divide the money available between service 1 and service 2 so that the amounts indicate your view of how Medicare should value the services"*. In principle, this means that a person could express their own judgement about the social purpose of Medicare and not be constrained to the "benefits of health". Particular health benefits which the person valued for themselves could, in principle, be excluded or devalued in relative importance. The second version also added the comment that, as a citizen, the person might one day need the services themself. A further option would be to include the instruction that *"at some future date you may/will require one of these services"*. This personalised version has not been tested to date.

The properties of the instrument are summarised in Table 1 along with other commonly used instruments.

Table 1 Characteristics of scaling instruments

	Instrumental variable	**Type**	**Perspective**	**Subject of measure**	**Opportunity cost of choice**
VAS	distance / score on a calibrated line	No trade-off	Individual	Health State	None
TTO	Years of life	Trade-off	Individual	Health State	Years of life
SG	Probability of death	Trade-off	Individual	Health State	Risk of death
PTO	Persons treated	Trade-off	Social	Change in health state	Number of lives
RS-WTP	$ money	Trade-off	Social	Change in health state	Other health program
WTP	$ money	Sacrifice	Individual	Change in health state	consumption

5 Empirical Methods

The empirical study sought to test the following characteristics:

1. The test-retest reliability of postal versions of the instrument;
2. The validity of the postal version using an interview version as the criterion;
3. The validity of the interview version using TTO and PTO interview data as criteria;
4. The sensitivity of results to the framing effect of the budget size; and
5. The sensitivity of results to the form of the key question.

The study used respondents who were recruited from selected postcodes to provide a cross-section of Australian SEIFA (Socio-Economic Indicator for Areas) groups (or SES groups), but from locations close to Monash University to facilitate an interview. No other attempt was made to obtain a representative cross-section of the population as the purpose of the work was to pilot the instrument and to perform standard tests of reliability and validity (test, re-test and comparison with other instruments).

After the initial pilot interviews the questionnaire was revised and significantly simplified after consultation with a professional ‘simple English’ writing company, Yellow Pencil[1]. The main results below relate to the subsequent interviews. Health states were constructed from the AQoL-8D descriptive system (Richardson, Iezzi et al. 2009). As indicated above, two versions of the instrument were used to test the sensitivity of results to the form of the question. In one version the size of the budget was raised from $40,000 to $100,000 to test for the existence of a framing effect.

Those agreeing to participate were sent an initial ‘Postal 1’ (P1) survey. Two weeks after the return of this they were sent ‘Postal 2’ (P2), the same instrument. After the return of ‘Postal 2’, subjects were asked to attend an interview during which the instrument was administered by a trained interviewer along with TTO and PTO assessments of the same health states. Subjects were offered a small financial remuneration for out-of-pocket costs such as transport and parking. Our social-willingness to pay was $30.

The postal questionnaires, the two forms of the instrument, the interview protocol and visual props are reported (on line) in (Richardson, Iezzi et al. 2007). This also reports results of a statistical comparison between the two forms of PTO described earlier . In both versions of the questionnaire respondents were reminded that each service lasts for one year but may be given again with new funding the following year. An important feature of all questionnaires, interviews and props was that elements of good health were drawn to the subject’s attention to overcome ‘focusing effects’, i.e. focusing upon negative and ignoring positive elements. Interviewers reminded subjects verbally to keep positive elements of the quality of life in mind as well as the negative elements.

The conventional test–re-test of reliability would be based upon the correlation between RS-WTP results for the same health states, and tests of validity upon the correlation with the TTO and PTO for the interview version of the RS-WTP. Regression analyses were also employed. An insignificant constant and slope coefficient of 1.00 indicate that the independent variable provides an unbiased estimate of the dependent variable. While the RS-WTP differs conceptually in several respects from the TTO and PTO these regression results would not necessarily be expected, but to the extent that the quality of life dominates all other considerations the regression coefficients should be close to these.

6 Results

The data collected are summarised in Table 2.

Table 2 Data collected

	Version 1 **Postal 1, then interview**			**Version 2** **Postal 1, Postal 2, then interview**	
Budget	$100,000	$40,000	Total	$40,000	Total
Persons	25	39	64	46	110
Health states	5	5	5	18	18
Observations	124	193	317	414	731

[1]. Yellow Pencil: http://www.yellowpencil.com.au/

The first version of the interview was completed by 64 people which resulted in a total of 317 sets of observations (P1, INT, TTO, PTO). Of these, 124 used a figure of $100,000 in the RS-WTP questions and 193 used $40,000. The second version which used 18 health states was completed by 46 individuals and provided 414 sets of observations (P1, P2, INT, TTO, PTO). The budget was $40,000. In total, 731 observations were obtained from 110 respondents.

