Willingness to Pay:
A Valid and Reliable Measure of Health State Preference?

BERNIE O'BRIEN, PhD, JOSE LUIS VIRAMONTES, MD, MSc

The development of methods to measure willingness to pay (WTP) has renewed interest in cost–benefit analysis (CBA) for the economic evaluation of health care programs. The authors studied the construct validity and test–retest reliability of WTP as a measure of health state preferences in a survey of 102 persons (mean age 62 years; 54% male) who had chronic lung disease (forced expiratory volume <70%). Interview measurements included self-reported symptoms, the oxygen-cost diagram for dyspnea, Short-Form 36 for general health status, rating scale and standard gamble for value and utility of current health state relative to death and healthy lung functioning, and WTP for a hypothetical intervention offering a 99% chance of healthy lung functioning and a 1% chance of death. WTP was elicited by a simple bidding game. To test for starting-point bias, the respondents were randomly assigned to one of five starting bids. All health status and preference measurements except WTP (controlling for income) showed significant (p < 0.05) differences between disease-severity groups (mild/moderate/severe). WTP was significantly (p = 0.01) associated with household income, but other health status and preference measures were not. The measure most highly correlated with WTP was standard gamble (r = -0.46). There was no association between starting bid and mean WTP adjusted for income and health status. The test–retest reliability of WTP was acceptable (r = 0.66) but lower than that for the standard gamble (r = 0.82). It is concluded that: 1) large variation in WTP responses may compromise this measure’s discriminant validity; 2) there is some evidence of convergent validity for WTP with preferences measured by standard gamble; 3) there was no evidence of starting point bias; 4) the test–retest reliability of WTP is comparable to those of other preference measures. Key words: willingness to pay; health state preferences; economics.

There are a variety of methods for the economic evaluation of health care programs. The main distinguishing feature between methods is the way in which health improvements are measured and valued. A method that is enjoying renewed interest is cost–benefit analysis (CBA), in which health benefits are valued in monetary terms. CBA has theoretical appeal to the economist because of its foundation in welfare economic theory, in contrast to techniques such as cost–effectiveness analysis where the underlying theoretical base is unclear. CBA data may also be appealing to decision makers because they quantify costs and benefits in monetary units, thus permitting the net benefit (i.e., benefit minus cost) of a program to be calculated to determine whether it is worth implementing.

Enthusiasm for the theoretical advantages of CBA must be tempered with an appreciation of the practical difficulties inherent in placing money values on health outcomes. A convention in early CBA studies of health care was the human-capital approach, where the values of population health gains or losses were computed in terms of production gains or losses, usually proxied by discounted future earnings streams for individuals in or out of employment as a consequence of their health status. In a landmark paper, Mishan argued that the human-capital approach in CBA studies was flawed for two reasons: 1) the method assumes that the primary goal of society is to maximize the gross national product, and this is highly questionable; 2) the method is inconsistent with the theoretical foundations of CBA, which are based upon a compensation test (the Potential Pareto Improvement criterion). As pointed out by Mishan and explained by Gafni, the practical extension of the theory underlying CBA is the estimation of individuals’ maximum willingness-to-pay (WTP) to secure implementation of a program (compensating variation) or the minimum individuals would need to be compensated (willingness-to-accept) to forgo a program (equivalent variation).
Two general approaches to the estimation of WTP values—indirect measurement and direct measurement—can be distinguished. The indirect approach examines previous real-world decisions that involve tradeoffs between money and expected health outcomes; for example, implied dollar values by wage premiums accepted by workers in occupations with known increased health risks. In contrast to inferring preferences from actual choices, the direct measurement of WTP uses survey methods to elicit stated dollar values for some non-marketed phenomenon produced or destroyed by the project being evaluated. In economics this second approach has been termed contingent valuation, because the respondent is being asked to consider the contingency of a market’s existing for the thing being valued. A number of CBA studies in environmental and transport economics have employed contingent-valuation techniques to value non-marketed phenomena such as improved air quality. Contingent valuation studies are also becoming more widespread in health care and have been undertaken in areas such as arthritis management, ultrasonography, care of the elderly, management of hypertension, blood transfusion, and the use of ionic versus non-ionic contrast media.

