ELSEVIER

Ecological Economics 42 (2002) 479–491

ECOLOGICAL ECONOMICS

This article is also available online at:
www.elsevier.com/locate/ecolecon

ANALYSIS

Judged seriousness of environmental losses: reliability and cause of loss

Thomas C. Brown [a,*], Dawn Nannini [b], Robert B. Gorter [c], Paul A. Bell [b], George L. Peterson [a]

[a] *Rocky Mountain Research Station, US Forest Service, 2150-A Centre Avenue, Fort Collins, CO 80526, USA*
[b] *Colorado State University, Fort Collins, CO 80523, USA*
[c] *Simon Fraser University, Burnaby, BC, Canada*

Received 18 March 2002; received in revised form 13 June 2002; accepted 21 June 2002

Abstract

Public judgments of the seriousness of environmental losses were found to be internally consistent for most respondents, and largely unaffected by attempts to manipulate responses by altering the mix of losses being judged. Both findings enhance confidence in the feasibility of developing reliable rankings of the seriousness of environmental losses to aid resource allocation and damage assessment. In addition, seriousness of loss was found to be sensitive to the cause of the loss, with human-caused environmental losses considered more serious than identical losses caused by natural events. This difference has important implications for assessment of environmental losses.

Keywords: Preferences; Environmental values; Decision making; Losses; Paired comparisons; Context

1. Introduction

This paper reports on a test of the reliability of comparative value judgments of environmental losses. We examined the extent to which such judgments were consistent from one judgment to the next and robust to changes in the mix of other losses being judged. Our motivation for the test was the difficulty of estimating the monetary values of environmental losses, and the consequent need for additional information on the value of those losses.

Good decisions about the public's natural resources often depend on the accuracy of monetary values of resource changes. Accurate values are, among other things, comparable; that is, more

* Corresponding author. Tel.: +1-970-295-5968; fax: +1-970-295-5959
E-mail address: tcbrown@lamar.colostate.edu (T.C. Brown).

0921-8009/02/$ - see front matter
PII: S0921-8009(02)00171-4

valuable resources are valued more highly than less valuable ones. In the case of non-pecuniary environmental resources, however, existing economic valuation methods cannot yet be assured to provide comparable values from one application to the next, in part because results are sensitive to methodological choices about which consensus is lacking. For example, with contingent valuation—a method for directly assessing individual monetary values using carefully worded surveys, and the only available method for measuring economic values of many resource changes—such a fundamental choice as that between an open-ended or dichotomous-choice response mode can have a large impact on measured values and is still being seriously debated (Brown et al., 1996; Green et al., 1998). And even when the methodology is held constant, the accuracy of contingent valuation remains in question (e.g. Diamond and Hausman, 1994; Cummings et al., 1997). For example, yea-saying and anchoring (Holmes and Kramer, 1995; Boyle et al., 1997; Green et al., 1998) are thought to affect dichotomous-choice contingent valuation. Until economic valuation methods more fully meet their objectives regarding non-pecuniary environmental resources, it may be wise to also utilize alternative sources of valuation information based on a more modest objective, that of achieving a value-based ranking of resources. Such a ranking could serve three purposes.

First, the ranking could serve as a check on the comparability of independent economic valuations. Economic valuation of non-pecuniary environmental resources typically occurs one resource at a time. For example, if the value of some resource loss is needed for a damage assessment, a separate study is commissioned to estimate that value. Multiple resources are rarely valued together, and values estimated in individual studies may never be compared. However, quality control on the economic valuation process would be enhanced if values that should be comparable were compared with some independent set of relative value judgments. Comparing the two sets of rankings would offer a check on the validity of the individual values. For example, consider three resource changes (a, b, and c) and their individually estimated economic values (WTPa, WTPb, and WTPc). If comparative judgments produce the ranking $a > b > c$, then independent economic valuations would be expected to satisfy the order $WTPa \geq WTPb \geq WTPc$. This condition allows for equality among two or more independently estimated economic values, such as might occur if budget constraints placed a cap on willingness to pay.[1]

Second, the ranking could facilitate resource assessments or allocation decisions that rely on valuation but do not require cardinal estimates of value. For example, if a decision has already been made to implement one of three options and if the options do not differ in cost, selection of the option to implement could be based on a simple ranking of the values of the three options. Or if an unavoidable action will require the loss of either of two resources, selection of the resource that is the smaller loss requires only that the values of the losses be ranked.

Third, relative value judgments for a set of resource losses could serve to provide step 1 of the two-step approach to valuation envisioned in the resource damage schedule. The 'damage schedule' has been proposed as an alternative source of valuation information, for use until more efficient and accurate non-market valuation methods are developed (Rutherford et al., 1998). An environmental damage schedule separates the process of monetary valuation into two steps. In the first step, random samples of public respondents are questioned about the relative importance or seriousness of a series of potential environmental losses. These responses are combined to approx-

[1] The option of using a ranking of goods as a check on individual economic valuations requires that certain constraints be placed on the individual economic estimates. Things that should affect economic value must be held constant, such as the consumer population, the descriptions of the goods, and the method of provision. However, things that should not affect economic value, such as the valuation method, the choice of statistical model, and the treatment of outliers, need not be held constant.

imate an interval scale measure of the seriousness of the losses. Then, in step 2, dollar amounts are mapped onto this measure of importance. The mapping is accomplished by the agency, courts, or elected representatives with whom the authority to protect the public resource resides, who are free to exercise their best judgment as to the level of deterrence and restitution they wish to impose. See Kahneman et al. (1998) for another proposal for a two-step valuation process, one dealing with assessing punitive damages.

