Evaluating the validity of stated-preference estimates of health values

F. Reed Johnson* and Matthew F. Bingham*

1. INTRODUCTION

The increasing importance of a demand perspective in health care also has increased the need for reliable monetary measures of value to patients. Obtaining credible measures of the economic value of morbidity, however, is one of the more difficult problems facing health economists. The existence of insurance, universal health-care systems, and market participants that are unrepresentative of the population of policy interest often obscure essential supply-and-demand relationships. Thus, revealed-preference information generally has proven to be an insufficient basis for obtaining policy-relevant values of human health states.

As a result, health economists have applied contingent-valuation (CV) methods to a number of health-outcomes valuation problems. However, there are well-documented problems with CV methodology (Hausman, 1993), and CV is not well-suited to valuing multiple health-state attributes. For these reasons, health economists increasingly are turning to stated-preference (SP) or conjoint-analysis approaches as an alternative.1 Trade-off based SP methods currently are the most powerful techniques available for quantifying patient preferences and satisfaction. In 1999, half of the published willingness-to-pay (WTP) studies used SP methods.

A useful valuation methodology should have several characteristics, including the ability to
- adapt task to relevant decision context
- provide opportunity for subject learning
- maximize information obtained for each subject
- support consistency and rationality tests of validity.

* Triangle Economic Research, Durham, USA

1. Terminology has not been standardized among various disciplines. We use the term “stated preference” to refer to a group of techniques used primarily in market-research studies to measure consumer preferences. The term “conjoint analysis” also has been used to describe some of these techniques. Although CV also could be called an SP technique, CV was developed independently by environmental economists and generally relies on a different set of elicitation formats and analytical approaches.
This paper specifically evaluates how consistency and rationality can provide tests of the validity of SP estimates for health valuation research.

SP estimates can provide both relative utility weights and monetary value-to-patient (VTP) measures. These estimates have obvious use in benefit-cost analysis of treatment options or public health policies. In addition, SP studies provide useful information for several areas of pharmaceutical product development and marketing, including pricing decisions, formulary approval, promotional programs, and allocating research resources for new pharmaceuticals.

An important advantage of SP is that it can provide a value for the commodity as a whole as well as for the individual attributes of the commodity. In effect, SP produces a total value for a given therapy as well as values for individual therapy attributes. SP recognizes that commodities have value because of their characteristics or attributes. People have preferences for each attribute and are willing to accept trade-offs among them. SP analysis examines these trade-offs to assess the weight people assign to various product attributes. The credibility of SP techniques for analyzing demand for market goods and services is well-documented (Wittink and Cattin, 1989). However, more recently, analysts have employed SP to quantify preferences for a variety of nonmarket goods and services. These include medical interventions, pharmaceutical treatments, and environmental health risks (Bryan et al. 1988; Johnson et al. 2000; Johnson et al. 1998; Johnson and Desvouges 1997; Ryan and Hughes, 1997a; and Viscusi, Magat and Huber, 1991). Despite the promise of using SP methods in this manner, quantifying nonmarket preferences using SP techniques raises a number of important questions regarding validity and reliability.

The most important question involves the inherently hypothetical nature of all SP surveys. Subjects do not have to commit real resources to their stated choices. Moreover, when subjects consider trade-offs among non-market as opposed to market goods, it is less likely that their answers express a set of defined, pre-existing preferences. When the survey choices available do not exist in typical markets, respondents must make decisions that are quite unlike those they face in everyday life. Follow-up studies of actual market behavior allow verifying the validity of SP studies based on market goods. However, conducting similar studies of revealed behavior with respect to non-market goods is difficult. In particular, price and information distortions in medical markets invalidate such comparisons.

