Respondent Consistency in a Tournament-Style Contingent Choice Survey

Blake Willmarth
University of Pennsylvania
Philadelphia, PA 19104
(215) 898-3016
blakew@wharton.upenn.edu

and

Robert W. TURNER
Department of Economics
Colgate University
13 Oak Drive
Hamilton, NY 13346
(315) 228-7529
rturner@mail.colgate.edu

February 2010

ABSTRACT

We present the results of an internet-based contingent choice survey about management options at North Cascades National Park, focusing on respondent consistency. A tournament-style contingent ranking design followed by a contingent rating exercise allows for tests of different kinds of consistency in survey responses. Many respondents give inconsistent responses, but these inconsistencies do not create large differences in estimated tradeoffs between scenario attributes.

Keywords: contingent ranking, contingent rating, survey respondent consistency

JEL codes: Q51, Q26

Introduction

Contingent ranking experiments are known to suffer from violations in the underlying axioms of utility theory. Their validity presupposes that respondents have fully-formed preferences and are fully able to transform those preferences to the setting of a complex survey instrument. They further assume that respondents do not introduce bias, neither wittingly nor unwittingly, into their stated outcomes. Existing literature demonstrates that these assumptions are frequently invalid, especially in the context of non-marketable goods, where respondents may be unfamiliar with pricing and appropriating trade-offs.

Various forms of violation to these underlying axioms suggest that preferences are not fully-formed or that respondents are less able to elicit them the deeper or more complex the rankings task. Sequencing effects – learning and fatigue – are phenomena that violate both areas. Respondents suffering from learning effects display a shift in preferences from their starting point over the course of the survey, suggesting their preferences are not fully-formed. Conversely, fatigue effects –whereby respondents' rankings become noisier as the survey progresses – suggest that respondents may have fully-formed preferences but are unable to sustain their elicitation. [Ben-Akiva et al; Chapman and Staelin; Hausman and Ruud]

Bias may be introduced by the instrument as well. Status quo bias results when respondents find the status quo scenario either systemically less or more preferable modulo the difference in attributes. It may be introduced just by the labeling of the scenario as such [ref] and has been found to bias results in either direction [Samuelson and Zeckhauser; Ben-Akiva et al; Foster and Mauroto]. Increased noise may result from respondents being indifferent between scenarios but forced to place an order on them if ties in ranking are not permitted [ref]. Complexity in various forms also increases the likelihood of non-logical responses, from the number of attributes and levels to the within-scenario variation in them [DeShazo and Fermo 2002].

Finally, contingent ranking experiments are frequently found to display differences in both efficiency and economic outcomes across ranks [ref]. Therefore violation of theory may differ based on the level of ranking or respondents may be displaying indifference between less preferred options.

This study introduces a novel survey design that aims to obviate some of the issues found in previous contingent ranking studies while simultaneously permitting a number of tests for logical fallacy. The next section introduces the survey; a section follows discussing the design; the following section presents results from a range of construct validity testing; and finally we present an econometric analysis.

North Cascades National Park Survey

The North Cascades Park General Management Plan (National Park Service 1988) identifies five attributes as the most relevant to park management and resource allocation: cultural preservation, wilderness preservation, threatened and endangered species protection, water quality, and visitation. Scenarios for the contingent choice survey were therefore constructed with these five attributes plus a compulsory, one-time tax change, included as an implicit cost mechanism. The varying levels of the attributes, shown in Figure 1, correspond to the current situation in the park and to plausible alternatives based on the management plan. Scenarios were constructed in a fractional factorial orthogonal matrix, with 47 remaining once clearly sub-optimal scenarios were removed. Each scenario represents a hypothetical description of the state of the park in five years.

Figure 1

Scenario Attributes and Their Levels

Attribute	Level (from lowest to highest)			
Cultural Preservation	60 (9% fewer) structures in good condition	66 (no change in) structures in good condition	72 (9% more) structures in good condition	80 (21% more) structures in good condition
Wilderness Preservation	60 acres disturbed and 963 acres unrestored (8% less restoration)	56 acres disturbed and 900 acres unrestored (no further restoration)	50 acres disturbed and 801 acres unrestored (12% more restoration)	45 acres disturbed and 720 acres unrestored (25% more restoration)
Threatened and Endangered Species Protection	No species protected and stable	Bald eagle protected and stable (status quo)	Bald eagle and grizzly bear protected and stable	Bald eagle, grizzly bear, and two other species protected and stable
Water Quality	65% unimpaired (10% less restoration)	75% unimpaired (no further restoration)	80% unimpaired (5% more restoration)	90% unimpaired (15% more restoration)
Visitation	390,000 (10% decrease)	430,000 (no change)	475,000 (10% increase)	530,000 (23% increase)
Tax	$20 decrease; no change; $20, $40, $55, $75, $100 increase			

Increases in cultural preservation, wilderness preservation, species protection, and water quality are expected to increase utility and thus the likelihood of a higher ranking, all else equal. An increase in tax is expected to have the opposite effect, *ceteris paribus*. *A priori*, the sign on visitation is unknown, since more visitation probably leads to more congestion, which might be thought of as deleterious even for those with only nonuse values, but on the other hand respondents might believe there are positive spillover effects of others' visits to society at large (Turner 2002).

