A Review of Methodological Issues in the Conduct of Willingness-to-Pay Studies in Health Care III:

Issues in the Analysis and Interpretation of WTP Data

Mr Richard Smith
Senior Lecturer, Health Economics Unit, Centre for Health Program Evaluation
Monash University

Professor Jan Abel Olsen
Visiting Professor, University of Tromso, Norway

Mr Anthony Harris
Senior Lecturer, Health Economics Unit, Centre for Health Program Evaluation
Monash University

April 1999
ISSN 1325 0663
ISBN 1 875677 98 4
The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of Public Health and Community Medicine at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre’s papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator  
Centre for Health Program Evaluation  
PO Box 477  
West Heidelberg Vic 3081, Australia  
**Telephone** + 61 3 9496 4433/4434  
**Facsimile** + 61 3 9496 4424  
E-mail CHPE@BusEco.monash.edu.au
ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE receives core funding from the National Health and Medical Research Council and Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR(S) ACKNOWLEDGMENTS

This paper forms part of a literature review commissioned by the Australian Department of Health and Aged Care to promote informed discussion of the measurement and valuation of health outcomes. It falls within the consultative process of revising the November 1995 edition of The Guidelines for the Pharmaceutical Industry on Preparation of submissions to the Pharmaceutical Benefits Advisory Committee: including major submissions involving economic analyses. The views expressed in the paper are not necessarily those of the Department, the Pharmaceutical Benefits Advisory Committee or its Sub-Committees and do not necessarily reflect the final outcome of this aspect of the next edition of these Guidelines.
Cost-effectiveness, and cost-utility, analyses have historically been the most widely used techniques of economic evaluation applied to the evaluation of health care programs. However, in recent years there has been renewed interest in the use of cost-benefit analysis, which requires the assessment of programme benefits in monetary terms. The emerging consensus is that such monetary valuation is most appropriately obtained using a survey of individual ‘willingness-to-pay’ (WTP) for the program of interest.

There are obviously a considerable number of methodological issues and potential biases to be considered in performing such a survey, which may be grouped into three main areas: (i) the construction and specification of the contingent market; (ii) the administration of the survey; and (iii) the analysis and interpretation of the WTP data. In addition, there are a few issues which also warrant consideration, such as assessing validity and the impact of ability to pay and income distribution issues. This paper is concerned with assessing the analysis and interpretation of WTP data, including the assessment of validity and distributional impact of WTP.

The paper considers the literature relating to these issues, and uses this to derive a set of ‘recommendations’ for current ‘state of the art’ conduct of WTP surveys with respect to market construction. WTP studies conducted to date in health and health care are then reviewed with such recommendations in mind, to assess the degree to which they reflect this ‘state of the art’. It is concluded that such studies perform poorly when judged in this manner.
## TABLE OF CONTENTS

### Abstract

1. Introduction

2. Literature Review Methodology
   2.1 Method for Selecting WTP Papers to Review
   2.2 Exclusion Criteria
   2.3 Review Process

3. Choice of welfare measure
   3.1 Recommendations
   3.2 Health Care WTP Studies

4. Income, Ability to Pay, and Distributional Issues Revisited
   4.1 Recommendations
   4.2 Health Care WTP Studies

5. Validity
   5.1 Recommendations
   5.2 Health Care WTP Studies

6. Summary Statistics
   6.1 Recommendations
   6.2 Health Care WTP Studies

7. Sample Size
7.1 Recommendations
7.2 Health Care WTP Studies

8. **Response Rate**
8.1 Recommendations
8.2 Health Care WTP Studies

9. **Transformation of Values**
9.1 Recommendations
9.2 Health Care WTP Studies

10. **Scale Effects and ‘Embedding’**
10.1 Recommendations
10.2 Health Care WTP Studies

11. **Presentation of Results**
11.1 Recommendations
11.2 Health Care WTP Studies

12. **Conclusions and Recommendations**

**REFERENCES**

**APPENDIX 1**
The complete reference list of papers reviewed

**LIST OF TABLES**

Table 1: Validity Testing by Pharmaceoeconomic Relevance

Table 2: Implications by Sample Size

Table 3: Sample Size by Data Collection

Table 4: Sample Size QST Format

Table 5: Response Size by Data Collection

Table 6: Response by QST Format
1 Introduction

Cost-benefit analysis (CBA) has been the major practical result of welfare economic theory over the last 50 years (Olsen & Smith, 1999). It has been used and developed particularly in the fields of environmental and transport economics\(^1\), and over the last decade interest in assessing the applicability of the technique to health care, and conducting CBAs of health care programs, has grown considerably\(^2\).

Within CBA the benefits of an intervention are assessed in units commensurate with the cost, typically monetary units, and there are two main methods of eliciting these values. The first uses market information for complementary or substitute goods. This so called ‘hedonic’ or ‘revealed preference’ technique uses, for example, labour or housing markets to value risk and thus place a money value on the risk of death or injury (Viscusi, 1993). The second involves an experimental

\(^1\) Carson et al (1993), for example, list over 1,400 contingent valuation studies conducted in environmental economics since 1964.

\(^2\) The first WTP study in health care is widely acknowledged to be one assessing the valuation of reductions in risk from mobile coronary care units (Acton, 1973). However, from then there were only another four studies conducted prior to 1990 (Thompson 1984, 1986; Berwick and Weinstein 1985; Reardon and Pathak 1989), with approximately 50 conducted since 1990.
survey, or contingent valuation (CV)\(^3\), approach to individual monetary evaluation of hypothetical changes in health and welfare. It asks individuals about either the amount of compensation required (their willingness to accept compensation) or their willingness to pay (WTP) to avoid an illness (Drummond et al, 1997).

The second of these techniques, contingent valuation, has a strong basis in the theory of constrained utility maximisation (Johansson, 1995) and has become the dominant technique used in the monetary valuation of benefits, in health care as well as other areas, in recent years; the main consensus being that individual ‘willingness to pay’ is the most appropriate means to value benefits in monetary terms (Olsen & Smith, 1999).

This paper is part of a series of papers undertaken to review the conceptual basis and methodological issues pertinent to the conduct of WTP studies in health and health care, and of pharmaceuticals in particular. The conceptual background, and its application in empirical studies, are discussed in two prior companion papers (Olsen et al, 1999; Olsen & Smith, 1999). The purpose of this paper is to review the methodological conduct of such studies. This includes an assessment of the current recommendations for “state of the art” contingent valuation studies, and an assessment of the manner in which CV studies in health care have been conducted to date, and how closely they match such recommendations.

Methodological issues are important at two levels. First, in the validity of the assessment made of the value of the benefits arising from a program. That is, does the WTP technique really measure what we think it does? Contingent valuation surveys (should) attempt to obtain precise and unbiased estimates of individual WTP for a program (Drummond et al, 1997). This means that CV surveys should be psychometrically robust (where apparently trivial changes in survey wording, information provided to the respondent, or question format, should not cause significant changes in the stated value), statistically reliable (stated values should be accurately estimated for the population, without significant bias), and economically sensible (stated values should correspond to other variables, such as income, in a manner predicted by economic theory).

Second, in order to use CBA, and therefore WTP, as a tool in decision making between various programs requires consistency in the use and reporting of analyses. Measurement should be driven by a coherent methodology to ensure we know what was valued, how it was valued, and the implications of changes in technique. It is therefore important that methodology in measurement is not left as an ad-hoc approach based on the disparate requirements of individual researchers. If methodology is not consistent then systematic biases\(^4\) will influence the result, giving an incorrect assessment of the benefits of a project.

\(^3\) So called because the answers to the valuation questions are contingent upon the particular hypothetical market described in the survey to respondents.

\(^4\) Biased responses are defined as those final WTP values that differ systematically from the ‘true’ values placed by respondents on the good in question.
Many such methodological issues and potential biases have been recognized and reviewed in a variety of publications over previous years, most notably in Mitchell & Carson (1989), Diamond & Hauseman (1993) and Beattie et al (1997). However, such reviews have been either general with respect to the WTP technique, or specific to the environmental or transport sectors. This review will not attempt to repeat such detailed reviews of issues where unnecessary, but summarize the relevant findings of these previous papers to assess recommendations for the conduct of ‘state of the art’ surveys. The contribution of this review is to: (i) collate this relevant information into a succinct review of the major issues in WTP surveys; (ii) to discuss and consider the relevance and impact of these issues in the application of WTP to the evaluation of health care programs, especially the evaluation of pharmaceuticals; and (iii) to assess the conduct of WTP studies in health care to date with respect to the recommendations apparent concerning each of these issues/biases.