It has long been maintained that comparative judgments are easier for people than are absolute judgments. As Nunnally (1976) states it, "People simply are not accustomed to making absolute judgments in daily life, since most judgments are inherently comparative... people are notoriously inaccurate when judging the absolute magnitudes of stimuli... and notoriously accurate when making comparative judgments". Nunnally's statement focused largely on judgments of physical phenomena such as the length of lines or the brightness of lights, but his notion can perhaps be applied as well to judgments of monetary value. The proposal for a ranking based on value presumes that people can more easily and consistently rank a set of resource changes in order of importance or value than they can indicate a cardinal value such as an amount they are willing to pay.

We propose that the most direct way to achieve a ranking of the values of a set of resource changes (either gains or losses) would be to ask people to make comparative judgments. If each person in a sample drawn from the relevant population were asked about the full set of resource changes at issue, a ranking could be produced for each respondent. Respondents' rankings could then be combined to form one set representing the population, or respondents could be sorted into groups representing whatever 'market segments' (different preference structures) were found and the rankings could be summarized for each group.

Reliability is a primary consideration in the development of such a ranking. In this study we test reliability of comparative judgments among sets of environmental resource losses. We chose losses instead of gains, both because of a wish to contribute toward the problem of adequately performing resource damage assessments and because of an interest in the idea of a resource damage schedule.

Various response formats—including paired comparisons, ratings, and ratio estimation—allow ranking of a set of items. In this study we use paired comparisons, which require each respondent to choose one item from each of various pairs of items. Respondents were asked to choose the more serious loss from each of several pairs of losses. Because paired comparisons allow for intransitive responses, they have the advantage of providing a more complete measure of respondent reliability than other formats (Peterson and Brown, 1998).

Two aspects of reliability seem most relevant and are investigated here. First and most fundamentally, judgments must be consistent from one choice to the next for a given respondent. If respondents are unable to provide consistent judgments within a given set of items, reliance on public judgments is of limited value. We extend past assessments of consistency in choices among environmental losses performed by Rutherford et al. (1998), Chuenpagdee et al. (2001).

Second, the judgments should be reliable across minor changes in the assessment context, such as changes in the mix of items being compared (Parducci, 1968; Brown and Daniel, 1987). Robustness to minor changes in the mix of items being compared is important because the items included in any given public assessment exercise are dependent on which of the many possible items researchers happen to include. The more robust that public judgments of the value of the items are found to be, the more freedom researchers would have to mix different types of items and the more flexible and useful the tool would become. Further, robustness is essential if judgments of value obtained from two or more groups of respondents from the same population are to be combined. Using more than one group of respondents may be necessary if the number of items to be assessed is larger than one group can manage in a given setting. If sets of scale values are to be

combined, the sets should not be significantly affected by item mix context effects.[2]

The basic experimental approach used here to test for item mix context effects was to obtain paired comparisons of seriousness for one set of losses, called the test losses, judged in the context of four alternative sets of other losses, called the treatment losses. Each pairing of the test losses with a set of treatment losses constituted an experimental condition that might produce different judgments of the test losses.

This experimental approach allowed examination of an additional issue, the effect of cause of loss on judgments of seriousness. A cause of loss was listed with each environmental loss. Two of the sets of treatment losses were identical, but their causes differed. One set listed natural causes and the other listed equally plausible human causes. Comparison of the judgments of these two sets of treatment losses allowed the effect of cause type to be observed.

2. Theory

What might cause a loss of reliability across pairs of items or item mixes when judging a set of complex items such as environmental losses? The possibilities may be considered in terms of a model proposed by Thurstone (1927) characterizing judgments as a stochastic process through which the location of an item along the dimension of interest at a given time falls along a range distributed about the item's true value. For item i the model is $U_i = V_i + \varepsilon_i$ where U_i is a momentary position along the judgmental continuum, V_i is its expected value, and ε_i represents dispersion attributed to random fluctuations in perception and judgment. Some degree of randomness is always expected, leading to inconsistency in a respondent's judgments. Inconsistency is expected to increase the closer are the Vs and the larger are the εs of the items being compared. V, however, is assumed to remain constant for a person in a given judgmental context, such as in a session during which paired comparison responses are provided among a set of randomly ordered pairs.[3] When used with a sufficient number of respondents, the paired comparison procedure, given certain assumptions about the εs, allows the Vs to be estimated for the population along an interval scale (Torgerson, 1958).