Results reported throughout the present paper are based upon the full dataset. The application of an edit algorithm based upon comparison of the two postal results reduced individual level discrepancies and improved statistical results (Richardson et al. 2007). However results of analyses of mean data were very similar. For reasons of methodological conservatism results reported in the present paper are based upon unedited data.

For the purposes of this study – which was to test reliability and validity – individual characteristics were not important. They are nevertheless summarised in Table 3, which indicates that respondents were drawn from across all SES ('SEIFA') age and gender groups.

Table 3 Respondent Characteristics

	Male	Female	Total		Male	Female	Total
Age				**SEIFA (SES)**			
18-24	8	6	14	1	5	2	7
25-34	9	15	24	2	6	12	18
35-44	6	10	16	3	5	11	16
45-54	8	13	21	4	16	15	31
55-64	10	18	28	5	14	24	398
64+	5	2	7				
Education				**INCOME**			
Trade/TAFE	8	11	19	<$30,000	8	9	17
University	14	25	39	$30-$60 K	18	14	32
Post grad	6	13	19	$60,000+	20	41	61

Table 4 reports the results of the comparison of the two versions of the instrument - version 1 asking for an evaluation of the 'benefits to health' and version 2 asking 'how Medicare should value' services. The table indicates that average values for individual health states are not significantly different and particularly close for mean interview values where the maximum discrepancy between the two versions for any health state was 10 per cent (HS 3) and the overall averages of the mean scores were identical (0.58). These results indicate an insensitivity to the (quite different) forms of the question used here.

Table 4 Comparison of results from 2 versions of RS-WTP: Mean values by health state

	V1: Version 1			V2: Version 2			V2/V1	
Health States	n	Postal (P1)	Inter-view (INT)	n	First Postal (P1)	Inter-view (INT)	Postal	Inter-view
HS 3	63	0.61	0.68	26	0.57	0.75	0.94	1.10
HS 4	64	0.45	0.50	20	0.40	0.46	0.89	0.92
HS 6	62	0.39	0.44	26	0.43	0.42	1.10	0.95
HS 7	64	0.49	0.56	20	0.41	0.52	0.84	0.93
HS 11	64	0.67	0.75	20	0.63	0.76	0.94	1.01
Average	64	0.52	0.58	22	0.49	0.58	0.94	1.00
Total n	317			112			429	

Table 5 reports the mean RS-WTP values of the 18 health states from the three RS-WTP instruments which were used in the main analysis. The two postal surveys resulted in very similar mean values. Postal 1 had higher and lower scores than Postal 2 in 8 and 9 cases respectively with the same overall average value. The average value of individual health states differed by more than 10 per cent in only 3 of the 18 cases. By comparison, results for the RS-WTP interview (INT), the 'proximate gold standard' for this postal instrument, were 0.07 or 14.3 percent higher and the average value from interviews exceeded the postal values by more than 10 per cent in 12 of the 18 cases. The average difference is significant at the 5 per cent level.

Table 5 Mean values of RS-WTP by health states and by type of instrument

Health states	N	1st postal (P1)	2nd postal (P2)	P2/P1	Interview (INT)	Int/P2
HS 1	20	0.49	0.53	1.08	0.57	1.08
HS 2	20	0.66	0.68	1.03	0.78	1.15*
HS 3	26	0.57	0.63	1.11**	0.75	1.19**
HS 4	20	0.40	0.44	1.10**	0.46	1.05**
HS 5	20	0.51	0.51	1.00	0.67	1.31**
HS 6	26	0.43	0.40	0.93**	0.42	1.05**
HS 7	20	0.41	0.46	1.12**	0.52	1.13**
HS 8	26	0.51	0.48	0.94	0.57	1.19
HS 9	26	0.48	0.42	0.88*	0.51	1.21*
HS 10	20	0.38	0.35	0.92	0.39	1.11
HS 11	20	0.63	0.68	1.08**	0.76	1.12**
HS 12	26	0.53	0.56	1.06	0.64	1.14
HS 13	26	0.49	0.46	0.94	0.50	1.09
HS 14	26	0.44	0.40	0.91	0.44	1.10
HS 15	20	0.50	0.55	1.10	0.60	1.09
HS 16	20	0.32	0.33	1.03	0.30	0.91
HS 17	26	0.54	0.50	0.93	0.60	1.20
HS 18	26	0.49	0.48	0.98	0.57	1.19
Averages		0.49	0.49	1.00	0.56	1.14
Total	414					

* difference significant at 5%

** difference significant at 1%

TTO and PTO instruments resulted in higher mean values than the RS-WTP interview for every health state with overall mean values for both instruments for the 18 states 0.09 (16.1 per cent) greater than the RS-WTP. Data are reported by health state in Appendix 1.

The standard test of association between instruments for validation and reliability is the simple correlation coefficient. These are reported in Table 6 for unedited mean data, which are most relevant for QALY calculations. Coefficients are very high and highly significant, particularly between INT and all other variables.