After respondents went through several informational web pages related to each attribute, an analysis of current park resource allocation, and a brief explanation of each attribute's levels, they were presented with several mandatory framing exercises before the contingent choice section. These served as a warm-up to the contingent choice task and also led respondents to consider basic tradeoffs between attributes. In line with the literature, they were also designed to force respondents to think about competing substitute public goods and their own budget constraints. This should help reduce hypothetical bias, though some authors argue that these lead-in questions have little effect on responses (Loomis *et al*. 1994, Kotchen and Reiling 1999, Whitehead and Blomquist 1999; Loomis *et al.* also have an interesting exchange with Whitehead and Blomquist in the November 1995 issue of *Land Economics*).

The survey was designed and pre-tested in stages from 2004 to the fall of 2005. In the spring of 2006 emails with a link to the survey's website were sent to a random collection of individuals in the U.S. 240 respondents gave answers to the contingent choice questions, though not all ranked every scenario group.

Contingent Ranking Design

We introduce a novel survey design that aims to obviate some of the issues found in previous contingent ranking studies while simultaneously permitting a number of tests for logical fallacy. Using a web-based instrument and with the assistance of a computer programmer, we employ a "tournament-style" format, where, starting from a pool of eight randomly allocated scenarios, favored alternate scenarios are sequentially ranked against each other and the status quo until a most-preferred scenario is revealed. In addition, respondents are asked to rate their last set of scenarios immediately following the rankings exercise and before a final round of post-survey and demographics questions. (Throughout the paper we use the terms "set" and "page" interchangeably.) The primary motivations of this design are two-fold: Firstly, by focusing respondents on their most-preferred alternate scenarios it should increase the precision of parameter estimates and reduce disparities across ranks; secondly, because of the high degree of repetition it permits a range of construct validity testing.[1]

The tournament format is illustrated in Figure 2. In the first round of ranking exercises, each respondent ranked four sets of scenarios, three at a time. In the second round the higher-ranking alternative scenarios from the first two sets were pitted against each other and the *status quo*; similarly, another set of scenarios to rank was formed from the *status quo* and the higher-ranking alternative scenarios from the third and fourth original sets. Finally, the higher-ranking alternative scenarios from the two sets of scenarios in the second round were grouped with the *status quo* for a third round consisting of one last ranking exercise. This was followed by a rating exercise on the same set of three scenarios.

Figure 2
Illustration of Tournament Format and Implied Orderings

Round 1	Round 2	Round 3
1 vs 2 vs SQ		
	1 vs 3 vs SQ	
3 vs 4 vs SQ	$(1 > 4)$	
		1 vs 5 vs SQ
		$(1 > 7,\ 1 > 6,\ 1 > 8)$
5 vs 6 vs SQ		
	5 vs 7 vs SQ	
7 vs 8 vs SQ	$(5 > 8)$	

SQ: Status Quo
Higher-ranked alternative scenarios in bold face
Implied rankings in parentheses

From the design we identify three broad categories of consistency tests: ranking, rating, and transitivity.[2] A scenario is said to be rank inconsistent if its ranking relative to the status quo

[1] We also investigate whether extrapolating implied orderings increases efficiency and test for difference between the contingent ranking and rating models, though not discussed in this draft.

changes between rounds. That is, if scenario 3 in the example were ranked higher than the status quo in round 1 but lower than the status quo in round 2, it fails its one test for rank inconsistency; the top two most preferred scenarios receive two tests for rank consistency. They further receive an additional test for rating consistency, which is applied in the same way as rank consistency by observing the underlying rankings. Note, however, that the exact same three scenarios – those already revealed to be their two most-preferred – are being immediately repeated, so respondents should have little difficulty in the underlying rankings task according to theory.

Finally, tests for transitivity are possible when the status quo is alternately ranked first and less-than-first in successive round-pairings; in the example, if 1 > SQ > 2 then SQ > 3 > 4. In these a full ordering for the next round is already implied, so a further test that the observed order matches the implied order was performed when testing for transitivity. A respondent fails transitivity if the expected order of the two alternate scenarios is not matched by observation. In the example, the respondent would pass the transitivity test if 1 > 3 and the full order test if 1 > SQ > 3 – that is, the respondent passed both *rank* and *transitivity* testing.