In implementing contingent valuation in health care, there is debate about who should be asked (e.g., current users of a program, ex-users, potential users); what they should be asked (e.g., willingness to pay for, or accept, certain or expected outcomes); and how they should be asked (e.g., single open-ended questions or multiple close-ended questions in a “bidding game”). As indicated by Froberg and Kane in their comprehensive review of health preference measures, very little is known about the psychometric properties of stated willingness-to-pay measures in health care. Available data are largely limited to two studies by Thompson et al. of arthritis patients, where test–retest reliability was found to be low (r = 0.25) and there was indirect evidence of a lack of agreement between WTP responses and health-state utility as measured by standard gamble. In their review, Froberg and Kane did not find any health care study in which WTP had been directly compared with other measures of health-state preference.

The purpose of the present study was to evaluate one approach to contingent valuation in terms of construct validity and test–retest reliability. The therapeutic context of the study was chronic lung disease (chronic bronchitis, emphysema, or asthma) and questions related to current health status and preferences for hypothetical therapies and outcomes. We sought to address four questions regarding WTP:

1. What is the association between a person’s current disease severity and his or her willingness to pay for a therapy that offers a specified expected health improvement? As a test of construct validity we predicted that, for a given income status, persons with greater disease severity will be willing to pay more for the same expected health improvement.

2. What are the associations between WTP for health improvement and other measures of health-state preference such as the standard gamble? What is the association between WTP and a generic measure of current health-related quality of life (Short-Form 36)? Associations between WTP and such health status indicators would be evidence of the convergent validity of WTP.

3. Does the use of a simple “bidding game” to measure WTP introduce a starting-point bias?

4. What is the test–retest reliability of WTP over a period of four weeks?

Methods

A convenience sample of patients who had chronic lung disease (defined as forced expiratory volume in 1 second [FEV₁] of less than 70% of predicted) who lived in the Hamilton area was identified from a respiratory clinic register held at McMaster University Medical Centre. Given that the interview was likely to be cognitively demanding, we excluded from survey all patients who were more than 70 years old, those unable to read or speak English, and those who had other health problems such as deafness or mental illness where understanding of the survey might have been a problem. Patients eligible for survey were approached by mail with the consent of, and a supporting letter for the study from, their physician. The patients were interviewed in their homes by a professional survey interviewer; the survey did not require any clinical examination. All patients gave written informed consent to be interviewed.

SYMPTOMS AND DISEASE SEVERITY

Self-reported symptom data were collected for dyspnea, cough, phlegm, and wheeze using the modified scales proposed by the Medical Research Council (MRC) of Great Britain. For dyspnea this questionnaire comprises a five-point scale based on degrees of physical activity that produce breathlessness. The MRC questions for cough and phlegm record symptom frequency on a four-point scale (none, mild, moderate, severe), and that for wheeze records frequency on a three-point scale (none, occasional, most days). We combined the four symptom responses into a simple classification of disease severity of mild, moderate, and severe. We defined severe disease as at least one of the four symptoms at its highest level (i.e., most limiting or frequent). Mild disease was defined as all four symptoms reported as mild. Moderate disease in-
cluded all conditions not classified as severe or mild (i.e., mixes of mild and moderate symptoms).

**OXYGEN-COST DIAGRAM (OCD)**

The OCD is a visual-analog scale for measuring dyspnea and disability in patients with respiratory disease. The OCD is presented as a vertical line with 13 everyday activities ranging from "brisk walking uphill" to "sleeping" marked against the line at intervals corresponding to their oxygen costs (i.e., metabolic equivalents required to carry them out). Respondents are asked to make a mark on the line corresponding to the most strenuous activity they could undertake before they became breathless. Responses are recorded in millimeters from the bottom of the 10-cm scale, with low scores indicating a higher propensity towards breathlessness for a given activity level. For example, a person who becomes breathless sitting or standing would score around 10 out of 100, while a person who becomes breathless only with "medium walking uphill" would score around 80 out of 100.