It is particularly important to assess the relevance/implication of these issues with respect to the health care sector. It is almost received wisdom in health economics that health care is a ‘peculiar’ commodity, and therefore any recommendations from other sectors should be considered in light of these ‘peculiarities’ (see for example Mooney (1986) and McGuire et al (1988)). For example, in the environmental literature it is thought that the questionnaire, to ensure ‘realism and credibility’ should resemble market conditions as closely as possible (for instance Brookshire et al, 1976). However, asking values in a market format in an area where consumers are not used to paying for the goods may give misleading responses and have different implications for the meaning of ‘realism’. Similarly, although the NOAA Panel guidelines\(^5\) are appropriate for the use they were targeted, assessing non-use (or existence) values of environmental amenities, it is reasonable to question the rationality of all CV studies, including those concerned with use-value in health care, strictly adhering to them, no matter what the policy making context. Thus, although these issues have general relevance across whatever sector WTP is applied, there may be some particular implications emanating from use in the health care sector.

Although there are a considerable number of methodological issues and potential biases to be considered, it is possible to group these into three main areas of consideration when reviewing or conducting a WTP survey: (i) the construction and specification of the contingent market; (ii) the administration of the survey; and (iii) the analysis and interpretation of the WTP data. In addition, there are a few issues which also warrant consideration, such as assessing validity and reliability, and the impact of ability to pay and income distribution issues. This paper is concerned with assessing the analysis and interpretation of WTP data, incorporating issues of ability to pay and distributional issues. The other two areas being considered in companion papers (Smith et al, 1999a; 1999b).

\(^5\) The National Oceanic and Atmospheric Administration called a consensus panel together to review the state of CV studies for non-use value following the Exxon Valdez oil spill in Alaska (NOAA, 1993, 1994). A substantial set of guidelines where developed from this exhaustive review, which have been suggested for adoption in the conduct of most CV studies. However, they have not been without their critics (Hauseman, 1993).
Obtaining valid and reliable estimates of individual WTP for a commodity is vital to the assessment of cost-benefit. However, so is how these results are analysed and presented. For example, whether there is likely to be bias in the values of those who have responded versus those who have not, in the choice of summary statistic, or in the transformation of values from one context to another.

These, and other, issues are discussed here, including:

(i) choice of welfare measure;
(ii) distributional issues, reflecting concerns around the impact of ability to pay on WTP;
(iii) issues of validity;
(iv) choice of summary statistic;
(v) sample size;
(vi) response rate and non-response;
(vii) transformation of values;
(viii) scale effects; and
(ix) presentation of results.

Following this introduction, Section 2 describes the literature review process used in this paper. Subsequent sections then consider each of the issues outlined above in turn. Assessment will be made of the issue, its characteristics, importance and remedial measures, by reference to published works, and a ‘state of the art’ approach discussed. Based on this ‘state of the art’ evidence, a recommendation concerning the handling of a particular issue/bias will be made. Within the discussion of each issue consideration will be made of the applicability and/or relevance of the issue/bias, and the recommendations concerning it, to the conduct of WTP in health care, specifically with respect to the evaluation of pharmaceutical products. Section 12 concludes with a subjective assessment by the authors on the ‘validity’ of CV surveys from a methodological viewpoint, and with a summary of recommendations concerning the review and conduct of CV studies of pharmaceuticals.
2 Literature Review Methodology

Within this paper are two literature ‘reviews’. The first concerns the relevant literature, from health care and beyond, concerning each of the major methodological issues dealt with in this paper. These papers were not collected or reviewed as part of a ‘systematic’ review, but on a more ad-hoc basis. Many were already in the possession of one of the authors (RS) who already had a substantial collection of ‘general’ CBA/CV papers.

The second ‘review’ was of WTP studies specifically applied in the health care sector. The process for this is described here. Within each subsequent section of this paper the two reviews are presented as follows: first a review of the wider WTP literature relevant to the issue at hand, which will predominantly be from non-health care studies, followed by a brief summary of how studies conducted in health care have handled this issue, with an (obvious) indication of the quality of such studies.

2.1 Method for Selecting WTP Papers to Review

We have reviewed papers reporting from actual contingent valuation surveys of health or health care programmes published during the period from January 1st 1985 until May 31st 1998 (see Appendix 1). The selected papers were identified from three sources. First, a computerised bibliographic database search was conducted for papers written in English. Databases used for this search were Medline and EconLit, searched using the ‘WinSpirs’ package. The search was conducted using keywords (singularly and in combination with health and health care), as follows: contingent valuation, willingness to pay, willingness to accept, and cost-benefit analysis. Full details, including abstracts, were downloaded and reviewed for appropriateness and relevance to this review, with initial selection made by one of the authors (RS). This reduced list was also considered by the other author (JAO) and the final list used to order papers.

Second, during the review of these papers (once collected) any papers which looked to be of importance, and had been neglected by the above review, were noted and ordered. These additional papers were mostly focused upon specific issues within the method of WTP, CV or CBA, rather than empirical studies. In addition a search was made of a comprehensive in-house ‘EndNote’ database held by one of the authors (RS) concerning CBA and CUA.

Third, once the complete list of empirical studies was decided upon, this was sent to the four main authors in this area (Cam Donaldson, Magnus Johannesson, Bernie O’Brien and Mandy Ryan) asking them to identify: (i) whether all relevant papers of their own had been included; and (ii) whether they were aware of any other papers, not on the list, which they thought should be included in the review. The major contribution at this stage was to be forwarded copies of papers which were forthcoming, but at that present time (April 1998) had yet to have been published.

It should be noted that, as with any literature review, there is a possibility that literature may be overlooked due to: (i) being unpublished, or not published in peer-reviewed journals (such as consultancy reports for government or industry) and hence not on the on-line databases; (ii)
databases used not covering all relevant literature. We would appreciate to be made aware of any papers of relevance which we have not considered.

2.2 Exclusion Criteria

Papers had to report from a CV-survey of health care programmes or of dimensions of health. Four ‘exclusion criteria’ were applied: First, those which did not report the results of a survey. The majority of papers found in the review did not report from any specific WTP survey conducted, but addressed methodological or theoretical issues which could be of relevance to health. Second, papers which reported on the same survey as had been more extensively reported in other publications we had already included (Miedzybrodzka et al 1994, 1995; Johannesson 1992).

Third, papers which reported from surveys on WTP for the size of the health care sector (Eckerlund et al 1995) or the value of reduced waiting time (Johannesson et al 1998, Propper 1990). These papers made no reference to any health outcomes, nor to any health care programmes. Fourth, papers which were tangential to survey-based specific WTP, particularly those considering conjoint analysis (e.g. Ryan 1997). We found these to be inappropriate to this review as this method does not explicitly ask for a monetary value, but implies such values in a more indirect way.

2.3 Review Process

There were several specific criteria which were used by the authors to structure the review, relating to both conceptual and methodological issues. In this paper we shall be concerned with the following characteristics of the surveys:

- Data collection method (face-to-face, telephone, postal, self-administered);
- Development of scenario description;
- Presentation of the scenario description (separate or as part of the question);
- Pharmacoeconomic relevance (explicit, implicit or no drug);
- Methodological and/or policy implications considered;
- Payment vehicle used (out-of-pocket, tax, insurance);
- Types of respondents in the survey (eg users, general population);
- Cost-benefit statistic presented (eg NPV);
- Welfare measure used;
- Time period of assessment and use of WTP (eg per year, per month);
- Questionnaire format (eg bidding, discrete);
- Order effects of questions;
- Starting point/range bias tested for;
- Interviewer bias tested for;
• Strategic bias tested for;
• Other biases assessed;
• Duration of interview;
• Summary statistic (eg mean, median);
• Use of confidence intervals;
• Type of statistical analysis used;
• Sample size;
• Response rate;
• Zero/high responses – frequency;
• Transformation of WTP values obtained;
• Income assessed;
• WTP adjusted for income effects;
• Presentation of income constraint to respondents;
• Validity and reliability assessment.

The above criteria are used as the empirical basis within this paper. Both authors (RS & JAO) independently classified each study along these criteria, and then conferred to assess the degree of agreement. Areas upon which the authors had differed were then reviewed and a consensus decision made.
3 Choice of Welfare Measure

There are two main money measures of the value of a change in health, based on the Hicksian (income-compensated) demand curve, termed compensating and equivalent variation. The former assesses the value of the change in health according to the individual remaining at the initial level of utility prior to the change in health, and the latter at the level of utility which would be attained following the change in health (Johansson 1995). Typically these are expressed in CV surveys as WTP for a change in health, representing the compensating variation (the individuals maximum WTP being equivalent to the amount which will leave them at least no worse off than before the change, ie bring them back to their original, pre-change, level of utility), and WTA representing the equivalent variation (the minimum amount the individual would be WTA to be no worse off if the change did not occur, with reference to the utility level which would be achieved if the change did occur).