Although health-related SP studies often cannot be directly validated in the standard fashion, researchers can increase the credibility of SP estimates in several ways. For example, potential hypothetical bias associated with health-valuation studies can be mitigated through thoughtful study design and careful data analysis by competent researchers. Skilled use of focus groups and survey pre-testing enhance the realism of the final instrument. Such efforts should lead to valid health-state preference measures. However, despite such efforts, the lack of behavioral data available for comparison with survey data calls for developing additional measures for assessing the performance of SP techniques in health-valuation applications.
SP encourages subjects to explore their preferences for various attribute combinations through a series of judgments. This process of explicitly trading off attributes encourages subject introspection. Non-market researchers in the stated preference arena currently take advantage of the repetitive, introspective nature of SP data by employing experimental designs that support a battery of statistical tests. These tests determine whether or not survey participants are responding in a rational, consistent, utility-theoretic manner.

Implementing a valid and reliable SP study requires accurate commodity definition (attributes and levels),\(^2\) attention to format selection (ratings, rankings, or choice), and efficient experimental design. The remainder of our discussion examines the validity implications of each of these aspects and provides examples from several recent health-outcomes valuation studies.

2. COMMODITY DEFINITION

Once identified, health states potentially associated with treatments must be defined in sufficient detail such that subjects can distinguish among them. In addition, to reduce hypothetical bias, these health-state definitions must be consistent with the ways that people think about their health. For instance, people often do not think of their health in terms of clinical outcome measures. Rather, they consider how the severity of symptoms associated with clinical outcomes limit or affect physical, social, and emotional functions. It is the job of survey developers to determine how subjects think about health outcomes for the intervention of interest and to identify salient attributes and levels.

Accurate commodity definition relies crucially on employing focus groups and careful pre-testing. However, researchers may rely on both clinical descriptions of relevant health states and existing health classification systems to provide an initial foundation for survey development. For example, a recent study (Johnson, Banzhaf and Desvouges, 2000) uses a simplified Quality of Well-Being (QWB) framework for defining health attributes. The QWB health-classification system defines health states in four dimensions: three function states (mobility, physical activity, and social activity) and the most severe symptom/problem complex. Instead of using the three function states as separate attributes, the authors combine them into a single attribute called “daily activity”.\(^3\) The authors then add the attributes of duration and annual subject costs to characterize the desired health-state.

As stated earlier, the additional hypothetical aspect of non-market SP studies can be mitigated through thoughtful study design by competent researchers. In the case of Johnson et al. (2000), the final survey instrument underwent extensive development,

---

2. An attribute is a qualitative characteristic of the commodity, while a level is one of several values the attribute may have. Color and price are attributes. Blue and $25 are levels.

3. The authors initially treated the QWB function states as separate attributes, but we found in pre-testing that subjects tended to view these function states as interdependent, which complicates the development of an orthogonal experimental design.
including two focus groups, three pretests, and a large-scale pilot test. These measures successively led to a final survey with clear, believable, trade-off scenarios. In addition, the survey incorporates several state-of-the-art features including a computerized format, an information treatment with quiz questions, and detailed health-history questions. The information treatment consists of a two-page article on heart and lung illnesses of concern in this study.\(^4\) The four quiz questions reinforce the key concepts presented in the article, such as the differences between acute and chronic conditions, and the notion of attacks or “episodes” associated with chronic illnesses.\(^5\) They also provide an empirical measure of how well the subjects understood and processed the information in the article. The health-history questions collect information on the subject’s personal health history, the health history of family members, and demographic information about the subject and the subject’s household. The information gathered is used in the analysis to determine whether health history and other demographic variables influence VTP for health. The relationship between variables such as income and VTP for improved health can be important indicators of a study’s validity. At a minimum, these variables should be included to help assess the validity of any SP study.

3. ALTERNATIVE SP FORMATS

SP question formats describe the attribute-level combinations for two or more health outcomes and ask the subject to perform a particular evaluation task involving these alternatives. Ranking, rating, and discrete-choice formats have been used in SP surveys. In a ranking study, subjects may be given cards, each showing a product profile. Subjects are asked to order these cards from most preferred to least preferred. There currently are no examples of the ranking format in the health-valuation literature. In a graded-pair comparison rating, subjects see pairs of commodity profiles and are asked to indicate how strongly they prefer one to the other. Viscusi, Magat and Huber (1991) used this approach to measure the value of avoiding an increase in the risk of contracting chronic bronchitis. Alternatively, discrete choice provides subjects with several different products or programs simultaneously and simply asks them to identify the most-preferred option in each choice set. Ryan and Hughes (1997b) used the discrete-choice format to value women’s preferences for miscarriage management.