Inconsistency among a respondent's paired comparisons occurs as circular triads, which indicate intransitive choice. For example, the following circular triad could result from the three paired comparisons resulting from all possible pairings of three items: A > B > C > A. We computed an overall measure of consistency for each respondent based on a comparison of the observed number of circular triads to the maximum possible number, and compared this measure across respondents to observe the distribution in consistency among the sample. The measure of

[2] An initial test of item mix context dependence was performed by Gorter (1997), who obtained paired comparison judgments of the seriousness of three environmental losses and three personal injury losses, with each set of three judged in the context of two different item mixes. The environmental losses were judged when mixed with other environmental losses and also when mixed with personal injuries. Similarly, the personal injuries were judged when mixed with other personal injuries and also when mixed with environmental losses. For both kinds of losses, judged seriousness appeared to be largely unaffected by the switches in the other losses with which they were mixed, suggesting that judgments of the seriousness of losses are quite robust to item mix context effects, but the mix of losses in Gorter's study was small. We extend his work by using a larger set of losses.

[3] Economists will note the similarity of this model to the random utility maximization (RUM) model (Hanemann and Kanninen, 1999). Although the development of the RUM model owes something to Thurstone's early work (see McFadden, 2001), the goals of economic choice modeling are quite different from those Thurstone was pursuing. Whereas Thurstone was modeling the variability in an individual's choices given data from repeated choices of that individual, the RUM model is intended to characterize the choices of a sample of individuals (with data typically consisting of only one observation per individual) in terms of explanatory variables (Ben-Akiva and Lerman, 1985). The random term in Thurstone's model characterizes the imprecision in each individual's attempts to maximize utility, and the random term in the RUM model characterizes the imprecision in the analyst's attempt to explain utility in terms of explanatory variables describing goods and/or individuals.

consistency is assumed to decrease monotonically as ε increases. Further, we averaged the consistency measures for each respondent group; because different groups responded to different kinds of losses or losses of different causes, comparison of mean consistency across groups indicates whether ε varies by kind of or cause of loss.

Turning now to the Vs, Tversky (1969) suggested that some of the variability observed in judgment may not be random. He hypothesized that from one comparative judgment to the next, people may weigh the attributes of the items differently depending on which attributes are present in the items being judged. This differential weighting would essentially cause the Vs to shift among comparisons. The relevance of this hypothesis was enhanced by Slovic's (1975) finding that, in choosing between closely matched dual-attribute options, some subjects tend to select the option that is superior in the more important attribute, thus largely ignoring the other attribute.

For paired comparisons, the mix of attributes may change from one pair of items to the next. If Tversky's hypothesis applies, this differential weighting of attributes could lead to systematically (i.e. non-random) intransitive judgments among a mix of items, and thus to a lack of consistency. Careful detection of systematic differential weighting from one choice to the next would require that respondents provide paired comparisons multiple times for a given set of items. Our experimental design required only one set of choices per respondent, precluding a test of Tversky's hypothesis within respondents, and thus we leave such variability, if it occurs, within the error term ε. Rather, we tested across sets of items for the differential weighting that Tversky contemplated. Our reasoning was that different sets of items might highlight different attributes by the frequency with which the attributes appear among the items of each set, leading respondents in each group to place more weight on the attribute(s) that their set of items happens to highlight. As mentioned above, we tested for such a context effect by obtaining paired comparisons of seriousness for a set of test losses judged when mixed with four alternative sets of treatment losses. One measure of a systematic effect across contexts would be indicated by the correlation of the Vs for the test losses that were obtained from the different contexts, when compared with the correlation of Vs for the test losses obtained from two replications of the same context.

The V for an item is estimated from the proportions of respondents who chose that item over each of the other items in the choice set. Proportions are obtained for each pairing of one item with another, and each proportion is a best estimate of the probability of choosing one item over the other, given the population of respondents and the context in which the respondents provided their choices. These individual probabilities offer another test of item mix context effects, one based on the IIA (independence from irrelevant alternatives) principle (Arrow, 1951).

An item mix context effect with binomial choices is analogous to a failure to meet the IIA assumption with multinomial choices. As Luce (1959) explains, if choices adhere to the IIA assumption, the ratio of the probabilities of selecting each of two alternatives from among a multinomial set of alternatives remains constant when one or more other alternatives are added to or subtracted from the choice set. For example, if the IIA property holds $P(x \mid xyz)/P(y \mid xyz) = P(x \mid wxy)/P(y \mid wxy)$. Thus, if IIA holds, the relative preferences between two alternatives of interest (e.g. x and y) are not affected by the presence of the other alternatives. Each different set of alternatives containing the two alternatives of interest is a unique context for judgment—if IIA holds, the ratio of the probabilities of choosing x to choosing y is the same for each context.