Received economic theory establishes that the amount of money that individuals are WTP for marginal increases in consumption states available to them should approximately equal the amount of money that they are willing to accept for an identical decrement in such consumption states. That is, in theory these two measures, apart from a slight income effect, will give the same monetary value of a change in welfare brought about by an intervention. This is shown by Willig (1976) for price changes and by Randall and Stoll (1980) and Takayama (1982) for quantity changes. As part of these theoretical arguments, income effects, typically viewed as ‘small’, are shown to drive a ‘small’ wedge between measures of WTP and WTA for a given individual.

Yet in contrast to the theoretical axioms predicting small variance in WTP and WTA, results from empirical and experimental studies have shown large persistent differences, with monetary valuations based on WTA invariably many times greater than those based on WTP (generally a factor of three or four) (Rowe et al. 1980; Bishop & Heberlein, 1986; Bishop et al, 1983; Brookshire et al, 1980; Knetsch & Sinden, 1984, 1987; Coursey et al, 1987; Garbacz & Thayer, 1983; Gregory 1986). There is also evidence that this disparity applies not just to the ‘point estimate’ of WTP/WTA, but also to the ‘confidence intervals’ around them (Dubourg et al, 1994). That is, when respondents were asked not just for a point estimate, but a ‘range’ within which they are confident their WTP/WTA value would lie, it was found that there was no overlap between the range of WTP values and range of WTA values (the lower bound on WTA exceeding the upper bound on WTP).

The problem of this potential disparity is simple. Any sizeable difference in the two measures could lead to ambiguity in the assessment of losses and gains. The benefits of a proposal might

---

6 Readers interested in other, less frequently used, measures are referred to Ng (1979).
7 A comprehensive theoretical exposition of these two money measures of change in utility can be found in Johansson (1995).
8 It has been argued that this disparity of WTP>WTA is predictable in theory if the commodity is a normal good (Johansson, 1995, p38).
exceed the costs if measured on one basis and fall short if measured on the other. The divergence can also give rise to the ‘Scitovsky paradox’, where a change and a subsequent return to the original position may both be warranted on the criterion that gainers are able to compensate losers: the latter presumably due to the effect of the change in wealth brought about by the initial redistribution of entitlements.

Although there is no single convincing explanation why these large disparities emerge, there have been a variety of reasons suggested, both conceptual and methodological. For example, Thayer (1981) argues that the divergence is due to a perceived distinction between out-of-pocket costs and opportunity costs, with the former coded as losses and the latter viewed as foregone gains, such that “once people have something it is very hard to take it away”. Some evidence is provided for this in the form of lottery games (Knetsch & Sinden, 1984). Here respondents appeared to be willing to spend actual or ‘realised’ income less readily than ‘opportunity’ income. If this is the case then evaluations will be affected accordingly, as it would take more dollars of the latter to be equivalent to any sum of the former. This may be due to some cognitive bias, such as to protect against a feeling of regret. Knetsch & Sinden (1984) further found that when respondents acted as agents for another individual, WTP approximately equalled WTA. Thus it appears that certain aspects of the asset or the nature of the entitlements are being overlooked or valued differently by people when they are acting as agents or on their own behalf. The motivations giving rise to the ‘endowment’ effects seem prime candidates.

A related conceptual cause, this time advanced by psychologists as distinct from economists, is the possible existence of what has been coined as “prospect theory” (Kahneman & Tversky, 1979; Kahneman et al, 1993), whereby people are more averse to a loss than attracted to an equivalent gain (ie loss aversion). In prospect theory the utility function is replaced by a value function that evaluates changes in income from the current level. Increases in income are weighted by a relatively small marginal utility. Decreases in income are weighted by a much larger marginal (dis)utility. In effect, the value function implies that a kink in the relationship between utility and income occurs at the initial income or reference point and that the slope of the utility function for losses in income is steeper than it is for gains. Thus, rather than evaluating outcomes in absolute terms as if from some neutral standpoint, respondents tend to code them as gains or losses relative to some reference point, and then treat them asymmetrically, with perceived losses looming larger than corresponding gains. This notion has quite recently been incorporated into another theory of behaviour termed ‘reference dependent utility’ (Tversky & Kahneman, 1981).

Economic theory, as distinct from psychological theory, would suggest that the difference in WTP and WTA is because people under-perceive gains and behaving irrationally to achieve a lower level of welfare then a true utility maximiser. In demonstrating this, Coursey et al (1987) use a Vickery auction process, where a number of iterations allows subjects to ‘learn’ that revealing their ‘true’ value is the dominant, best, strategy. Here opening bids showed the disparity as

---

mentioned, but the final, or end, bids did not. Thus economic theory is seen as predicting that such a disparity will not exist in a mature market, but may persist outside of markets, or where inexperienced market decisions are made. The obvious implications of this for studies in health care are that the market is to a large extent immature as the amount of information is severely limited, and there may be little incentive to learn to behave rationally if the consumer faces zero price.

As another explanation, Hanemann (1991) reanalysed Brookshire et al’s (1980) work on showing that Willig’s (1976) conclusions will carry over for a quantity setting. Here Hanemann’s analysis shows that the implications have been misunderstood. For quantity changes there is no presumption that WTP and WTA must be close in value and, unlike price changes, the difference depends not only on an income effect but also on a substitution effect. ie the ease with which other privately marketed commodities can be substituted for the given public good, while maintaining the individual at a constant level of utility. Hanemann shows that the smaller the substitution effect (the less substitutes are (perceived to be) available) the greater the disparity, citing health as an example of such a case. The paper offers an explanation for the disparity by showing that the theoretical presumption of approximate equality is misconceived. This is because the relation between the two measures depends on the substitution effect as well as income effect. Thus the large empirical differences may be indicative not of some failure in the survey methodology, but of a general perception on the part of individuals surveyed that their private market goods available in their choice set are, collectively, a rather imperfect substitute for the public good under consideration. Recent evidence for this comes from an experiment with real payment/compensations. Shogren et al (1994) found a convergence of WTP and WTA when the commodity considered was a private market good with reasonable levels of substitutes, yet found that there was a persistent disparity between WTP and WTA for a non-market commodity (reduced health risk) with imperfect substitutes.

On a more practical level, Gafni and Ravid (1989) suggest that the disparity between WTP and WTA can be explained by the use of bounded and non-bounded questions. That is, the individual faced with WTP will face a ‘bounded’ question, bounded by the maximum amount they can afford to spend - ultimately their lifetime earnings. However, those asked WTA are not bounded in asking a vast sum as compensation. ie WTP must be finite, yet WTA may be infinite. This seems perhaps an intuitively valid explanation.

However, whatever the reason, the observed disparities do not appear to be random, nor in many ways irrational. Due to this there has been an emerging consensus that WTP is the more ‘valid’ and reliable indicator of value. That is, that hypothetical WTP values are closer to true market values than WTA values. In particular in health care, where it would be unusual to be looking at removing some program, WTP also appears the most ‘sensible’ estimate of value as we are looking to the implementation of a program which currently does not exist.
3.1 Recommendations

In common with the growing consensus, and guidelines such as those by the NOAA Panel (1994), we recommend that WTP (representing compensating or equivalent variation) be the welfare measure used for the valuation of pharmaceutical products.

3.2 Health Care WTP Studies

The issue of choice of welfare measure was invariant between studies, with 51 (95%) opting for WTP, two for both and one solely for WTA. This is consistent with received wisdom within CV that WTP is the more appropriate measures, particularly in health care.
4 Income, Ability to Pay, and Distributional Issues Revisited

WTP, as an aggregate of individual values, is a function of the value placed on a commodity and the individuals ability to pay (ATP) for it (eg Gafni & Feder, 1987; Appel et al, 1990; Thompson et al, 1982, 84, 86; Berwick & Weinstein, 1985; Garbacz & Thayer, 1983). A frequent objection to WTP is, therefore, that “one dollar, one vote” inevitably accords the rich more votes in the allocative or regulatory decision process than the poor. There are, broadly speaking, three reactions to this issue.

First, if one takes the view that the current income/wealth distribution is optimal, then it is quite appropriate that those with greater command over resources should have potentially greater influence on the way they are allocated. The acceptance of ATP as a problem is bringing an important value judgement to bear on the WTP methodology, which may be seen as running counter to economic history in the use of the free market as the ideal allocative mechanism. As Gafni & Feder (1987) state: “equitable distribution of wealth does not exist in most (if any) places, and does not seem to be the goal of Western societies” (p16). One may presume that Gafni & Feder are implicitly asking whether, as equity is not a goal of western society, why do we wish to have it in this particular instance?

