Chapter 7

MULTIPLE GOOD VALUATION

With Focus on the Method of Paired Comparisons

Thomas C. Brown and George L. Peterson
U.S. Forest Service, Rocky Mountain Research Station

1. INTRODUCTION

Assume for the moment that you are the supervisor of the Roosevelt National Forest (it's located near our town of Fort Collins, Colorado). You have been asked by the Chief of the Forest Service how you would spend a specified increase in your budget. You could spend the increase on such projects as improved campgrounds, better roads for reaching the backcountry, reduction in forest fuels to lower the risk of wildfire, and watershed management to lower erosion and thereby improve fish habitat. You know your chances of getting an increase depend on how well you support your proposal. You can design options that use up the budget increase, but you don't have good information about the benefits of, or even the public preferences for, the options. You know what the vocal interest groups want, but you would like your proposal to have some quantitative justification that reflects the values of the wider citizenry who care about the National Forest. You examined the economic analyses that had been done about forest resources in the area and found no studies for most of the major resources you manage. You don't have time for separate valuation studies of each of the options. However, you could commission a single study directly comparing the values people place on the options—a multiple good valuation study. Furthermore, because the options cost the same amount, a preference ordering of the options is all you need to

choose the best one. That option might then become the focus of an economic valuation study, which would allow a benefit-cost comparison.[1]

At the very least, a multiple good valuation study provides a reliable ranking of the goods—that is, it provides an ordinal scale measure of the values. A next step would be to provide an interval scale measure (defined in Chapter 4) of the values of the goods. Both ordinal and interval scales are preference orderings, but the latter has useful properties not found in the former, as seen later in this chapter. A further step would be to obtain a set of values that each have meaning in an absolute (i.e., ratio scale) sense, as would a set of economic values. In this chapter "value" refers to assigned value in general, which includes but is not restricted to monetary value (Brown 1984).

Stated preference methods for ordering preferences—such as ranking, rating, binary choice, and multiple choice—rely on asking each respondent to consider numerous items. Because the items are compared by each respondent, the resulting values are by definition comparable—a condition that is not necessarily assured for values that are each obtained from individual studies, where uncontrolled differences between studies may affect results.

Given the limited space available here, we will focus on only one multiple good valuation method, paired comparisons, which is a form of binary choice. Paired comparisons provide a rich set of data that allows estimation of interval or, in certain applications, ratio scale values as well as estimates of individual respondent reliability. Paired comparisons require more questions of respondents to evaluate a given set of goods than do, say, ratings, but the measures of respondent reliability that paired comparisons provide may justify the extra effort.

With paired comparisons, items in a choice set are presented in pairs and respondents are asked to choose the item in each pair that is superior on a specified dimension. Early use of paired comparisons, such as by Fechner (1860), focused on perception of physical dimensions (e.g., weight, length). The method was later extended by psychologists, marketing researchers, and others to measure preference dimensions such as product attractiveness (Ferber 1974), landscape preference (Buhyoff and Leuschner 1978), and children's choice of playground equipment (Peterson et al. 1973). In the national forest example above, respondents might be asked to select from each pair the project that they would prefer be funded. Most recently, paired comparisons have been used in economic valuation of attributes (Chapter 6) and goods (section 6 of this chapter).

We will present two applications of paired comparisons to value multiple goods. The first application achieves a preference ordering among a set of resource losses, an ordering that could support a schedule of costs to be imposed if the resources were to be damaged. The second application estimates economic values of a mix of public and private goods. Before we present these applications, we explain the theoretical model upon which the analysis of paired comparisons is based, provide some basics about the method, and summarize the steps to follow in implementing a paired comparison study.

In describing the mix of items presented to respondents when using the paired comparison method, we use "goods" generally to indicate public or private goods, resource conditions, or resource losses. "Items" may include such goods and any other stimuli included in the mix, such as monetary amounts.

2. A CHOICE MODEL FOR PAIRED COMPARISONS

Psychologists in the 19th century found that comparative judgments of physical stimuli (such as weight of objects, or loudness of noises) became less accurate the more alike were the stimuli. For example, the closer the weights of paired objects, the greater was the proportion of subjects who misjudged which object of a pair was heaviest, and the greater was the likelihood that a given subject would misjudge the relative weights some of the time in repeated trials (Guilford 1954). Differences in consistency of judgment were later also observed with qualitative judgments, such as of the excellence of hand writing samples or the seriousness of offenses (Thurstone 1927b). Subjects had more difficulty consistently judging some pairs than others. It was assumed that, as with physical stimuli, increasing inconsistency was associated with increasing similarity of the items on the dimension of interest.

In an effort to explain inconsistent paired comparison responses, Thurstone (1927a) proposed a model characterizing judgment as a stochastic or random process, wherein a stimulus, or item, falls along a "discriminal dispersion" around the modal value for the item on the "psychological continuum." The dispersion was attributed to random errors in judgment. This model was soon applied to preference as well as judgment, with the dispersion attributed to random fluctuations in preference.

The characterization of preference as a stochastic process was formalized as a direct random utility function (U) consisting of systematic (deterministic)

and random (error) components. For example, the utility (U) of item i to respondent n can be represented by the following relation:

$$(1) \quad U_{in} = E(U)_{in} + \varepsilon_{in}$$

U is a momentary relative magnitude internal to the person.[2] The systematic component, E(U), represented, for Thurstone, the expected value of U; the error component, ε, represented momentary variability about the expected value due to unobservable influences within a given respondent.[3]

More recent research has indicated that preference and judgment are sensitive to minor changes in the decision context. Contextual influences include phrasing of the questions and the characteristics of the survey setting (e.g., time of day or week, personality of the interviewer) (summarized by Payne, Bettman, and Johnson 1992; Slovic 1995). Some contextual influences are inherent to multiple-item stated preference methods. For example, in judging a set of multi-attribute items, the utility of an item may depend in part on characteristics of the other items in the choice set (Tversky 1996; Brown et. al. 2002) or on the order in which the items appear. One explanation of this behavior is that people simplify their decisions by more heavily weighting certain attributes in certain contexts (Tversky, Sattath, and Slovic 1988). In paired comparisons, for example, the weights for a given comparison may depend on which attributes are shared by the items of the pair; as the items change from one comparison to the next, the weights may change slightly, perhaps as respondents focus on the most obvious differences between the items, causing dispersion in preference or judgment for a given item across the several times it appeared in the pairs presented to the respondent.[4] Such minor contextual influences are difficult to detect and are typically left within the error term ε of equation 1.

Whereas attempts to model responses in psychology have often focused on understanding intra-personal variation, the use of the random utility model in economics has focused on explaining inter-personal variation. In modern demand modeling, the systematic component represents measured attributes of items (e.g., size, color, availability) and measured characteristics of people (e.g., income, age, education) (Ben-Akiva and Lerman 1985). Error is attributed to the analyst's incomplete ability or effort to measure and model.[5] A case for modeling is typically an individual or household—not a given

respondent at a specific moment during a survey—and the variability of interest is across people.

Both interpretations of the random utility model—one from psychological scaling and decision making and the other from economic demand modeling—recognize the influence of unmeasured variables; both interpretations model that influence as a random component. However, the two interpretations have historically served different purposes, in keeping with the different objectives of the analyses. The psychological approach has served as a model of respondent behavior, allowing an interpretation of inconsistent responses and forming the theoretical basis for scaling the responses to order preferences for the items of interest. The economic demand interpretation has provided the theoretical structure for modeling preferences as a function of explanatory variables by utility maximization.[6] It is the psychological interpretation that provides the theoretical structure of methods for multiple good valuation, to which we return.

In the absence of strategic behavior, items of higher $E(U)$ will tend to be chosen above other items in a paired comparison exercise. However, because of the randomness inherent in preference or judgment, responses may not always match the order of the expected values of items. Consider the utility of two items i and j:

$$(2) \qquad U_{in} = E(U)_{in} + \varepsilon_{in}$$
$$U_{jn} = E(U)_{jn} + \varepsilon_{jn}$$

If the error distributions of the items (ε_{in} and ε_{jn}) overlap, the order of the utilities of the items at a given moment (U_{in} and U_{jn}), and therefore the response, may be inconsistent with their respective expected values ($E(U)_{in}$ and $E(U)_{jn}$). The probability (P) that item i will be considered to be of greater utility than item j is:

$$(3) \qquad P(U_{in} > U_{jn}) = P(E(U)_{in} + \varepsilon_{in} > E(U)_{jn} + \varepsilon_{jn})$$

This probability increases as the difference between $E(U)_{in}$ and $E(U)_{jn}$ becomes larger, and the distributions ε_{in} and ε_{jn} become narrower.

