

Discussion Papers report work in progress. They are written to share ideas, concepts, and theories as well as preliminary empirical data, and have not been peer reviewed or approved for publication. Comments are welcome.

## RMRS-RWU-4851 Discussion Paper

### **Paired comparisons of public and private goods, with heteroscedastic probit analysis of choice consistency**

**Thomas C. Brown,<sup>a</sup> David Kingsley,<sup>b</sup> George L. Peterson,<sup>c</sup>  
and Nicholas E. Flores<sup>d</sup>**

Rocky Mountain Research Station, U.S. Forest Service  
Fort Collins, Colorado 80526

10 October 2007

<sup>a</sup> Tom Brown is an economist at the Rocky Mountain Research Station in Fort Collins. He can be reached at: [thomas.brown@colostate.edu](mailto:thomas.brown@colostate.edu)

<sup>b</sup> Dave Kingsley was at the University of Colorado when working on this paper. He is now an assistant professor at Westfield State College, Westfield MA 01086. He can be reached at: [dkingsley@wsc.ma.edu](mailto:dkingsley@wsc.ma.edu).

<sup>c</sup> George Peterson retired in 2004 after 22 years with the Rocky Mountain Research Station. Previously he was a full professor at Northwestern University. He can be reached at: [glpskibum@aol.com](mailto:glpskibum@aol.com)

<sup>d</sup> Nick Flores is Chairman of the Economics Department at the University of Colorado, Boulder. He can be reached at: [floresn@stripe.colorado.edu](mailto:floresn@stripe.colorado.edu).

## Abstract

We examined the reliability of a large set of paired comparison value judgments involving public goods, private goods, and sums of money. The consistency of the respondents' choices was examined in three ways: computation of the coefficient of consistency, using a simple preference-score-based approach that attempts to isolate individual inconsistent choices, and with heteroscedastic probit analysis. As respondents progressed through a random sequence of paired choices they were each given, their response time decreased and they became more consistent, apparently fine-tuning their responses, suggesting that respondents tend to begin a hypothetical value exercise with relatively imprecise preferences and that experience in expressing preference helps reduce that imprecision. Consistency was greater for private than for public good choices, and greater for choices between a good and a monetary amount than for choices between two goods. However, the consistency for public good choices was only slightly lower than for the private goods.

## Table of Contents

|   |    |
|---|----|
| Introduction.....   | 3  |
| Conceptual Model.....                                     | 4  |
| Methods.....  | 6  |
| The Data .....  | 6  |
| Data Analysis.....  | 7  |
| Results.....  | 9  |
| Coefficient of Consistency .....                          | 9  |
| Change in Consistency with Sequence.....                  | 9  |
| Response Time .....                                       | 10 |
| Inconsistency and Preference Score Difference .....       | 10 |
| Separating Public and Private Goods .....                 | 11 |
| Choice Switching.....                                     | 11 |
| Discussion.....   | 12 |
| Appendix. Heteroscedastic Probit Analysis of Results..... | 14 |
| Method.....   | 14 |
| Results .....   | 15 |
| References.....   | 16 |
| Tables.....   | 18 |
| Figures.....  | 20 |

## Paired comparisons of public and private goods, with heteroscedastic probit analysis of choice consistency

### Introduction

How precisely do people know their preferences? It is common in applications of utility theory to assume that people know their preferences perfectly. In modeling people's choices we therefore assume that error in estimation is attributable to missing variables or errors in measurement. But is this a reasonable assumption, especially when the preferences are about unfamiliar goods?

According to McFadden (2001), Thurstone (1927) was the first to propose a choice model that allowed for, indeed expected, errors in human judgment and preference. Thurstone had the advantage of years of prior research by fellow psychologists into people's paired judgments of physical stimuli, such as the weights of objects or the loudness of sounds, which demonstrated that people's accuracy in judging the relative magnitudes of such stimuli decreased as the difference between the stimuli lessened (Brown & Peterson, 2003). He extended this finding to the domain of preferences (e.g., Thurstone & Chave, 1929) and developed a theory and method for estimating the relative magnitudes of the stimuli, using the amount of inconsistency in preference as an indication of the closeness of the stimuli.

The possibility of error—or, more precisely, imprecision or uncertainty—in individual preference as proposed by Thurstone did not go unnoticed over the years. As McFadden explains, Marschak (1959) brought the idea to the attention of economists. Although economists' adaptation of the random utility notion focused on exogenous sources of error, endogenous error has been mentioned occasionally over the years by economists (e.g., Hausman & Wise, 1978) and is gradually becoming more accepted (e.g., Ben-Akiva & Lerman, 1985; Bockstael & Strand, 1987; Li & Mattsson, 1995; Rieskamp et al., 2006).

Survey methods such as contingent valuation or conjoint (multi-attribute) analysis have emerged as primary methods for estimating the economic value of public goods. With these methods we rely on people's responses to questions about their willingness to pay (WTP) for a certain good or on their choices among goods that are each available at a price. These methods were initially used to value quasi-private nonmarket goods, such as an individual recreation opportunity on public land, or, in the case of conjoint analysis, consumer goods, but the methods have been extended to value public goods such as air quality protection and wilderness preservation. Many contingent valuation studies (see Carson et al., 1994) and several conjoint studies have valued public goods. However, the extension to public goods incurs potential problems related to, among other things, respondents' lack of familiarity with purchasing such goods.

The question about how well people know their preferences has been addressed, within contingent valuation, in part via reliability studies. Some of these studies used the test-retest approach, where a sample is asked the same question on two different occasions. Time periods between tests have varied from a few weeks to several years, and both public and private goods have been studied. In all cases significant correlations were found between the two occasions; most correlations fell in the 0.5 to 0.8 range (see Jorgensen et al., 2004; McConnell et al., 1998 for summaries of these studies). Other studies used different samples at different times and

compared estimates of mean WTP (Jorgensen et al., 2004; McConnell et al., 1998 list some of these), generally finding no significant difference between the estimates.

These studies provide a general sense of confidence in the contingent valuation method, but they do not deal directly with the question posed above, about the precision with which a given subject is able to respond to a WTP question. They also do not help us understand whether subjects learn about their preferences in the course of forming an answer. The studies are unable to address these questions because they ask each respondent only one or a few monetary valuation questions. Our methods, because we ask each respondent a large number of valuation questions, help answer these questions. Most importantly, our methods allow us to observe and test for changes in the consistency of preferences as respondents proceed through the multiple valuation questions they are given.

Despite the fact that the reliability of stated WTP has been found to be adequate in the studies referred to above, it has long been thought that reliability would be substantially less for public than for private goods (Cummings et al., 1986), principally because public goods lack market prices and people have less experience valuing them. Only one study, by Kealy et al. (1990), tested this conjecture. The authors obtained WTP estimates for a private good and a public good from separate samples, and then replicated those estimates with the same two groups two weeks later. They found that the test-retest reliability of the estimate for the public good was only slightly less than, and not significantly different from, that of the private good estimate, thereby rejecting the Cummings et al. conjecture. We took a second look at this issue by including both public and private goods in our surveys.