Table 6 Correlation between instruments, mean values of health states

	P1	P2	INT	TTO	PTO
P1	1.00	0.95	0.96	0.81	0.86
P2		1.00	0.96	0.91	0.94
INT			1.00	0.92	0.93
TTO				1.00	0.96
PTO					1.00

Results of linear regressions are reported in Table 7 and plotted in Figures 2 to 6.

Regression 1 (Figure 2) indicates a close to, but less than perfect, fit between postal 2 and postal 1 values (the ideal being indicated by a regression 'P2' = 0.00 + 1.00 P1). As P1 rises from 0.00 to 1.00 predicted P2 rises from 7 points below to 7 points above P2. The discrepancy between P1 and the interview values (Figure 3) is greater with predictions rising from 13 points below to 20 points above (regression 2) . The corresponding discrepancy for regression 3 (Figure 4) in which RS-WTP (interview version) is predicted from P2 is from 1 point below to 16 points above. However to keep these differences in perspective they are very significantly less than the discrepancies found between any two instruments in the Hawthorne et al. (2001) five instrument study – one of the more comprehensive comparative studies.

In regressions 4 and 5 results suggest that while RS-WTP (INT) predicts lower scores than the TTO and PTO, incremental changes in the instruments are very similar with 'b' coefficients of 1.01 and 1.09 (Figures 5 and 6).

Finally, Table 8 reports a single regression designed to test the importance of reframing the RS-WTP instrument with a budget of $100,000. In the regression the dependent variable is the value obtained by each of the instruments with the exception of the interview RS-WTP; that is, the dependent variable includes values for P1, P2, TTO, PTO and the initial version of the questionnaire. The chief independent variable is the corresponding value for the health state found by the interview version of the RS-WTP. Dummy variables are included for the use of the postal questionnaires (Dum P1, Dum P2) and for the TTO and PTO (Dum TTO, Dum PTO). Finally a variable Dum 100 is included which indicates that the budget in the question was $100,000, not $40,000.

Results are broadly consistent with the tabulated data. The postal questionnaires obtain lower scores, the TTO and PTO higher scores. Importantly, the dummy variable for the inclusion of the $100,000 budget was not significant in this or variations of this equation.

Table 7 Regression results (Mean values)

Independent variable	Dependent variable				
	RS-WTP				
	1	2	3	4	5
	P2	Int	Int	TTO	PTO
P1	1.14	1.39			
(t)	(13.1)	(13.81)			
P2			1.16		
(t)			(13.25)		
Int				1.01	1.09
(t)				(9.12)	(10.27)
Constant	-0.07	-0.13	-0.01	0.09	0.06
(t)					
R^2	0.92	0.92	0.92	0.84	0.87
Interviews	414	414	414	731	731
n / HS	23	23	23	41	41

Table 8 Regression results: Dependent variable 'utility'

Independent variable	1		2		3	
	b	(t)	b	(t)	b	(t)
Interview (Int)	0.73	10.77	0.83	14.56	0.83	14.54
Dum P1	-0.18	-7.12	-0.18	-6.86	-0.17	-7.32
Dum P2	-0.18	-7.02	-0.18	-6.77	-0.17	-7.23
Dum TTO	-0.17	-1.96	-0.02	-0.79	-0.01	-0.23
Dum PTO	-0.21	-2.35	-0.18	-0.69	0.00	-0.11
Dum 100	-0.03	-1.18	-0.03	-1.15		
Int * (Dum PTO)	0.34	2.23				
Int *(Dum TTO)	0.27	1.80				
Constant	0.26	6.14	0.21	5.45	0.0	19
n	102		102		102	
R^2	0.77		0.76		0.76	
RMS	0.07		0.08		0.07	
F	44.8		55.7		66.4	

Table 9 Regression result: The framing effect of the budget. Independent variable 'value'

Independent variable	b	(t)	Statistics	
RS-WTP	0.83	(14.56)	R^2	0.76
Dum P1	-0.18	(-6.86)	RMS	0.08
Dum P2	-0.18	-6.77	F	55.7
Dum TTO	-0.02	-0.79	n	=102
Dum PTO	-0.18	-0.69		
Dum 100 constant	-0.03	-1.18		

Key:

RS-WTP = RS-WTP interview score; Dum P1; Dum P2; Dum TTO; Dum PTO = dummy variables; = 1 when data gathered by Postal 1, Postal 2, TTO, PTO respectively; = 0 otherwise; Dum 100 = dummy variable = 1 when budget = $100,000 = 0 otherwise