The format of the SP question has several cognitive effects:\(^6\)

- Focuses attention on specific attributes
- Alters perceptions of substitute possibilities

\(^4\) These illnesses included minor lung illnesses, lung infections, chronic lung conditions, coronary-artery disease, heart failure, and heart-rhythm problems.

\(^5\) The computer format was an advantage in this part of the survey because if subjects answered the quiz question incorrectly, the computer provided them with the correct response.

\(^6\) See Huber (1997).
STATED-PREFERENCE ESTIMATES OF HEALTH VALUES

- Induces task-specific evaluations
- Involves learning and fatigue processes

Choosing an elicitation format thus is an important step in developing an SP survey. Huber et al. (1993) argue that each format has relative merits because each relies on a different cognitive process for eliciting preferences. Different formats cause subjects to focus on different evaluation strategies. Huber et al. cite a well-established result in marketing that the choice format is a better predictor of short-term market behavior in familiar, repeat-purchase situations, while graded-pairs is a better predictor of longer-term market adaptations to new products with relatively unfamiliar attributes. Health outcomes often have attributes of both kinds of situations in different degrees for different subjects. Furthermore, real-world behavior often involves a mixture of valuation strategies. Thus, there may be no a priori basis for choosing between graded pairs or choice format as the better measure of health preferences in a given context. Huber et al. have found that employing more than one elicitation format may predict behavior better than using one format alone, simply because each format requires different heuristics and may provide only a partial picture of preferences.

**Figure 1: Example of Graded-Pair Question**

<table>
<thead>
<tr>
<th>Category</th>
<th>Condition A</th>
<th>Condition B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of episode</td>
<td>5 days</td>
<td>1 day</td>
</tr>
<tr>
<td>Symptoms</td>
<td>Eye irritation (itching, burning, redness)</td>
<td>Pain in chest or arm</td>
</tr>
<tr>
<td>Daily activities</td>
<td>– CAN go to work, go to school, do housework</td>
<td>– CANNOT leave your house, go to work, go to school, do housework, and participate in social or recreational activities</td>
</tr>
<tr>
<td></td>
<td>– CANNOT participate in social or recreational activities</td>
<td>– Have SOME physical limitations</td>
</tr>
<tr>
<td></td>
<td>– Have SOME physical limitations</td>
<td>– CAN care for yourself</td>
</tr>
<tr>
<td></td>
<td>– CAN care for yourself</td>
<td></td>
</tr>
<tr>
<td>Total costs of this episode to your household</td>
<td>$50</td>
<td>$200</td>
</tr>
</tbody>
</table>

1 2 3 4 5 6 7

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A is much better</td>
<td>A is somewhat better</td>
<td>A is slightly better</td>
<td>A and B are about equal</td>
<td>B is slightly better</td>
<td>B is somewhat better</td>
<td>B is much better</td>
</tr>
</tbody>
</table>

Please press a number from 1 to 7 that best reflects your rating.

7. See also Huber (1997); Tversky, Sattath and Slovic (1988); Payne (1976, 1982); Huber and Klein (1991); and Huber et al. (1993) for discussion of the different cognitive processes involved in graded-pair versus choice tasks.
Figure 1 illustrates an example of a graded-pair question. In this example, the price is expressed as illness-related costs. The subjects indicate their preferences for Condition A versus Condition B. Because there is no unambiguous baseline reference point in the graded-pair questions and no health conditions with zero cost, the graded-pair questions are designed to obtain information on marginal trade-offs among health attributes and costs.