Our approach uses the paired comparison (PC) method, which dates back to Fechner (1860) and has been studied extensively (e.g., Bock & Jones, 1968; David, 1988; Thurstone, 1927; Torgerson, 1958). The method yields an individual respondent's preference order among items of a choice set by presenting the items in pairs and asking respondents to choose the item in each pair that best meets a given criterion, such as being the more preferred. Importantly for our purposes here, the method yields individual estimates of reliability.

The PC method asks each respondent to make numerous binary choices. When all possible pairs of the items are presented and the number of items is large enough that respondents have difficulty recalling past choices, the method offers abundant opportunities for inconsistency in preference, which provide our first measure of reliability. In addition, we repeated some of the pairs at the end of the session, offering a short-term test-retest measure of reliability. We report on these two approaches for assessing reliability for public and private good choices, and also examine how response time and consistency of preference changes over the course of a session and how response time varies by type of choice (public or private good) and by consistency of the choice.

## **Conceptual Model**

As proposed by Thurstone (1927) in his exposition on paired comparisons, preferences are probably best described by a stochastic function. This function is now commonly known as a random utility function in recognition of the belief that the true utilities of the items are the expected values of preference distributions. The random utility function ( $U$ ) consists of

deterministic ( $V$ ) and random ( $\varepsilon$ ) components.<sup>1</sup> For example, the utility of items 1 and 2 to respondent  $i$  on choice occasion  $j$  can be represented by the following two relations:

$$\begin{aligned} U_{ij1} &= V_{i1} + \varepsilon_{ij1} \\ U_{ij2} &= V_{i2} + \varepsilon_{ij2} \end{aligned} \tag{1}$$

The randomness signified by  $\varepsilon$  in the current application is inherent in the individual choice process, which is subject to fluctuations and disturbances that are beyond our ability to model deterministically. Among a respondent's choices in a PC exercise, this variation has at least three potential sources. First, choices are subject to measurement error, which occurs, for example, when a respondent mistakenly pushes the wrong key or checks the wrong box to record a choice. Second, preferences may be subject to momentary fluctuations resulting from preference imprecision (Thurstone, 1927). Third, preference for multi-attribute items may vary with the mix of attributes of the pair of items being compared. This may occur, for example, when respondents weight attributes differently when making different comparisons (Tversky, 1969), make different assumptions about incompletely described items when making different comparisons, or have multiple objective functions (Hicks, 1956) and focus on different objectives when making different comparisons.

The probability of an individual selecting item 1 in a comparison of items 1 and 2 is given by:

$$P(U_{ij1} > U_{ij2}) = P(V_{i1} + \varepsilon_{ij1} > V_{i2} + \varepsilon_{ij2}) \tag{2}$$

This probability  $P$  is greater the larger is  $V_1$  B  $V_2$  and the narrower are distributions of  $\varepsilon_1$  and  $\varepsilon_2$ . An inconsistent response occurs when  $U_{ij2} > U_{ij1}$  although  $V_{i1} > V_{i2}$ , and can happen when the distributions of  $\varepsilon_{ij1}$  and  $\varepsilon_{ij2}$  about their respective  $V$ s overlap.

We suspect, in line with the Cummings et al. (1986) conjecture, that  $\varepsilon$  will tend to be wider when the items lack market prices and when respondents have little experience valuing them. Because these conditions are more often the case with public than with private goods, we hypothesize that reliability will be lower for public good choices than for private good choices. Similarly, we hypothesize that respondents will take more time to make public than private good choices and that inconsistent choices will take more time than consistent choices.

In each PC session, respondents made over 100 choices among all possible pairs of the items. As respondents work through the sequence of choices they may fine-tune their preferences. Perhaps the most likely change with sequence is a narrowing of the disturbance term as respondents become more settled in their preferences. We hypothesize such a narrowing, and thus a drop in inconsistency with sequence.

---

<sup>1</sup> Thurstone thought of  $\varepsilon$  a normally distributed about mean  $V$ , leading to a random utility function that is now characterized using the binary probit model. Other distributional forms for  $\varepsilon$  are feasible and commonly assumed in modern discrete choice analysis (Ben-Akiva & Lerman, 1985).

## Methods

For this paper we combined data from three studies. Each study focused on specific valuation issues apart from the reliability concerns of this paper, but all three studies used the same basic procedure for eliciting PC responses, enabling us to aggregate the data to evaluate two measures of respondent reliability, one based on isolation of inconsistent choices among the original choices and the other relying on a retest of the original choices. In this section, the PC procedure and the data it provides are described and our analysis procedures are summarized.

### *The Data*

The three studies that provided the data for this paper (Table 1) each obtained judgments for a mixture of public goods, private goods, and amounts of money. Some goods were used in more than one study, but most were unique to a specific study. Across the three studies, a total of 979 respondents, all university students, provided PC judgments. All public goods were locally relevant, involving the university campus or its surrounding community. All private goods were common consumer items with specified prices. In total, these three studies yielded 129,984 respondent choices, each between a pair of items.

As an example of the methods of these three studies, we summarize the methods used by Peterson and Brown (1998). The choice set consisted of six public goods, four private goods, and eleven sums of money (21 items in total).<sup>2</sup> Each respondent made 155 choices consisting of 45 choices between goods and 110 choices between goods and sums of money. They did not choose between sums of money (it was assumed that larger amounts of money were preferred to smaller ones), but did choose between all other possible pairs of the items. The 327 respondents yielded a total of 50,685 binary choices.

The four private goods of the Peterson and Brown (1998) study were: a meal at any local restaurant, not to exceed a cost of \$15; a \$200 certificate useable at any local clothing store; a \$500 certificate for air travel on any airline; and two tickets for one of four listed events (e.g., a professional football game) with an estimated value of \$75. The six public goods were of mixed type. Two were pure public environmental goods, in that they were nonrival and nonexcludable in consumption. One was the purchase by the university of 2,000 acres of land in the mountains west of town as a wildlife refuge for animals native to the area; the other focused on clean air and water. The remaining four public goods were excludable by nature but stated as nonexcludable by policy, and were nonrival until demand exceeds capacity. One was a no-fee library service that provides video tapes of all course lectures; the others involved campus parking capacity, a campus music festival, and student dining facilities. The eleven sums of money were \$1, \$25, \$50, \$75, and \$100 to \$700 in intervals of one hundred. These amounts were derived from pilot studies in order to have good variation and distribution across the values of the goods.

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 respondent simply chose the preferred item in each pair. If respondents were indifferent between the two items in a pair, they were still

---

<sup>2</sup> Complete lists of the goods are found in chapter 2 of the Discussion Paper “An enquiry into the method of paired comparison: Reliability, scaling, and Thurstone’s law of comparative judgment” available at <http://www.fs.fed.us/rm/value/discpapers.html>.

asked to make a choice; indifference across respondents was later revealed as an equal number of choices of each item.<sup>3</sup>

The surveys were administered on personal computers that presented the pairs of items on the monitors in random order for each respondent to control for order effects. The items 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, as long as a subsequent choice had not yet been entered, could correct a mistake by pressing the backspace key and then selecting the other item. Review of prior choices was not possible. At the end of the original paired comparisons, the computer presented some of the pairs a second time, without a break in the presentation and without a prior announcement that some pairs would be repeated. The pairs presented for retrial were of two kinds, those pairs for which the individual's choice was not consistent with the dominant preference order as defined by the preference scores (see the definition of preference score below), and ten randomly selected consistent pairs. The individual pairs in these two sets of repeated choices were randomly intermixed. The computer also recorded the time taken to enter each choice. Respondents were not told that response time was recorded.