**SHORT FORM 36 (SF-36)**

The SF-36 is a generic questionnaire for measuring health-related quality of life in eight domains: general health perception; physical functioning; role limitations due to physical health problems; role limitations because of emotional problems; social functioning; bodily pain; vitality (i.e., energy); general mental health. These eight domains comprise a total of 36 items. For each domain, item scores are summed and transformed onto a scale from 0 (worst) to 100 (best). SF-36 has been developed over a number of years from the Medical Outcomes Study and has undergone validity and reliability testing in a number of populations, including patients with chronic airways obstruction. Its breadth of coverage and ease of use have made it a popular measure of health-related quality of life in treatment-evaluation studies.

**HEALTH-STATE VALUES AND UTILITIES**

The preference value of each respondent's current health state in the interval from "death" (= 0) to "healthy lung functioning" (= 100) was elicited by two different methods: rating scale and standard gamble.

**Rating scale.** As described by Feeny and Torrance, our rating scale is a vertical and calibrated visual-analog scale (sometimes called a "feeling thermometer") with labeled anchors of "death" (= 0) and "healthy lung functioning" (= 100). Consistent with the guidance given in Furlong et al., the respondents were first asked to use this scale to rate two marker health-state descriptions for chronic lung disease (one good and one bad); they were then asked to rate their own current health states on the scale. Following terminology in the decision sciences, we refer to this measure as a value rather than a utility because the latter is measured under conditions of uncertainty.

**Standard gamble.** This is the classical method from economics and decision theory for measuring preferences for uncertain outcomes; the method stems directly from the continuity axiom of the theory of decision making under uncertainty proposed by von Neumann and Morgenstern. The respondents were asked to compare two hypothetical therapy options: option Y is a certain outcome (the status quo) where the person remains in the current health state for the rest of his or her life; option X is to take a hypothetical new medication that has an uncertain binary outcome—it will either return the person to healthy lung functioning (with probability p) or result in immediate death (with probability 1 - p).

Using colored pie charts with percentage slices as visual aids for probability concepts, the interviewers systematically varied the probabilities in option X to find the risk of death where the respondents would regard options X and Y as being equivalent. Consistent with the theory underlying this measurement, the observed indifference probability (in this case expressed as a percentage) is the utility value of the respondent's current health state in the interval from death (= 0) to healthy lung functioning (= 100). The intuitive logic of the standard gamble is that the extent to which a person will accept the risk of a worse outcome (i.e., death) to achieve a better outcome provides information about the individual's utility for his or her current health state.

**WILLINGNESS TO PAY (WTP)**

Given uncertainty about the health outcomes associated with a procedure, it has been argued that WTP questions should be asked about expected health outcomes. Therefore, the WTP questions were designed to flow from the previous questions based on the standard gamble and utilized the same scenario of a hypothetical medication, using the same construct of probabilistic binary outcomes. For the WTP questions the outcome probabilities were fixed at 99% chance of healthy lung functioning and 1% chance of immediate death. The respondents were told that their health insurance did not cover this medicine and they would be required to pay some amount out-of-pocket (see the appendix for verbatim questions). To elicit the maximum each individual would be willing to pay to secure this medication, we employed a simple "bidding game" method.

As illustrated in figure 1, for a risk-averse individual with a concave utility function over lung functioning, the utility value of the (99% healthy, 1% dead) lottery for WTP is 0.99. For each individual we predict that willingness to pay for this lottery will be negatively
correlated with the utility value of his or her current health status (p* by the previous standard gamble) and, by extension, with the gain in utility associated with the lottery.