Test Results

Table 1 presents an overview of the results. Just over half of respondents were ever rank inconsistent while one in five tests failed. This is roughly in line with Foster and Maurato (2002) who also found about half of respondents occasionally failing and a third of the tests failing overall in a similar study. Even at a lower ratio there were a surprising number of failures in the ratings category and one in five respondents failed at least once.

[2] Because a few clearly sub-optimal scenarios were removed after generating an orthogonal set, dominance testing was not possible.

There were 127 tests for transitivity out of a possible total of 618 round-2-or-3 ranking sets where 45% of the sample (92 respondents) had at least one test for transitivity.[3] Of these, one in four respondents tested had at least one failure and one in five tests failed. Foster and Maurato (2002) observed 13% of their sample failing transitivity tests but a majority of non-testing in our survey precludes a straightforward comparison. Additionally, there was a much higher rate of failure of rank consistency when transitivity was tested for, with over one half of both subjects and tests failing full-order consistency when observable.

Table 1 - Incidence of Test Failures

Test	Tests	Failure	Test %	Sample	Failure	Sample %
Rank	1236	238	19%	206	115	56%
Transitivity	127	27	21%	92	23	25%
Full-Order	127	68	54%	92	52	57%
Rate Consistency	412	58	14%	206	46	22%
All (Rank or Rate)				206	123	60%

Sample failures include respondents failing at least one test.

Table 2 breaks down the frequency distributions of each broad category of test failure. The shapes reflect Foster and Mourato (2002) and in both surveys the modal number of failures is once for each test.

Table 2 - Frequencies of Test Failures

Test	0	1	2	3	4	5	6
Rank	91	45	35	22	9	3	1
Transitivity	69	19	4	0	--	--	--
Full-Order	40	38	12	2	--	--	--
Rate	160	34	12	--	--	--	--

[3] The status quo was ranked first 26% of the time in the first two rounds. 62 respondents (30%) had one transitivity test, 25 (12%) had two tests, and 5 (2%) had 3 tests.

Because tests for ranking and rating are primarily against the status quo, perhaps consistent respondents are simply faced with an easier task if they either totally prefer or totally dislike the status quo. From the design, one would expect that the number of top rankings (1) declines and the number of lowest rankings (3) increases as the rounds progressed and respondents honed in to their alternate scenarios of choice. Then together with the former observation, respondents who rank the status quo last most of the time will, independent of other factors, be observed with a higher rate of consistency. Figure 3 shows the different distributions of status quo rankings across rounds by consistent and inconsistent respondents and reveals just this.