### *Data Analysis*

Given a set of  $t$  items, the PC method presents them independently in pairs as  $(t/2)(t-1)$  discrete binary choices. These choices yield a *preference score* for each item, which is the number of times the respondent prefers that item to other items in the set. 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 21-item choice set, an individual preference score vector with no circular triads contains all 21 integers from 0 through 20. Circular triads (i.e., choices that imply  $A > B > C > A$ ) cause some integers to appear more than once in the preference score vector, while others disappear.

For a given respondent, a pair's *preference score difference* (PSD) is simply the absolute value of the difference between the preference scores of the two items of the pair. This integer, which can range from 0 to 20 for a 21-item choice set, indicates on an ordinal scale the difference in value assigned to the two items.

The number of circular triads in each individual's responses can be calculated directly from the preference scores. The number of items in the set determines the maximum possible number of circular triads. The individual respondent's *coefficient of consistency* is calculated by subtracting the observed number of circular triads from the maximum number possible and

---

<sup>3</sup> Providing subjects with an indifference option may have its advantages. If it worked well, it would allow us to separate real indifference from other sources of inconsistency. However, providing an indifference option runs the risk of allowing subjects to avoid all close calls. If, as Torgerson (1958) argues, the probability of true indifference at any given time is "vanishingly small," forcing a choice maximizes the amount learned while still allowing indifference to be revealed in data from multiple subjects or from a given subject responding multiple times.

dividing by the maximum.<sup>4</sup> 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.

When a circular triad occurs, it is not unambiguous which choice is the cause of the circularity. This is easily seen by considering a choice set of three items, whose three paired comparisons produce the following circular triad: A>B>C>A. Reversing any one of the three binary choices removes the circularity of preference; selection of the one to label “inconsistent” is arbitrary. However, with more items in the choice set, selection of *inconsistent choices* (i.e., choices where  $U_{ij2} > U_{ij1}$  although  $V_{i1} > V_{i2}$ ), though still imperfect, can be quite accurate. For each respondent, we selected as inconsistent any choice that was contrary to the order of the items in the respondent's preference score vector, with the condition that the order of items with identical preference scores was necessarily arbitrary. Simulations show that the accuracy of this procedure in correctly identifying inconsistent choices increases rapidly as the PSD increases. In simulations with a set of 21 items and assuming normal dispersion distributions, the accuracy of the procedure rises quickly from 50% at a PSD of 0 to nearly 100% at a PSD of 5. Simulations also show that the procedure is unbiased and thus can be used to compare consistency across sets of choices, such as sets representing different classes of goods or different points along the sequence of choices (e.g., first pair versus second pair).<sup>5</sup>

The proportion of choices that were selected as inconsistent by the double-sort procedure, across all respondents, provides one of our three measures of reliability. We computed this measure for all choices together, and then for three partitions of the data. First, the measure was computed for each choice in the order the choices were made (i.e., it was computed for all the first choices, all the second choices, etc.). Plotting the proportion inconsistent for each choice, with the choices in sequence, shows how inconsistency changes as respondents gain more experience with the choice task and the items being compared. Because the pairs were presented in a unique random order to each respondent, this measure is independent of pair order. Second, the measure was computed for the five different types of choices: private good versus money, public good versus money, private good versus private good, public good versus public good, and private good versus public good. Third, the measure was computed for each level of PSD to show how inconsistency changes as the choices become easier to make. We performed numerous statistical tests on these measures of inconsistency, either to compare results for public goods with those for private goods, or to evaluate trends over time. Proportions were compared using a test based on the normal approximation to the binomial distribution. Trends were evaluated using linear regression. For all tests, we used a 0.05 probability level to test for significance.

---

<sup>4</sup> The maximum possible number of circular triads,  $m$ , is  $(t/24)(t^2-1)$  when  $t$  is an odd number and  $(t/24)(t^2-4)$  when  $t$  is even, where  $t$  is the number of items in the set. Letting  $a_i$  equal the preference score of the  $i^{\text{th}}$  item and  $b$  equal the average preference score (i.e.,  $(t-1)/2$ ), the number of circular triads is (David, 1988):

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

The coefficient of consistency (Kendall & Smith, 1940) is then defined as:  $1 - c/m$ .

<sup>5</sup> A thorough explanation of the procedure for specifying inconsistent choices, called the double-sort procedure, is found in chapter 4 of the Discussion Paper “An enquiry into the method of paired comparison: Reliability, scaling, and Thurstone's law of comparative judgment” available at <http://www.fs.fed.us/rm/value/discpapers.html>.

Because our procedure is uncommon, we also used a more well-recognized method, a heteroscedastic probit implementation of the random utility model, to evaluate the reliability of respondents' choices. The probit analysis uses the variance of the disturbance term ( $\varepsilon$ , equation 1) as the measure of inconsistency (DeShazo and Fermo, 2002; Swait and Adamowicz, 1996). This approach confirmed the findings obtained based on our simple decision rule for isolating inconsistent choices, both in examination of changes in inconsistency with sequence and in comparing inconsistency of public and private good choices. The details of the heteroscedastic probit analysis are found in the Appendix.

Finally, our third measure of reliability uses the repeats of originally consistent choices. (For completeness, we also present the data on the repeats of inconsistent pairs.) *Choice switching* (i.e., choosing one item initially and the other item upon retrieval) when an originally consistent choice was made indicates a lack of reliability.

## Results

Results from the original choices are presented in subsections 4.1 – 4.5, followed by results from the repeated choices in subsection 4.6. All results shown in figures or reported in tables are based on the full set of 979 respondents.

### *Coefficient of Consistency*

Coefficients of consistency were computed for each respondent. The mean coefficients of the three studies range from 0.908 to 0.915, and the median coefficients range from 0.927 to 0.939. In each set, the median exceeds the mean, as the means are sensitive to the relatively low coefficients of a minority of respondents in each set (Figure 1). The overall median coefficient is 0.93; 95% of the respondents had a coefficient of at least 0.77.

### *Change in Consistency with Sequence*

Across the three studies, 7.2% of the choices were inconsistent with respondents' dominant preference orders as determined using the double-sort procedure. To examine how this inconsistency varies over time, the proportion of choices that were inconsistent was computed for each choice in sequence (i.e., all first choices, all second choices, etc.). As Figure 2 shows, inconsistency drops from 15% for the first choice to about 7% by the 30<sup>th</sup> choice, and then drops only slightly more after that (all three studies show the same pattern). The drop in inconsistency that occurs over the first 30 pairs is highly significant ( $df = 29$ ,  $F = 116.06$ ,  $p < 0.01$ ), but the slight drop from the 31<sup>st</sup> to the 100<sup>th</sup> choice is not significant ( $df = 29$ ,  $F = 2.95$ ,  $p = 0.09$ ).<sup>6</sup> A minimum level of inconsistency, which is roughly 6.5% for these data (Figure 2), is always expected, largely because some of the choices are between items of roughly equal desirability (i.e., they are close calls). Because each respondent encountered the choices in a unique random order, Figure 2 is not dependent on the nature of the particular items that were first encountered. The drop in inconsistency over time suggests that respondents were fine-tuning or firming up the

---

<sup>6</sup> Figures of choices by sequence show only the first 100 choices, which is sufficient to depict the trend and avoids including later choices, some of which are based on only a subset of the respondents (as shown in Table 1, the total number of choices varied by study). In any case, we found no trends in proportion inconsistent past the 100th choice (e.g., no evidence of fatigue, which could cause an increase in inconsistency).

preferences with which they began the PC exercise.