**BIDDING GAME**

A technique commonly employed in contingent-valuation studies is a simple bidding game where a respondent is bid up or down by the interviewer in an iterative fashion to converge upon his or her maximum WTP. An advantage of the bidding game is that it requires only yes/no responses to each bid and thus has more market realism than single open-ended questions asking respondents for their maximum WTPs. An important disadvantage of the bidding game is the threat of starting-point bias, where the respondent’s final WTP value is not independent of the first bid prompted by the interviewer. Our study provided an opportunity to test the hypothesis of starting-point bias. Accordingly, the patients were randomly assigned to one of five starting bids ($10, $25, $50, $75, $100). For each starting bid the interviewers followed predetermined bidding algorithms; the bidding algorithm for the $50 starting bid is presented in figure 2. Mindful not to frustrate or bore respondents with a lengthy bidding game, we designed all the algorithms to have no more than three bids.

**STATISTICAL METHODS**

Means and standard deviations (SDs) are reported for all continuous data. Given that household income (ability to pay) is expected to be associated with WTP, our tests of hypotheses need to allow for such confounding. Accordingly we undertook analysis of covariance (ANCOVA) for two of our main hypotheses: variation in measurement scores by disease severity as the main effect (three groups) with household income as a covariate; variation in mean WTP by starting bid (five groups) with household income and health status (SF-36 domain of Health Perception) as covariates. Associations between measures are analyzed by Pearson product–moment correlation. Test–retest reliability for instruments is reported by intra-class correlation coefficient.

**Results**

From a sample of 133 names of eligible subjects with chronic airways limitation identified from the clinic register, a total of 102 interviews were conducted. Reasons for not interviewing were 1) that subjects could not be contacted (16/31); 2) that subjects refused to participate (13/31); 3) exclusion due to language or hearing problems (3/31). Details of respondent characteristics are presented in table 1. Fifty-three percent of the subjects reported that they had been diagnosed by a physician or having chronic bronchitis; 45%, emphysema; and 53%, asthma; more than half of the sample (57%) reported more than one lung disorder. The most frequent symptom reported (at the level of mod-
erate to severe) was phlegm (53%), followed by cough (50%), wheeze (36%), and dyspnea (31%). Based on symptom severity, the respondents where classified as having mild (21/102), moderate (22/102), or severe (59/102) disease.

FEASIBILITY AND COMPLIANCE

The interviewers rated the subjects' compliance and understanding of the interview as generally good; the WTP questions had a response rate of 94% (96/102). The question with the greatest number of refusals was household income (10/102 refused). No interview was terminated prematurely by a respondent. The mean duration of the interviews was 50 minutes (SD 13).

RELATIONSHIPS WITH DISEASE SEVERITY

Table 2 presents the measurements for all the respondents and also compares the responses by subgroups of mild, moderate, and severe disease; all means are adjusted for income. The mean response on the oxygen-cost diagram (OCD) was 53 mm (SD 19), which indicates that the average respondent became breathless at an activity level of "medium walking." There was a clear and significant gradient for OCDS between grades of disease severity in the expected direction.

On all eight domains of Short Form 36, the mean scores for the combined sample were lower (indicating poorer health status) than published reference scores for healthy persons from the general public. In six of the eight domains there was a consistent gradient, in the expected direction, between disease-severity groups; these differences were statistically significant for health perceptions, physical functioning, physical role, and energy.