Figure 2 Second postal *vs* first postal

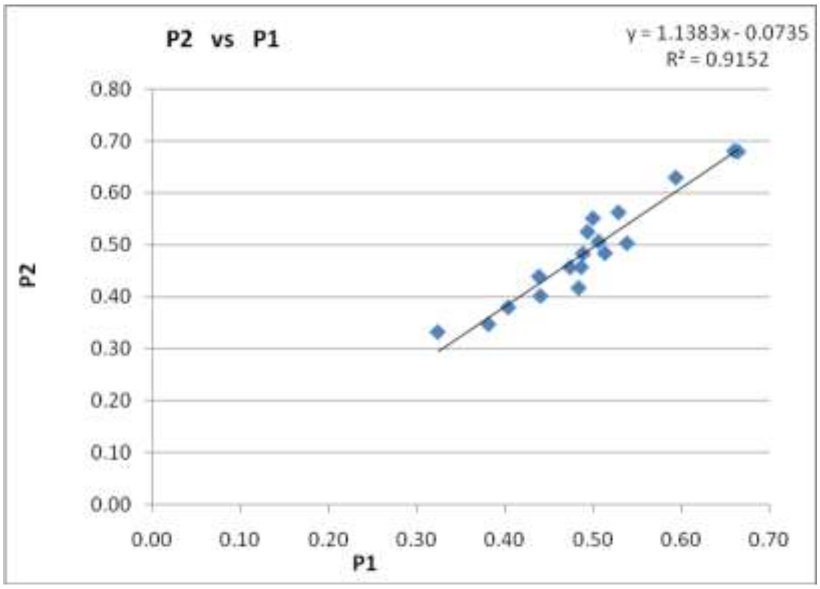


Figure 3 Interview *vs* first postal

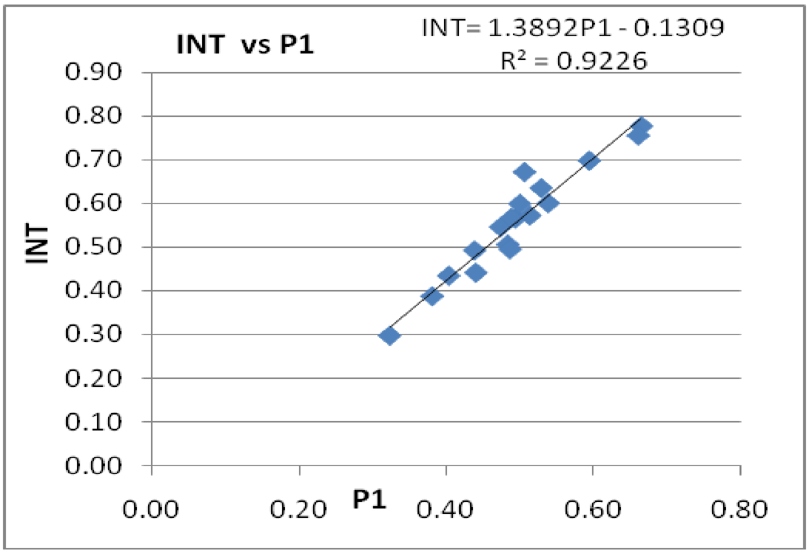


Figure 4 Interview *vs* second postal

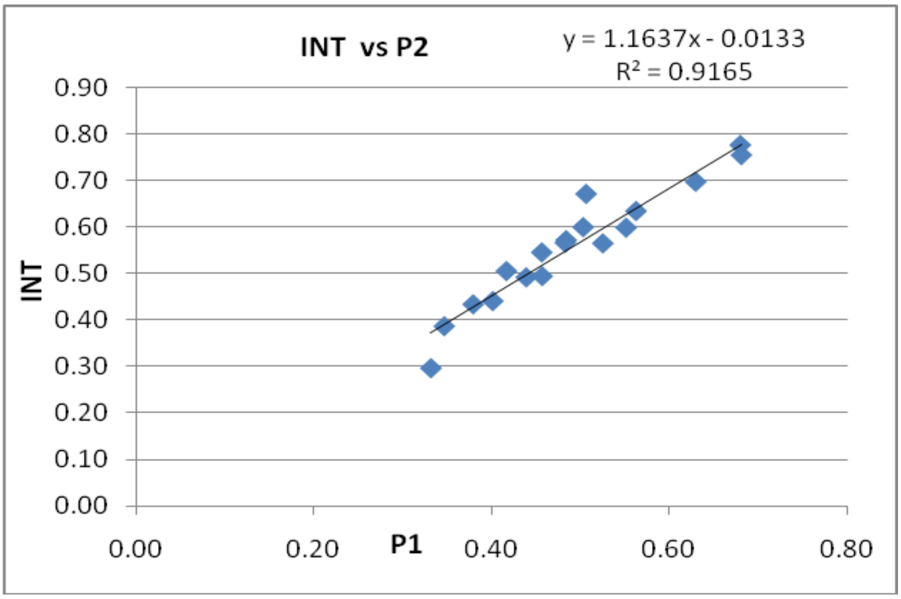


Figure 5 Time trade off *vs* interview

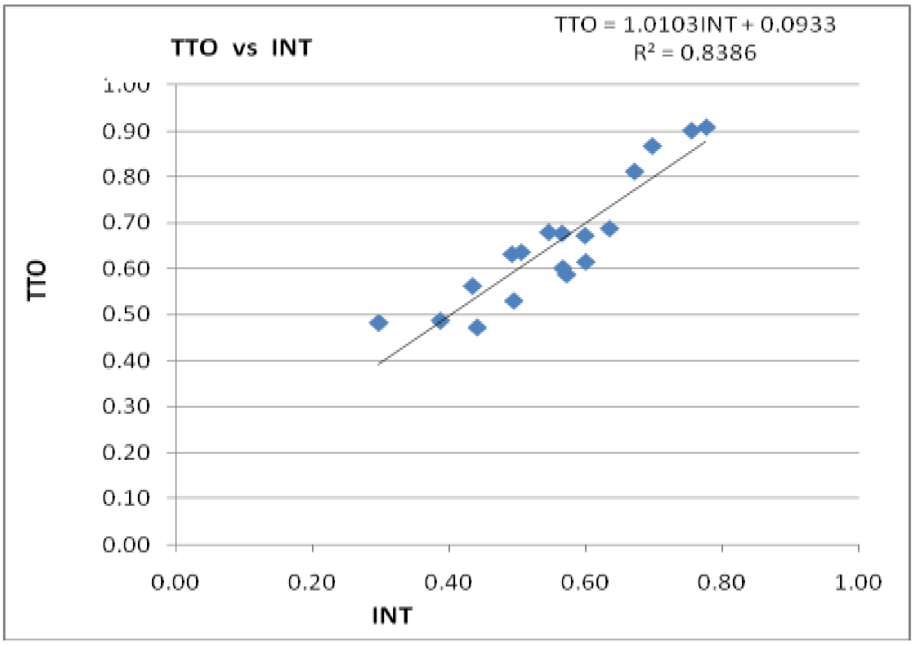
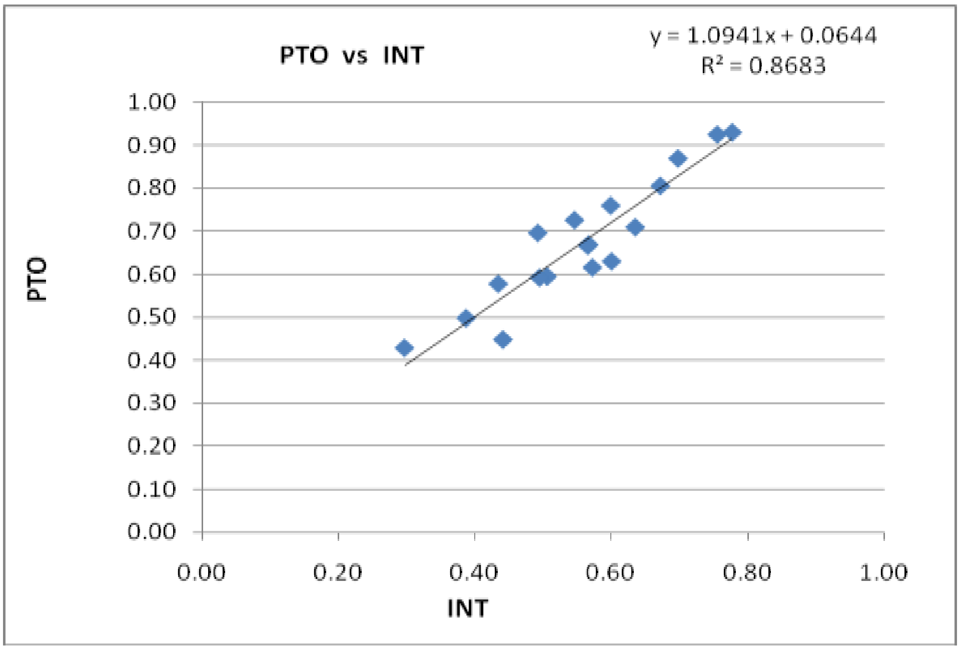


Figure 6 Person trade off *vs* interview



7 Discussion

This study commenced by arguing that there is no empirically ‘correct’ – i.e. completely validated – instrument for measuring the weights that should be used to measure QALYs. Correlations with disease-specific or generic psychometric instruments fall short of ensuring that an instrument provides a valid number for calculating QALYs. Furthermore, as argued here, there are theoretical grounds for concern with respect to each of the commonly used instruments.