The changing consistency with sequence depicted in Figure 2 suggests that respondents were narrowing the magnitudes of the random components of their preference distributions ( $\epsilon$ ) over the course of the session, but it does not eliminate the possibility that respondents' preferences ( $V$ ) were also changing with sequence. The sample sizes of the three studies are insufficient for computing average preference scores for each item for each choice in the sequence of choices, but we can compare  $V$ s from sets of choices, such as sets of early versus late choices. A comparison of the  $V$ s estimated from the first 30 choices with those estimated from choices 71-100 produced correlations of early versus late  $V$ s ranging from 0.96 to 0.99 across the three studies, indicating very little shift in preferences with sequence, thus supporting the claim that only the disturbance terms were changing over the course of the sessions.

### *Response Time*

Mean response (decision) times for the first six choices were 29, 10.0, 8.0, 6.7, 6.3, and 6.0 seconds. Response time continued to drop at a decreasing rate until about halfway through the choices, when it stabilized at about 2.4 seconds. Response times were longer for inconsistent than for consistent choices. Indeed, for every one of the first 100 choices, inconsistent choices took more time on average than did consistent choices (Figure 3), a result that is extremely unlikely to occur by chance alone. On average, consistent choices took 3.4 seconds and inconsistent choices took 4.7 seconds.

### *Inconsistency and Preference Score Difference*

Figure 4 shows that inconsistency decreases rapidly with PSD.<sup>7</sup> Fifty percent of the choices are inconsistent at a zero PSD, as expected. Inconsistency drops to 1% by a PSD of 8. Seventy-two percent of all inconsistent choices occurred at a PSD  $\leq 2$ . The fact that inconsistency drops to near zero for choices of high PSD indicates that mistakes, which would not be restricted to choices of low PSD, are not a major cause of inconsistent choices. The fact that 72% of the inconsistent choices occur at a PSD  $\leq 2$  suggests that most inconsistent choices result from indifference or near indifference.

Further, mean response time dropped monotonically with mean PSD, from 5 seconds at PSD = 0 to about 2 seconds at PSD = 20, indicating that people labor more over close calls than obvious ones.<sup>8</sup> That close calls are the most difficult ones is, of course, not news to anyone who has fretted over a choice between two equally good options.

Although most inconsistent choices involve close calls, a substantial amount (28%) of the inconsistent choices occurred at PSD  $> 2$ , indicating a rather high degree of imprecision. The data show that many of these inconsistent choices were produced by a minority of respondents. For example, one-half of the inconsistent choices of PSD  $> 2$  were produced by the 20% of the respondents with the lowest coefficients of consistency, as might be anticipated from Figure 1.

---

<sup>7</sup> Because the Clarke (1999) data set includes only 18 items, the contribution to Figure 4 from that data set is limited to PSDs of from 0 to 17.

<sup>8</sup> Response time increases not only with the closeness of the items being compared. In a related study, of choices between lotteries, Wilcox (1993) found that decision time increased with the complexity of the lotteries.

The finding that over one quarter of the inconsistent choices involve pairs with a rather large PSD, and that mistakes account for very few of the inconsistent choices, raises the question: when do these high PSD inconsistent choices occur? Are they sprinkled evenly over the sequence of pairs, or are they concentrated among the first few pairs? As shown in Figure 5, inconsistent choices of higher PSD are more common early in the sequence. The mean PSDs of the first few pairs are near 3, mean PSD drops over the first 30 pairs or so, and after the 30<sup>th</sup> pair the mean PSDs are nearly all below 2, averaging 1.73. Regression shows that the slope of the straight line fitted for first 30 points is significantly negative ( $df = 29$ ,  $F = 52.54$ ,  $p < 0.01$ ), and that the subsequent points show no slope ( $df = 69$ ,  $F = 1.65$ ,  $p = 0.20$ ). If we accept the conceptual model of the choice process represented by equation 1, this finding suggests again that respondents begin the exercise with relatively imprecise values (i.e., relatively large disturbance terms  $\varepsilon$ ) and that the disturbance gradually narrows as respondents gain experience with the items in the course of comparing them.

Also shown in Figure 5 are the mean PSDs of the consistent choices. These mean PSDs are considerably larger than those of the inconsistent choices, averaging 6.78 across the first 100 pairs in comparison to 1.90 for the inconsistent choices. Mean PSD is larger for consistent choices because choices for pairs with a large PSD tend to be consistent. There is no trend in mean PSD over sequence for the consistent choices.

### *Separating Public and Private Goods*

Separating the public good choices from the private good choices shows that both classes of goods follow the pattern shown in Figure 2, but that the degree of inconsistency is slightly greater for public good than for private good choices (Figure 6). The public good choices are those involving only public goods or public goods and dollars, and the private good choices are those involving only private goods or private goods and dollars (thus, this breakdown ignores the choices comparing a public good with a private good). For 62 of the first 100 choices, public good choices were less consistent than private good choices (0.62 is significantly greater than 0.5). The greatest differences occur among the first 14 choices, but the tendency for inconsistency to be greater for public than for private goods persists throughout the sequence of choices. Ignoring sequence, 7.0% of the 64,946 public good choices and 6.5% of the 43,394 private good choices were inconsistent (the two percentages are significantly different).

Response times were also longer for public good choices, taking an average of 3.5 seconds compared with 3.3 seconds for private good choices. Although the difference is small, it was persistent; for example, mean public good response times were longer than mean private good response times for 86 of the first 100 choices.

Comparison of consistency across all five types of choices (Table 2) yields the following two observations. First, choices involving monetary amounts were the most consistent. Second, the greater inconsistency for public good choices than for private good choices, shown in Figure 6, holds for choices involving money (the 6.7% is significantly greater than the 6.2% in Table 2) but not for choices involving only goods (the 8.4% is not significantly smaller than the 8.6%).

### *Choice Switching*

Recall that respondents repeated inconsistent choices and a random sample of ten consistent choices after they completed the initial set of paired comparisons. We compare choices for these

repeated pairs with the choices made when the pairs were first presented. Across the three studies, 9,424 originally consistent choices were repeated, and 817 (8.7%) were switched on retrial. Similarly, 9,312 choices were originally inconsistent, and 5,524 of those (59%) were switched on retrial. Thus, the overall impression is that respondents tended to correct inconsistencies when repeating inconsistent pairs and apply consistent decision criteria when repeating consistent choices.