In the health-state–preference interval from 0 (death) to 100 (healthy lung functioning), the standard gamble

Table 2  •  Mean (SD) Health-related Quality of Life, Utility, and Willingness-to-pay (WTP) Responses: ANCOVA by Disease Severity Controlling for Income

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Total Sample</th>
<th>Mild</th>
<th>Moderate</th>
<th>Severe</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 102*</td>
<td>n = 21</td>
<td>n = 22</td>
<td>n = 59</td>
<td></td>
</tr>
<tr>
<td>Oxygen-cost diagram</td>
<td>53 (19)</td>
<td>69 (17)</td>
<td>57 (14)</td>
<td>46 (18)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Short Form 36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health perception</td>
<td>41 (25)</td>
<td>55 (26)</td>
<td>49 (25)</td>
<td>33 (21)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>48 (26)</td>
<td>70 (18)</td>
<td>52 (22)</td>
<td>39 (24)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Physical role</td>
<td>52 (41)</td>
<td>74 (35)</td>
<td>52 (41)</td>
<td>44 (40)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Emotional role</td>
<td>75 (40)</td>
<td>75 (41)</td>
<td>76 (36)</td>
<td>73 (41)</td>
<td>0.09</td>
</tr>
<tr>
<td>Social functioning</td>
<td>70 (33)</td>
<td>81 (31)</td>
<td>73 (32)</td>
<td>65 (34)</td>
<td>0.2</td>
</tr>
<tr>
<td>Pain</td>
<td>68 (28)</td>
<td>76 (23)</td>
<td>69 (30)</td>
<td>65 (29)</td>
<td>0.3</td>
</tr>
<tr>
<td>Energy</td>
<td>45 (22)</td>
<td>58 (20)</td>
<td>48 (22)</td>
<td>39 (21)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Mental health</td>
<td>72 (23)</td>
<td>70 (24)</td>
<td>79 (21)</td>
<td>70 (23)</td>
<td>0.03</td>
</tr>
<tr>
<td>Health state utility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating scale</td>
<td>63 (23)</td>
<td>76 (13)</td>
<td>69 (18)</td>
<td>50 (25)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Standard gamble</td>
<td>83 (19)</td>
<td>91 (8)</td>
<td>87 (12)</td>
<td>78 (23)</td>
<td>0.02</td>
</tr>
<tr>
<td>Willingness-to-pay (mean)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per month</td>
<td>$113 (205)</td>
<td>$56 (112)</td>
<td>$63 (109)</td>
<td>$150 (249)</td>
<td>0.09</td>
</tr>
<tr>
<td>As % of income</td>
<td>5%</td>
<td>3%</td>
<td>3%</td>
<td>6%</td>
<td>0.22</td>
</tr>
</tbody>
</table>

* WTP was missing for six respondents, reducing the sample to 96 for this item.
Table 3 – Mean (SD) Health-related Quality of Life, Utility, and (WTP) Responses: ANOVA by Household Income*

<table>
<thead>
<tr>
<th>Instrument</th>
<th>&lt;$20,000 (n = 35)</th>
<th>$20,000 to $39,999 (n = 35)</th>
<th>$40,000 to $59,999 (n = 15)</th>
<th>&gt; $60,000 (n = 7)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen-cost diagram</td>
<td>54 (18)</td>
<td>48 (18)</td>
<td>58 (17)</td>
<td>54 (26)</td>
<td>0.43</td>
</tr>
<tr>
<td>Short Form 36 (overall score)</td>
<td>57 (33)</td>
<td>63 (32)</td>
<td>63 (14)</td>
<td>77 (7)</td>
<td>0.09</td>
</tr>
<tr>
<td>Health state utility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating scale</td>
<td>82 (19)</td>
<td>81 (23)</td>
<td>89 (14)</td>
<td>85 (15)</td>
<td>0.57</td>
</tr>
<tr>
<td>Standard gamble</td>
<td>60 (23)</td>
<td>63 (22)</td>
<td>68 (21)</td>
<td>76 (20)</td>
<td>0.29</td>
</tr>
<tr>
<td>Willingness-to-pay (unadjusted mean per month)</td>
<td>$34 (31)</td>
<td>$165 (234)</td>
<td>$102 (136)</td>
<td>$343 (450)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

* Ten respondents refused to answer the income question.