If, by contrast, the distribution of income/wealth is viewed as suboptimal, then, in principle, one could argue that the way to effect a redistribution is not by tampering with the results of CBA, but by direct redistributional measures of taxation and subsidy. However, it seems to have been accepted by many researchers that the distribution of income is suboptimal, and that (for whatever reason) such redistributional taxes and transfers are unfeasible. In this case it is argued that WTP results can be adjusted by some means. For example, Thompson et al (1984) argue that “this problem might be avoided if willingness to pay is expressed as a proportion of personal income, if the mean proportional willingness to pay is calculated, and if this proportion is multiplied by total income to determine total, adjusted, societal willingness to pay” (p200).

However, this raises several questions. The use of WTP expressed as a proportion of income implies that the author is of the opinion that an egalitarian (equal) distribution of income is more appropriate than the present distribution. Who is he/she to claim this? On what basis and evidence? What are the implications for comparison with other sectors of the economy with respect to public investment where ATP is not ‘controlled’ for? Can the simple use of this measure really ensure that a dollar for the rich will represent the same perceived amount of utility that it represents to the poor? These questions are bound up in the wider discussion on the social welfare function and equity notions embodied in WTP and economics as a whole, to which we now turn.

In order for individual WTP values to be used to guide decision making they require aggregation and summary. This paper has already considered issues in providing a summary measure of WTP above. However, a crucial issue not addressed was that using such an average (or other summary) statistic hides those who are gaining and losing, and by how much they are gaining.
and losing. Although, in theory, a positive aggregate WTP implies that those who gain can, hypothetically, compensate those who lose, in reality the picture is less clear cut, as illustrated by Boadway (1974).

Issues in the aggregation of individual values, and the underlying foundation of welfare economics based on Pareto welfare improvements, are based on the ‘Social Welfare Function’ (SWF). This is simply a function of the utility levels of all individuals within society, such that a higher value of the function is preferred to a lower value. Such functions are typically assumed to have four characteristics: (i) it is ‘welfarist’ in orientation, in that social welfare is simply determined by the sum of individual utility; (ii) social welfare increases as each person’s utility level increases (with no loss of another’s if it is to satisfy the strict Pareto criterion, or a net increase if it satisfies the weak Pareto criterion); (iii) social indifference curves are convex to the origin, such that society is not indifferent to the degree of inequality in society (as suggested by strict utilitarianism), but is only willing to accept a decrease in utility of the poor if the rich are made substantially better off (or alternatively accept a slight increase in utility of the poor for a substantially larger decrement in utility of the rich); and (iv) it does not matter who it is who experiences an increase or decrease in utility (Johansson, 1995).

Although the money value resulting from CBA provides individual values of changes in utility, it does not confer an ethical rule for weighting of gains and losses according to whom they accrue. That is, in the presence of convex SWF’s, in order to determine whether or not a program is a societally beneficial one is required to estimate the magnitude of the marginal social utility of income to be assigned to each individual involved (according to their existing income level prior to the proposed intervention taking place). It is this which establishes the ‘fairness’ or equity of an intervention in practice, and is inescapable as one either accepts these weights as currently indicated in the current societal distribution of wealth, or one makes an adjustment to the WTP values to reflect the degree of wealth of individual concerned10. Johansson (1995) provides a simple example illustrating this point, whereby although net WTP may be positive, accounting for the relative wealth of the individuals concerned results in the project being socially undesirable. Johansson concludes that “this example highlights the fact that it is dangerous to draw far reaching conclusions about a project’s desirability from the unweighted sum of money measures” (p122). The only exception to this is when the income distribution faced by the CBA has already been adjusted to reflect the equitable or fair SWF desired by society. However, arguably this is rarely, if ever, the case in practice.

If the distribution in society is considered optimal, or ‘fair’, then monetary gains and losses (across small programs) could be simply summed to determine an unambiguous improvement in social welfare. However, if this is not the case then some form of weighting is required. However, since these are not observable in practice, some other means must be established to assess them, which creates considerable difficulty. There are a number of procedures which have been

---

10 Note that most of the discussion here of aggregation also applies to QALYs or any other means to value outcomes.
outlined to try to overcome this difficulty, such as using SWF’s assessed for countries as a whole (for a review of such measures see Stoker (1993)).

An option more generally undertaken is simply to show how different distributional assumptions may affect the outcome of the CBA (such as using the extremes of Rawlsian and strict utilitarian SWF’s) (see Boadway & Bruce (1984) for a review of these functions). Such differences in SWFs can have a significant impact on the result of a CBA. For example, Loomis 1987 compared five different aggregation techniques, with total value of a human life varying from $153 to $394 million per annum according to the aggregation approach used. Another possibility is to outline the unweighted CBA results, and supplement this with a more disaggregated analysis of gains and losses to different groups within the population, such as across income and age\textsuperscript{11}.

As Johansson (1995) states concerning this issue “an important part of a properly conducted cost-benefit analysis is to show the distributional consequences of the project, i.e. to point out who wins and who loses from the project……The underlying assumption, ie that social welfare is optimally distributed, must however be stressed by authors using this approach [of presenting the profitability of the project by summing benefits and costs across individuals]” (p158).

Perhaps the best that occurs in most situations is to assume that the program is sufficiently small, assume that we are close enough to an optimal SWF and assume that other goods and services are in general equilibrium. Such assumptions certainly appear to be the case in recent WTP studies in health care.

4.1 Recommendations

It is likely that income will be correlated with WTP, but with programs expected to have a slight impact on overall expenditure, and therefore overall income, one might expect this income effect to be slight. In this sense it could be argued that it should be ignored. However, this ignores the impact that ATP undoubtedly has. The NOAA Panel (1994), for example, have recommended that respondents are reminded of their income constraint during the valuation procedure to ensure that they are aware that if they state an amount X which they are WTP for this commodity then they will have X less income to spend on other goods and services. One would therefore expect income to be positively associated with ATP (and as we see in the next section is used as an assessment of validity of the WTP values themselves).

In practice we would recommend that unadjusted results are provided, but that these are \textit{disaggregated} according to income group and any other relevant subgroup. It is our opinion that

\textsuperscript{11} Note also that there is an anomaly termed the ‘Boadway paradox’ (Boadway, 1974). Boadway showed that, when the vector of prices and compensated incomes used in defining WTP/WTA do not correspond to a general equilibrium (as is arguably the case in virtually all nations at all times) then one cannot assume that a net WTP figure necessarily represents a passing of the Kaldor-Hicks PPI compensation test (see Boadway for more detail). A further and final point is of course that such compensation is always only considered to be hypothetical, with winners never actually compensating losers.
most health care programs, especially if WTP is expressed as a form of increased taxation, are unlikely, for most people, to require a substantial expenditure and therefore income effects are likely to be minor.

4.2 Health Care WTP Studies

Forty studies (75%) recorded respondents income, although not all of these used that information to consider the distributional impact. Only three studies reported WTP values according to ‘proportion of income’. Thirty-three studies considered ATP in another, less direct, fashion, mainly in regression, or other correlation, analyses. However, it is worth noting that this was not primarily to consider the impact of income, but to assess validity according to construct validity (that WTP rises as income rises). This issue is discussed further below. In fact only 11 studies (20%) considered distributional issues in their WTP studies.

Of the 11 considering distributional issues, only two were primarily concerned with policy (three with methodology, and six with both), which means that most studies aimed at policy did not take account of distributional effects. Four of the 11 studies assessed validity by using income, and nine contained an assessment of factors predicting WTP, including income.
5 Validity

Assessment of the validity of WTP responses is difficult. The ‘gold standard’ against which to compare predicted WTP values from CV surveys is obviously what individuals would be WTP in a market context. However, the very fact we are undertaking WTP surveys indicates that such a market is not available. In health care an additional difficulty is that there may not even be a close market against which to make any comparisons (or at least a market which is not severely distorted through, for example, moral hazard or supplier-induced demand). In this case criterion validity may not be established\(^\text{12}\).

In light of this, there have been attempts to establish tests of construct validity. That is, to assess whether the WTP responses accord to predetermined theoretical constructs that should be present if the WTP responses are accurately assessing the value we intend. There have been two proposals made for such constructs in the assessment of WTP: (i) that (most) goods have a positive income elasticity, meaning that, \textit{ceterus paribus}, higher respondent incomes should be associated with higher expressed WTP by those respondents; and (ii) that the more of a (positively valued) good which is supplied, the higher the WTP will be, subject to diminishing marginal returns, meaning that as more of a commodity is valued in a CV survey, the higher the WTP for it should be, increasing at a diminishing rate (Drummond et al, 1997).