Figure 7 shows the proportion of choices switched on retrial as a function of the sequence in which the choices were originally encountered for the first 100 choices. For both originally inconsistent choices and originally consistent choices, the proportion switched drops with sequence. This drop is obvious in the figure for inconsistent choices, but it also exists for consistent choices, as separate regressions produced significant negative slopes for both sets of choices (for inconsistent choices  $df = 99$ ,  $F = 121.17$ ,  $p < 0.01$ ; for consistent choices  $df = 99$ ,  $F = 17.91$ ,  $p < 0.01$ ). The negative slopes indicate that switching becomes less likely the more recently the original choice was made. Assuming independence between original and repeated choices, this finding suggests that respondents become more fixed in their preferences as they proceed through the exercise.<sup>9</sup>

Comparison of switching behavior of public good and private good choices provides another look at the relative reliability of such choices. As with Figure 6, the public good choices are those involving only public goods or public goods and dollars, and the private good choices are those involving only private goods or private goods and dollars. On average, 30.2% of the private good choices and 34.4% of the public good choices were switched on retrial, a small but significant difference.

## Discussion

We examined the reliability of a large set of PC choices involving public and private goods. The private goods included a variety of consumer items, and the public goods included both pure and congestible public goods. There are five principal findings.

First, inconsistency, resulting in intransitivity, is common in binary choices, whether comparing goods with other goods or goods with money amounts. Inconsistency varies substantially both across choices for a given respondent and across respondents. Inconsistency across choices is perhaps bothersome to the economic valuation practitioner but it is also useful, for it indicates the distance between the items being compared. Across respondents the degree of inconsistency varies from those who are remarkably consistent to those few who are very inconsistent. These are not new findings, but they add weight to the claim that preference imprecision or uncertainty is a legitimate source of error within the random utility maximization model, and that the standard assumption in utility theory that the consumer knows her preferences precisely is unrealistic.

---

<sup>9</sup> Based on the change with sequence found for the original choices (see Figs. 2-4), we must question the assumption of independence between original and repeated choices. Relaxing the independence assumption allows for the possibility that memory affected the retrial choices, if we also assume that more recent original choices are more likely to be remembered than earlier ones. Although this possibility cannot be ignored, the observation (Figure 7) that respondents typically reversed previously inconsistent choices (i.e., they did not simply remember and repeat the earlier choice) suggests that reliance on memory was at least not a dominant strategy during the retrials.

Second, response time varies systematically across types of choices, being longer for early as opposed to later choices, inconsistent as opposed to consistent choices, and close calls as opposed to more obvious choices. If we accept response time as an indication of choice difficulty, we have some evidence that difficulty increases with closeness of the items in the pair, and that choices become less difficult with experience choosing among the goods.

Third, reliability improved dramatically over the course of the first 30 or so choices. Further, we found no evidence that values ( $V$ s) were changing. Respondents were apparently firming up their preferences as they considered additional choices involving the same items. In terms of equation 1, the variances of the disturbance distributions ( $\varepsilon$ ) diminished with sequence. This finding supports the hypothesis that most respondents begin a hypothetical value exercise with relatively imprecise preferences about the items presented and that experience in expressing preference about the items helps to reduce that imprecision. This increasing precision is consistent with the notion of “value learning,” a component of preference learning, proposed by Braga and Starmer (2005), but only in the limited sense of firming up of preferences, not in the sense of changing preferences.<sup>10</sup> The evidence implies that a valuation study that relies on only one choice per respondent, as is common in contingent valuation, for example, may be unduly limiting the reliability of its value estimate.

Increasing precision of preference (i.e., narrowing of  $\varepsilon$ ) as respondents become more familiar with the items being compared is not the only explanation for the observed improvement in consistency with sequence. It is also possible that respondents gradually develop simplifying rules for making difficult choices and then use those rules for all choices except those for which there are overwhelming reasons to ignore the rule. One example of such a rule would be that people tend to cue on specific attributes. For example, a person might tend to choose a public good over a private good whenever the values of the two goods are similar. We suspect that some people do use simplifying rules, but we have no way of knowing how common such behavior is. Future research should strive to determine the relative importance of the different possible explanations for the improving consistency with experience choosing among the items.

The fourth finding is that reliability was generally greater for private good than for public good choices, and generally greater for choices involving a monetary amount than for choices comparing two goods. This is in contrast to the study by Kealy et al. (1990), who found a similar but insignificant difference for private versus public goods. We hypothesize that our findings reflect the degree to which people have experience trading such items. Monetary amounts were the most commonly traded items of the choice sets, and private goods are more commonly traded than public goods. Further, the private goods had specified prices, which may help to anchor their values. Lack of experience valuing an item may tend to widen the distribution of its value, leading to more close calls, and thus to more circular triads, inconsistent choices, and switches

---

<sup>10</sup> Our findings on value learning bring to mind the controversy regarding the learning about one’s preferences that occurs over repeated trials of some choice task (Braga and Starmer, 2005, cite many of the relevant papers). Some authors adhere to the *discovered* preference hypothesis proposed by Plott (1996), which maintains that stable (i.e., context independent) underlying preferences are discovered through sufficient repetition of a choice task that provides relevant feedback. Others counter, especially when dealing with unfamiliar goods, that labile (i.e., context dependent) preferences are *constructed* in the course of choosing (Gregory et al., 1993), and that repetition of the choice task in the same context will not alter the effects on preference of contextual cues. Unfortunately, because the task we presented to respondents involved no consequences and feedback, and because we did not specifically test for the effect of contextual cues, we can offer no insight on the controversy.

on retrieval.

Finally, the reliability for public good choices was only *slightly* lower than for the private good choices. Although we found significant differences in consistency between the two classes of goods, those differences were small. Based on this and earlier evidence, reliability of public good choices does not appear to be a major concern. The validity of the estimates of economic value of public goods is a separate issue.

## Appendix. Heterscedastic Probit Analysis of Results

Analyzing the choices using a heteroscedastic probit model provides a check on the results obtained using the double-sort procedure. We examined both changes in consistency with sequence and differences between public and private goods.

### *Method*

The probit analysis uses the variance of the disturbance term ( $\varepsilon$ , equation 1) as the measure of inconsistency (DeShazo and Fermo, 2002; Swait and Adamowicz, 1996). As the data are arranged in rows ( $r$ ) and columns ( $c$ ),  $P_{rc}$  represents the probability that the row item is chosen over the column item;  $P_{cr}$  represents the reverse. The disturbances across individuals are assumed to be independent and identically distributed (i.i.d.) normal random variables with a mean of zero and a constant variance  $\sigma_\varepsilon^2$ , and heteroscedastic across choice occasion  $j$ . They are also assumed to be uncorrelated with  $V$ . Given these assumptions:

$$Var(\varepsilon_1 - \varepsilon_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2 = 2\sigma^2 \quad (3)$$

Particular attention was paid to the standard deviation of the disturbance term,  $\sigma_\varepsilon(j)$ , and to the types of goods (public or private) involved in the choice.