Figure 1 depicts the utility of three items along the psychological continuum, assuming the errors are normally distributed. Given the preferences of Figure 1, responses will nearly always indicate that item k is preferred to

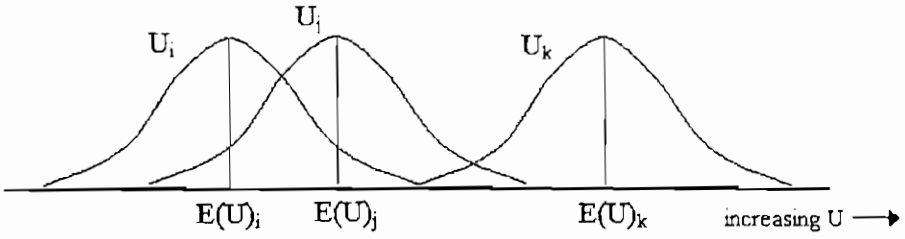

Figure 1. Utility of Three Items Along the Psychological Continuum

item *i*. However, as indicated by the over-lapping error distributions of items *i* and *j*, item *i* will sometimes be preferred to item *j*. Similarly, item *j* will sometimes, although relatively rarely, be preferred to item *k*. A preference for item *i* over item *j*, or of *j* over *k*—although perfectly reasonable given the utility function of Figure 1—would be inconsistent with the expected values of the preference distributions. A successful valuation method will estimate the expected values despite the variability of the response process.

Aside from these item-by-item concerns, multiple item valuation methods are also susceptible to two types of systematic changes that may occur over the course of a series of preference responses: the expected values of the items may change, and the error distributions may change. A change in the expected values may occur, for example, because the order in which the pairs of items are presented affects preference. Such an order effect, to the extent it occurs, cannot be avoided at the individual respondent level, but it can be neutralized across the sample by randomizing for each respondent the order in which the pairs appear.

Regardless of the order in which pairs of items are presented, systematic changes in the error distributions may occur over the course of a series of responses. Two such changes seem plausible. First, fatigue may cause responses to become erratic, leading to increasing inconsistency in the data. Second, the processing of multiple valuations may lead respondents to become more certain about their preferences, leading to decreasing εs and thus decreasing inconsistency with sequence.

3. PAIRED COMPARISON BASICS

With paired comparisons, respondents choose the item in each pair that has the greater magnitude on the given dimension. For valuation applications, the items are either all gains or all losses. The simplest approach, which we will use here, is to present all possible pairs of the items. With t items, there are $t(t-1)/2$ pairs in total. Each pair results in a binary choice that is assumed to be independent of all other choices. The choices allow calculation of a set of scale values indicating the position of the items along the specified dimension.

When presented with a pair of items, respondents typically are not offered an indifference option. This practice is supported by the theory of stochastic preference, wherein the probability of true indifference at any one moment is assumed to be very small. The practice also has the practical benefit of maximizing the amount of information potentially learned about the respondent's preferences. Although the lack of an indifference option may sometimes force respondents to make what seems like unwarranted distinctions, this is not worrisome because, across many comparisons, indifference between two items will be revealed in the data as an equality of preference scores.

3.1 Preference Scores

The full set of choices yields a *preference score* for each item, which is the number of times the respondent prefers an item to other items in the set. Preference scores are easily calculated by creating a t by t matrix and entering a 1 in each cell where the column item was preferred to the row item, and a 0 otherwise. Column sums give the preference scores. For example, Figure 2 contains a hypothetical matrix for a five-item choice set.[7] Preference scores, at the bottom, indicate that, for example, item j, with a preference score of 2, was chosen two of the times it appeared.

A respondent's vector of preference scores describes the individual's preference order among the items in the choice set, with larger integers indicating more preferred items. In the case of a five-item choice set, an individual preference profile with no circular triads (defined in section 3.2) contains all five integers from 0 through 4 (Figure 2).[8] For a given respondent,

	i	j	k	l	m
i	--	1	1	1	1
j	0	--	0	1	1
k	0	1	--	1	1
l	0	0	0	--	1
m	0	0	0	0	--
total	0	2	1	3	4

Figure 2. Five-item Response Matrix for One Respondent, without Intransitivity

the difference between the preference scores of items in a pair is that pair's *preference score difference*. This integer can range from 0 to 4 for a five-item choice set.

A response that is inconsistent with the expected values of a respondent's utility function will not necessarily be detected as inconsistent in the paired comparison data. For example, the responses of Figure 2 for items i, j, and k could have been produced from the utility function depicted in Figure 1. However, because of the theoretically inconsistent choice for the pairing of items j and k, the preference scores of Figure 2 misidentify the preference order of the Figure 1 utility function. The perfectly correct preference order for a single respondent could only be obtained with certainty if the respondent could be independently re-sampled several times and the several data sets were combined.

3.2 Circular Triads and Reliability

If an individual respondent is sampled only once, inconsistency is detectable only if it causes a lack of internal consistency in the data, which appears as one or more *circular triads*. A circular triad is an intransitive preference order among three items. For example, Figure 3 shows a five-item data set where one inconsistent choice causes the following circular triad: $l > k > j > l$. Such an inconsistent choice can occur because of the stochastic nature of preference, because of a context effect, or if the respondent makes a mistake in recording the choice. Circular triads cause some integers to appear more than once in the vector of preference scores (as a preference score of 2 does in Figure 3), while others disappear. In general, the greater the preference score difference of the items involved in the inconsistent response, the greater the number of circular triads that result.[9]

	i	j	k	l	m
i	--	1	1	1	1
j	0	--	1	0	1
k	0	0	--	1	1
l	0	1	0	--	1
m	0	0	0	0	--
total	0	2	2	2	4

Figure 3. Five-item Response Matrix for One Respondent, with Intransitivity

Because respondents enter choices for all pairs of items, reliability can be assessed individually for each respondent. A primary measure of reliability for a set of paired comparisons is the *coefficient of consistency*, which relates the number of circular triads in the respondent's choices to the maximum possible number. The maximum possible number of circular triads, m, is determined by the number of items in the choice set, t; m is equal to $(t/24)(t^2 - 1)$ when t is an odd number and $(t/24)(t^2 - 4)$ when t is even. The number of circular triads in each individual's responses can be calculated directly from the preference scores. Letting a_i equal the preference score of item i (i.e., the number of items in the choice set dominated by the i^{th} item) and b equal the average preference score (i.e., $(t - 1)/2$), the number of circular triads for an individual respondent, c, is (David 1988):

$$(4) \qquad c = \frac{t}{24}(t^2 - 1) - \frac{1}{2}\sum (a_i - b)^2$$

The coefficient of consistency is then $1 - (c/m)$ (Kendall and Smith 1940). The coefficient varies from one, indicating that there are no circular triads in a person's choices, to zero, indicating the maximum possible number of circular triads. Brown et al. (2001) found, based on paired comparisons from 1,230 respondents, that the median coefficient of consistency was 0.93, and that 95 percent of respondents had a coefficient of consistency of at least 0.77 (Figure 4). Over one-half of the range in coefficient of consistency was contributed by the remaining 5 percent of respondents. Clearly, a small minority of respondents had highly variable preferences or, more likely, made little or no effort to answer carefully.

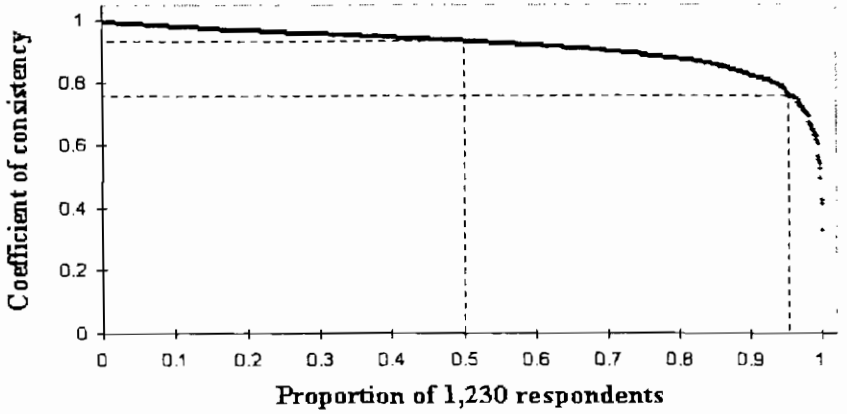

Figure 4. Distribution of Coefficient of Consistency

A short-term test-retest measure of reliability is made possible by repeating a random selection of pairs at the end of a paired comparison session. Some switching of choice over multiple presentations of the same pair is expected because of the stochastic nature of preference, whereby the utilities (U) of items change slightly from one moment to the next, as in equation 1.[10] Switching is more likely the closer are the $E(U)$s of the items of the pair. The closer are the $E(U)$s of the items, the smaller is the preference score difference, all else equal. Figure 5 shows the relation for the large set of paired comparisons summarized by Brown et al. (2001). For their data, the likelihood of switching drops from about 0.4 at a preference score difference of 0 to near 0 at a preference score difference of 12 or more.[11] Overall, 12 percent of the pairs were switched on retest.