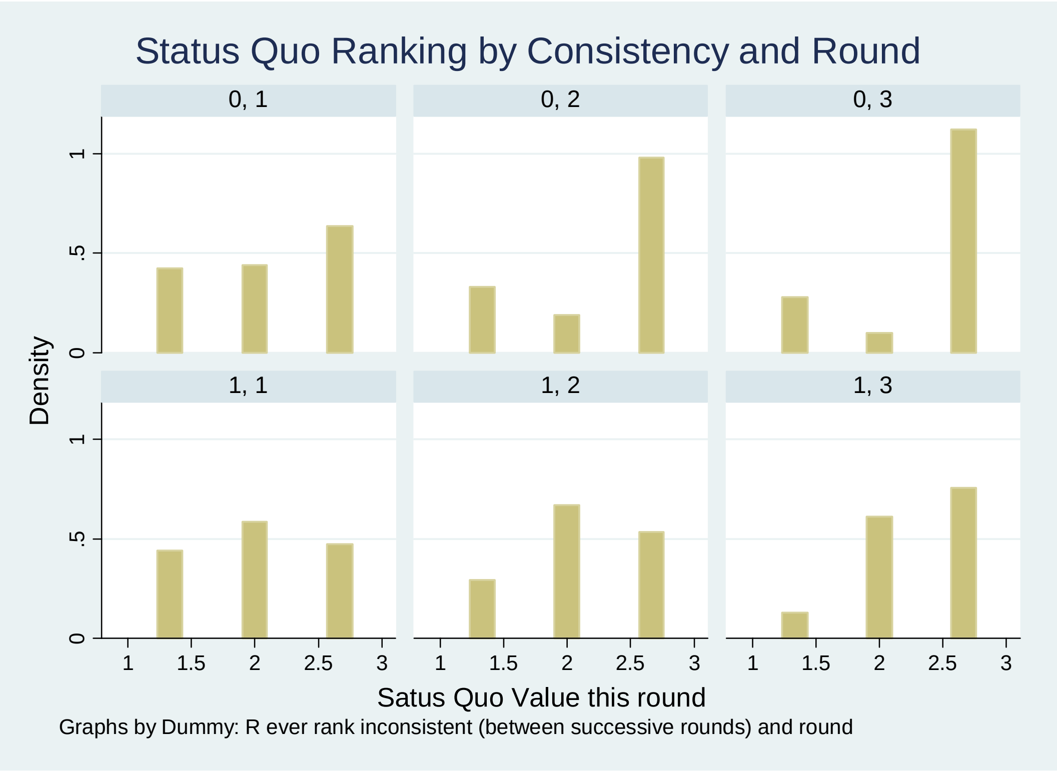

Table 3 tabulates status quo rankings by their lags and further clarifies the relationship of status quo movement with consistency. First, note the volume of passing tests tilts far to the end of low lagged status-quo rankings being preserved as low. Second, the failure rate of last-ranked lagged status quo is much smaller than higher lagged ranks due to the decreasing ambiguity of preferred scenarios to the status quo.

Table 3 - Patterns of Status Quo Ranking by Test Failure

	Lagged by Current Status Quo Ranking											
	1				2				3			
	1	2	3	Total	1	2	3	Total	1	2	3	Total
Pass	169	75		244		97	223	320		58	376	434
Fail		24	55	79	30	63		93	25	41		66
Failure Rate				24%				23%				13%
Total	169	99	55	323	30	160	223	413	25	99	376	500

Rank Inconsistency Patterns

To investigate the role of sequencing effects, we next break down inconsistencies into four distinct categories of rank consistency tests permitted by the three-tiered tournament design. Final round scenarios ranked differently relative to the status quo than the first two rounds we call *fatigue* failures because of this last-minute change in preference of a highly-ranked scenario. The two other cases of final round inconsistency – rounds 2 and 3 disagreeing with round 1, round 2 disagreeing with rounds 1 and 3 – which we refer to as *shift* and *noise*, respectively. Finally, we label the remaining category of test failure *round2-losing*, where the “losing” scenario in the second round has a change in preference relative to the status quo, to denote a distinction in sequence from the other three.[4]

Table 4 presents the results of these sequencing tests. Of all respondents, 41% were ever round2-losing, 20% were ever fatigued, 13 % ever shifted their preference and 11% ever gave noisy rankings. These add up to well over the number of respondents who were ever inconsistent, suggesting considerable overlap. Indeed over half of ever round2-losing

[4] It should be noted that these labels do not necessarily connote a precise interpretation of the behavior of inconsistency. It is entirely possible that the same behavior psychology is manifested across different categories.

respondents also failed another test. In total 38% of respondents at some point had trouble ranking even their two most preferred scenarios.

The distribution of inconsistent patterns suggest round2-indifference and fatigue were the primary causes of inconsistency in the survey. Round2-losing can only be tested in the second round and it accounts for 68% of all test failures in that round (with shift and noise sharing the rest). Fatigue explains two-thirds of round3-failing scenarios (the noise category making up the rest). Ratings failures can also be thought of as a fatigue test since respondents are implicitly re-ranking their two most-preferred scenarios immediately after ranking them explicitly. Recall that 22% of respondents failed at least one such ratings test. Including these failures to the fatigue category now has 31% of all respondents and half of all inconsistent respondents. It therefore appears that most fallacies are a manifest of a change of preference or not being able to express indifference between less-preferred scenarios and the status quo, and fatigue for most-preferred scenarios.[5]

Incidence of Test Failures

	Tests	**Failure**	**Test %**	**Sample**	**Failure**	**Sample %**
Rank	1236	238	19%	206	115	56%
Round 3 Tests						
Shift	412	28	7%	206	26	13%
Noise	412	25	6%	206	23	11%
Fatigue	412	47	11%	206	42	20%
S+N+F	412	100	24%	206	91	44%
Round 2 Test						
Round2-Losing	412	113	27%	206	85	41%
Round2-Losing-Only				206	37	18%
S+N+F-Only				206	78	38%

Sample failures include respondents failing at least one test. The S+N+F category includes all three round-3 test failures.

[5] The inability to express ranking ties is a condition of the tournament instrument, although one could permit respondents to tie so long as they don't occur between the two alternate scenarios.

In combination with the *sequence* of test failures we also look at the *direction* of change relative to the status quo in Table 5, which yields an interesting fact: inconsistencies were twice as likely to be the result of a *decrease* in the relative ranking of "losing" alternate scenarios. Fatigue and round2-failures overwhelmingly resulted from lower rankings relative to the status quo from the previous round, while preference-shift failures were more likely to yield higher rankings. This combined with the fact that respondents were more likely to be consistent when giving the status quo low ranks shows that inconsistent respondents are shifting their preferences downwards, while consistent respondents will appear to shift theirs upward because their favorite scenarios are consistently being repeated in the experiment. If respondents are willing to pay for changes to the survey attributes one would therefore expect a magnitude of difference between consistent and inconsistent respondents.