A paired comparison exercise involves a multinomial set of alternatives (i.e. items), but requires only binomial choices among them. The influence of other items is indirect, in that they are present in the overall choice set but not available for a given choice. With paired comparisons, each pairing of items yields a ratio of probabilities (which sum to 1.0). For example, with items x and y the ratio is $P(x \mid xy)/P(y \mid xy)$. With t test items we have $t(t-1)/2$ test pairs (assuming all possible pairs are judged), and thus that many ratios to be compared across contexts, providing the second test of item mix context effects. For example, given item mixes

a and b, one of the $t(t-1)/2$ comparisons would be $P(x \mid xy; a)/P(y \mid xy; a)$ versus $P(x \mid xy; b)/P(y \mid xy; b)$. We perform a test that examines all $t(t-1)/2$ comparisons for each pair of item mix contexts.

3. Method

3.1. Approach and design

To assess the robustness of choices to changes in the mix of losses being judged, our approach was to try to cause a context effect. We sought to design different sets of treatment losses that, when mixed with the losses of the test set for presentation to respondents, might systematically influence respondents' judgments of the seriousness of the test losses, altering the choice probabilities of the test items and producing sets of scale values for the test losses that do not correlate highly with each other. There is little guidance in the literature about item mix context effects, but one factor that seemed likely to shift standards for loss judgments is the cause of the loss (Walker et al., 1999). Thus, we examined how judgments of the test losses varied when those losses were mixed with naturally caused environmental losses versus when they were mixed with human-caused environmental losses. And in our other attempt to cause a context effect, we used two very different kinds of losses, environmental losses and personal injury losses, and examined how judgments of the test losses varied when mixed with the environmental versus the injury losses.

We began by obtaining rating judgments from 40 undergraduate students about a set of 46 losses (36 environmental losses and ten personal injury losses).[4] Each loss was rated on nine-point scales for seriousness, and the environmental losses were also rated for likely cause. The endpoints of the seriousness rating scale were 'not very serious' and 'very serious.' The endpoints of the cause scale were 'human caused' and 'naturally caused.' The results were used to select 24 environmental and eight personal injury losses for further study.

One hundred and fifty-three other students from the same subject pool were then randomly assigned to one of four conditions, resulting in sample sizes of 36, 42, 38, and 37 for conditions 1–4, respectively. In each condition, respondents chose the 'more serious' loss in each of 120 pairs of losses, consisting of all possible pairs of 16 different losses. The 16 losses in each condition consisted of eight test losses that were included in each condition, and eight treatment losses that differed across the conditions (Table 1). Two of the four sets of treatment losses (sets 1 and 3) contained naturally caused environmental losses, one contained human-caused environmental losses (set 2), and the other contained personal injury losses (set 4). The test losses were all environmental losses, four naturally caused and four human-caused. Environmental loss sets 1 and 2 were identical except for cause of the losses, either a natural event or human action. For example, two comparable losses were 'Loss of all aquatic life in a 20-mile stretch of Poudre River due to a drought' and 'Loss of all aquatic life in a 20-mile stretch of Poudre River due to contamination by a chemical spill.' Table 2 lists the individual losses of the different sets.

Table 1
Experimental design

Condition	Losses
1	Treatment set 1: eight environmental losses (naturally caused)
2	Treatment set 2: eight environmental losses (human-caused; same losses as set 1)
3	Treatment set 3: eight environmental losses (naturally caused; different losses from set 1)
4	Treatment set 4: eight personal injury losses (no cause listed)
1–4	Test set: eight environmental losses (four naturally caused, four human-caused)

[4] The students were recruited from introductory psychology classes, and thus represent a variety of majors. Although they cannot be assumed to constitute a random sample of the general public, whether in terms of preferences or response consistency, neither do they represent a narrow field of interest or an academically advanced group.

Table 2
Treatment and test losses

Treatment set 1: Naturally caused environmental losses
(1) Loss of a large herd of elk on the Roosevelt National Forest 20 miles west of Fort Collins due to starvation caused by drought
(2) A 1/4-square mile of mature Douglas fir trees along the banks of the Poudre River 35 miles from Fort Collins, lost by wildfire caused by a lightning strike
(3) Loss of 30% of the native trout in the Poudre River due to drought
(4) A 1/2-square mile of mature ponderosa pine in the Roosevelt National Forest 30 miles west of Fort Collins lost from extremely rare high winds
(5) Loss of all of the mature trees growing on the CSU oval caused by Dutch Elm Disease
(6) A square mile of mature ponderosa pine on the Roosevelt National Forest 50 miles west of Fort Collins, lost by wildfire caused by a lightning strike
(7) Loss of all prairie dogs in the prairie dog colony located across the road from Hughes Stadium due to natural predators
(8) Loss of all aquatic life in a 20-mile stretch of the Poudre River due to a drought