The probability contribution to the likelihood function can now be constructed. For the choice between two goods, the probability of choosing the good in row  $r$  over the good in column  $c$  is:

$$P_{rc} = P(U_{ijr} > U_{ijc}) = P(V_r + \varepsilon_{ijr} > V_c + \varepsilon_{ijc}) = P(\varepsilon_{ijr} - \varepsilon_{ijc} > V_c - V_r) = \Phi(V_r - V_c / \sqrt{2}\sigma_\varepsilon(j)) \quad (4)$$

where  $\Phi$  is the standard normal cumulative density function and  $\sqrt{2}\sigma_\varepsilon(j)$  is the standard deviation of  $\varepsilon_{ijr} - \varepsilon_{ijc}$ . Similarly, when the item in column  $c$  is a monetary amount  $x$  the probability of choosing the good is:

$$P_{rc} = P(U_{ijr} > U_c) = P(V_r + \varepsilon_{ijr} > x_c) = P(V_r - x_c > \varepsilon_{ijr}) = \Phi(V_r - x_c / \sigma_\varepsilon(j)) \quad (5)$$

The heteroscedastic probit is assumed to change over choice occasion  $j$ ; thus the standard deviation, or scale, is in general  $\sigma_\varepsilon(j) = \lambda + \beta(1/j)$ . The scale is different depending whether the choice involves public or private goods. For a choice between either two public goods ( $b$ ) or a public good and a dollar amount the scale is  $\sigma_\varepsilon(j) = (\lambda_b + \beta_b(1/j))$ . Similarly, for a choice

involving private goods the scale is  $\sigma_\varepsilon(j) = (\lambda_p + \beta_p(1/j))$ . These specifications assume that i.i.d. holds within choice type so all public goods have the same scale and that choices between public goods have no covariance between them. Finally, for the choice between a public good and a private good the scale is  $\sigma_\varepsilon(j) = \sqrt{(\lambda_b + \beta_b(1/j))^2 + (\lambda_p + \beta_p(1/j))^2}$ , which maintains the independence assumption but relaxes the identical assumption.

The sample is pooled over individuals  $i$  and choice occasions  $j$  for each item. The likelihood function takes the form:

$$L(y_{ijk}; V_k, \lambda_b, \beta_b, \lambda_p, \beta_p) = \prod_i^N \prod_j^J P_{rc}^{1-y_{ijk}} P_{cr}^{y_{ijk}} \quad (6)$$

where the dependent variable  $y_{ijk}$  equals 0 if the row item is chosen and 1 if the column item is chosen.<sup>11</sup> Note that the item index  $k$  equals  $r$  or  $c$ .

Interpretation of these parameters is as follows. Researcher error,  $\lambda$ , is hypothesized to be constant. This parameter will also pick up any constant error generated by the respondent. A significant  $\beta$  coefficient implies a significant magnitude of preference imprecision or uncertainty, as it represents a significant change in the scale of the model through choice sequence. As no other aspects of the session are changing, a positive  $\beta$  represents learning while a negative  $\beta$  represents fatigue or boredom. The hypothesis to be tested is  $H_0: \beta = 0$ , which implies that choice sequence has no effect the scale of the model.

The hypothesis of equal scale for public and private goods ( $H_0: \lambda_b = \lambda_p = \lambda; \beta_b = \beta_p = \beta$ ) was tested with a likelihood ratio test comparing a model that assumes only a single parameterization is necessary with the alternative that scale differs between public and private good choices.

## Results

The coefficient  $\beta$  of the log-likelihood function was found to be significantly positive for the full set of choices and for both the public and private subsets (Table A1). Figure A1 shows the parameterized values of the scale for each choice occasion. Because a significant reduction has been found in the scale of the model, representing the dispersion of the sample valuation distribution, preference refinement (reduction in preference imprecision or fine tuning) is supported.

The levels of preference imprecision measured by the standard deviation of the disturbance term suggests that consistency is lower for choices involving public goods than those involving private goods (see Figure A1). The unrestricted likelihood function allows the public and private parameters to be estimated separately. Conversely, the restricted likelihood function only allows

---

<sup>11</sup> Identical items appearing across the three studies are combined into a single item. For example, the Clean Air Arrangement is described identically in each study, therefore in the likelihood function only a single alternative specific constant is estimated for this item rather than four. As the estimation of these parameters is not the primary interest of this paper, this is used to reduce the necessary computational requirements for the estimation of the heteroscedastic probit.

a single parameterization estimating a single scale for public and private goods (this is the pooled model in Table A1). The unrestricted log likelihood is  $\ln L_u = -67,608$  and the restricted log likelihood is  $\ln L_r = -69,173$ ; thus  $-2[\ln L_r - \ln L_u] = 3130$ . This easily exceeds the 95% confidence level necessary to reject the null hypothesis,  $\chi^2 = 5.99$ . Thus, using a very different procedure we confirm the earlier findings that consistency improves with sequence and that choices involving public goods are less consistent than choices involving private goods.

## References

- Ben-Akiva, M., & Lerman, S. R. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: Massachusetts Institute of Technology.
- Bock, R. D., & Jones, L. V. 1968. *The Measurement and Prediction of Judgment and Choice*. San Francisco, CA: Holden-Day, Inc.
- Bockstael, N. E., & Strand, I. E., Jr. 1987. The effect of common sources of regression error on benefit estimates. *Land Economics* 63(1): 11-20.
- Braga, J., & Starmer, C. 2005. Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics* 32(1): 55-89.
- Brown, T. C., & Peterson, G. L. 2003. Multiple good valuation. In P. A. Champ & K. Boyle & T. C. Brown (Eds.), *A Primer on Non-market Valuation*, pp. 221-258. Norwell, MA: Kluwer Academic Publishers.
- Carson, R. T., Wright, J., Carson, N., Alberini, A., & Flores, N. 1994. *A Bibliography of Contingent Valuation Studies and Papers*. La Jolla, CA: Natural Resource Damage Assessment, Inc.
- Clarke, A., Bell, P. A., & Peterson, G. L. 1999. The influence of attitude priming and social responsibility on the valuation of environmental public goods using paired comparisons. *Environment and Behavior* 31(6): 838-857.
- Cummings, R. G., Brookshire, D. S., & Schulze, W. D. (Eds.). 1986. *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Totowa, NJ: Rowman and Allanheld.
- David, H. A. 1988. *The Method of Paired Comparisons* (Second ed. Vol. 41). New York, NY: Oxford University Press.
- Fechner, G. T. 1860. *Elemente der Psychophysik*. Leipzig: Breitkopf and Hartel.
- Gregory, R., Lichtenstein, S., & Slovic, P. 1993. Valuing environmental resources: A constructive approach. *Journal of Risk and Uncertainty* 7: 177-197.
- Hausman, J. A., & Wise, D. A. 1978. A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46(2): 403-426.
- Hicks, J. R. 1956. *A Revision of Demand Theory*. London: Oxford University Press.
- Jorgensen, B. S., Syme, G. J., Smith, L. M., & Bishop, B. J. 2004. Random error in willingness to pay measurement: a multiple indicators, latent variable approach to the reliability of contingent values. *Journal of Economic Psychology* 25(1): 41-59.
- Kealy, M. J., Montgomery, M., & Dovidio, J. F. 1990. Reliability and predictive validity of contingent values: Does the nature of the good matter? *Journal of Environmental Economics and Management* 19: 244-263.
- Kendall, M. G., & Smith, B. B. 1940. On the Method of Paired Comparisons. *Biometrika* 31:

- 324-345.
- Li, C.-Z., & Mattsson, L. 1995. Discrete choice under preference uncertainty: an improved structural model for contingent valuation. *Journal of Environmental Economics and Management* 28: 256-269.
- Marschak, J. 1959. *Binary-choice constraints and random utility indicators*. Paper presented at the Mathematical Methods in the Social Sciences.
- McConnell, K. E., Strand, I. E., & Valdes, S. 1998. Testing temporal reliability and carry-over effect: the role of correlated responses in test-retest reliability studies. *Environmental and Resource Economics* 12(3): 357-374.
- McFadden, D. 2001. Economic choices. *American Economic Review* 91(3): 351-378.
- Peterson, G. L., & Brown, T. C. 1998. Economic valuation by the method of paired comparison, with emphasis on evaluation of the transitivity axiom. *Land Economics* 74(2): 240-261.
- Plott, C. R. 1996. Rational individual behavior in markets and social choice processes: the discovered preference hypothesis. In K. Arrow & E. Colombatto & M. Perleman & C. Schmidt (Eds.), *Rational foundations of economic behavior*, pp. 225-250. London: Macmillan and St. Martin's.
- Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. 2006. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature* 44(3): 631-661.
- Thurstone, L. L. 1927. A law of comparative judgment. *Psychology Review* 34: 273-286.
- Thurstone, L. L., & Chave, E. J. 1929. *The Measurement of Attitude*. Chicago, IL: University of Chicago Press.
- Torgerson, W. S. 1958. *Theory and Methods of Scaling*. New York, NY: John Wiley & Sons.
- Tversky, A. 1969. Intransitivity of preferences. *Psychology Review* 76(1): 31-48.
- Wilcox, N. T. 1993. Lottery chance: incentives, complexity and decision time. *The Economic Journal* 103: 1397-1417.

## Tables

Table 1

Data sources

| Study                     | Number of respondents | Number of goods |         | Number of money amounts | Number of choices per respondent |
|---------------------------|-----------------------|-----------------|---------|-------------------------|----------------------------------|
|                           |                       | Public          | Private |                         |                                  |
| Birjulin (1997)           | 189                   | 6               | 4       | 11                      | 155                              |
| Clarke et al. (1999)      | 463                   | 5               | 4       | 8                       | 108                              |
| Peterson and Brown (1998) | 327                   | 6               | 4       | 11                      | 155                              |
| Total                     | 979                   |                 |         |                         |                                  |

Table 2

Occurrence of inconsistency by type of choice, for pooled data

| Type of choice                | Number of choices | Percent inconsistent |
|-------------------------------|-------------------|----------------------|
| Private good vs. money        | 37,520            | 6.2                  |
| Public good vs. money         | 52,576            | 6.7                  |
| Private good vs. private good | 5,874             | 8.6                  |
| Public good vs. public good   | 12,370            | 8.4                  |
| Private good vs. public good  | 21,644            | 9.0                  |
| All choices                   | 129,984           | 7.2                  |

Table A1

Probit estimates (t-statistics in parentheses)

| Coefficient | Private       | Public        | Pooled         |
|-------------|---------------|---------------|----------------|
| $\lambda$   | 230<br>(86.6) | 530<br>(83.3) | 393<br>(121.7) |
| $\beta$     | 327<br>(6.6)  | 394<br>(5.2)  | 379<br>(8.0)   |

**Figures**

Figure 1. Distribution of respondent reliability

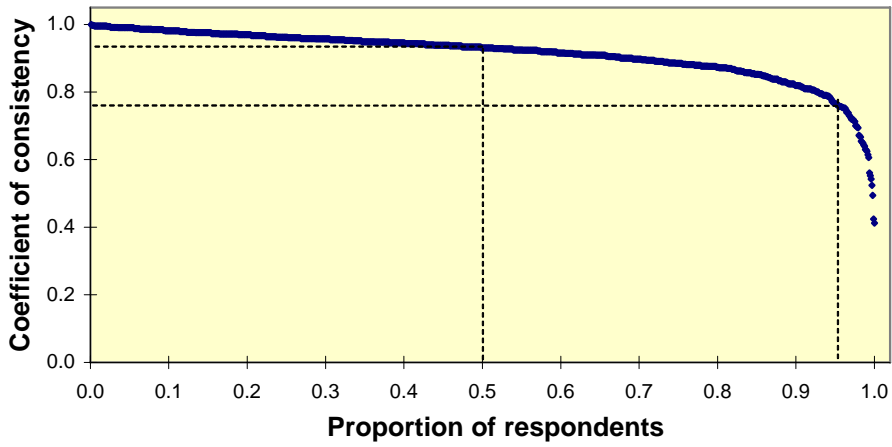


Figure 2. Proportion of respondents making an inconsistent choice, first 100 choices