This second construct in particular has been advocated by the NOAA Panel, under the term ‘scope tests’ (the proposition being that WTP should vary with the scope of the benefit arising from the hypothetical program). Such ‘scope test’ have also be advocated by others (Johannesson, 1996). However, both these constructs (income and scope effects) have begun to be used in WTP studies in health care (eg Kartman et al 1996; Stalhammer and Johannesson 1996; O’Brien et al 1998).

One of the most significant issue in validity, from a psychological perspective, is the belief that respondents will have a single underlying preference for a commodity. Fischhoff (1991) points to a possible scale of “articulated values” in which one extreme is characterised by well-formed preferences concerning any and every topic which may be directly retrieved in response to an appropriate elicitation question. A belief in this extreme of preference function would then entail a focus upon methodology of elicitation to ensure that the appropriate question, and therefore response, is being addressed. CV literature seems to be focussed on this stage, as most of the papers referenced in this review testify. However at the other extreme it is assumed that respondents will only have (reasonably) well defined preferences, or basic values, for very few and familiar topics. In this case, when faced with unfamiliar topics, or preference decisions, respondents must infer their valuation from these underlying values. That is, rather than

\[^{12}\text{To the authors knowledge only one study has set out with the aim of ‘validating’ the contingent method. Brookshire et al (1987) who performed an experiment in Los Angeles on air pollution to validate the survey approach by direct comparison with hedonic property values. However, the fundamental problem with this type of study is that we do not have a ‘Gold Standard’ measure to tell us the true value, and so we can never be sure of what we are measuring - we are in essence comparing two indirect measures and as such assessing con\textit{vergent validity}.}\]
retrieving a pre-defined fixed value they construct a value there and then in response to the question (Slovic et al, 1990).

Thus, if such values in CV surveys were constructed one might expect great variance and sensitivity, by chance as well as systematic factors concerning construction and expression of the contingent market. Perhaps the most significant conclusion of such constructed, rather than retrieved preferences, is that the value expressed need not be generated by a consistent, invariant, algorithm, such as an expected utility calculation (Tversky et al, 1988). On the contrary, it might well be, and indeed has been shown, that individuals have a great many methods by which they will identify and develop their preferences (Fong et al, 1986; Kruglanski, 1989; Larrick et al, 1990). Whereas in CV studies it is assumed that values may differ across individuals (due to different utility functions), or across different situations (some things are worth more than others), the process by which these values are reached does not differ. The essential problem may well be that it is not measuring preferences which is the major difficulty but simply that “human beings have unstable, inconsistent, incompletely evoked, and imprecise goals, at least in part because human abilities limit preference orderliness” (March, 1978: p598). This would explain the sensitivities of CV responses to theoretically irrelevant features of question format, as well as insensitivities to factors that may be implicit rather than explicit in CV surveys. Under such circumstances it is not at all surprising that such questionnaire-based methods will generate values that are inconsistent with one another from the standpoint of standard economic theory. It may therefore be the case that in future we should be considering assessing the ‘range’ of values people will be WTP, together with the ‘probability’ that they will pay any particular one of these at one particular time.

One method which could be used to assess the validity of WTP response is the verbal-protocol analysis, or more commonly termed “think-aloud” method, which is a type of psychological “process tracing technique” (Schkade & Payne, 1993). These are used to map the choice rules people deploy when faced with complex choice situations (Payne et al, 1978; Harte & Koele, 1995). This is a relatively new procedure in WTP studies, but shows some promise in understanding how people come to the WTP value expressed. In their study, Schkade & Payne (1993) elicited WTP values for the protection of migratory waterfowl from drowning in waste-oil polluted ponds. During the survey individuals were encouraged to “think aloud” as they considered the information presented and to report all that went through their minds in coming to the WTP value expressed. Many different considerations where reported, but the overall impression was that individuals had great difficulty in coming to defining their underlying preferences and how these should be expressed in monetary amounts.
5.1 Recommendations

WTP should be treated with some scepticism, although other measures of value, such as QALYs, are equally suspect and there is, unfortunately, no means of establishing the most valid. However, it has to be recognised that this lack of demonstration of the validity of WTP values throws serious doubt upon the validity of this method in the estimation of ‘definitive’ values for commodities. It is therefore recommended that WTP studies should always be comparative, as it is likely that the difference in WTP values expressed is (at least more) valid.

Notwithstanding the above, studies should undertake the tests of construct validity outlined (testing for positive price elasticity of demand and scope tests).

5.2 Health Care WTP Studies

We classified studies according to whether they had addressed the validity of the WTP values estimated comprehensively, vaguely or not at all. We interpreted ‘comprehensive’ in this situation to mean that more than one test for validity had been undertaken (eg two or more of content, construct, criterion or convergent validity). Vague tests for validity were, in contrast, considered when only one such test was undertaken.

We found that, according to this classification, only three studies had comprehensively tested for validity. However, 25 were found to have conducted vague tests, comprising an assessment only of construct validity, by assessing the relationship between WTP and income and/or size of the benefit offered (scope tests). Overall, 26/47 studies (48%) still made no attempt to assess the validity of the WTP values presented. The profile of validity testing for studies focussed upon drugs is however slightly better, as illustrated in Table 1.

<table>
<thead>
<tr>
<th>Pharmaco-economic relevance</th>
<th>VALIDITY TESTING</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprehensive</td>
<td>Vague (income)</td>
</tr>
<tr>
<td>Explicit drugs</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Implicit drugs</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Non-drug</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3</td>
<td>25</td>
</tr>
</tbody>
</table>

The only studies conducting comprehensive tests for validity were also those whose primary purpose was methodological, and in these cases an assessment of validity was an aim of the studies. It is alarming that 12/18 studies focused primarily on policy did not have any assessment of the validity of the values presented. However, on the positive side, there has been an increase
in validity assessment: over the period 1985-89 25% of papers reported any (vague or comprehensive) assessment of validity, compared with 80% for 1990-93, 70% in 1994-96 and 55% for 1997-98. There does not seem to be a ready explanation of the drop off in over the last 18 months, however.
6 Summary Statistics

In general, the main means of presenting summary WTP data is to use the arithmetic mean (occasionally the geometric or truncated means) and/or the median. However, as these measures often do not (in fact rarely) coincide, the issue becomes which measure to use. Such an issue has long been recognised. Hanemann (1984), for example, acknowledges that the mean will be sensitive to slight changes in the shape of the distribution resulting from different estimation methods or incorporation/exclusion of outliers, whilst the median is relatively robust. He also acknowledges that the choice between them entails a value judgement as to the appropriate method of conducting welfare evaluations. This has led to a (renewed) debate in the literature concerning the most appropriate summary statistic to use as a point estimate of WTP (Hanemann, 1984; Johansson et al, 1989; Hanemann, 1989).

It is generally accepted that there is only one theoretically correct measure (according to welfare economic foundations), which is the mean (Smith et al, 1986; Johansson et al, 1989). The ultimate goal of CBA is to compare a project’s expected aggregate costs and benefits. That is, compare the total WTP for the project against the projects total costs (leaving equity and income distribution issues aside for the moment). The mean is the correct statistic which, when multiplied by the number of persons in the population who receive the program under evaluation, yields the total population (societal) WTP (Johannesson, 1996). If, for example, there are 100 individuals concerned, then total WTP is simply 100 times the sample mean value. However, often the results of CV surveys are presented as a median, which might be considered as a sort of ‘referendum’ approach (particularly for CV surveys using discrete-format questionnaires), where the median is that point at which 50% of the sample are in favour, and 50% against, the proposed project at a specific ‘cost’, represented by the median.

The problem with the median voter approach is that it is known to not, in general, produce a Pareto-efficient outcome (Johansson et al, 1989), as median voters prefer more/less public expenditure than is consistent with Pareto efficiency. Two points arise from this. First, the median appears to be a pedagogical device which may help the respondent understand the intuition behind the WTP questionnaire. However, it is not the correct statistic to use in the calculation of cost-benefit. Second, if one is simply interested in the outcome of a referendum then one could simply ask ‘voters’ if they would plan to vote yes or no. It would be unnecessary to perform a full CV experiment concerning actual WTP (Johansson et al, 1989). What we therefore witness is the use of a theoretically incorrect measure because of its practical advantages. A counter argument to this is that the mean only differs from the median in such studies due to an artefact of the CV study itself skewing the WTP data due to a few individuals being WTP very high amounts. Thus using the median could be argued to be closer approximating the true mean WTP.