The method generated a higher preference score for the respondent's current health state (utility = 83) than did the rating scale (value = 63). However, both methods showed clear and statistically significant gradients between disease-severity subgroups in the predicted direction. For example, persons who had mild disease stated that, on average, the maximum risk of death they would accept from therapy that relieved them of their lung disease would be 9% (utility = 0.91), compared with those in the severe-disease group, who would accept up to 22% risk of death (utility = 0.78).

The mean WTP for a therapy that offered a 99% chance of healthy lung functioning but a 1% risk of death for all respondents was $113 (SD 205) per month; the median WTP was $65. Expressed as a percentage of respondents' household incomes, the mean WTP was 5%. The large difference between the mean and median values indicates a skewed distribution and the large standard deviation for mean WTP indicates a wide spread of responses. When WTP was analyzed by respondent disease severity there appeared to be some gradient, as predicted, but it was not significant (p = 0.09).

**RELATIONSHIPS WITH INCOME**

Table 3 presents ANOVA by four income groups and indicates no obvious difference or trend with respect to any of the health measures. In contrast, there appeared to be a marked and significant gradient, in the predicted direction, between income and unadjusted mean WTP.

**ASSOCIATIONS BETWEEN MEASURES**

Pearson correlations between instruments are presented in table 4. The strongest and most consistent agreement appeared to be that between OCD and SF-36 domains (range r = 0.27 to r = 0.65). Preference value for current health as measured by rating scale had a greater association with OCD and SF-36 than did health state utility by standard gamble. On the other hand, the standard gamble was the method that correlated most highly with WTP (r = -0.46), with higher WTP associated with lower current health state. Associations between WTP and SF-36 were small and ambiguous (with some unpredicted signs to coefficients).

**STARTING-POINT BIAS**

Table 5 presents data on final maximum WTP analyzed by starting bid. There appeared to be no obvious trend or heterogeneity between groups defined by starting bid when the final bids were analyzed as mean responses adjusted for income and health status. When the medians were computed for the individual groups there was a suggestion of the predicted trend, but this did not reach conventional statistical significance (p = 0.07).
TEST–RETEST RELIABILITY

Table 6 presents the test–retest reliability data for the 20 respondents who were interviewed a second time, approximately four weeks after their first interviews. Intraclass correlations were acceptable for all the instruments, including WTP ($r = 0.66$), which was better than the rating scale ($r = 0.61$) but worse than the standard gamble ($r = 0.82$).

MULTIVARIATE MODELS

Variation in WTP was also analyzed by multiple linear-regression models to determine whether other factors (e.g., age, sex, education) were confounding the univariate relationships explored in the main analysis. These models (not reported here) confirmed that the main association was between WTP and income and that very little of the variation in WTP, even when controlling for income, was explained by health status or symptom measures.

Discussion

We undertook this study being supportive of the principles underlying the general concept of WTP but having concerns about the practical measurement of WTP. In particular, we sought to apply to WTP principles of validity and reliability testing that have been used to assess the measurement properties of instruments for health-related quality of life.18–21 Responses to health-related questions are not verifiable because no “gold standard” exists, and this absence of a single criterion has led analysts to principles of construct validation. In some circumstances, for preference measures such as the standard gamble and WTP there may exist a simple test of response validity: observing whether the individual, if actually faced with the experimental choice, acts in a way consistent with what he or she stated in the survey. In the present study we were not able to compare stated and revealed preferences, so we attempted to explore the construct validity of stated WTP in two general ways: 1) does WTP appear to measure what we think it should be measuring, in an unbiased way, and 2) does WTP have convergence with other measures, such as the standard gamble, that measure similar constructs of health state preference?