**Figure 2: Example of Discrete-Choice Question**

<table>
<thead>
<tr>
<th>Category</th>
<th>Initial Condition</th>
<th>Alternative A</th>
<th>Alternative B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of episode</td>
<td>5 days</td>
<td>1 day</td>
<td></td>
</tr>
<tr>
<td>Symptom</td>
<td>Shortness of breath and swelling in ankles and feet</td>
<td>Shortness of breath and swelling in ankles and feet</td>
<td></td>
</tr>
<tr>
<td>Daily activities</td>
<td>• Are in hospital</td>
<td>• CANNOT leave your house, go to work, go to school, do housework, or participate in social or recreational activities</td>
<td>Your level of health as it is today.</td>
</tr>
<tr>
<td></td>
<td>• Need help caring for yourself</td>
<td>• Have SOME physical limitations</td>
<td>You do not experience this episode.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Need help caring for yourself</td>
<td></td>
</tr>
<tr>
<td>Additional costs to your household</td>
<td>$0</td>
<td>$300</td>
<td>$500</td>
</tr>
</tbody>
</table>

1. Prefer Initial Condition  
2. Prefer Alternative A  
3. Prefer Alternative B

Please press a number from 1 to 3 that best reflects your choice if faced with these options.

Figure 2 illustrates the discrete-choice format used in the Johnson et al. (2000) study. The discrete-choice format directly elicits total values for movements from a given diminished health state to the subject’s current health state. The initial condition profile consists of a relatively severe hypothetical initial condition with zero cost. Alternatives A and B represent two courses a subject could choose if experiencing the initial condition. Alternative A portrays a condition of intermediate severity and intermediate cost. Alternative B is described as the subject’s current health-state on the day of the survey with a relatively high cost. In essence, subjects can choose to remain in a relatively severe condition and not pay any additional costs for treatments outside the government.

8. The payment vehicle for both formats is described as illness-related costs that are not covered by the government health system or a company insurance plan. These costs are associated with items that reduce discomfort or the length of illness (such as vitamins, medicines, air filters or humidifiers, special foods or liquids, or other optional treatments). These costs also may include such costs as child care while sick or transportation to the doctor. Subjects were instructed to assume that any missed time from work would be covered by paid sick leave.
health plan or insurance plan, or they can choose to pay for additional treatments to improve their health to their current health state.

Unlike the graded-pairs format, the conditions that subjects evaluate must be related to obtain meaningful values. Therefore, while this format has the advantage of obtaining values for changes in health relative to an identified reference point, it potentially has the disadvantage of reducing the salience of the symptom attribute if subjects focus on changes in other attributes. Furthermore, the structure of the choice experiment assumes that subjects can exchange money for specified health improvements. For most people in the general population in good health, the choice sets include two options (the Initial Condition and Alternative A) worse in health attributes and better in cost than current health, as intended. However, some subjects’ current health may be worse than one or both of the hypothetical alternatives. In such cases, at least one of the options would be preferred, or dominant, in all of the attributes, and thus choices would reveal less information about trade-off relations. This situation does not invalidate the design, but introduces some degree of statistical inefficiency.

Figure 3: Marginal Utility of Money Estimates by Question Sequence and Question Format

In Johnson et al.'s (2000) study, the authors apply both graded-pair and discrete-choice formats to estimate joint VTP estimates. Each subject evaluated eight graded pairs and eight choices. The sequence was randomized across respondents, so it was possible to estimate cross-section models for each question in the 16-question sequence. Figure 3 shows estimates of the marginal utility of money parameter. The two formats induce quite different cognitive responses with respect to the cost attribute. The graded-pairs estimates vary quadratically during the course of the elicitation, with a mode of about
In contrast, the choice estimates are nearly constant over the 8-question sequence at about 0.06.

4. EXPERIMENTAL DESIGN

Full-factorial experiments generate data based on all possible combinations of attribute levels. Such designs typically are impractical for SP surveys because subjects' cognitive limitations and time constraints do not allow consideration of a large number of profiles. For example, a full-factorial design of a treatment containing 5 attributes each with 3 levels leads to 243 (3^5) possible combinations. In addition, subjects do not rate these options individually. Rather, subjects compare two or more options at a time. Considered in pairs, the number of possible combinations is 29,403, clearly an impossible task.