In addition to the challenge of evaluating health states per se in a way that satisfies social objectives, evidence in the literature indicates that QALYs themselves may need to be adjusted according to the personal or social characteristics of patients (e.g. age, number of dependents, own contribution to illness) or the context in which illness occurs (e.g. the “rule of rescue”). The initial objective of the present project was to devise an instrument which permitted these elements to be measured easily and economically without loss of reliability and validity, in much the same way as the TTO was developed by Torrance as an easy way of producing numbers similar to the standard gamble (Torrance, Thomas et al. 1972).

We have argued here that, like the TTO, there are reasons for accepting the RS-WTP as an independent candidate for producing QALY weights. Depending upon whether or not an individual or social perspective is required it is arguable that the question asked in the RS-WTP questionnaire is more suitable that the alternatives for QALYs intended to be used for the evaluation of services in an NHS. By adopting a social perspective it asks respondents to indicate preferences for spending in an NHS, with the budget amount realistically tied to NHS thresholds. Further, if the severity of a health state is of independent importance to the public, as evidence suggests (Nord 1993; 2001), then like the PTO, the RS-WTP will take account of this.

Stronger claims for the RS-WTP cannot be made at present.

The empirical study reported here sought to test the reliability and validity of the new RS-WTP and the self-administered (postal) version of it. This led to the five study questions outlined earlier. With respect to the first, the test-re-test reliability of the postal version of the RS-WTP was very high. The average values of the 18 health states was identical in the questionnaires and the correlation of 0.95 was extremely high. Patrick (1993) cites a coefficient of 0.7 as being satisfactory for mean data and Froberg and Kane (1989) report correlation coefficients for mean scores between utility instruments within the range 0.77 and 0.8. The correlations here were obtained with unedited data and confirm individual reports that respondents to mail questionnaires did not encounter significant difficulty understanding them. Nevertheless, as noted earlier, data reliability at the individual level were significantly increased using a simple edit program. This left mean values unchanged but increased the correlation at the individual level between P2, TTO and PTO by 41.0 and 70.2 per cent respectively (see Richardson et al. 2007).

The correlation of 0.96 between both postal results and the RS-WTP interview indicates that both of the postal instruments have been satisfactorily 'validated' as this term is used in the psychometrics literature, i.e. the evidence suggests that they measure the same or a very similar construct or latent variable (study questions 2, 3). However, this form of validation is a necessary but not sufficient condition for validation as QALY weights. The regression results indicate that mean postal values under-report the 'true' RS-WTP (INT) values and this is reflected in the tabulated values where the average postal results are 7 points below the average of the mean interview values.

This suggests that the postal version of the instrument may also be used, as initially intended, to test the effect of patient attributes, and context, but that scores obtained in this way should be adjusted before being interpreted as measuring the real RS-WTP. The regression results represent a reliable transformation function for doing this.

The correlation of 0.93 and 0.95 between mean values of RS-WTP and TTO and PTO scores represents very strong psychometric 'validation' of the new instrument (study question 4). However, 'validation' (in the psychometric sense) operates 'both ways'. To the extent that the RS-WTP is intrinsically plausible and it asks the correct question in a sufficiently simple way that its answers may be taken seriously, it also 'cross validates' the numbers that have been obtained from the TTO and PTO.

The RS-WTP produces lower values than either the PTO or TTO. This is not due to known bias in the former. The interview was the third occasion on which respondents had encountered the health states to be evaluated and they were familiar with them and the form of the question. There was little evidence, at this stage, of difficulty with the task. The result suggests the possibility of upward bias in both the PTO and TTO as discussed earlier.

An ongoing problem with measurement for QALY weights has been the effect of framing upon scores. With respect to the fourth study question, the regression results indicated that changing the size of the budget had no effect. The more significant change in framing occurred with the alteration in the form of question from an assessment of 'the value of health benefits' to 'your view of how Medicare should value the services'. Despite the potential for considerable discrepancies, the detailed analysis of results found no systematic variation with the form of question (study question 5). It remains possible, of course, that with more controversial services or health states this result would not be replicated. By contrast, with the insignificance of these framing effects the severe editing of the questionnaire (not data) described earlier did have a quantitatively important effect. Results from the initial pre-edit version of the postal questionnaire have not been reported but were revealing. The revised questionnaire resulted in correlations between instruments which were significantly higher (reliability) although mean values did not change significantly. The difference between the questionnaires was that the first emphasised the ‘completeness’ of the scenario, a full explanation of the task, context and examples which would ‘stand up in a court of law’ when our colleagues challenged people’s interpretation of the question. Despite its minimalism as we saw it, the subsequent (main) version eliminated most of this background and emphasised simplicity and brevity. Our empirical results support the superiority of this approach implying that explanatory notes, context, etc. added complexity not clarity.