However, much of the argument becomes ethical, based on whether the criterion of Pareto efficiency is one which is considered the most valid in the decision of whether a project should proceed or not. For example, Hanemann (1989) advances the example that “out of a community
of 1,000 individuals, 999 value some environmental amenity at $1 each, while one person values it at $1,000. If there were some project to protect this amenity at an average cost of $1.98 per person, and if the cost were to be shared equally among all the members of the community, I would regard it as a bad project - even though it passed the Kaldor-Hicks test" (1989, p1060). Hanemann recognises that his position would disenfranchise those who have the largest stake, but suggests that this may occur every time a majority rule is taken.

Perhaps then the issue rests on more ethical grounds of income distribution, for example in assuming that one individual in Hanemanns example can afford to pay $1,000, whereas the other cannot. If this is not the case, that for example income distribution is equal, then the case for not accepting the Pareto efficient position seems hard to defend, as the issue in interpersonal comparison of utility is that the person representing $1,000 values the good 1,000 times the others (the argument then of course returns to the issue of interpersonal comparison (Olsen et al, 1999)). Let's phrase this another way - suppose the good in question is one persons life. Now let's run the example with that person WTP $1,000, but the others only WTP $1 to save his life. What would be the optimal criterion now? We would suggest Pareto efficiency. However, this argument becomes more conceptual and philosophical than methodological, and will therefore not be considered further here.

From a methodological viewpoint, the main interest is in the consistency of the summary measure used. It seems clear that presentation of a single summary figure is not sufficient, but that some form of measure of variance is required, such as by placing confidence intervals around the mean at a specified level of significance. Confidence intervals have been estimated a number of ways, including the delta method (Adelbasit & Plackett, 1983), bootstrapping (Duffield & Patterson, 1991) and Monte Carlo simulations (Langford & Bateman, 1993; Adamowicz et al, 1989). For discrete questionnaires, this issue of particular concern as the performance of each is variable.

For example, Cooper (1994) provides a comparative analysis of the reliability and precision of four techniques for estimating confidence intervals: the bootstrap (Dorfman et al, 1990), the jackknife (Efron, 1979), and those proposed by Krinsky and Robb (Krinsky & Robb, 1986; Park et al, 1991) and Cameron (1991). Using Monte Carlo analysis, Cooper concludes that although all four measures perform well on average, they differ in the frequency with which they do well, and that the best choice depends on the sample size, distribution of the WTP estimate and choice of functional form for estimating WTP. eg bootstrap being the preferred choice where sample size is over 1,000, but the Cameron method is preferred where the sample size is less than 100 (provided the distribution is logistic; the Krinsky and Robb performs best with a Weibull distribution). The results of this study again highlight the complexities of estimating precise WTP estimates from discrete data. Similarly, Bowker and Stoll (1988) also indicate that “models with fairly similar statistical fits can lead to very disparate measures of economic value, regardless of whether the mean or median is chosen” (Bowker & Stoll, 1988, p379).
6.1 Recommendations

We would maintain an argument for the use of the mean as the summary welfare measure in WTP studies. However, we would not recommend that this mean be reported alone, or without qualification. It is critical to be provided with an estimation of the range of values (particularly those of zero and of a high level to consider protest bids, as discussed) and confidence levels for the mean values given. We would also recommend that researchers report the analysis adjusted for any values considered to be ‘outliers’ and present results in a systematic fashion to consider the influence of extreme values, whether these are likely to be protest or ‘true’ WTP values and how they influence the WTP, and therefore CBA, result.

6.2 Health Care WTP Studies

Of the 54 studies, 39 reported arithmetic mean values, two geometric or truncated means, 27 median values and 11 ‘other’ summary measures (mainly the modal values). Of these, the majority of studies reported both means and medians, and some form of confidence interval (30 reported some form of confidence interval or standard error).

Medians were approximately twice as likely to be reported if the questionnaire format was discrete (nine versus four for open ended, four for bidding and five for payment cards of the 27 reported). The chances of reporting arithmetic means are more equal, with five open-ended, six bidding, nine payment card and nine discrete (of the 39 reported). Similarly, for reporting of confidence intervals, these were similar across question format (six open-ended, five bidding, seven payment cards and seven discrete, from the 30 reported). Out of the nine explicit drug studies, seven reported arithmetic mean values, four reported medians, and six reported confidence intervals.

Of those studies using discrete format questionnaires, eight (out of 15 studies which performed discrete surveys) used parametric statistical analysis only to derive the WTP result. These studies did not, however, assess the impact of differing functional forms which, as outlined, can significantly alter the WTP value. No study performed non-parametric analyses alone, and seven performed both.
7 Sample Size

What is the appropriate sample size for a CV study? How long is a piece of string is frequently the answer. Economic evaluation in general is very poor at assessing adequate sample sizes, more frequently relying on the most that can be gained, or what is provided by a clinical trial. It is the attitude of ‘getting as large a sample as one can’ that led to the use of mail-out questionnaires. CV surveys may be administered by one of three main means: direct face-to-face interview, by telephone interview or by postal questionnaire. The cost of the survey to achieve any particular sample size generally will fall as one moves along this range of possible means.

However, there is no reason why an estimate cannot be made of appropriate sample size required to assess the difference in WTP between interventions, or difference from zero, for studies on a similar basis to that for other studies. The real problem is that there is no guidance for an acceptable standard error around the mean value observed. A further complication is that sample sizes are likely to differ according to whether the valuation is of a single commodity or close substitutes. With the latter it is likely that, to show a statistically significant difference in WTP value, the sample size will have to be larger than for the former, where a value greater than zero is required.

7.1 Recommendations

We would recommend that sample sizes are used which are sufficient to provide statistically significant estimates of WTP. It is therefore the researchers responsibility to ensure that their sample size is sufficient to achieve this.

7.2 Health Care WTP Studies

Sample size varied dramatically across studies, from a minimum of 13 to a maximum of 3,486. The mean was 558 (with a standard error of 95 and a deviation of 700), modal values of 50 and 400, and quartiles of 121, 406 (median) and 655. The majority of studies (over 50%) are between 200 and 1000, with few higher than 1000, although a significant amount have samples less than 100 (10 studies or 18%).

Sample size did not show a discernable trend with the purpose of the study being primarily methodological or policy orientated, although there was a slightly high representation of few respondents within those aimed at policy than those aimed at methodology, which seems counter intuitive. This can be seen in Table 2.
Table 2: **Implication by Sample Size**

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Methodological</th>
<th>Health policy</th>
<th>Both</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;50</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>51-100</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>101-200</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>201-500</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>501-1,000</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>1,001-2,000</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2,001-3,000</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>&gt;3,000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>21</strong></td>
<td><strong>18</strong></td>
<td><strong>15</strong></td>
<td><strong>54</strong></td>
</tr>
</tbody>
</table>

One point of interest is that studies sampling the general population used higher total sample sizes, with none below 200. In contrast, the smaller samples, less than 200, were only found in samples of users or convenience samples (which also had no samples greater than 2,000). Face-to-face interviews were spread more evenly across sample sizes from 13 to 2000, whereas postal and other self-administered questionnaires tended to be concentrated in the mid range from 200 to 1000. This is illustrated in Table 3.

Table 3: **Sample Size by Data Collection**

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Face2face</th>
<th>Telephone</th>
<th>Postal</th>
<th>Self adm</th>
<th>Not available</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;50</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>51-100</td>
<td>3</td>
<td></td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>101-200</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>201-500</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>501-1,000</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>1,001-2,000</td>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2,001-3,000</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>&gt;3,000</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>19</strong></td>
<td><strong>6</strong></td>
<td><strong>12</strong></td>
<td><strong>16</strong></td>
<td><strong>1</strong></td>
<td><strong>54</strong></td>
</tr>
</tbody>
</table>

Looking at variation in sample size according to questionnaire format, we can see from Table 4 that the techniques of open-ended, bidding and payment card, tend to be concentrated in samples of less than 1000, with most less than 500. Discrete surveys however, tend to be found in larger samples of over 100, and a greater representation in samples over 500 and 1000.
### Table 4: Sample Size QST Format

<table>
<thead>
<tr>
<th>Samp. Size</th>
<th>Open ended</th>
<th>Bid’g card</th>
<th>Discrete</th>
<th>Discrete +FU</th>
<th>Other</th>
<th>N/A</th>
<th>Open +pay card</th>
<th>Bid’g+discrete</th>
<th>Open + other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;50</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>51-100</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>101-200</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>201-500</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>501-1,000</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>1,001-2,000</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>2,001-3,000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>&gt;3,000</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>7</td>
<td>12</td>
<td>12</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>54</td>
</tr>
</tbody>
</table>
8 Response Rate

Non-response causes problems in any survey if the sample not responding is likely to be systematically different from those responding. That is, in the case of contingent valuation if the non-responders would have a significantly different WTP compared to those who did respond. Ensuring a high response rate is therefore important for any survey, and considerable efforts are expended in trying to ensure a survey will yield as high a response rate as possible. For example, this was a major reason behind the development of the bidding and payment card approaches to assisting respondents in answering open-ended contingent valuation surveys.