Treatment set 2: Human-caused environmental losses
(1) Loss of a large herd of elk on the Roosevelt National Forest 20 miles west of Fort Collins due to a virus introduced by ranching cattle that roam freely throughout the area
(2) A 1/4-square mile of mature Douglas fir trees along the banks of the Poudre River 35 miles from Fort Collins, lost by wildfire caused by a careless camper
(3) Loss of 30% of the native trout in the Poudre River due to the introduction by the Game and Fish Department of a more aggressive fish species
(4) A 1/2-square mile of mature ponderosa pine in the Roosevelt National Forest 30 miles west of Fort Collins lost from timber harvest
(5) Loss of all of the mature trees growing on the CSU oval caused by an accidental application of the wrong pesticide by grounds keepers
(6) A square mile of mature ponderosa pine on the Roosevelt National Forest 50 miles west of Fort Collins, lost by wildfire caused by a careless hiker
(7) Loss of all prairie dogs in the prairie dog colony located across the road from Hughes Stadium due to housing development
(8) Loss of all aquatic life in a 20-mile stretch of Poudre River due to contamination by a chemical spill

Treatment Set 3: Naturally caused environmental losses
(1) Loss of 200 elk in remote areas of the mountains west of Fort Collins due to lack of food during a particularly harsh winter
(2) Loss of 200 of the Colorado Blue Spruce on campus due to a species-specific naturally occurring disease
(3) One of the six known populations of a rare flowering plant, lost by an invasion of more successful, natural vegetation
(4) Blackening of 50% of the forested areas in the hills visible from campus by a lightning-caused wildfire
(5) Blackening of 40% of the forested areas in Rocky Mountain National Park by a lightning-caused wildfire
(6) Loss of all large trees along the Poudre River through Fort Collins due to flooding
(7) Loss of 60% of the Poudre River trout population to whirling disease, caused by naturally occurring bacteria
(8) Loss of 40% of the elk population in Rocky Mountain National Park due to a naturally occurring disease

Treatment set 4: Personal losses
(1) Permanent loss of mobility of both legs
(2) Permanent loss of a thumb
(3) Complete loss of mobility in both legs that at a gradual recovery rate will return to pre-injury function in 5 years
(4) Complete loss of mobility in one hand that at a gradual recovery rate will return to pre-injury function in 1 year
(5) Permanent loss of sight in one eye
(6) Complete loss of mobility in one arm that at a gradual recovery rate will return to pre-injury function in 2 years
(7) Permanent loss of hearing in one ear
(8) Complete loss of mobility in one ankle that at a gradual recovery rate will return to pre-injury function in 6 months

Test set (included with all four of the treatment sets)

Naturally caused environmental losses
(1) Loss of 300 elk in the Roosevelt National Forest west of Fort Collins, due to a periodically occurring virus affecting only elk
(2) Loss of 10 of the bald eagles in the Fort Collins area due to lack of prey, resulting from unusually dry weather
(3) Loss of 5 old cottonwood trees on the CSU campus due to old age
(4) Loss of 20 deer in the area near Horsetooth Reservoir due to a rare virus that affects only deer

Human-caused environmental losses
(5) Loss of all life in a 40-mile stretch of the Poudre River upstream of Fort Collins due to a toxic chemical spill
(6) Twenty-five percent of the view of the mountains from campus marred by housing development in the foothills
(7) Loss of 20% of the days in which a clear view of the mountains exists from campus due to increased air pollution
(8) Loss of all fish in City Park Lake due to a toxic chemical spill

All 40 losses (four sets of eight treatment losses plus eight test losses) were different from each other in some respect, either in the loss or its cause, or both. Each of the environmental losses listed a cause, either a natural event (e.g. strong winds, disease, lightning strikes, flooding) or a human action (e.g. timber harvest, chemical spill, intro-

duced species, air pollution). Personal injury losses did not list a cause.[5]

The use of eight test items and eight treatment items was based on a compromise among conflicting objectives. The total number of items was constrained by concerns about respondent attention; 16 items produce 120 comparisons, which previous experience has shown can be judged without a drop in inconsistency (Peterson and Brown, 1998). Use of an equal number of test and treatment items was based on a balance between a desire to maximize the ratio of treatment to test items (to increase the likelihood of a context effect) and a desire to maximize the number of test items (to improve the statistical power of tests for a context effect).

Respondents recorded their choices on individual computers. The losses of each pair were presented to each respondent on the left and right side of the computer screens. To randomize any order effects, the pairs were randomly ordered for each subject, and the left-right position of the two losses in a pair was also randomly determined. Respondents indicated their choices using the left and right cursor keys, and could undo their choice with the backspace key in order to make a correction.

3.2. Analysis

Each loss was compared by respondents with the 15 other losses in the condition, and thus could be chosen as more serious a maximum of 15 times. A respondent's vector of the numbers of times each loss was chosen above the others is the set of preference scores. If a respondent's choices were perfectly consistent, the set of preference scores would in this study contain each integer from 0 to 15. Inconsistency in a respondent's choices causes some integers to disappear and others to appear more than once. An individual respondent's coefficient of consistency relates the observed number of circular triads to the maximum possible number (David, 1988). The maximum possible number of circular triads, m, is $t(t^2-4)/24$ when t is an even number, where t is the number of items in the set. A maximum of 168 circular triads is possible for choices among all possible pairs of 16 items. Letting a_i equal the preference score of item i (i.e. the number of items in the choice set dominated by the ith item) and b equal the average preference score (i.e. $(t-1)/2$), the number of circular triads for an individual respondent, c, equals $t(t^2-1)/24-0.5\ \Sigma(a_i-b)^2$. The respondent's coefficient of consistency is then $1-c/m$. The coefficient varies from 1.0, indicating that there are no circular triads in a person's choices, to 0, indicating the maximum possible number of circular triads. Our null hypothesis was that average coefficient of consistency would not differ across conditions.