3.3 Scale Values—Estimating the $E(U)$s

The response matrices of all respondents in the sample can be summed to provide a frequency matrix for the sample. For example, Figure 6 shows a hypothetical frequency matrix for a sample of ten respondents who each judged all possible pairs of five items. Column sums (next to last line) give the

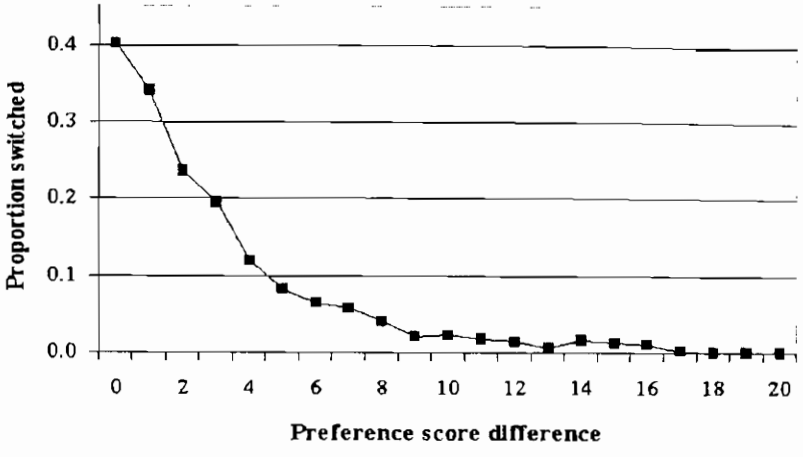

Figure 5. Likelihood of Choice Switching on Retrial

aggregate preference scores for the sample, specifying the number of times each item was chosen across all paired comparisons made by the respondents, and indicting the ordinal position of the items. It is also common to convert aggregate preference scores into percentages expressing the number of times that an item was chosen over the number of times it could have been chosen (Dunn-Rankin 1983). In the example of Figure 6, ten respondents judged each item four times, giving a maximum possible frequency of 40. The scale values (SV) vary from 15 percent to 90 percent of the maximum (last line of Figure 6). Such scale values are a linear transformation of the aggregate preference scores; thus, a zero on this scale indicates only that the item was never chosen, not that the item is of zero utility.

The aggregate preference scores, and thus the scale values, not only indicate the ordinal position of the items; they also approximate an interval scale measure of preference, revealing the sizes of the intervals between items. The approximation, given a sufficient sample size, tends to be very accurate except near the ends of the scale (i.e., except close to the most and least preferred items). The accuracy of the interval information is less near the ends of the scale because the paired comparison choice data are less rich there (i.e., the lack of items beyond those of minimum and maximum *E(U)* limits the number of choices that could help map out interval sizes in the regions of those items).[12]

	i	j	k	l	m
i	--	7	8	9	10
j	3	--	9	7	9
k	2	1	--	8	9
l	1	3	2	--	8
m	0	1	1	2	--
total	6	12	20	26	36
SV	15	30	50	65	90

Figure 6. Example of Frequency Matrix, with Aggregate Preference Scores and Scale Values

Thurstone (1927c) proposed a more involved scaling procedure based on what he called the "law of comparative judgment" that in theory corrects for the problem near the ends of the scale.[13] His approach models $(E(U)_{in} - E(U)_{jn})$ as a function of the probability of $(U_{in} > U_{jn})$. His scaling model relies on the assumption that U is normally distributed about $E(U)$, along with other assumptions regarding the standard deviations of and correlations between the disturbance terms. The most easily applied version of his approach, which assumes independent and identically distributed errors, leads to the probit model (Ben-Akiva and Lerman 1985).

McFadden (1974) proposed another approach to analyzing binary response data. If Thurstone's assumption about the error term is replaced by one of independent double-exponential random disturbances, we have the basis of the logit model, as the difference distribution of two independent double-exponential random variables is the logistic distribution. It is important to note, however, that Thurstone's and McFadden's methods each require data sets that comply with some rather demanding assumptions, and that contain enough cases to allow capture of the full extent of the variability of preference within the population. Without a rich data set, the improvements over aggregate preference scores promised by their methods are not likely to be achieved—a conclusion that would be of more concern if it were not for the fact that scale values computed using their approaches typically correlate very highly with the aggregate preference scores. We do not describe how to apply Thurstone's or McFadden's methods here. The logit model is described in chapter 5 for dichotomous choice contingent valuation and in many other books and articles. Thurstone's method is described by Guilford (1954) and Torgerson (1958).

These concerns about the degree to which a true interval scale measure is achieved become moot if well-chosen anchors of known value, such as monetary amounts, are included among the items to be judged. If anchors of known value are included, ratio scale values of the goods can be approximated from the individual vectors of preference scores (explained in section 6 of this chapter). These approximations can then be averaged to form sample estimates of the values of the goods.

When the known anchors are monetary amounts, each respondent makes multiple binary choices between goods and sums of money. Such paired comparison responses can be analyzed using discrete choice methods, such as binary logit analysis.

4. STEPS IN A PAIRED COMPARISON VALUATION STUDY

The first step in using the method of paired comparisons is to decide on the goal of the study (Table 1). As mentioned above, the method of paired comparisons can be used to order preferences among a set of goods or to estimate monetary values of goods. In keeping with the goal of the study and the types of goods involved, the dimension that respondents are instructed to use in making their choices is also specified. There are many options. For example, if the goal is a preference order among certain public goods, respondents may be asked to choose the good of each pair that they prefer or that they think is the most important for society. If the goal is economic valuation, the analyst must decide whether a payment or compensation measure is desired before the response dimension is specified, as described in the economic valuation application at section 6.1.

Second, the items to be valued are specified. If only a preference order is desired, monetary amounts do not need to be included among the items to be presented to respondents. In this case, specification involves deciding precisely what goods will be valued and how they will be described to respondents. The goods must be carefully defined in terms that respondents will understand, a process that may require focus groups and pre-testing involving persons selected from the likely respondent population. The decision about how much

Table 1. Steps in Using Paired Comparisons to Value Multiple Goods

1.	Decide what is to be measured (a preference order, WTPc*, or WTAc) and what dimension is to be used by respondents in making their choices.
2.	Specify the items to be presented (the goods and, if necessary, the monetary amounts).
3.	Select the respondent population and sample frame.
4.	Choose the method of administration and design the instrument for presenting the pairs of items.
5.	Apply the instrument with the sample.
6.	Analyze the data, including assessing reliability and scaling the choices.
7.	Interpret the results for the application at hand.

*WTP = willingness to pay; WTA = willingness to accept

information to present about the goods is not unlike the similar decision faced in contingent valuation: the analyst wants to present a rich description of the good but not so much information that respondents become bored or confused. However, with multiple good valuation the decision is faced multiple times, once for each of the goods that must be described. The larger the number of goods, the shorter must be the descriptions, all else being equal, to avoid overburdening the respondent.

We have used two methods of presenting descriptions of the goods when they are complex, which is commonly the case with nonmarket goods. The first is to describe all of the goods before respondents are presented with the paired comparisons. Then for the comparisons, each good is indicated by an abbreviated description or just a title, with the full descriptions still available at all times on one or more separate sheets of paper. With the other approach the goods are not described before the paired comparisons begin, but for each comparison the full description is presented and an abbreviated description is presented just below it. Once respondents become familiar with the goods—which they encounter many times in the course of making their choices—they tend to rely on the short descriptions, lessening the respondents' burden.

The set of goods need not include only those of primary interest. It may be useful to also include goods selected to help respondents choose between the

goods of primary interest. Two approaches are useful. First, close substitutes for the goods of primary interest may be included, to help bring to respondents' attention the characteristics of the goods of primary interest. Second, familiar, commonly available market goods may be included. These may serve two purposes: to help familiarize respondents with the paired comparison task by having them make choices between goods with which they are more familiar, and to provide the researcher with data to test whether respondents are making sensible choices.[14] Additional private goods were included in the item mix in the monetary valuation application presented in section 6.

When the paired comparison approach is being used to estimate economic values, monetary amounts (called "bid levels" in the contingent valuation literature) must be included in the mix of items to be compared. Because the paired comparison approach to economic valuation is still being developed, less thought has gone into the details of bid level selection than in the case of dichotomous choice contingent valuation. However, two basic considerations can be offered. First, the bid levels should span the values of the goods for the bulk of the respondent population so that respondents' values for the goods are bracketed by their values for sums of money, allowing the values of the goods to be estimated. Second, increasing the number of bid levels within the specified range allows the values of the goods to be more precisely estimated, but the bid levels must not be so numerous that the pairs of items cannot all be judged within a reasonable amount of time for the respondent population.