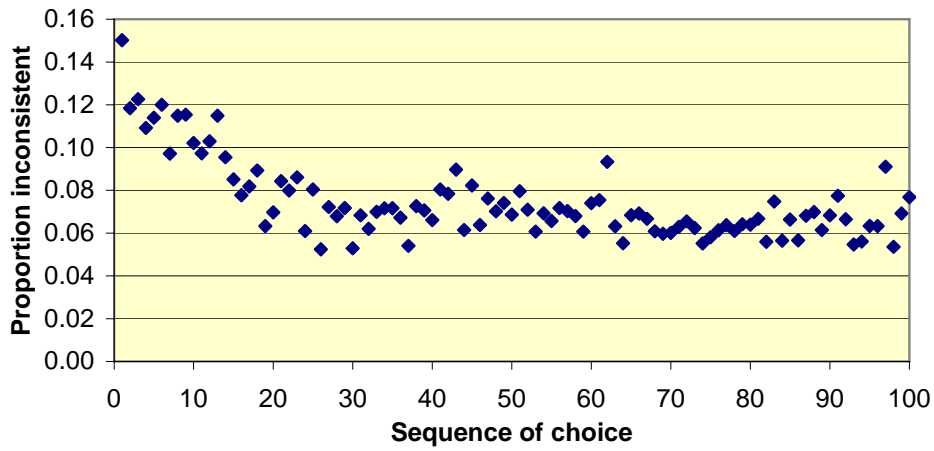


Figure 3. Mean response time, 2<sup>nd</sup> through 100<sup>th</sup> choice

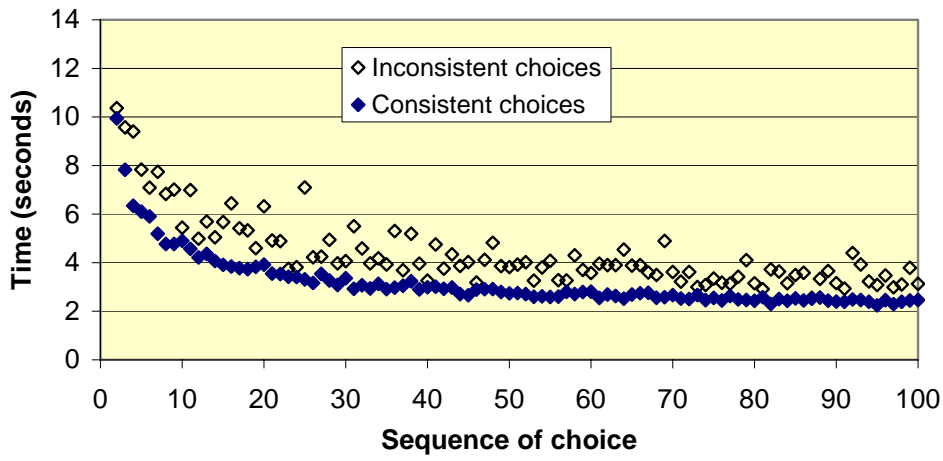


Figure 4. Inconsistency versus preference score difference

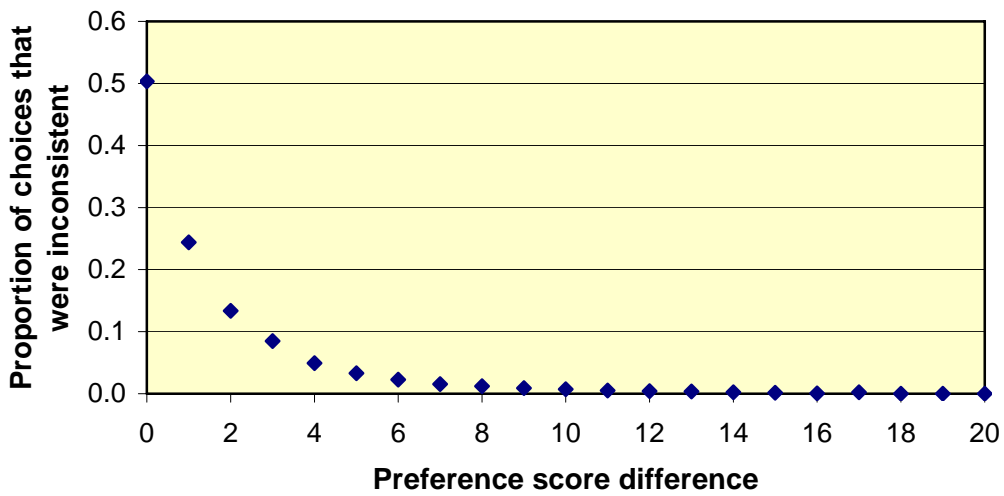


Figure 5. Mean preference score difference, first 100 choices

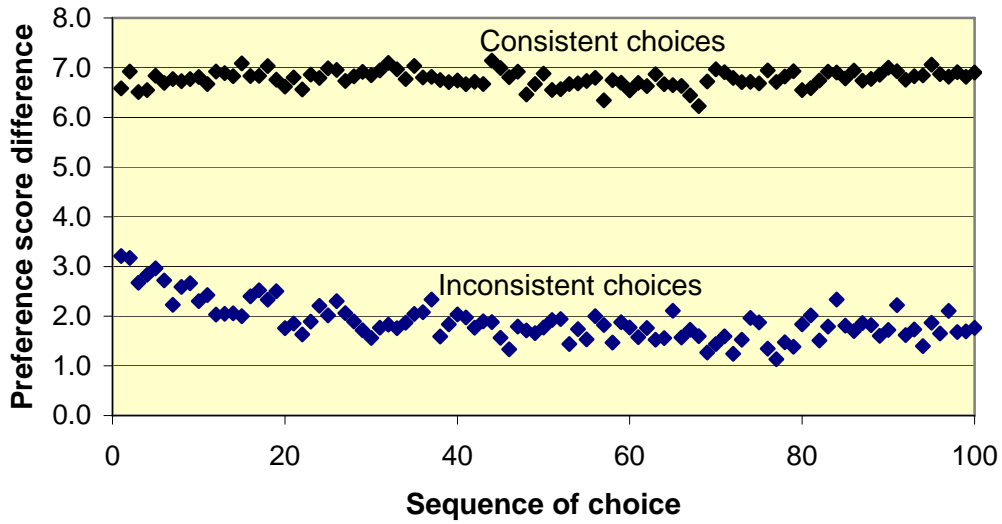


Figure 6. Proportion of respondents making an inconsistent choice, for public and private good choices, first 100 choices

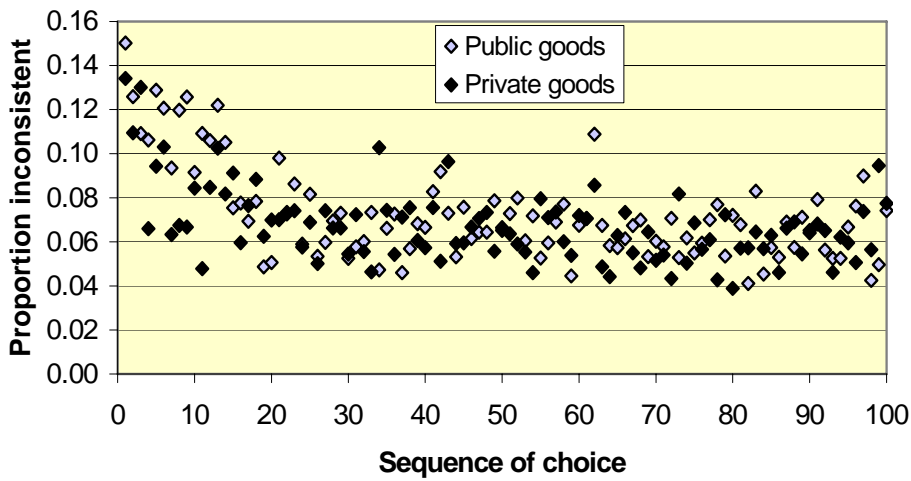


Figure 7. Proportion of choices switched upon retrial, first 100 choices

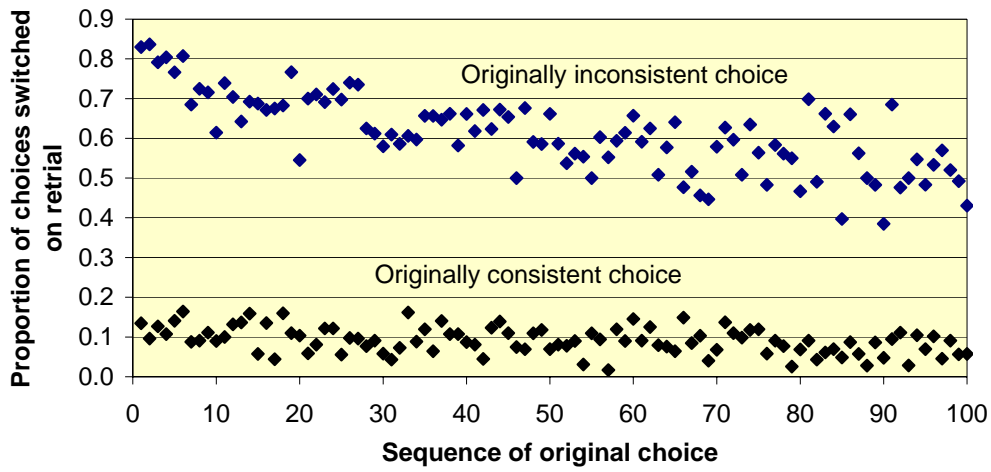


Figure A1. Change in scale with sequence of choice