Data from this study offer something to both WTP antagonists and WTP protagonists. The antagonists will conclude that, unlike other health preference and health-related quality-of-life measures, WTP did not offer good discrimination. Controlling for income, there was no significant association between severity of disease and willingness to pay for a given health improvement. Antagonists probably would also draw attention to the suggestion (for median responses) of starting-point bias as a threat to the validity of this method of interviewing. In summary, the case against WTP from these data is that the elicitation method may introduce bias and the resultant data have large variance such that the validity of the measure as a discriminative tool is compromised. Furthermore, these cross-sectional data can tell us nothing about the longitudinal issue of an instrument’s ability to detect change (responsiveness).

Protagonists for WTP probably would emphasize different aspects of this study and the data. In terms of discrimination between categories of disease severity, the “severe” group had a WTP that was double (in absolute $ and percentage-income terms) that of the “mild” and “moderate” groups; but this difference was not significant, the standard deviation ($249) being twice the mean ($141) for “severe” disease. This non-

Table 6 — Test–Re test Reliabilities (Four-week Interval) of the Instruments for a Subsample of Respondents ($n = 20$)*

<table>
<thead>
<tr>
<th>Instrument</th>
<th>First Interview</th>
<th>Second Interview</th>
<th>Intraclass Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen-cost diagram</td>
<td>60 (12)</td>
<td>56 (16)</td>
<td>0.66</td>
</tr>
<tr>
<td>Short Form 36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health perception</td>
<td>52 (24)</td>
<td>49 (20)</td>
<td>0.88</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>58 (21)</td>
<td>52 (22)</td>
<td>0.85</td>
</tr>
<tr>
<td>Physical role</td>
<td>49 (44)</td>
<td>58 (48)</td>
<td>0.59</td>
</tr>
<tr>
<td>Emotional role</td>
<td>84 (34)</td>
<td>89 (27)</td>
<td>0.87</td>
</tr>
<tr>
<td>Social functioning</td>
<td>84 (26)</td>
<td>86 (16)</td>
<td>0.49</td>
</tr>
<tr>
<td>PAin</td>
<td>70 (22)</td>
<td>72 (25)</td>
<td>0.66</td>
</tr>
<tr>
<td>Energy</td>
<td>56 (20)</td>
<td>56 (22)</td>
<td>0.86</td>
</tr>
<tr>
<td>Mental health</td>
<td>79 (19)</td>
<td>83 (16)</td>
<td>0.91</td>
</tr>
<tr>
<td>Health state utility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating scale</td>
<td>70 (19)</td>
<td>71 (21)</td>
<td>0.61</td>
</tr>
<tr>
<td>Standard gamble</td>
<td>89 (11)</td>
<td>88 (11)</td>
<td>0.82</td>
</tr>
<tr>
<td>Willingness-to-pay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mean per month)</td>
<td>$129 (241)</td>
<td>$89 (117)</td>
<td>0.66</td>
</tr>
</tbody>
</table>

* All correlations were significant at the 5% level.
significant finding might lead the WTP protagonist to criticize two aspects of study design: 1) the sample size may have been too small to detect WTP differences between groups because of its variability; 2) the experimental part of the survey, with its random assignment to starting bids, may actually have induced greater variation in the WTP data than would otherwise have existed. The WTP protagonist would also point out that the test-retest reliability of WTP was better than that of the rating scale for health utility and as good as that of a popular disease-specific scale, the oxygen-cost diagram.

A person’s willingness to pay is clearly limited by his or her ability to pay; this is appropriate because the method is attempting to elicit what the person would be prepared to forgo from current (and future) consumption to achieve an expected health improvement. In this way WTP is similar to the standard gamble, where the value of a health state is measured in terms of what a person is prepared to forgo or trade off (i.e., risk of death). Our data show a clear positive association between household income and WTP; this was a predicted association, and therefore offers partial evidence that we are measuring the desired construct. The association between income and WTP is not a source of measurement bias but it does reflect an important equity assumption underlying cost–benefit analysis—that the current income distribution is acceptable. In practice, a CBA study using WTP data might incorporate explicit distributional weights for outcomes based upon income or other factors. From this study two practical issues arose with income: 1) non-response to income questions was relatively high at 10%; 2) the definition and measurement of household income are open to interpretation due to issues regarding the definition of the household and also the influence on WTP of accumulated wealth and assets, the distribution of which may not be identical to income.