The number of items (goods, or goods plus bid levels) to be included in the mix should be large enough to help maintain independence among the choices (i.e., to make it difficult for respondents to remember prior choices or make their task easier by memorizing a ranking of the items), but not so large that respondents become fatigued or lose interest and answer without care. Research has not been done on what number of items is optimal. From our experience we suggest that at least ten items be used. As for the maximum number, when the pairs are presented and responses are recorded on the computer, we have found that respondents can judge at least 200 pairs without loss of reliability. With a 200-pair maximum, if all items are goods, this allows for a maximum of twenty goods ($20(20-1)/2=190$). If, say, ten of the items are bid levels, the 200-pair limit allows at most twelve goods, not ten, because the bid-level-by-bid-level pairs are not presented (since choices between amounts of money are obvious).

In the third step, the respondent population and sample frame are selected. Considerations in making these selections are no different in concept from those with the other stated preference methods. See Chapter 3 of this book for details.

Fourth, the instrument for presenting the pairs and recording the choices is designed. Perhaps the simplest approach is to gather respondents in a room and present the pairs to all of them at the same time, using whatever visual aids are appropriate. However, this approach does not allow the order of presentation to be randomized for each respondent, which can be done by presenting the pairs on paper. Presenting each pair on a separate half sheet of paper helps maintain independence among the choices; however, one cannot be assured that respondents will abide by the request to not look back at prior choices, which would compromise the attempt to maintain independence among choices. Independence is more likely to be maintained if respondents are contacted in person, rather than by mail, because the interviewer is present to observe if respondents ignore the request to not look back at prior choices. Perhaps the most satisfying approach is to use the computer, either via the Internet or by bringing computers (e.g., laptops) to the respondents or bringing respondents to the computers.

Use of a computer program for presenting the pairs of items and recording respondents' choices has several advantages. First, the order of the pairs can be randomized easily for each respondent. Second, independence among an individual's responses is more likely maintained, because respondents cannot go back to see how they responded to prior pairs. Third, if some pairs are to be repeated to check for reliability, the retest pairs can easily and quickly be selected based on the initial responses. Fourth, the computer can keep track of the time required to enter each response; these data can later be used to see if, for example, respondents take more time for early versus later pairs, or more time for difficult (i.e., small preference score difference) versus more easy pairs (Peterson and Brown 1998).

The fifth step is application of the instrument, ideally over a short enough time period that events or news stories do not change respondents' knowledge and experience over the course of data collection. To help maintain independence among the choices, respondents are instructed at the start that each choice is a change from their situation when they began the experiment (i.e., that each choice is made as if it were the first and only choice).

Sixth, the data are analyzed. Methods of assessing reliability and scaling the choices were discussed section 3 and are demonstrated in the two applications that follow in sections 5 and 6.

Finally, the results are interpreted for the policy maker. If only a preference ordering was produced, the interpretation must include the warning that economic benefits have not been estimated and that such benefits may not exceed the costs of providing the goods that were assessed. It may also involve the mapping of monetary values onto the preference ordering, as envisioned with the damage schedule approach described in the next section. If monetary values were produced, the interpretation must include an explanation of the nature of the values that were estimated, which are not necessarily identical to those estimated with more traditional nonmarket valuation methods. The differences are explained in the monetary valuation application in section 6. Also, because independence among choices is assumed, the monetary values are not additive.

5. APPLICATION 1: ORDERING LOSSES AND THE DAMAGE SCHEDULE

It has long been maintained that comparative judgments are easier for people than are absolute judgments. As Nunnally (1976) states, "People simply are not accustomed to making absolute judgments in daily life, since most judgments are inherently comparative...people are notoriously inaccurate when judging the absolute magnitudes of stimuli...and notoriously accurate when making comparative judgments" (p. 40). Nunnally's statement focused largely on judgments of physical phenomena such as the length of lines or the brightness of lights, but his notion can perhaps be applied as well to judgments of monetary value. If so, in contingent valuation we would prefer to use a binary choice question to an open-ended question, because the latter would require an absolute judgment of willingness to pay (WTP). However, even a dichotomous choice contingent valuation question demands some measure of quantification, because to know whether one would pay a specified bid amount, say \$30, for an item, one's maximum WTP must be compared to the \$30. It may simply be more difficult for people to know if they would pay \$30 for an item than to decide which good in a pair they would pay more for.

Binary valuation questions involving a distinct monetary amount cannot avoid requiring the respondent to quantify their values to some degree. This difficulty of binary choice WTP questions perhaps contributes to the sensitivity of such questions to anchors provided by the bid amounts (Boyle, Johnson, and McCollum 1997; Green et al. 1998). It may also account for the recent findings of Breffle and Rowe (2002) that the randomness in paired comparisons of resource conditions was less than that for two other kinds of binary choices involving monetary amounts, referendum contingent valuation and an attribute-based approach. Thus, simplifying the judgments required in binary choices by removing the need for the respondent to quantify his or her values in monetary terms has potential benefits. Such simplification also incurs a cost, in that the researcher receives only a preference ordering of the goods, not quantified values. However, in some cases, such as the situation described in the introduction, a preference ordering is sufficient.

A preference ordering may also be sufficient is when it is used as input for a damage schedule. A damage schedule, as described by Rutherford, Knetsch, and Brown (1998), is a predetermined set of sanctions against resource losses. It relies on a community-based preference ordering with respect to deteriorations in environmental conditions. Only after citizens' judgments about the relative importance of a series of resource losses are obtained, and then scaled to provide an interval scale measure of importance of loss, are monetary damage payments and other sanctions mapped onto the loss scale. The damage schedule is described at the end of this section. First, we summarize how a preference ordering was obtained for coastal resources in Thailand.

5.1 Importance Judgments for Natural Resources of Phangnga Bay, Thailand

The viability of a damage schedule based on citizens' judgments of importance of loss depends critically on the reliability of those judgments. One measure of reliability is the level of agreement among different groups of respondents. Chuenpagdee, Knetsch, and Brown (2001a) tested this agreement for importance judgments from paired comparisons of individuals familiar with the natural resources of Phangnga Bay, a coastal area of southern Thailand.[15]

Phangnga Bay, like other Thai coastal regions, is rich in natural resources. Rivers flowing into the bay supply nutrients and minerals, making it an

important spawning ground, nursery area, and habitat for many commercially important species including marine shrimps, lobsters, crabs, clams, Indian mackerel, and pomfret. Several species of molluscs and crustaceans inhabit the remaining old-growth stands of mangroves. During the past decade, coastal aquaculture—involving black tiger prawns, cockles, oysters, and cage culture of snapper and groupers—has joined traditional fishing and gathering as an important economic activity. Furthermore, the coast is rapidly being developed for housing, tourism, and a variety of other industries. All this growth has enhanced conflicts among resource users and increased the probability of resource losses.

The eight resource losses used in Chuenpagdee, Knetsch, and Brown's test of the paired comparison approach focused on four ecosystems at risk in Phangnga Bay: sandy beaches, mangrove forests, coral reefs, and seagrass beds.[16] The eight losses were developed from personal visits to the area, interviews of resource users and other residents, discussions with local resource managers, and the results of an extensive pre-test of the survey. The losses (listed in the left-hand column of Table 2) included two levels of damage for each of the four ecosystems.

Respondents were given information about the nature and productivity of the resource, the extent of the human-caused damage at issue, the expected changes in the level of productivity due to the losses, and the length of the likely recovery time for the resource losses where recovery was possible. For example, in the case of coral reefs, the reefs were shown on a map of the bay; the importance of the reefs to marine organisms, recreation, and natural beauty were outlined; and the possible sources of damage (pollution, sedimentation, boat anchoring, discarded fishing nets, and tourist activities) were listed. "Partial" damage was then specified as a 50 percent reduction in resource productivity of the reefs, with a recovery period of six to ten years, whereas "serious" damage was specified as productivity being reduced to almost nothing and requiring twelve to fifteen years to recover.

The respondent population, depending on the objectives of the damage schedule, may consist of persons very knowledgeable about the resources, or the general public. Two sets of respondents were sampled, experts and resource users. A comparison of the preferences of these two populations is perhaps most instructive, because it shows whether, and to what extent, they differ. The

Table 2. Scale Values[1] of Resource Losses in Phangnga Bay, with Two Measures of Consistency

Resource Loss	Experts (51)[2]	All Resource Users (170)	Fishers (45)	Shrimp Farmers (40)	Tourism (39)	Others (46)
Clear-cutting mangroves	85	83	84	81	80	84
Severe damage to coral reefs	83	76	73	76	79	76
Severe damage to mangroves	62	67	72	67	64	65
Partial damage to coral reefs	59	53	51	53	56	51
Severe damage to seagrass beds	51	42	42	41	45	41
Severe damage to sandy beaches	31	43	41	44	41	47
Partial damage to seagrass beds	24	19	18	20	23	16
Partial damage to sandy beaches	6	17	19	18	12	19
Average number of circular triads	1.5	3.7	3.5	3.9	3.5	3.9
Average coefficient of consistency	0.93	0.81	0.82	0.81	0.82	0.81

[1]Scale values are presented as the percent of the total number of times the loss could have been considered more important.
[2]Sample size in parentheses.

former included researchers, academics, administrators, and other government officials with experience and knowledge of the area and the resources at issue. The resource users were people living in the area and dependent on the resources. Four groups of resource users were sampled: (1) fishers, (2) shrimp farmers, (3) people in tourism-related businesses, and (4) others living in the area whose dependence on coastal resources was less specific. Convenience

samples of these four groups of resource users in the Phangnga Bay area were selected. Sample sizes for the five groups (four resource user groups plus the experts) are listed in Table 2, as is the sample size for the entire set of resource users. Altogether 221 people were surveyed.