The response rate, however, is a complex function of many factors. These include the overall level of difficulty of the questions posed, the amount of information provided which may lead to cognitive overload in response, the physical length of the survey leading to respondent fatigue, the use of personal face-to-face, or telephone, interviews versus postal questionnaires and the use of financial or other incentives to complete, such as entry in prize draw or promise of lottery tickets (Chapman & Wong, 1991).

It is often argued that, by being shorter and more ‘realistic’, the discrete choice format approach is more likely to yield a higher response rate (Donaldson et al, 1995). However, it is not clear that this has been the case empirically, and there is evidence that discrete surveys have produced (in directly comparative studies) lower response rates than other formats (eg Herriges & Shogren, 1996; Kristrom, 1993; Holmes & Kramer, 1995; Johansson et al, 1995).

Clearly then it is desirable to get as high a response rate as possible. However, what level of response is deemed ‘acceptable’? The NOAA panel (1994), for example, recommend a minimum response rate of 70% for environmental CV studies, although this is rather an arbitrary figure. Of more importance is to assess the likely difference between non-responders and responders. For example, there might be an assessment made of other characteristics, providing, of course, that the remainder of the survey is completed and just the WTP questions incomplete. Where there is no data available, one might try to assess a ‘surrogate’ non-responder, such as those who make protest bids. Perhaps all that can be said for certain is that the higher the response rate the more confidence we can have in the results, and that at the very least sensitivity analysis should be performed on the values to try to reflect potential differences in non-responders. That is, if there is no response then some assumption has to be made concerning the WTP of this sample.

13 For this reason as well as to ensure survey expenditure is kept to a minimum.
14 Note also that double and triple bounded surveys have typically been used in phone or personal surveys, where the flow of questions is controlled by the interviewer. The use of these in mailout surveys raises similar issues to that of postal bidding surveys of the ability and willingness to respondents to follow the inevitably more complex survey instructions, and it has been shown that the addition of more bounds leads to correspondingly lower response rates (37).
8.1 Recommendations

Non-response is a problem if we cannot be sure that the values from this group do not differ systematically from those of respondents. It is therefore important to have as high a rate as possible, or to establish that these values are unlikely to differ systematically. Of these options, the former is the simplest approach. In the absence of other guidelines, we would suggest that a level, such as 70% as proposed by the NOAA Panel, be set as the minimum response rate required.

8.2 Health Care WTP Studies

The studies reviewed here reported response rates of between 6% and 100%, with a mean of 74% (standard deviation of 26%). The majority of studies reported response rates of greater than 61%, with 50% greater than 80%. However, a significant number failed to gain responses of over 40% (six studies) which is cause for concern.

It is interesting to see what correlates, and may therefore explain, differences in response rates. Response rates varied according to the purpose of the study. Those with a primary methodological aim experience a greater variation, being evenly spread over the 41% to 100% range. In contrast, those primarily aimed at policy where heavily concentrated (12/17) in the 81-100% range.

Studies using users were mostly within the 81-100% response rate (19/37), with 9/37 in the 61-80% range. A similar pattern was found for convenience samples and general population samples, although the total numbers here were very small (total of six and seven respectively). The variation in response rates with mode of collection of data are presented in Table 5.

Table 5: Response Size by Data Collection

<table>
<thead>
<tr>
<th>Response size</th>
<th>Face2face</th>
<th>Telephone</th>
<th>Postal</th>
<th>Self adm</th>
<th>Not available</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;20%</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>21-40%</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>41-60%</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>61-80%</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>81-100%</td>
<td>11</td>
<td>6</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>TOTAL</td>
<td>19</td>
<td>6</td>
<td>12</td>
<td>14</td>
<td>1</td>
<td>52</td>
</tr>
</tbody>
</table>
One can see that around 60% of face-to-face studies yielded response rates between 81% and 100%, compared with 8% of postal and 57% of other self-complete questionnaires. Telephone surveys all yielded high response rates. Interestingly, self-complete questionnaire did not yield less than 41% response, probably because these mainly involved respondents self-completing at a doctor’s surgery or where the intervention was to be offered. We can conclude from this data that interviews, or at least having people complete surveys whilst within the vicinity of the intervention, is preferable in yielding higher response rates than postal questionnaires, as expected.

We also considered response rates tabulated against questionnaire format, and these results are presented in Table 6.

**Table 6: Response by QST Format**

<table>
<thead>
<tr>
<th>Samp. size</th>
<th>Open ended</th>
<th>Bid’g</th>
<th>Paycard</th>
<th>Discrete</th>
<th>Discrete+FU</th>
<th>Other</th>
<th>Open+discrete</th>
<th>Open+paycard</th>
<th>Bid’g+discrete</th>
<th>Open+other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;20%</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>21-40%</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>41-60%</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>61-80%</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>81-100%</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>7</strong></td>
<td><strong>7</strong></td>
<td><strong>11</strong></td>
<td><strong>12</strong></td>
<td><strong>3</strong></td>
<td><strong>3</strong></td>
<td><strong>2</strong></td>
<td><strong>1</strong></td>
<td><strong>2</strong></td>
<td><strong>1</strong></td>
<td><strong>52</strong></td>
</tr>
</tbody>
</table>

Here one can see that discrete surveys do seem to yield higher average response rates, although payment cards and bidding surveys are comparable. The problem here is disentangling all other effects, such as whether interviewer or postal etc, from the data to state categorically that one format appears to offer higher response rates than another. However, from this data they appear roughly comparable.

We would expect response rate to correlate with the time taken to complete surveys, or the size of the task asked of respondents. However, only five studies provided an indication of the duration of the survey, which is too few to make any conclusion concerning this issue.
9 Transformation of Values

There is an issue involved in the use of values gained for one place, intervention or benefit, being applied to another. An obvious example of this comes in the transformation of values across size of benefit, such as used in the value of life literature, using the WTP for a small change in risk to value a large change in risk. However, this issue may be broadened to consider potential bias in values gained from elsewhere being ‘recycled’, and thus the validity of doing this.

The issue is whether the value an individual places on a commodity is simply a function of the benefits obtained from consumption of the commodity or a combination with other considerations, or ‘contextual’ factors. For example, is the WTP for a reduction in risk of facial scarring independent of the circumstances under which such injuries may be sustained, such as a road accident versus playing a contact sport? There is nothing which rules out the possibility that such contextual factors may enter into individual tastes and preferences and cause an individual’s WTP for a commodity to vary between different circumstances. The problem is that, to these authors knowledge, there has been no systematic attempt to establish the importance of such factors.

9.1 Recommendations

WTP values should only ever be used and applied to the specific circumstances from which they were obtained and not used to value alternative sizes of the benefit, or different levels of risk or probability.

9.2 Health Care WTP Studies

As alluded to above, some studies undertook to transform values from the context in which they were valued to another (eg from values given across a one month period to values which are inferred across a 12 month period). We considered whether studies had used the values within the context they were derived, whether they had been transformed, and whether the authors had considered the issue of transformation of values, whether or not they had undertaken such transformation. The results were that 45 had used the values in the context in which they were gained, 12 embarked on some form of transformation (some of whom also reported the values in context, hence some overlap), and for three it was not possible to tell. However, only one study explicitly considered issue involved in the transformation of values (which was one of the studies undertaking such transformation).
10 Scale Effects and ‘Embedding’

It has been discovered in the environmental literature that WTP values do not necessarily vary with the scale of the commodity being valued. This phenomena is known as ‘embedding’. This refers to the respondent viewing the proposed program as being symbolic of a larger program. For example, WTP for the prevention of decline in fish populations in the whole of Ontario, Canada, was found to be virtually identical to the WTP for the same measures in lakes within a small area of Ontario (Cummings et al, 1986). Similarly, another study found no significant difference in WTP values to save 2,000, 20,000 and 200,000 waterfowl from dying (Desvousges et al, 1993). However, although this has been found to be problem in environmental economics, it has yet to be well tested in health care (Kahneman and Knetsch 1992a, 1992b; Carson and Mitchell 1994a, 1994b; Smith 1992).