Table 5 - Change in Relative Ranking to the Status Quo

	Round 2			**Round 3**			**Outcome**		
	higher	**lower**	**% lower**	**higher**	**lower**	**% lower**	**win**	**lose**	**% lose**
Rank	54	112	67%	25	47	65%	74	164	69%
Round2-Losing	26	87	77%	--	--	--	0	113	100%
Shift	18	10	36%	--	--	--	14	14	50%
Noise	10	15	60%	15	10	40%	12	13	52%
Fatigue	--	--	--	10	37	79%	8	39	83%

Outcomes pertain to the round(s) of the test: Rank is both rounds 2 and 3, Round2-Losing is round 2, and Shift, Noise, and Fatigue are round 3.

Econometric Analysis

The contingent ranking method has been used to value a variety of environmental goods (for example Beggs et al., 1981; Lareau and Rae, 1989; Garrod and Willis, 1997; Caplan, et al. 2002. Most researchers use a rank-ordered logit model; we briefly summarize the underlying theory here. First, utility U_{ij} (where *i* indexes the :individual and *j* the scenario) is assumed to be divided into a measurable component V_{ij} and a random component e_{ij} which is assumed to be independent and identically distributed with a type 1 extreme value distribution. Rankings indicate relative utility levels for a respondent, for example $U_{11} > U_{1} > U_{13}$. V is an indirect utility function with each park attribute ($a_k, k = 1,\ldots,5$) plus cost (c; the tax attribute here) as arguments. An alternative-specific constant (ASC) representing the *status quo* scenario is often added. Personal characteristics can be added using interaction terms. For the simple, attributes-only case, the probability of a particular complete ordering of a group of scenarios for individual *i* is

$$P\left(U_{i1} > U_{i2} > U_{i3}\right) = \frac{e^{V_{i1}}}{e^{V_{i1}} + e^{V_{2}} + e^{V_{i3}}} \cdot \frac{e^{V_{i2}}}{e^{V_{i1}} + e^{V_{2}}} \text{ where } V_{ij} = ASC_j + \sum_{k=1}^{5} \beta_k a_{jk} + \beta_6 c_j . \quad (1)$$

Increases in cultural preservation, wilderness preservation, species protection, and water quality are expected to increase utility and thus the likelihood of a higher ranking, all else equal, so their β s should be positive. An increase in tax is expected to have the opposite effect, *ceteris paribus*, so β_6 should be negative. *A priori*, the sign on visitation is unknown, since more visitation probably leads to more congestion, which might be thought of as deleterious even for those with only nonuse values, but on the other hand respondents might believe there are positive spillover effects of others' visits to society at large (Turner 2002).

Equation (1) assumes that each ranking of three scenarios is independent. Each respondent generates multiple sets of rankings, so some might question this assumption. It is consistent, though, with the simple, attributes-only case we are using here which assumes that respondent characteristics do not affect utility. In any case, we follow the standard practice of assuming that (1) gives a good approximation of the true likelihood function, choosing coefficients to maximize (1), and then when estimating the variance-covariance matrix of the estimators taking into account the possible correlation of different observations from the same respondent. We use the Stata® *rologit* command with the *cluster* option, which gives a heteroskedasticity-consistent variance-covariance matrix adjusted for clusters of correlated observations.[6]

Marginal rates of substitution between pairs of attributes are, by the implicit function theorem, the negatives of ratios of coefficients in the specification of *V*. So, for example, for the basic specification shown in (1), the marginal willingness to pay for a change in attribute a_k is the ratio $-\beta_k/\beta_c$. We use the Krinsky and Robb procedure (1986) to append simulated non-linear confidence intervals.

When estimating the rank-order logit model, we removed from the sample all respondents who reported that they were residents of a foreign country, on the grounds that U.S. national park policy should reflect primarily American preferences. In another paper we also consider two subsamples: respondents who say they have never been to North Cascades National Park and never expect to go there—our *nonusers* group—and the respondents who either have been to the park or expect to go there—our *users* group. If the nonusers have any preferences about the park's management, those preferences must reflect nonuse values. The responses of the users will reflect both use and nonuse values. A few respondents did not answer the question about whether they had been or planned to go to the park, so we removed those observations as well. This left us with 206 respondents and 1,442 sets of rankings.

We examine the effect of inconsistency splitting the sample along patterns of inconsistency and pooling them to test for differences in parametric estimates and economic

[6] Most results are unchanged if the nonrobust (and nonclustered) estimator of the variance-covariance matrix is used, except that standard errors are all smaller.

outcomes. Recall that we in general expect to see noisier parameter estimates and smaller willingness-to-pay values for inconsistent subsamples.