Use of a computer to present the pairs of losses was not practical with some of these samples, in part because of respondents' lack of familiarity with computer technology. After being given detailed descriptions of the eight losses, plus a map and a reference table summarizing the magnitudes and recovery times of the losses, each participant was given a set of paired losses, with the losses of each pair presented side-by-side on a separate half sheet of paper. The pairs were arranged in random order and the losses in each pair were randomly placed (left versus right) to randomize order effects. For each paired comparison, participants were asked to choose "the more important loss, not only to yourselves, but also to the environment, to the economic and social values of the community, and to the future of the area."

Of the 28 possible pairs of the eight losses, three were not included in the questionnaires because they compared a more severe loss to a less severe loss of the same resource; the assumed answers to these three were included in the analysis. All participants answered all 25 of the remaining paired comparison questions.

5.1.1 Reliability

The 221 respondents averaged 3.2 circular triads of a possible maximum of 20, yielding an average coefficient of consistency of 0.84. Thirty-nine percent of the respondents had no circular triads, 68 percent had 3 or less, but 9 percent had more than 10, indicating that a minority of respondents was responsible for much of the range in coefficient of consistency, as in Figure 4. As reported in the bottom two rows of Table 2, the experts were more consistent in their choices than were the user groups. The experts averaged only 1.5 circular triads per respondent, whereas the user groups averaged roughly 3.7; corresponding mean coefficients of consistency were 0.93 and 0.81, respectively.

5.1.2 Scale Values

The scale values for the eight losses, presented as the percent of the total number of times the loss could have been considered more important, are listed

in Table 2 for each sample and for all resource users together.

The close correspondence of the scale values across the different samples is apparent in Table 2. Not only did resource users generally agree with the experts, but the scale values among the groups of users did not vary widely. All samples, for example, considered clear-cutting of mangrove forests to be the most important loss, followed by severe damage to coral reefs. The overall level of agreement is indicated by the correlation coefficients comparing the scale values across the five samples, which range from 0.934 to 0.999, with a median of 0.976. The general agreement about the rankings of resource losses suggests that the groups did not act strategically in favoring resources of particular interest to them.

The high level of agreement among groups supports the combination of responses from the various groups to form a single importance scale for the 221 respondents. The aggregate preference scores, expressed as before in terms of percent of times, in relation to the maximum possible number of times, the loss was considered most important (as in Table 2), are in parentheses in Figure 7.

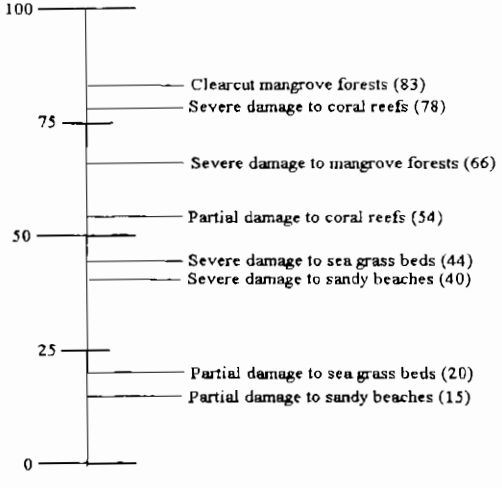

Figure 7. Importance Scale of Selected Losses in Phangnga Bay

5.2 The Damage Schedule

Once scale values are computed, they may be used to support a damage schedule. As mentioned earlier, a damage schedule is a set of sanctions, perhaps including monetary payments, that are mapped onto a preference ordering such as that of Figure 7. If monetary values are to be mapped onto the scale, the degree of measurement represented by the importance scale must be determined. If it is assumed to be only ordinal, a separate monetary amount must be specified for each loss. However, if interval properties are assumed, the mapping requires that only two points along the scale be specified in monetary terms; all other points along the scale follow from these two in a linear fashion.

The specification of sanctions is more a political than a technical task. Sanctions would be specified by the elected or appointed officials with statutory authority for protecting the public's resources, perhaps in consultation with economists and others knowledgeable about the full range of evidence on the economic value of the resources. Such a damage schedule is not intended to provide accurate monetary measures of value. However, it is based on a carefully derived community-based ordering of the importance of alternative resource losses. To the extent that the ordering provides a consistent set of comparable judgments of the importance of alternative resource losses, it may provide a well-founded basis for damage payments (Rutherford, Knetsch, and Brown 1998).

Damage schedules in general have several advantages, compared to approaches that require post-incident valuation. Two are mentioned here. First, because damage schedules specify remedies in advance rather than after an event (such as an oil spill or degradation of wildlife habitat) has occurred, they can provide more effective deterrence incentives, because those responsible for potential losses would be more fully aware of the cost to them of their actions, allowing them to undertake appropriate levels of precaution. Second, enforcement of sanctions would likely be easier and acceptance greater, because once liability is established in any particular case, the consequence is predetermined from the schedule rather than subject to the uncertainties of contentious adjudication.

Schedules, or their equivalent, are accepted in other areas in which specific assessments of the value of losses are difficult or expensive. An example is the schedule of awards for injuries used in many workers' compensation schemes.

Although the specified sums are not taken to actually reflect the value of such losses to individuals, they do reflect relative values and are therefore widely accepted and achieve many of the goals of sanctions such as efficiency enhancement. Damage schedules, or replacement tables, have also been used for environmental losses, especially for small oil spills. However, nearly all instances of such use have based sanctions on notions of replacement costs or on fairly arbitrary legislative directives rather than on an empirical assessment of community preferences regarding the importance of different losses.

A schedule for dealing with unauthorized loss is not the only use of the importance scale. It could also be used to guide further development and use of an area's resources. For example, absolute prohibitions or more onerous sanctions might be adopted to severely restrict uses that could cause losses judged to be of the highest importance, such as clear-cutting of mangrove forests and severe damage to coral reefs. Uses that could cause somewhat less serious losses, such as partial damage to seagrass beds and sandy beaches, might be subjected to somewhat less stringent restrictions or payments to discourage the loss but to allow compromise and accommodation in cases of extremely valuable alternative uses. Uses that could cause losses considered to be increasingly less serious might be made subject to successively more lenient restrictions and smaller payments. And in the cases of losses judged to be trivial, an absence or near absence of sanctions could reflect this valuation.

6. APPLICATION 2: ECONOMIC VALUATION

Economic valuation of multiple goods using the method of paired comparisons is made possible by including sums of money among the items in the choice set. Choices between a good and a monetary amount indicate which is preferred. A series of choices involving a range of carefully chosen monetary amounts allows us to estimate the break point—where the respondent switches from the good to the money. The inclusion of several different goods in the choice set potentially allows the values of each of those goods to be approximated for each respondent.

This approach has several attractive features related to the inclusion of multiple goods in the choice set. First, it produces individual respondent vectors of preference scores involving several goods, allowing multiple tests of individual reliability. Second, requiring respondents to compare several goods

may increase the likelihood that respondents will think more carefully about the characteristics and relative worth of the goods. Third, including several public goods in the choice set may reduce the tendency of respondents to exaggerate the importance of a single good to which attention has been drawn. In addition, the method uses a range of monetary amounts, perhaps lessening the tendency to anchor on a given amount. However, the method faces limitations as well, related to the quantification of value implied by the good versus money choices (as described in the introduction to section 5) and to the need to describe numerous goods to each respondent. Because several goods must be described, the tendency is to shorten the descriptions, compared with contingent valuation (where typically only one good is described), so as to not over-burden the respondent. The challenge of any stated preference method, to maintain content validity by adequately specifying the item(s) of interest, is magnified when the items are numerous.

6.1 Theory

To understand what is being measured when monetary amounts are included in the choice set of a paired comparison survey, consider Figure 8, which shows three indifference curves. The horizontal axis measures quantity of the good of interest (i.e., one of the goods included in the choice set), which we will call good X, and the vertical axis measures money and all other goods. Assume in all cases that the individual is at point A along U_2, that point B falls on a higher indifference curve U_3, that point D falls on a lower indifference curve U_1, that moving from A to B represents a zero price change increase in X, that moving from A to D represents a zero price change decrease in X, and that AB=AD (see Freeman (1993) for background on the economics of quantity changes). Because the items included in the paired comparisons can be losses or gains, both WTP and WTA measures are obtainable, as will be seen shortly.