There are a number of potential sources of measurement bias in WTP studies. We believe out study was the first to explore starting-point bias in WTP for health improvement, but our data are not conclusive; previous studies in environmental economics have detected such bias while others have not. Other sources of bias are more difficult to detect with hypothetical preference data in the absence of the “gold standard” of actual decisions. For example, a form of strategic bias would be a very high or low WTP bid as a protest against the interview. In his study of arthritis patients, Thompson decided to trim his data so that he might incorporate explicit distributional weights for outcomes based upon income or other factors. From this study two practical issues arose with income: 1) non-response to income questions was relatively high at 10%; 2) the definition and measurement of household income are open to interpretation due to issues regarding the definition of the household and also the influence on WTP of accumulated wealth and assets, the distribution of which may not be identical to income.

We suspect that a major difficulty and source of response heterogeneity with WTP for health improvements is what has been called hypothetical bias. Asking the respondent to conceive of a hypothetical market for health improvements is, at best, cognitively demanding. The mechanism of the bidding game has more market realism than earlier health studies that used one-shot open-ended questions about the maximum a person would pay. But we believe our respondents, in the environment of universal health insurance in Canada, found it difficult and perhaps unrealistic to focus on what they would pay for an expected health improvement of such magnitude. Further research on methods for explaining the measurement task and presenting additional data (e.g., current per-capita tax expenditures on health care items) may be useful. However, we suspect that WTP methods will generate the most reliable data in studies of minor diseases where the respondents already have some familiarity with consumer purchases such as over-the-counter medicines, for example, in the treatment of allergies or migraines.

Our discussion of the measurement issues arising from this study leads us to the broader consideration of the theoretical validity of the measurement we undertook. Recently Gafni has argued that the relevant respondents for WTP studies in publicly funded health care systems (such as Canada’s) are the tax-paying general public, with the relevant question being how much they would pledge as an additional insurance premium to their taxes so that a therapy would be available if they ever needed it.

Even in a predominately private market insurance-based system such as that of the United States, the relevant dollar–health tradeoffs can be elicited by offering individuals a choice between hypothetical health maintenance organizations (HMOs) that offer different annual premiums but alternative packages of health care and health benefits. This tax-premium–insurance-premium approach has considerable conceptual merit because it could capture the important aspect of WTP for benefit to others (externalities)—something usually omitted from economic evaluation studies. But the measurement challenges are likely to be large because respondents would be asked to value probabilistic health outcomes associated with therapies and diseases with which they may not be familiar. We speculate that such public surveys will generate data even more heterogeneous than those we obtained in our study, which may limit their usefulness for policy making; but ultimately this is an empirical question. We conclude that WTP still holds conceptual promise but that the empirical agenda before it can become recognized as a valid and reliable measurement tool will be large.
The authors are grateful to the study interviewers, Joyce Brown, June Hogg, and Diane Woikowski; and to the physicians and patients who agreed to participate. Valuable comments on an earlier draft of this paper were received from Stephen Birch, David Feeney, George Torrance, Chris Woodward, Greg Stoddart, and other members of the Polinomics Seminar group.

References


APPENDIX

Willingness-to-Pay Scenario

Now let’s assume that choice Y offers a 99% chance of restoring you to healthy lung functioning, and there is only a 1% chance of immediate death. That is, of 100 people who take the medicine in choice Y, 99 will have healthy lung functioning restored and one person will die.

Now assume that the medication in choice Y is expensive and is not fully covered by your health insurance, so you are required to pay some amount out of pocket, each month and for the rest of your life, for this medication. Thinking about the value of this medication to you and how much you could, realistically, afford to pay each month: would the maximum amount you would be willing-to-pay be bidding game begins