The means (across respondents within a condition) of the preference scores for each item are commonly used in paired comparison assessments as scale values of the items, and are considered to approximate an interval scale measure (Dunn-Rankin, 1983). However, this measure has the disadvantage of being in terms of the number of items in the set. We used another measure, the mean estimated probability of choosing the item, computed for an item by dividing its mean preference score by the number of pairs in which the item appears. To test for context effects, scale values for the test items were based on comparisons among the eight test items only. Restricting the pairs on which computation of the scale values of the test items was based to the choices between pairs of the test items (i.e. excluding test item by treatment item pairs) provides the cleanest test of context effects. To test for the effect of cause, scale values for the treatment items were based on comparisons among the full set of 16 items. Expressing scale values in terms of probabilities allows the two sets of values to be presented in the same (probability) units.

Obtaining judgments of the seriousness of the test losses in the context of the four different sets

[5] We did not include two different levels of the same loss, such as 'Loss of all aquatic life in a 20-mile stretch of Poudre River' and 'Loss of all aquatic life in a 40-mile stretch of Poudre River,' because the choice of which is more serious would be obvious to participants. Not including such pairs of losses precludes a 'scope' test (Smith and Osborne, 1996), but then comparative judgments—by making the difference so obvious—more or less assure that scope is satisfied.

of treatment losses allows computation of six correlation coefficients comparing mean preference scores of the test losses, one for each pairing of the four conditions. The correlation comparing conditions 1 and 3 will not reflect a systematic item mix difference, because the treatment sets of the two conditions contain only naturally caused environmental losses. Any drop in correlation below 1.0 for this comparison should reflect random influences due to interpersonal differences and the different naturally caused losses that we happened to include in the two treatment sets. A significant item mix context effect would be possible for the other five pairwise comparisons among the four conditions. If item mix context effects were minimal, the correlation coefficient comparing conditions 1 and 3 would not differ significantly from those of the other five comparisons. Formally, our null hypothesis was that the correlation for conditions 1 and 3 would not significantly exceed those from each of the other comparisons.

The other test for context effects was based on the ratios of the choice probabilities of the pairs of test items. Each condition produced a set of 28 ratios, one for each pair among the eight test items. The sets were compared to determine whether the ratios differed significantly across conditions. Our null hypothesis was that they would not differ.

As reported above, environmental loss sets 1 and 2 were identical except for cause of the losses, either a natural event or human action. Comparison of the mean preference scores for these two sets of treatment losses allows a measure of the importance of cause in determining seriousness of environmental loss. Formally, our null hypothesis was that cause would not matter.

It should be noted that our tests of the effect of item mix context and of cause are dependent on the particular losses and causes that we chose to include. Although we attempted to design losses and causes that were roughly comparable across treatments, our results could be to some extent an artifact of the particular losses and causes that we used.

4. Results

4.1. Consistency

The coefficient of consistency ranges across all 153 respondents from 0.99 for the most consistent respondents to 0.30 (Fig. 1). The mean and median coefficients are nearly identical, at 0.81 and 0.82, respectively. The distribution of the coefficient is linear among the more consistent 80% of the respondents (those with a coefficient above 0.73, Fig. 1), indicating a uniform distribution. Below that point, the drop in coefficient is more precipitous. Eighty-eight percent the respondents (135 of 153) have coefficients above 0.645, the midpoint of the range in coefficient, and 95% of the coefficients are above 0.58. The least consistent 5% of the respondents are responsible for 44% of the coefficient range.

The mean coefficients of consistency range from 0.77 to 0.88 among the four conditions (Table 3). The mean coefficients of the three environmental loss conditions are all close to 0.80, whereas the mean coefficient for the personal loss condition is 0.88. For comparison, among the five groups of respondents contributing to a coastal resources damage schedule in Thailand (Chuenpagdee et al., 2001), the mean coefficient of consistency ranged from 0.81 to 0.82 for the four resource user groups and was 0.93 for the expert (resource manager) group.

Coefficient of consistency is not normally distributed (Fig. 1), suggesting a non-parametric test. A Kruskal–Wallis test shows a significant differ-

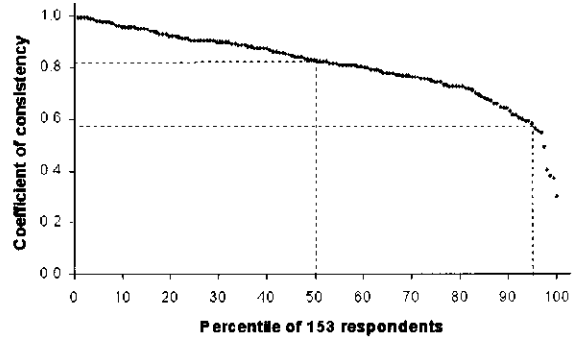

Fig. 1. Distribution of coefficient of consistency.