Most current marketing SP applications use an approximately orthogonal design to reduce the number of paired comparisons to the smallest number necessary for efficient estimation of utility weights (DEY, 1985). HUBER and ZWERINA (1996) list four properties of efficient designs:

- Level balance: levels of an attribute occur with equal frequency
- Orthogonality: the occurrences of any two levels of different attributes are uncorrelated
- Minimal overlap: cases where attribute levels do not vary within a choice set should be minimized
- Utility balance: the probabilities of choosing alternatives within a choice set should be as similar as possible.

Unfortunately, it is often not possible to achieve both level balance and orthogonality in small designs. Thus, design optimality generally requires trading off potential incompatibilities between these criteria. However, KUHFELD, TOBIAS and GARRATT (1994) show that it is possible to produce relatively efficient designs that are neither balanced nor orthogonal. Such efficient designs can be produced using an iterative computer algorithm. Health economists have begun adapting these market-research procedures to the special requirements of health-outcomes valuation.

The experimental design obtains multiple responses from each subject. The observed patterns of these responses in the context of an appropriate design facilitates testing a variety of hypotheses about consistency with welfare theoretic principles, use of judgment heuristics, learning, and fatigue effects. Demonstrating that subjects generally adhere to utility-theoretic principles is important in establishing the validity of a particular SP study.

In particular, recovering valid welfare measures from SP data requires that subjects' preferences are complete, monotonic, and transitive. Completeness requires that for any two choices A and B, individuals can rank whether they prefer A to B, prefer B to A, or are indifferent between A and B. Monotonic preferences require that, holding
costs constant, individuals should prefer more to less of any normal good. For preferences to be transitive, if a subject prefers A to B and B to C, that individual also should prefer A to C. In addition, we expect subject preferences to be stable, at least within the experimental controls of the SP survey itself. For convenience, we will refer collectively to monotonicity, transitivity, and stability as preference consistency.

If subjects’ stated preferences violate welfare-theoretic principles, we cannot impute valid welfare values for these subjects, even if we can estimate models using those observations whose econometric performance appears acceptable. Moreover, correct point estimates are important in empirical work, so that even statistically insignificant biases may influence policy conclusions. Nevertheless, SP tasks are cognitively challenging. Even the most attentive subjects with well-behaved preferences may report some inconsistent responses. Thus, the challenge of applied valuation research is to evaluate when consistency failures are serious enough to invalidate the welfare-theoretic validity of a subject’s responses.

Depending on design, SP experiments may allow determining whether subject responses conform to the basic tenets of utility theory. There are two possible tests for monotonicity. The first is a dominant-pair comparison. This test requires that all the attributes of one profile in a choice set are unambiguously better than all the attributes of the other profile in the comparison. Including a dominant-pair comparison in an SP survey provides a simple test of subject consistency. However, including this simple test reduces overall design efficiency since a dominant pair provides no information on subjects’ willingness to accept trade-offs among attributes.

A second possible test of monotonicity involves comparing subjects’ ratings across two sets of graded pairs. This test requires that a subject see a particular profile at least twice. In addition, it requires that one of the profiles compared to the repeated profile is either unambiguously better or worse than the other comparison profile. For example, suppose that a subject sees two sets of pairs, Option X versus Option Y, and Option X versus Option Z. Further suppose that Options Y and Z are identical in all attributes but cost, and Option Z costs less. Given that a subject prefers Option Y to Option X in the first pair, that individual should prefer Option Z to Option X at least as strongly because Option Z provides the same attributes as Option Y, but at a lower cost.

It is not possible to construct valid welfare estimates for inconsistent subjects. However, if inconsistent subjects’ preferences are the same as consistent subjects’ preferences, then including data from inconsistent subjects merely adds noise to VTP point estimates. On the other hand, if the two groups’ preferences are dissimilar, then including inconsistent responses biases welfare estimates and undermines the validity of aggregate point estimates.