Such ‘embedding’ is important for WTP surveys as it is hypothesised in economic theory that if the characteristics of a good changes (for example more of the good is offered) then the WTP value for it will also change (for instance increase). If this is not the case then there are implications for validity and for the practical usefulness of such value. In particular, that in assessing the value of a specific program we may be vastly overestimating it’s value, as the respondent is simply valuing it as a proxy for any program for that condition. In many studies to date in health care, there has been an assessment of an intervention versus no intervention. In these cases one might suggest that the WTP may reflect the value of an intervention per se, rather than any one specific intervention.

It has also been argued that individuals, in providing their WTP values, refer to “mental accounts” concerning groups of commodities, rather than individual commodities (Tversky and Kahneman, 1981). For example, instead of allocating a set amount to attending opera, another for cinema and another for theatre, an individual may allocate one figure to “entertainment” and distribute it differently at different times between these three different possible claims on that “entertainment” account. Thus decisions are two stage: first is the allocative decision to entertainment versus other use for the money, and the second is an evaluation of any particular possible use for that account at any specific time. It is therefore important to ascertain which value is being obtained in the process - the first or second decision.

Further, there is an additional issue of whether we can simply combine willingness to pay results across individuals over time. If we asked the everyone in the community what they would be willing to pay for every drug on the PBS separately, the total might exceed current total government expenditure.
10.1 Recommendations

We would recommend that WTP studies are conducted in a comparative sense and individuals made aware of any close substitutes to avoid WTP for any intervention being substituted for the specific intervention we are interested in. We would also, where different sizes of commodities are possible, recommend that WTP values are obtained for each rather than inferred and that these values be shown to accord with theory (see validity section later).

10.2 Health Care WTP Studies

There was no assessment made of an ‘embedment’ effect.
11 Presentation of Results

The results of CV studies may be presented in a number of ways, such as individual WTP values, by category, with a range, as mean, median and modal values, in tabular and graphical form, and in conjunction with costs. There are two issues of relevance here: (i) the physical presentation of the WTP data; and (ii) how it is combined with cost to assess a full CBA.

First, the physical presentation of WTP data. As alluded to, the more information which is presented the better. Presenting raw WTP data, ranges, median and mean values, quartiles, confidence intervals etc allows one to assess issues such as protest bids, skewed data etc.

Second, how the (monetary) benefits and costs are to be accounted for in a CBA is an important issue. Results may generally be expressed in one of two ways: either as Net Present Value (NPV) (benefit minus cost) or as a Benefit-Cost Index (BCI) (benefit divided by cost) (Drummond, 1980). There is an important distinction between these two statistics, resulting from how benefits and costs are handled. Consider for example, the situation presented in below.

<table>
<thead>
<tr>
<th>Program costs</th>
<th>$100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program benefits (WTP)</td>
<td>$120</td>
</tr>
<tr>
<td>Program cost savings</td>
<td>$ 50</td>
</tr>
</tbody>
</table>

1 BCI: ($120+$50)/($100) = $170/$100 = 1.7
2 BCI: ($120)/($100-$50) = $120/$50 = 2.4
3 NPV: ($120+$50) - ($100) = $170 - $100 = $70
4 NPV: ($120) - ($100-$50) = $120 - $50 = $70

Here a program costs $100, and has benefits of $120 (calculated as WTP). However, there are also medical cost savings of $50. How could we present these figures in a CBA? If we were looking to assess a BCI we can choose to have the savings as a benefit, or to put them as negative costs (cost savings). Which we do will alter the results from being a ratio of 1.7 dollars of benefit for every dollar to 2.4.

However, in calculating NPV it makes no difference how these are calculated as the result is that we have a net dollar benefit of $70.
11.1 Recommendations

We would recommend that WTP studies present as much data as possible, including raw data if possible, but certainly mean, median and modal values, range of values, coefficient of skew, quartiles, confidence intervals and standard errors. Data should be presented in tabular and graphical form to ease comprehension. If a C-B figure is to be calculated, we suggest this be based on a NPV basis.

11.2 Health Care WTP Studies

No studies presented a BCI, and only 16 (30%) presented NPV figures (or presented cost data in addition to WTP to allow a NPV to be inferred), with 38 only presenting WTP values and no cost information. Of the nine explicit drug studies six reported WTP values only, and three NPV (of 15 implicit studies, 11 presented WTP values only and four NPV).

The profile of studies reporting WTP values or NPV followed what would be expected when compared with the primary purpose of such studies. That is, 19 of the 21 focussed primarily on methodology reported only WTP values, although 11 (of 18) focused on policy (and 8/15 focused on both) also only reported WTP values, and it could be arguable whether they are truly answering a policy issue if cost, and therefore NPV, cannot be presented. The majority of studies reported an assessment between the WTP derived and various ‘explanatory’ factors, such as age and income (41 studies (75%)), with the majority of these performing regression analyses (35 studies (65%)).
12 Conclusions and Recommendations

This review provides a comprehensive overview of the major issues involved in the analysis and interpretation of WTP data. It represents the growing consensus, or lack of it, where appropriate, and reviews the conduct of WTP studies in health care to date. The review is necessarily subjective however, and there may be disagreements on the authors conclusions, interpretations and recommendations.

One caveat to note, however, is that many of the issues discussed here with respect to WTP should be seen in the wider context of trying to assess how individuals might behave in practice, where there is not an observable situation. Such problems associated with hypothetical surveys must affect all techniques whether WTP, QALYS or some other technique of assessing value. It is therefore important to try to distinguish between problems which are due to the inherent situation of trying to explain human behaviour in a hypothetical model, against those problems unique to the WTP method, compared to those specific to QALYS for example. No method is perfect so the choice becomes which level of imperfection one is content to live with.

Although it is difficult to provide rigid recommendations for how WTP data should be analysed and interpreted, several key features do present themselves from our reading of the literature.

1 In the choice of welfare measure, we would recommend that (along with the growing consensus of others) WTP (representing compensating or equivalent variation) be used.

2 We would not recommend that WTP values, or results, are adjusted for income in any way. However, we would recommend that these results are disaggregated in presentation according to income group, as well as any other relevant subgroup.

3 There is some difficulty in establishing that WTP values per se are valid. As such we would recommend that all WTP values are treated with some scepticism. It is therefore recommended that WTP studies should always be comparative, as it is likely that the difference in WTP values expressed is (at least more) valid. In addition, all studies should undertake tests of construct validity as a minimum.

4 The summary statistic presented should be the mean, which should be substantiated by presentation of measures of precision and dispersion. For example, presentation of the range of values (particularly those of zero and of a high level to consider protest bids) and confidence levels. We would also recommend the reporting of values adjusted and unadjusted for any ‘outliers’.

5 Sample size has to be sufficient to provide statistically significant estimates of WTP. It is the researchers responsibility to ensure that their sample size is sufficient to achieve this.

6 A high response rate is clearly desirable, and in the absence of other guidelines, we would suggest that a level, such as 70% as proposed by the NOAA Panel, be set as the minimum response rate required.
In terms of using values obtained elsewhere, or in transforming values within a study, we recommend that WTP values should only ever be used and applied to the specific circumstances from which they were obtained and not used to value alternative sizes of the benefit, or different levels of risk or probability.

Scale effects and ‘embedding’ may be a potential issues, and as such we recommend that WTP studies are conducted in a comparative sense and individuals made aware of any close substitutes to avoid WTP for any intervention being substituted for the specific intervention we are interested in. We would also, where different sizes of commodities are possible, recommend that WTP values are obtained for each rather than inferred and that these values be shown to accord with theory.

The presentation of WTP results with cost data should be done on a NPV basis, and BCI’s should not be used.

Reviewed against these recommendations, it is clear that WTP studies conducted to date have performed reasonably in their analysis and interpretation of WTP data, although not sufficiently that we would advocate their use in determining priority-setting across program areas. There is an urgent need for a set of consistent guidelines concerning the construction of a contingent market to be debated and agreed upon if such studies are going to become a useful addition to the economic evaluation ‘toolkit’.
REFERENCES


Donaldson C, Thomas R, Torgerson DJ 1995, Open-ended versus payment scale approaches to eliciting willingness to pay, Discussion paper 01/95, Health Economics Research Unit, Departments of Public Health and Economics, University of Aberdeen, Aberdeen, Scotland.


Garbacz C & Thayer MA 1983 'An experiment in valuing senior companion program services', *Journal of Human Resources*, vol 18, no 1, pp 147-153.


Stoker TM 1993, 'Empirical approaches to the problem of aggregation over individuals', Journal of Economic Literature, vol 31, pp 141-149.


Viscusi WK 1993, 'The value of risks to life and health', *Journal of Economic Literature*, vol 1, no 4, pp 1912-19??.

APPENDIX 1

The Complete Reference List of Papers Reviewed