Table 3
Coefficients of consistency

Measure	Condition			
	1	2	3	4
Mean	0.81	0.78	0.77	0.88
Median	0.82	0.79	0.80	0.90
Minimum	0.37	0.30	0.40	0.60
Maximum	0.99	0.98	0.98	0.99

ence among the conditions in coefficient of consistency, $\chi^2 = 17.13$, df = 3, $P < 0.001$. A multiple comparisons test for Kruskal–Wallis ranks (a Tukey–Kramer modification of the Dunn procedure, Hochberg and Tamhane, 1987) shows significant differences (at $\alpha = 0.05$) between condition 4 and conditions 2 and 3. Thus, consistency about the seriousness of the personal losses significantly exceeds that for two of the three sets of environmental losses.

4.2. Item mix context effects

Table 4 lists the mean probabilities of choice of each of the conditions. The first test of an item mix context effect is based on the correlations comparing the mean probabilities of the test losses obtained from different conditions. The correlation coefficients for the test losses for pairs of conditions range from 0.95 to 0.99 (Table 5). The correlation comparing the two conditions that do not systematically differ in context (conditions 1 and 3) is 0.98. The three correlations comparing conditions involving environmental losses range from 0.97 to 0.99, suggesting that cause of environmental loss caused little or no context effect. Correlations comparing an environmental loss condition with the personal loss condition were slightly lower, ranging from 0.95 to 0.96. The test statistic of a difference between two correlations, for the 0.98 (conditions 1 and 3) versus the lowest correlation of 0.95 (conditions 3 and 4), is 0.825, to be compared with the one-sided z-value of 1.645 at the 95% level. Thus, we cannot reject the null hypothesis of no significant drop in correlation even when substantially different treatment losses are included in the item mix.

Table 4
Mean estimated choice probabilities

ID[a]	Condition			
	1	2	3	4
Treatment sets[b]				
1	0.50	0.60	0.42	0.92
2	0.34	0.46	0.41	0.56
3	0.49	0.47	0.39	0.68
4	0.36	0.47	0.55	0.32
5	0.40	0.41	0.54	0.74
6	0.48	0.54	0.53	0.47
7	0.28	0.38	0.51	0.56
8	0.72	0.79	0.52	0.23
Test set[c]				
1	0.53	0.56	0.55	0.63
2	0.58	0.56	0.53	0.60
3	0.10	0.09	0.09	0.15
4	0.32	0.29	0.26	0.31
5	0.90	0.91	0.93	0.94
6	0.36	0.46	0.44	0.36
7	0.61	0.61	0.64	0.50
8	0.61	0.52	0.56	0.51

[a] ID numbers are listed in Table 2.
[b] Based on the full set of choices.
[c] Based on choices between test items only.

Table 5
Correlations of test loss mean estimated choice probabilities

Conditions	Correlation
1 vs. 2	0.975
1 vs. 3	0.981
1 vs. 4	0.952
2 vs. 3	0.994
2 vs. 4	0.960
3 vs. 4	0.947

Our other test of a context effect compared conditions based on the ratios of choice probabilities for the test losses. A ratio was computed for each of the 28 pairs among the eight test losses from each condition. A non-parametric test, the Friedman (randomized block design) test, was used to compare the four sets of ratios, as the distributions of ratios were significantly skewed. The test shows a significant difference among the conditions in probability ratio, $\chi^2 = 12.32$, df = 3, $P = 0.006$. A multiple comparisons test for the

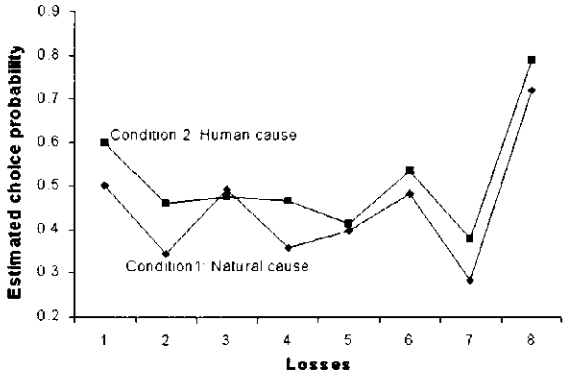

Fig. 2. Seriousness of treatment losses: effect of cause.

Friedman test (Hochberg and Tamhane, 1987, equation 2.26) shows a significant difference (at $\alpha = 0.05$) only between conditions 1 and 4. Thus we reject the null hypothesis of no difference between the sets of ratios for one of the six pairs of conditions, a comparison of naturally caused environmental losses with personal injury losses. This significant difference occurred despite a 0.95 correlation between the mean choice probabilities of conditions 1 and 4.