9. Such surveys typically do not allow testing for completeness. Rather, subjects must provide a rating or indicate a choice before being allowed to progress to the next question in the survey. Therefore, the survey basically forces completeness of preferences. Of course, requiring a subject to provide answers does not necessarily ensure that such answers are meaningful indicators of actual preferences.
In practice, the availability of economic consistency tests depends heavily on the amount of profile repetition within a series of rating pairs or choice sets seen by a given subject. Profile repetition is related to both the number of choice sets seen by each subject and the size of the full design. Thus, when SP experiments are not constructed specifically for consistency testing, those with comparatively small design spaces and long choice series will yield more opportunities for consistency testing. Designing SP experiments that specifically allow for consistency testing requires trading off statistical power of preference estimation for availability of information for validity testing.

In a study of electricity customers' WTP for so-called “green” electricity, Johnson and Desvousges (1997) estimate WTP for various categories of environmental benefits. The study includes two health-benefit categories, reduced incidence of respiratory disease and reduced incidence of cancer. Subjects were divided into two groups: those whose stated preferences were relatively monotonic and those whose preferences failed a number of monotonicity tests. Figure 4 compares WTP for the various commodities by group. In several cases, the estimated WTP is quite different between the two groups, although WTP for the two health commodities appears similar.

Figure 4: WTP by Commodity and Monotonicity Test of Consistency

Subjects' ratings of three sets of pairs may provide the ability to check for transitivity. Testing for transitivity requires that a subject see a particular profile at least twice. In addition, it requires that the profiles compared with the repeated profile are shown together in a paired comparison. For example, suppose a subject sees Option A versus Option B and Option B versus Option C. If so, a comparison of Option A versus Option C provides the ability to check transitivity.

10. More efficient designs typically do not repeat profile pairs. Thus, the dependence of consistency tests upon profile repetition limits the availability of such tests as design efficiency increases.
Also, if the experimental design permits, preference stability can be tested in either graded-pairs or choice formats. For example, questions at the beginning of the series can be repeated at the end. If questions are randomized over the sequence so that cross sections contain the complete experimental design, preferences can be estimated over cross sections of the sequence to test for stability of both preferences and variance.

A recent survey elicited evaluations of longevity and quality-of-life trade-offs from a sample of current smokers between the ages of 50 and 64 (JOHNSON, SMITH and SMITH, 2000). The experimental design includes three attributes: life extension, quality of life, and cost of a pack of 20 hypothetical cigarette filters that lower the health risks of smoking. Subjects saw six choice sets of three alternatives generated by an efficient choice experimental design algorithm by ZWERINA, HUBER and KUHFELD (1996). The third profile in each choice set allowed subjects to choose not to purchase a filter, giving them no life extension and no cost in addition to the cost of cigarettes. The quality-of-life attribute in this combination is male subjects' father's or female subjects' mother's quality of life in the three months prior to the parent's death or in the three months prior to the survey. In addition to these six repetitions, the survey included a dominated choice in the seventh repetition and repeated the first and second questions at the end of the sequence. The nine-repetition sequence was randomized across subjects.

We estimated VTP for a year of increased longevity using the repeated questions from the beginning and end of the SP question sequence. Figure 5 shows that VTP estimates are quite consistent, with differences in the two highest levels not being statistically significant.

![Figure 5: Older Smokers' VTP to Extend Life One Year at the Specified Quality of Life Level](image)

The simplest and most direct test of monotonicity is whether subjects are able to detect a case where one profile yields unambiguously higher utility than the other profile. The smoking survey included a dominated pair where all the attributes of one program are
better than all the attributes of the other program. The dominant pair was contrasted with the status quo alternative in the smoking survey. Of the 244 subjects, 30 selected the status quo alternative, providing no data on the dominance test. Among the remaining 214 subjects, only one incorrectly selected the dominated alternative.