Unlike the standard measures of WTP and WTA, which are computed by holding utility constant (thus providing compensating surplus measures), the paired comparison measures involve moving to a new utility level (providing equivalent surplus measures). The two familiar economic measures for a zero price quantity change are represented in Figure 8 as BC, the maximum WTP to obtain AB of good X, and DE, the minimum WTA to give up AD of good X.

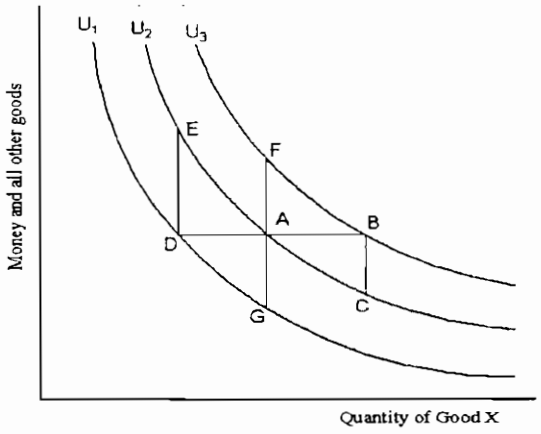

Figure 8. Measures of Economic Value

Paired comparisons offer two additional measures—a WTP measure obtained by offering the respondent a choice between losses, and a WTA measure obtained by offering the respondent a choice between gains. If offered a choice between losing AD of good X or losing various sums of money, the respondent will choose to retain the quantity of good X at small sums of money but will switch to retaining the money once the amount passes the maximum WTP to avoid losing AD of good X, which is equal to AG. If offered a choice between gaining (at no cost) AB of good X or gaining various sums of money, the respondent will choose the quantity of good X at small sums of money but will switch to taking the money once the amount passes the minimum WTA to forego AB of good X, which is equal to AF.

Equality of the two measures of WTP, or the two measures of WTA, is unlikely. For example, for WTA, DE (the standard measure) is not necessarily equal to AF (the paired comparison measure). (Of course, the standard measure of WTA assuming the individual is at point B, which is equal to AF, would be identical to the paired comparison measure of WTA assuming the individual is at point A, which is again AF.) Given the individual is at point A, the standard measure DE will equal the paired comparison measure AF only if the marginal rate of substitution (MRS) of good X for money and other goods is constant for all levels of good X at a given level of money and other goods (i.e., if the indifference curves are horizontally parallel).

In addition to effects of changing MRS, the standard measure may be affected by loss aversion (Knetsch 1989), an effect not depicted in Figure 8.

The potential effect of loss aversion is seen by noting the differences in questions put to the respondent. With WTP, the standard question asks WTP to *obtain* the good, whereas the paired comparison question essentially (though not literally) asks WTP to *keep* (to not lose) the good. If, as the loss aversion notion suggests, WTP to avoid losing a good is greater than WTP to obtain the good, the paired comparison measure will, all else being equal, exceed the standard measure. And with WTA, the standard question asks WTA to *give up* (to lose) the good, whereas the paired comparison question essen-tially (though not literally) asks WTA to *forego* (to not gain) the good. If, as loss aversion suggests, WTA to give up a good is greater than WTA to forego a good, the paired comparison measure will, all else being equal, fall below the standard measure as enhanced by loss aversion. Loss aversion causes a kink in the indifference curve at the current reference point, which causes a divergence from the expectation of standard economic theory depicted in Figure 8.

Both a diminishing MRS (of good X for money and other goods as amount of X increases at a given level of money and other goods) and loss aversion cause: (1) the paired comparison measure of WTP to exceed the standard measure, and (2) the standard measure of WTA to exceed the paired comparison measure. Because of the potential differences between the standard measures and the measures available with use of paired comparisons, we designate the paired comparison measures WTPc and WTAc, where the "c" signifies the *chooser* reference point. Thus, in the presence of either diminishing MRS or loss aversion, we would expect the magnitudes to order as follows: $WTP \leq WTPc \leq WTAc \leq WTA$.

The sums of money included in the choice set must be chosen with care to bound the values of the goods of interest. Pre-testing may be necessary to select the monetary amounts. If the amounts are properly chosen, the goods of interest are bounded for each respondent by monetary amounts along the vector of preference scores that specify upper and lower bounds on maximum WTPc or minimum WTAc for each good.[17] For WTPc, if respondents were asked to choose the item they preferred to keep, the upper and lower bounds are the proximate sums of money of higher and lower preference score, respectively. For WTAc, if respondents were asked to choose the item they preferred to gain, the upper and lower bounds are again the proximate sums of money of higher and lower preference score, respectively.[18]

These upper or lower bounds for a good, or an interpolated value between these bounds, can be aggregated across respondents to estimate mean or median

values for the good. Empirical bid curves can also be estimated for a good based simply on the proportion of respondents rejecting the good (i.e., accepting the money) at each bid level. As mentioned in section 3.3, point estimates of the value of a good may also be approximated from the aggregate preference scores, estimated using binary discrete choice methods common in dichotomous choice contingent valuation (based on respondents' choices between a good and the various sums of money), or estimated using scaling based on Thurstone's (1927c) "law of comparative judgment."

When respondents are presented with a series of paired choices, either gains or losses, they are assumed, as with contingent valuation, to take their current situation as the status quo. Respondents are assumed for each choice to be at point A in Figure 8, such that each choice is a mutually exclusive change from their current condition. Instructions to the respondent should make this clear.

People often consider choices between items and thus are familiar with the chooser reference point. Such situations may involve private items (e.g., a choice between chocolate bars) or public items (e.g., a choice between candidates, or between rapid transit options). However, economic decisions about public goods are not commonly made from this reference point. That is, citizens are not typically asked to choose between receiving a public good and receiving a sum of money, or between giving up a public good and giving up a sum of money, so such choices tend to appear distinctly hypothetical. For example, it is not common to be asked: "Which would be the greater gain to the citizens of this city, the previously described improvement in air quality or each receiving $100?"

Because of the hypothetical character of the chooser reference point when valuing pubic goods, and because numerous public goods are included in the choice set, respondents to a paired comparisons survey may not be inclined to believe that their choices will directly affect policy. Thus, use of the method of paired comparisons to value public goods may interfere with establishing the kind of "incentive compatibility" for which some contingent valuation practitioners strive (Carson, Groves, and Machina 1999).

6.2 Valuing Public Goods in Fort Collins, Colorado

Peterson and Brown (1998) made a first attempt to use paired comparisons to estimate monetary values of multiple public goods. Their study evaluated the

reliability and transitivity of respondent choices; the study was not intended to provide values for benefit-cost analysis, but it does demonstrate the method.

The choice set in this experiment consists of six locally relevant public goods, four private goods, and eleven sums of money. Each respondent made 155 choices, 45 between goods and 110 between goods and sums of money. They did not chose between sums of money. Three hundred thirty students at Colorado State University, located in Fort Collins, participated in the study. Three were dropped because of incomplete data, leaving a total of 327 respondents.

The four private goods were familiar market goods with suggested retail prices: a restaurant meal not to exceed $15, a nontransferable $200 certificate for purchase of clothing, two tickets to a cultural or sporting event worth $75, and a nontransferable $500 certificate for purchase of airline tickets. The private goods were included to encourage respondents to consider a wide range of goods and trade-offs, to avoid inducing value by focusing too much attention on any one good, and to examine WTAc for familiar private goods with suggested prices.

The six public goods were of mixed types. Two of the goods—a 2,000 acre wildlife refuge in the mountains west of Fort Collins purchased by the university (Wildlife Preserve) and an improvement in the air and water of Fort Collins (Clean Arrangement)—were pure public environmental goods (i.e., environmental goods that are nonrival and nonexcludable in consumption). The remaining four public goods—a no-cost on-campus weekend festival of music and other events (Spring Festival), a no-fee service of videotapes of all course lectures (Video Service), parking garages to eliminate parking problems on campus (Parking Capacity), and expansion of the eating area in the student center (Eating Area)—were excludable by nature but stated as nonexcludable by policy. Wildlife Preserve and Clean Arrangement benefit all people in the broader community, whereas the other goods benefit only the students. Respondents had a table describing each of the goods in front of them during the experiment and were free to refer to it at any time.

The eleven sums of money were $1, $25, $50, $75, and $100 to $700 in intervals of $100. The public and private goods used in the experiment were derived from pilot studies in order to have good variation and distribution across the dollar magnitudes. Respondents were asked to choose one or the other item under the assumption that either would be provided at no cost to the respondent.

The experiment was administered on laptop computers that presented the items on the monitor in random order for each respondent. The goods had short names which appeared side-by-side on the monitor, with their position (right versus left) also randomized. The respondent entered a choice by pressing the right or left arrow key and could correct mistakes by pressing "backspace."