8 Conclusion

In his summary comments on an ISPOR (International Society for Pharmacoeconomics and Outcomes Research) consensus development workshop Reed Johnson argues that ‘it is remarkable that three decades of QALY research have yielded so little substantive methods development.... QALY research appears to be stuck in a methodological traffic jam’ (Johnson 2009 p S38). As noted earlier there has been a recent upsurge of interest in the properties of the basic metrics, but over a longer period the quotation is apposite. The present study represents an experimental new approach to health state evaluation, which attempts to meet this.

The case for an additional instrument for transforming health state descriptions into value indices rests upon the limited testing of the instruments presently used. While these other instruments have been very widely employed, few studies have attempted to empirically demonstrate that they satisfactorily meet social objectives. As indicated by the ongoing discussion of social versus individual perspective (to mention only one of several examples), the extent to which they meet social objectives remains problematical. In these circumstances experimentation appears to be desirable.

The RS-WTP was initially conceived as a method for testing the importance of other elements that might be relevant in selecting health services for an NHS: age, severity, etc, and the purpose of the empirical study reported here was to determine whether the instrument and the postal version produced sufficiently reliable results for this task. The results presented here suggest that this goal has been achieved. In its final form the instrument has many of the attributes required of an ideal QoL metric while overcoming some of the existing shortcomings. There is, therefore, a strong case for treating the instrument weights as appropriate for QALY analysis, and worthy of further investigation.

A further important conclusion is that the close relationship between the mean values obtained from the different instruments, and particularly the RS-WTP and TTO, increases confidence that the numbers that have been used in QALY analyses correspond with peoples’ values even when these are accessed in quite different ways.

Appendix 1

Table A1.1 Mean values by health state, RS-WTP, TTO, PTO

	Observations	RS-WTP Interview (int)	TTO	PTO
HS 1	20	0.57	0.68	0.67
HS 2	20	0.78	0.91	0.93
HS 3	26	0.75	0.86	0.81
HS 4	20	0.46	0.62	0.57
HS 5	20	0.67	0.81	0.81
HS 6	26	0.42	0.46	0.47
HS 7	20	0.52	0.65	0.65
HS 8	26	0.57	0.59	0.62
HS 9	26	0.51	0.64	0.59
HS 10	20	0.39	0.49	0.50
HS 11	20	0.76	0.91	0.90
HS 12	26	0.64	0.69	0.71
HS 13	26	0.50	0.53	0.59
HS 14	26	0.44	0.50	0.45
HS 15	20	0.60	0.67	0.76
HS 16	20	0.30	0.48	0.43
HS 17	26	0.60	0.62	0.63
HS 18	26	0.57	0.60	0.67
Averages		0.56	0.65	0.65
Total n	414			

References

Abellan-Perpinan, J.-M., J.-L. Pinto Prades, et al. (2006). "Towards a better QALY model." Health Economics **15**: 665-676.

Abellan-Perpinan, J., H. Bleichrodt, et al. (2007). Testing the Predictive Validity of the Time Trade-off and the Standard Gamble, WP 07.17. Seville, Department of Economics, Universitad Pablo de Olavide.

Allais, M. and O. Hagen (1979). Expected Utility Hypotheses and the Allais Paradox. Dordrecht, D Reidel.

Attema, A. E. and W. B. F. Brouwer (2009). "The correction of TTO-scores for utility curvature using a risk-free utility elicitation method." Journal of Health Economics **28**: 234-243.

Bleichrodt, H. (2002). "A new explanation for the difference between time trade-off utilities and standard gamble utilities." Health Economics **11**: 447-456.

Bleichrodt, H., J. L. Pinto-Prades, et al. (2003). "A consistency testof the time trade-off." Journal of Health Economics **22**: 1037-1052.

Coast, J., T. N. Flynn, et al. (2008). "Valuing the ICECAP capability index for older people." Social Science & Medicine **67**(5): 874-882.

Cummins, R. A., A. L. D. Lau, et al. (2009). Subjective Wellbeing Homeostasis. Handbook of Social Indicators and Quality of Life Studies. K. C. Land. New York, Springer.

Doctor, J. N., J. Miyamoto, et al. (2009). "When are person trade-offs valid?" Journal of Health Economics **28**: 1018-1027.

Dolan, P. (2008). "Developing methods that really do value the "Q" in the QALY." Health Economics, Policy and Law **3**: 69-77.

Dolan, P. (2008). "In defence of subjective well-being." Health Economics, Policy and Law **3**: 93-95.

Dolan, P. and D. Kahneman (2008). "Interpretations of utiliy and their implications for the valuation of health'." Economic Journal **118**(525): 215-234.

Froberg, D. and R. Kane (1989). "Methodology for measuring health state preferences: II Scaling methods." Journal of Clinical Epidemiology **42**(5): 459-71.

Grewal, I., J. Lewis, et al. (2006). "Developing attributes for a generic quality of life measure for older people: Preferences or capabilities?" Social Science & Medicine **62**(8): 1891-1901.