Table 6 begins by splitting the sample among ever rank-inconsistent respondents, ever rank-or-rate-inconsistent respondents, and pages containing a scenario that is ever rank-or-rate-inconsistent. The page-level estimations enforce a clean separation between inconsistent and consistent scenarios and indeed demonstrate a loss of signal in the cultural and wilderness preservation attributes. Wald tests on the overall difference in coefficients for inconsistent observations are all significant at the 5% level. In general, all cultural and environmental attributes have lower estimates, especially the highest level of species protection – 4 species protected – which is significantly different for inconsistent respondents in all models at the 1% level. The tax attribute is marginally more negative for inconsistent respondents though this difference is insignificant in all models. Correspondingly, willingness-to-pay values for a unit increase in each cultural and environmental attribute are depressed for inconsistent samples, especially on the cultural and species attributes. This is in line with *a priori* expectation. Also note that the status quo intercept is insignificant in all consistent regression samples.

We next apply the same technique in Table 7 to examine the effects of sequencing patterns previously found to be important: round2-failing and fatigue plus rating inconsistency. Again we use both respondent- and page-level data recalling that the latter minimizes overlap between categories. Round2-failing respondents display only marginal difference in estimates and willingness-to-pay values though narrowing the sample to the page-level reveals an overall difference significant at the 5% level according to the Wald statistic; however individual willingness-to-pay estimates do not display a notable difference. Both fatigued respondents and pages, on the other hand, display significantly different estimates at the 1% level. Wilderness and species are the only two attributes to show significant differences in willingness-to-pay for unit increases in their protection, however, and again fatigued samples are more likely to be depressed. Lastly, the status quo displays an interesting duality: it is significantly more negative for inconsistent respondents in both categories yet loses its power in pages displaying that inconsistency.

A final round of estimations was done in Table 8 to examine whether the tournament design obviated the common problem of difference in estimates and outcomes across ranks, and further whether excluding inconsistent samples would help stabilize these differences. This was done by transforming the rank-ordered logit model to a series of conditional logit models reflecting the top choice against a set of otherwise indifferent scenarios. Note that for the top choice estimation consistency failures are now only observed when the winning scenario's ranking changes against the status quo – indeed this transformation eliminates much of the observed inconsistency in the survey.

We find that there does appear to be some gain in achieving across-rank stability employing the tournament scheme. Though there is a difference in estimates and outcomes across ranks (columns 1 and 2), it is not large economically. Reducing the sample to "clean" respondents – those respondents never rank, transitive, nor rate inconsistent – seems to exacerbate the differences. Reducing the sample into clean pages, however, appears to narrow the differences, especially economically. Throughout the analysis and given the high number of ranking sets in the survey, page-level stratification in general appears more appropriate in targeting violations of the underlying theory.

Table 8 - Comparison of Consistent Samples Across Ranks

	Full sample		Clean Respondents		Clean Pages	
	First Choice	Second Choice	First Choice	Second Choice	First Choice	Second Choice
Tax	**-0.007****	**-0.010****	**-0.007****	**-0.012****	**-0.007****	**-0.010****
	(0.002)	**(0.002)**	**(0.002)**	**(0.003)**	**(0.002)**	**(0.003)**
Cultural preservation	**0.010****	0.006	**0.014****	0.010	**0.012****	**0.017***
	(0.004)	(0.003)	**(0.005)**	(0.007)	**(0.004)**	**(0.007)**
Wilderness preservation	**0.018****	0.006	**0.021****	**0.018****	**0.023****	0.007
	(0.004)	(0.004)	**(0.005)**	**(0.007)**	**(0.005)**	(0.007)
Water quality	**0.040****	**0.025****	**0.044****	**0.022***	**0.045****	**0.024****
	(0.005)	**(0.006)**	**(0.007)**	**(0.009)**	**(0.006)**	**(0.009)**
Visitation	-0.003	-0.003	0.002	0.004	-0.001	0.008
	(0.004)	(0.004)	(0.005)	(0.007)	(0.005)	(0.007)
Bald Eagle protected	**0.986****	**0.569****	**1.237****	**0.774****	**1.187****	0.421
	(0.167)	**(0.139)**	**(0.225)**	**(0.224)**	**(0.202)**	(0.245)
Bald Eagle and Grizzly Bear protected	**1.467****	**1.101****	**1.950****	**1.702****	**1.794****	**1.920****
	(0.165)	**(0.149)**	**(0.231)**	**(0.264)**	**(0.205)**	**(0.288)**
Bald Eagle, Grizzly Bear and 2 others protected	**1.966****	**0.970****	**2.549****	**1.511****	**2.393****	**1.604****
	(0.187)	**(0.140)**	**(0.257)**	**(0.277)**	**(0.226)**	**(0.309)**
Status Quo Dummy	-0.150	**-0.626****	-0.236	**-1.136****	-0.079	**-0.605***
	(0.172)	**(0.138)**	(0.257)	**(0.234)**	(0.210)	**(0.254)**
Observations	4,326	2,884	2,982	1,162	3,573	1,246
Wald chi2	180.8	103.2	176.9	114.3	196.5	113.2
Prob > chi2	0.000	0.000	0.000	0.000	0.000	0.000
Wald test (p)			0.000	0.000	0.000	0.000
cultpres WTP	1.59 (0.41,4.08)	0.61 (-0.09,1.33)	1.94 (0.54,5.63)	0.83 (-0.30,2.20)	1.59 (0.39,4.16)	1.76 (0.28,4.92)
wildpres WTP	2.67 (1.57,4.68)	0.60 (-0.16,1.17)	2.91 (1.63,5.96)	1.46 (0.44,2.57)	3.08 (1.93,5.60)	0.72 (-1.00,2.04)
water WTP	6.01 (3.77,12.53)	2.49 (1.50,3.81)	6.07 (3.42,14.95)	1.82 (0.46,3.97)	6.16 (3.80,13.05)	2.38 (0.69,6.15)
species 2 WTP	223 (138,484)	108 (74,166)	272 (158,699)	140 (84,269)	245 (149,552)	193 (106,494)