6.2.1 Reliability

Across the 327 respondents, the coefficient of consistency ranged from 1 to 0.51, with a mean of 0.92. As suggested by the larger median, 0.94, a few respondents were particularly inconsistent; indeed, only 10 respondents (3 percent) had a coefficient of consistency below 0.75, the midpoint of the range.[19]

6.2.2 Scale Values

Application of various scaling options—averaging interpolated values from respondents' individual vectors of preference scores, plotting empirical bid curves, interpolation based on the aggregate preference scores, discrete choice analysis,[20] and Thurstone scaling—yielded similar results. We present results for the former two, more straightforward approaches.

Table 3 displays a vector of preference scores for a typical respondent. As described above, each respondent's monetary values can be obtained for each good from the preference profile by using the lower bound, upper bound, or an interpolation between these bounds. The conservative approach is to use the lower bounds. In this case, for the respondent in Table 3, Parking Capacity would be assigned a value of $100 and Video Tape Service and Wildlife Refuge would each be assigned a value of $400. However, when the preference score of a good coincides with that of a monetary amount, the good can be assigned that amount; thus, Clothing would be assigned a value of $200. Using the linear interpolation option, Parking Capacity would be assigned a value of $150 and the Video Tape Service and Wildlife Refuge would each be assigned a value of $433. Clothing would again be assigned a value of $200.

Table 4 shows mean and median estimates for the ten goods calculated from interpolated estimates of WTAc from each respondent's preference scores. The median is the value that is acceptable to at least 50 percent of the sample

Table 3. Preference Scores of One Respondent

Item	Preference Score
$700	20
$600	18
Air travel worth $500	17
Clean arrangement	17
$500	16
Video tape service	14
Wildlife preserve	14
$400	13
$300	12
Clothing worth $200	10
$200	10
Parking capacity	9
$100	8
$75	7
$50	6
Tickets worth $75	5
A meal worth $15	5
$25	4
Eating area capacity	3
Spring festival	1
$1	0

and therefore identifies the value that a majority will accept. The mean is an estimate of the expected value of the response. The means and medians differ substantially in some cases because the distributions are highly skewed. The reader must not generalize the values reported here because they are merely

illustrative and do not necessarily represent any population beyond the sample observed. To generalize such values beyond the sample requires rigorous sample design and more rigorous examination of the estimates.[21]

Table 4. Monetary Values (WTAc) Estimated by Interpolation of Individual Vectors of Preference Scores

	Median	Mean
Wildlife preserve	$400	$390
Air travel worth $500	400	382
Clean air arrangement	220	328
Clothing worth $200	150	179
Video tape service	88	190
Tickets worth $75	88	144
Parking capacity	81	165
Spring festival	63	151
Eating area capacity	36	94
A meal worth $15	17	44

Figure 9 shows empirical bid curves for Wildlife Preserve and Clothing, and another way to estimate medians. Each dot along the curves indicates the proportion of respondents who chose the good over the respective monetary amount. The straight lines connecting the dots allow estimates of the medians, which are the monetary amounts that would be rejected by 50 percent of the respondents. Using this approach, the median WTAc is $382 for Wildlife Preserve and $152 for Clothing.

Inclusion of familiar private goods with known prices in the choice set offers some information about the validity of the choices. Air Travel and Clothes, for example, have listed values of $500 and $200, respectively; their medians from Table 4 are $400 and $150, respectively. Their mean values are also lower than the suggested retail prices. The other two private goods, Tickets and Meal, have medians and means higher than their stated retail prices, of $75 and $15, respectively. The medians of the lower bounds for these two goods ($75 and $1, respectively) are equal to or lower than the retail prices.

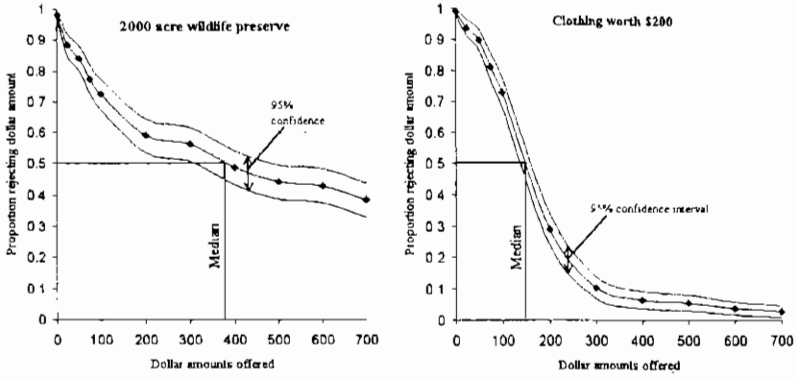

Figure 9. Empirical Bid Curves for Two Goods

Excess values for some private goods occurred for two reasons. First, some respondents sometimes actually chose the good over a dollar amount of greater magnitude than the market value of the good. Removing these respondents from the sample lowers all the median and mean values. For example, the median and mean interpolated values for the $15 Meal drop to $15 and $17, respectively, and the values of Wildlife Preserve drop to $200 and $312, respectively. Second, the good versus good choices also affect the preference scores from which the values reported in Table 4 were computed. Some respondents tended to value some private goods more highly when comparing them with public goods than when comparing them with dollar amounts, which elevates the standing of these goods in the frequency matrix relative to the dollar amounts.

A minority of the respondents appears to have placed considerably more value on the public goods than the other respondents. For example, 33 percent chose Wildlife Preserve over $700. We cannot know for sure whether such responses are valid, but the presence of such high values suggests that paired comparisons may be subject to hypothetical bias similar to that affecting dichotomous choice contingent valuation. That is, some respondents may be treating the choice as an opportunity to express attitudes towards certain kinds of goods rather than as an actual choice between a good and money.

7. FINAL COMMENTS

Multiple good valuation can provide a reliable ordering among goods, as well as a set of values that place the goods along an interval-level scale. An ordering of goods can support, for example, a ranking of environmental conditions along some dimension of interest or a ranking of desirability of equally costly public projects. An interval scale of value can support additional policy-relevant situations, such as development of a damage schedule or the grouping of goods of similar value.

For some decisions, a preference ordering will not be adequate or appropriate, and the analyst must estimate the economic value of one or more goods. Several options are available, as described elsewhere in this book. In addition to those more thoroughly tested methods, the inclusion of monetary amounts among the items to be judged in a multiple good valuation study allows an estimation of monetary values for a set of goods. Multiple good economic valuation has the advantage of encouraging respondents to compare goods in terms of their characteristics and desirability, but the validity of this procedure has not been thoroughly tested.

NOTES

1 Only by estimating the economic benefits of the preferred option and comparing them with the cost can one know whether implementing it will improve aggregate social welfare. However, benefit-cost analyses are expensive and, therefore, are not always realistic. Indeed, they may be counter productive if the cost of the analysis is substantial compared to the cost of the project. For relatively small-scale decisions, such as allocating a modest increment in funding at a national forest, something short of an economic benefit analysis may be appropriate.

2 Strictly speaking, utility is an ordinal concept and has no cardinal expected value; all that matters is that the relative positions of the $E(U)$s are maintained, indicating the order of the items. However, it is convenient to describe U as made up of a central tendency and a disturbance, which requires an assumption of an interval scale measure of utility.

3 Thurstone thought of ε as normally distributed about mean V_n. Other distributional forms for ε are feasible and commonly assumed in modern discrete choice analysis (Amemiya 1981).

4 Inconsistent responses due only to momentary fluctuations in preference are not systematically repeatable, but inconsistent responses resulting from subtle context changes are potentially repeatable. At the individual respondent level, and assuming the respondent completes the exercise only once, inconsistency due to context is indistinguishable from completely random momentary fluctuations in preference. Context dependent inconsistency is detectable only by having a respondent repeat the exercise enough times, or by combining responses across enough respondents of homogenous preferences, to accurately measure the subject's or group's choices for each pair of items. As Tversky (1969) argues, detecting systematic intransitivity in paired comparisons requires a carefully designed study. Detection is complicated because intransitive responses, like transitive responses, are subject to momentary disturbances, as in equation 1.

5 Strictly speaking, variations in U that depend on measurement effort—or, for that matter, context—are not truly random. Rather, they have observable causes and are repeatable. However, from the standpoint of the analyst who fails to notice errors or define and measure the variables causing the variation, the convention is to assume randomness.

6 Economic demand modeling, of course, requires knowing the functional form of the relation of preference to the explanatory variables, and measuring those variables.

7 A five-item choice set is used for illustrative purposes only. In practice, ranking would be the more efficient method of ordering preferences for such a small set of items. However, when the choice set contains more than about ten items, ranking becomes less effective and other methods, such as rating and paired comparisons, must be considered.

8 A response matrix with no intransitive responses can be reordered so that the upper right triangle contains all 1s and the bottom left triangle contains all 0s. In Figure 2, the order of the items would be: $i < k < j < l < m$.