Hausman, D. (2008). "Valuing health properly." Health Economics, Policy and Law **3**: 79-83.

Hawthorne, G., J. Richardson, et al. (2001). "A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments." Annals of Medicine **33**: 358-370.

Henry, D., S. Hill, et al. (2005). "Drug prices and value for money: The Australian Pharmaceutical Benefits Scheme." JAMA **294**(20): 2630-2632.

Johnson, F. R. (2009). "Editorial: Moving the QALY forward or just stuck in traffic?" Value in Health **12**(S1): S38-S39.

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An analysis of decision under risk." Econometrica **47**(2): 263-291.

Kahneman, D., P. Wakker, et al. (1997). "Back to Bentham? Explorations of experienced utility." Quarterly Journal of Economics **112**: 375-405.

Luce, R. D. (2000). Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches. London, L. Erlbaum.

Mann, R., J. Brazier, et al. (2009). "A comparison of patient and general population weightings of EQ-5D dimensions." Health Economics **18**: 363-372.

Marschak, J. (1950). "Rational behavior, uncertain prospects, and measurable utility." Econometrica **18**(2): 111-141.

Morgenstern, O. (1974). Some reflections on utility. Expected Utility and the Allais Paradox: Contemporary Discussions of Decisions under Uncertainty with Allais' Rejoinder. M. Allais and O. Hagen. Boston, Dordrecht Reidel.

Morgenstern, O. (1979). Some reflections on Utility. *Expected Utility Hypotheses and the Allais Paradox*. M. A. O. Hagen. London, UK, D Reidel Publishing Company**:** 175-184.

Nord, E. (1993). "The trade off between the severity of illness and treatment effect in cost-value analysis of health care." Health Policy **24**: 227-238.

Nord, E. (1995). "The person trade-off approach to valuing health care programs." Medical Decision Making **15**: 201-208.

Nord, E. (1999). Cost Value Analysis in Health Care. Cambridge, Cambridge University Press

Nord, E. (2001). "Severity of illness versus expected benefit in societal evaluation of health care interventions." Expert Review of Pharmacoeconomics Outcomes Research **1**(1): 85-92.

Parkin, D. and N. Devlin (2006). "Is there a case for using visual analogue scale valuations in cost-utility analysis?" Health Economics **15**: 653-664.

Patrick, D. and P. Erickson (1993). *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*, Oxford University Press.

Pope, R., J. Leitner, et al. (2007). The Knowledge Ahead Approach to Risk: Theory and Experimental Evidence Berlin, Heidelberg, Springer-Verlag.

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

Richardson, J. (2002). Evaluating summary measures of population health. Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications. C. J. Murray, J. Salomon, C. D. Mathers and A. D. Lopez. Geneva, World Health Organization**:** 147-160.

Richardson, J., A. Iezzi, et al. (2009). Data used in the development of the AQoL-8D (PsyQoL) Quality of Life Instrument, Research Paper 40. Melbourne, Centre for Health Economics, Monash University.

Richardson, J., A. Iezzi, et al. (2007). The Relative Social Willingness to Pay Instrument: Justification and Initial Results, Research Paper 22. Melbourne, Centre for Health Economics, Monash University.

Richardson, J. and J. McKie (2009). The Impossibility of an Ideal Metric for Health Service Benefit Measurement, Research Paper 33. Melbourne, Centre for Health Economics, Monash University.

Richardson, J. and R. D. Smith (2004). "Calculating society's willingness to pay for a QALY: Key questions for discussion." Applied Health Economics and Health Policy **3**(3): 125-126.

Savage, L. (1954). The Foundations of Statistics. New York, John Wiley & Sons.

Schoemaker, P. (1982). "The expected utility model: Its variants, purposes, evidence and limitations." *Journal of Economic Literature* **XX**: 529-563.

Schwarzinger, M., J.-L. Lanoe, et al. (2004). "Lack of multiplicative transivity in person trade-off responses." Health Economics **13**: 171-181.

Sen, A. (1993). Capability and Well-being. The Quality of Life. M. Nussbaum and A. Sen. Oxford, The Clarendon Press**:** 30-53.

Sen, A. (2009). The Idea of Justice. London, Allen Lane.

Smith, D. M., S. L. Brown, et al. (2008). "Are subjective well-being measures any better than decision utility measures?" Health Economics, Policy and Law **3**: 85-91.

Smith, R. D. and J. Richardson (2005). "Can we estimate the 'social' value of a QALY: Four core issues to resolve." Health Policy **74**: 77-84.

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

Torrance, G. and D. Feeney (1989). "Utilities and quality-adjusted life years." *International Journal of Technology Assessment in Health Care*.

Torrance, G. H., W. H. Thomas, et al. (1972). "A utility maximisation model for evaluation of health care programs." Health Services Research **7**: 118-33.