Standard errors clustered on respondent ID shown in parentheses

**=Significant at 1%; *=Significant at 5%

Clean samples remove all logical inconsistencies (rank, transitivity, and rate). First choice considers consistency of top-ranked scenarios and second choice includes bottom-ranked scenario c

The Wald test-statistic represents the probability of equal consistent and inconsistent coefficients

Conclusions

This study employed a novel online tournament-style ranking instrument where respondents successively ranked their preferred alternate scenarios until a most-preferred scenario was revealed, followed by a ratings task on the top set of scenarios. In doing so we hoped to both obviate a number of violations to economic theory in contingent ranking surveys found in the literature and test for a range of consistency patterns.

The test results indicate a rate of consistency failure similar to that found in Foster and Mourato (2002). Examining various sequencing effects shows that respondents are relatively poor at sustaining consistency for less-preferred scenarios across rounds, and that these failures are largely manifested in a drop in preference relative to the status quo. It appears many respondents suffered from fatigue and possibly an inability to express indifference between scenarios and the status quo (we suggest correcting this in future instances). It is also possible, for a preponderance of inconsistent within-round losing scenarios, respondents were appropriating downwards their preference for scenarios that were at one point preferred but superseded by a better alternative. If true this suggests either (misguided) strategic bias or, more likely, noisy or incomplete preferences on the part of inconsistent respondents.

Splitting the sample by various consistent subsamples generally yielded lower and occasionally noisier parameter estimates for the cultural and environmental attributes, which did translate to marginally lower willingness-to-pay values. Despite these differences, the design does appear to reduce the disparity in across-rank differences in estimates and economic outcomes, and especially when discarding pages containing ever-inconsistent scenarios.

Bibliography

Adamowicz, Wiktor, Peter Boxall, Michael Williams, and Jordan Louviere. 1998. Stated Preference

Approaches for Measuring Passive Use Values: Choice Experiments and Contingent Valuation.

University of Alberta, Department of Rural Economics Staff Papers.

Baarsma, Barbara E. 2004 The Valuation of the Ijmeer Nature Reserve using Conjoint Analysis.

Beggs, S., S. Cardell, and J. Hausman. 1981. Assessing the Potential Demand for Electric Cars. *Journal of Econometrics* 17(1): 1-19.

Berrens, Robert P., Alok K. Bohara, Hank C. Jenkins-Smith, Carol L. Silva, and David L. Weimer. 2004. Information and effort in contingent valuation surveys: application to global climate change using national internet samples. *Journal of Environmental Economics and Management* 47(2): 331-363.

Boyle, et al. 2001. A Comparison of Conjoint Analysis Response Formats. *American Journal of Agricultural Economics* 83(2): 441-54.

Boyle, Kevin J. 2002. Experimental Designs Affect Preference Measures. *Journal of Environmental Economics and Management* 44.

Cameron, et al. 2002. Alternative Non-Market Value-Elicitation Methods: Are the underlying preferences the same? *Journal of Environmental Economics*.

Caplan, Arthur J. et al. 2002. Waste not or want not? A contingent ranking analysis of curbside waste disposal options. *Ecological Economics* 43: 186-197.

DeRuiter, D. S. and G. E. Haas. 1995. National Public Opinion Survey on the National Park System: Executive Summary Report. Washington, DC and Fort Collins, CO: National Parks and Conservation Association and Colorado State University.