9 For example, if all choices reflected the preference order $m > l > k > j > i$ except for the choice between i and k, one circular triad results and the vector of preference scores contains one three-way tie among items i, j, and k. If the only exception is between i and l, two circular triads result and the vector of preference scores contains two two-way ties (one between i and j and the other between k and l). And if the only exception is between i and m, three circular triads result and the set of preference scores contains two two-way ties (one between items i and j and the other between items l and m).

10 Switching is also expected in the case where the respondent made a mistake (e.g., pushed the wrong key) in recording the original choice.

11 A preference score difference of 0, when it occurs, reflects the full set of choices made by the respondent. It does not necessarily mean that the respondent is indifferent between the two items. When the respondent is not indifferent, the likelihood of the respondent making different choices on different occasions, and thus of switching, is something less than 0.5.

12 Space does not allow us to fully develop this point here, but we have verified it via extensive testing using a model that simulates choices assuming the random utility function of equation 1. See also Torgerson (1958).

13 Thurstone's approach was only the first attempt to improve upon the use of aggregate preference scores to summarize paired comparisons. See David (1988) for other methods.

14 Such an observation is possible when monetary amounts are included among the items to be compared. A nonsensical choice would be one indicating a value for the good well above its market price. Although respondents may have a maximum willingness to pay in excess of the market price, they would not choose the good much above a sum of money equal to its market price because they could always choose the money and then buy the good.

15 A similar application for another Thai coastal area is found in Chuenpagdee, Knetsch, and Brown (2001b). Resource damaging activities as well as resource losses were assessed in this study.

16 A set of eight items is slightly smaller than we would now recommend, as mentioned in section 4.

17 An exception occurs if a respondent would choose any amount of money, no matter how small, over the good (precluding a monetary amount with a lower preference score), or would choose the good over any amount of money (precluding a monetary amount of higher preference score). In the former case, the lower bound is presumed to be $0 unless the good is in fact a bad.

18 Alternatively, for WTPc, if respondents were asked to choose the item they preferred to give up, the upper and lower bounds are the proximate sums of money of lower and higher preference score, respectively. And for WTAc, if respondents were asked to choose the item they preferred to forego, the upper and lower bounds are the proximate sums of money of lower and higher preference score, respectively.

19 As explained in Peterson and Brown (1998), after presenting the 155 paired comparisons, the computer program administering the experiment selected those pairs for which the respondent's choices were not consistent with his or her dominant preference order established from all 155 choices. The computer also randomly selected ten consistent pairs. These two sets of selected pairs were randomly ordered and presented again to the respondent, with no break in the presentation, so that the respondent was unlikely to notice when the 155 original pairs ended and the repeats began. For the originally inconsistent pairs, respondents switched many of their earlier choices, suggesting that respondents' earlier choices were mistakes or that their preferences became more defined in the course of considering the various pairs. When the original choices for the originally inconsistent pairs are replaced by the new choices, the number of circular triads drops dramatically, causing both the mean and median coefficient of consistency to rise to 0.99, compared with 0.92 and 0.94, respectively.

20 Discrete choice analysis relies on the assumption of independent and identically distributed error terms. This assumption was not met in the Peterson and Brown experiment, especially for the choices involving dollar amounts and private goods. As indicated in Figure 9, later in this section, the variances of the disturbance distributions for private goods were relatively narrow compared with those of the public goods.

21 Updating the choices for the originally inconsistent pairs, as described in footnote 19, had only a small effect on the estimated monetary values. For example, using the method of interpolation of individual vectors of preference scores, updating lowered the mean values an average of $6 per good, compared with those for the original choices reported in Table 4.

REFERENCES

Ben-Akiva, M., and S. R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand.* Cambridge, MA: Massachusetts Institute of Technology. 390 p.

Boyle, K. J., F. R. Johnson, and D. W. McCollum. 1997. Anchoring and Adjustment in Single-bounded, Contingent-valuation Questions. *American Journal of Agricultural Economics* 79(5):1495-1500.

Breffle, W. S., and R. D. Rowe. 2002. Comparing Choice Question Formats for Evaluating Natural Resource Tradeoffs. *Land Economics* 78(2):298-314.

Brown, T. C. 1984. The Concept of Value in Resource Allocation. *Land Economics* 60(3):231-246.

Brown, T. C., G. L. Peterson, A. Clarke, and A. Birjulin. 2001. Reliability of Individual Valuations of Public and Private Goods. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Brown, T. C., D. Nannini, R. B. Gorter, P. A. Bell, and G. L. Peterson. 2002. Judged Seriousness of Environmental Losses: Reliability and Cause of Loss. *Ecological Economics* 42(3):479-491.

Buhyoff, G. J., and W. A. Leuschner. 1978. Estimating Psychological Disutility from Damaged Forest Stands. *Forest Science* 24 (3):424-432.

Carson, R. T., T. Groves, and M. J. Machina. 1999. Incentive and Informational Properties of Preference Questions. Paper presented at the Meeting of the European Association of Resource and Environmental Economists, Oslo, Norway: June, 1999.

Chuenpagdee, R., J. L. Knetsch, and T. C. Brown. 2001a. Environmental Damage Schedules: Community Judgments of Importance and Assessments of Losses. *Land Economics* 77(1):1-11.

Chuenpagdee, R., J. L. Knetsch, and T. C. Brown. 2001b. Coastal Management Using Public Judgments, Importance Scales, and Predetermined Schedule. *Coastal Management* 29:253-270.

David, H. A. 1988. *The Method of Paired Comparisons.* Second ed. Vol. 41, *Griffin's Statistical Monographs and Courses.* New York, NY: Oxford University Press. 188 p.

Dunn-Rankin, P. 1983. *Scaling Methods.* Hillsdale, NJ: Lawrence Erlbaum Associates. 429 p.

Fechner, G. T. 1860. *Elemente der Psychophysik.* Leipzip: Breitkopf and Hartel.

Ferber, R. 1974. *Handbook of Marketing Research.* New York: McGraw-Hill. 1,358 p.

Freeman, III, A. M. 1993. *The Measurement of Environmental and Resource Values: Theory and Methods.* Washington, D.C.: Resources for the Future. 516 p.

Green, D., K. E. Jacowitz, D. Kahneman, and D. McFadden. 1998. Referendum Contingent Valuation, Anchoring, and Willingness to Pay for Public Goods. *Resource and Energy Economics* 20:85-116.

Guilford, J. P. 1954. *Psychometric Methods*. Second ed. New York: McGraw-Hill. 597 p.

Kendall, M. G., and B. B. Smith. 1940. On the Method of Paired Comparisons. *Biometrika* 31:324-345.

Knetsch, J. L. 1989. The Endowment Effect and Evidence of Nonreversible Indifference Curves. *American Economic Review* 79(5):1277-1284.

McFadden, D. 1974. Conditional Logit Analysis of Qualitative Choice Behavior. Pages 105–142 in *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press. 252 p.

Nunnally, J. C. 1976. *Psychometric Theory*. New York: McGraw Hill. 640 p.

Payne, J. W., J. R. Bettman, and E. J. Johnson. 1992. Behavioral Decision Research: a Constructive Processing Perspective. *Annual Review of Psychology* 43:87-131.

Peterson, G. L., R. L. Bishop, and R. M. Michaels. 1973. Children's choice of playground equipment: Developing methodology for integrating preferences into environmental engineering. *Journal of Applied Psychology* 58:233-238.

Peterson, G. L., and T. C. Brown. 1998. Economic Valuation by the Method of Paired Comparison, with Emphasis on Evaluation of the Transitivity Axiom. *Land Economics* 74(2):240-261.

Rutherford, M. B., J. L. Knetsch, and T. C. Brown. 1998. Assessing Environmental Losses: Judgments of Importance and Damage Schedules. *Harvard Environmental Law Review* 22:51-101.

Slovic, P. 1995. The Construction of Preference. *American Psychologist* 50(5):364-371.

Thurstone, L. L. 1927a. A Law of Comparative Judgment. *Psychology Review* 34:273-286.

Thurstone, L. L. 1927b. The Method of Paired Comparisons for Social Values. *Journal of Abnormal Social Psychology* 21:384-400.

Thurstone, L. L. 1927c. Psychophysical Analysis. *American Journal of Psychology* 38:368-389.

Torgerson, W. S. 1958. *Theory and Methods of Scaling*. New York, NY: John Wiley & Sons. 460 p.

Tversky, A. 1996. Contrasting Rational and Psychological Principles in Choice. In *Wise Choices: Decisions, Games, and Negotiations*, edited by R. J. Zeckhauser, R. L. Keeney and J. K. Sebenius. Boston, MA: Harvard Business School Press.

Tversky, A., S. Sattath, and P. Slovic. 1988. Contingent Weighting in Judgment and Choice. *Psychological Review* 95(3):371-384.