Testing for transitivity requires manipulation of three sets of pairs for each survey respondent. The smoking design did not include enough repetition to construct a sufficient number of such tests. However, the smoking survey repeated the first two questions at the end of the sequence, allowing testing for consistency. Over 60 percent of the subjects provided consistent answers on both questions. Only 9 percent of respondents fail both checks.

The smoking study focuses on VTP for increased longevity in a given health state at the end of a subject’s life. The study defines subjects' reference health-state limitation as that experienced by their same-sex parent just prior to death, while the reference life extension is zero months beyond the gender-specific life expectancy. VTP estimates for an additional one-year life extension relative to the reference condition of being bedridden with no life extension beyond the life expectancy of the average subject are not significantly noisier at the beginning of the survey than at the end.

Figure 6 shows the effect of the linear time trend on VTP for each attribute. The net effect of time-trend significantly affects only the “Can Leave Home” and “Can Drive a Car” health states. Nevertheless, Figure 6 indicates that all but one of the point estimates is somewhat lower at the end of the question sequence than at the beginning, suggesting a possible dynamic cognitive process at work.

5. CONCLUSIONS

SP surveys can provide a demand perspective in health care. Unfortunately, evaluating trade-offs among health-outcome attributes may be an unfamiliar task for many subjects, and hence preferences often are not fully established. Instead, to some extent, preferences evolve as the experiment progresses as subjects have time to think about and react to the trade-offs with which they are faced. Multi-response, multi-attribute SP methods are well suited to capturing these dynamic and cross-sectional effects and providing meaningful measures of VTP.

Nevertheless, the validity of such surveys depends heavily upon the care and competence with which they are administered. In particular, the judicious use of focus groups and pre-testing can enhance the ultimate validity of SP estimates. Choice of an elicitation format also is an important decision. Because format triggers different cognitive responses, the format should simulate the characteristics of market choice as closely as possible. Where there is no clear market analog to the SP task, multiple formats may help ensure that estimates account for task-specific differences. Finally, careful construction of the experimental design can facilitate tests of whether subjects’ responses provide a valid basis for constructing welfare values.
Figure 6: VTP by Question Sequence and Quality of Life, Smoking Survey

REFERENCES


**SUMMARY**

Economists have long been faced with the challenge of valuing goods and services that are not traded in markets. A variety of multiple-response stated-preference (SP) methods, which evolved independently in the area of consumer market research, also have been applied to nonmarket-valuation problems such as health-state value measurements. Unfortunately, distortions of the essential supply-demand relationships in such markets make it difficult to verify the validity of SP surveys using market data. However, because SP data include multiple observations for each respondent, it is possible to test a variety of hypotheses about consistency with welfare-theoretic principles and the use of judgment heuristics and learning and fatigue effects. This paper provides
some results from two SP surveys eliciting health-state valuations. It focuses on the economic consistency of stated preferences. We show that with a few exceptions, subjects' stated preferences generally agree with welfare-theoretic requirements.

ZUSAMMENFASSUNG


RESUME

Les économistes sont depuis longtemps confrontés à la difficulté d’évaluer des biens et services qui ne sont pas échangés sur les marchés. Diverses méthodes de préférences exprimées à réponses multiples, développées dans le domaine du marketing, ont été appliquées à l’évaluation de biens non marchands tels que la santé. Malheureusement, les distorsions dans la relation offre – demande qui caractérisent ces marchés rendent difficile la vérification de la validité des enquêtes de préférences exprimées à partir des données du marché. Puisque les données recueillies à l’aide des méthodes de préférences exprimées contiennent plusieurs observations pour chaque enquêté, il est cependant possible de tester différentes hypothèses relatives à la cohérence des principes de la théorie économique du bien-être, à l’utilisation de jugements heuristiques, ainsi qu’aux effets d’apprentissage et de fatigue. Cet article fournit quelques résultats issus de deux enquêtes qui évaluent différents états de santé à l’aide des préférences exprimées. Il se concentre sur la validité économique des méthodes fondées sur les préférences exprimées. Nous montrons, qu’à quelques exceptions près, les préférences que les individus expriment sont cohérentes avec les postulats de la théorie économique du bien-être.