DeShazo, J.R. and Fermo, G. 2002. Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. *Journal of Environmental Economics and Management* 44(1): 123-143.

Foster, Vivien and Mourato, Susana. 2002. Testing for Consistency in Contingent Ranking Experiments. *Journal of Environmental Economics and Management* 44: 309-328.

Garrod, G.D. and Willis, K.G. 1997. The non-use benefits of enhancing forest biodiversity: A contingent ranking study. *Ecological Economics* 21: 45-61.

Hanley, Maurato, and Wright. 2001. Choice Modeling Approaches: A Superior Alternative for Environmental Valuation? *Journal of Economic Surveys* 15(3): 435-62.

Harrison, Glenn W. 2005. Experimental Evidence on Alternative Environmental Valuation Methods. *Environmental and Resource Economics* (January).

Harrison, R Wes, Stringer, Timothy, Prinyawiwatkul, Witoon. 2002. Conjoint Analysis of Groundwater Protection Programs. *Agricultural and Resource Economics Review* 26(2): 229-236.

Irwin, Julie R., Buying/Selling Price Preference Reversals: Preference for Environmental Changes in Buying Versus Selling Modes (February 9, 2009). Organizational Behavior and Human Decision Processes, Vol. 60, pp. 431-457, 1994. Available at SSRN: http://ssrn.com/abstract=1340207

Johnson, F. Reed and Matthews, Kristy E. 2001. Sources and Effects of Utility-Theoretic Inconsistency in Stated-Preference Surveys. *American Journal of Agricultural Economics* 83(5):1328-1333.

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion. 2006. An Econometric Method of Correcting for Unit Nonresponse Bias in Surveys. Development Research Group, World Bank: Washington DC.

Kotchen, Matthew J. and Stephen D. Reiling. 1999. Do Reminders of Substitutes and Budget Constraints Influence Continent Valuation Estimates? Another Comment. *Land Economics* 75(3): 478-82.

Krinsky, I. and Robb, L. 1986. On approximating the statistical properties of elasticities. *Review of Economic Statistics* 68: 715-19.

Krutilla, John. 1967. Conservation Reconsidered. *American Economic Review* 57(4): 777-86.

Lareau, Thomas J. and Rae, Douglas A. 1989. Valuing WTP for Diesel Odor Reductions: An Application of Contingent Ranking Technique. *Southern Economic Journal*: 728-.

Layton, David F. 2000. Random Coefficient Models for Stated Preference Surveys. *Journal of Environmental Economics and Management* 40.

Loomis, John, Armando Gonzalez-Caban, and Robin Gregory. 1994. Do Reminders of Substitutes and Budget Constraints Influence Contingent Valuation Estimates? *Land Economics* 79 (November): 499-506.

Lvarez-Farizo, A. Begon, N. Hanley, and R.N. Barbera. 2001. The Value of Leisure Time: A Contingent Rating Approach. *Journal of Environmental Planning and Management* 44(5): 681–699.

Mackenzie, John. 1993. A Comparison of Contingent Preference Models. *American Journal of Agricultural Economics* 75(3).

Marshall, Pablo and Bradlow, Eric T. 2002. A Unified Approach to Conjoint Analysis Models. *Journal of the American Statistical Association* (September).

National Park Service.1988.General Management Plan, North Cascades National Park, Ross Lake National Recreation Area, and Lake Chelan National Recreation Area. http://www.nps.gov/noca/gmp.htm, accessed on March 2, 2004.

Roe, B., K.J. Boyle, and M.R. Teisl.1996. Using Conjoint Analysis to Derive Estimates of Compensating Variation. *Journal of Environmental Economics and Management* 31: 145-159.

Smith, V. Kerry. 1996. “Pricing what is Priceless: A Status Report on Non-Market Valuation of Environmental Resources” Yale School of Management’s Economics Research Networks.

Tano, Kaudio, et al. “Using conjoint analysis to estimate farmer’s preferences for cattle traits in W. Africa.” June 3, 2002.

Turner, Robert.W. 2002. Market Failures and the Rationale for National Parks. *Journal of Economic Education* 33 (Fall): 347-356.

Turner, Robert W., Alita Giuda, and Laura Noddin. 2005. Estimating Nonuse Values using Conjoint Analysis. *Economics Bulletin* 17(7): 1-15.

Walsh, R., J. Loomis, and R. Gillman. 1984. Valuing Option, Existence and Bequest Demands for Wilderness. *Land Economics* 60 (1): 14-29.

Whitehead, John and Glenn Blomquist. 1999. Do Reminders of Substitutes and Budget Constraints Influence Contingent Valuation Estimates? Reply to Another Comment. *Land Economics* 75(3): 483-84.