4.3. Cause of loss

Fig. 2 (plotted from the treatment set choice probabilities of conditions 1 and 2) shows the effect of cause of loss on the judged seriousness of each of the eight treatment losses that differed only in cause. All but one of the eight treatment losses was judged as more serious if caused by a human action than if caused by a natural event. The exception, treatment loss number 3, was a 30% loss of native trout in the Poudre River due to either drought or introduction by the Game and Fish Department of a more aggressive fish species.[6]

The correlation between the two sets of mean preference scores is 0.94, suggesting that cause had little effect on judgments of the relative seriousness of the eight losses. However, cause did shift the magnitude of the seriousness judgments. An analysis of variance including main effects and two-way interactions found significant differences among losses, $F(7, 338) = 27.43$, $P < 0.001$, and between types of cause, $F(1, 35) = 26.62$, $P < 0.001$, but no significant interactions. Thus, preference scores for the treatment losses of condition 2 (human cause) are significantly larger on average than those of condition 1 (natural cause).

5. Discussion and conclusions

Our main objective was to test whether respondents could provide judgments of the seriousness of environmental losses that were largely unaffected, both in average respondent consistency and in scale value, by changes in the mix of losses being assessed. We obtained paired comparison judgments from four groups of respondents about four mixes of losses, where each mix contained a set of losses common to each mix and a set of losses unique to the mix. We designed alternative mixes of treatment losses in a concerted attempt to cause the context effects we were testing for.

Regarding consistency, we found that although people differed widely in the consistency with which they judged the seriousness of losses (coefficients of consistency ranged across all respondents from 0.99 to 0.30), relatively few respondents fell in the lower half of the coefficient range. A small minority of respondents were either extremely inconsistent in their judgments or simply did not take the exercise seriously. More importantly, we found that consistency varied depending on the type of losses being compared, as median coefficient of consistency was close to 0.80 for the three environmental loss conditions but was 0.90 for the personal injury condition. This finding suggests that respondents had more clearly defined judgments of the personal injuries than of the environmental losses, which could reflect a greater familiarity with personal injuries than with environmental losses, or the more complex (i.e. multidimensional) nature of the environmental losses.

Regarding item mix context effects, correlations between sets of test loss scale values were at least

[6] We can only speculate about why this anomaly occurred. One possible explanation is that drought was thought to cause numerous other problems in addition to loss of native trout, but the impact of the more aggressive species would be limited to aquatic life in the river. Another possibility is that drought was assumed by some to have a human cause.

0.95, and were only that low when comparing the effects of very different types of treatment losses (i.e. environmental vs. personal injury losses). Interval-level scale values of seriousness of losses appear to be quite robust to item mix differences, thus ameliorating a major concern following from the likelihood that the specific mix of losses included in any one assessment of seriousness is likely to be somewhat arbitrary. However, the IIA assumption was not satisfied for one of the six comparisons of conditions (again a pair comparing environmental vs. personal injury losses), suggesting caution in mixing different kinds of losses. Given this evidence, if judgments of seriousness of losses from different assessments are to be combined in the course of developing a more complete set of losses for a comprehensive damage schedule, it would be prudent to include a small set of test losses in each assessment, the judgments of which could be used to check for item mix effects.

Turning now to the effect of cause of loss on judged seriousness, environmental losses were considered more serious when they were caused by human actions than when caused by natural events, all else equal. Thus, judgments of seriousness appear to reflect not only the magnitude of the loss but also the reason for the loss. This finding has implications for construction of a damage schedule, for it suggests that any one loss may have a series of levels of seriousness associated with its various possible causes.

The effect of cause on people's assessments of losses may also reflect concerns about responsibility, preventability, etc. For example, Kahneman et al. (1993) found that human-caused losses were more upsetting than natural losses and engendered greater support for and willingness to pay for intervention. Walker et al. (1999), however, found that the relation of the respondent to the persons who caused the loss also affects willingness to pay judgments; unlike the Kahneman et al. result, willingness to pay for an environmental cleanup program was less if the pollution was human-caused than if it was naturally caused. A key difference between the studies was that in the Walker et al. study the human cause was a corporation that was dumping waste, whereas in the Kahneman et al. study the blame for the human-caused losses of most of the scenarios was less easily placed. For willingness to pay judgments, Walker et al. emphasize the importance of whether respondents feel moral responsibility for the loss or can project that responsibility to others.

Clearly the identification of a negligent party that has the ability to pay for some sort of restitution (such as a corporation) will lower the general public's willingness to pay, possibly even below willingness to pay if the loss were caused by a natural process. Thus, depending on who the negligent party is, judgments of seriousness (or importance or upset) may not correlate well with judgments of willingness to pay. Although it may be a safe guess that seriousness of human-caused loss correlates positively with strength of feeling that someone should pay for restitution, the role of cause in assessments of loss is complex and perhaps not fully understood.

Due to the complex effect of cause on people's assessments of losses, perhaps the most straightforward approach to constructing a damage schedule would be to leave cause unspecified during step 1, in which losses are submitted for public judgments of seriousness. To the extent that cause plays a role in the damage schedule, it could be a consideration during step 2 of damage schedule construction, when damage payments and other injunctions are specified. In any case, users of the ranking of losses would be wise to remember the importance of cause when considering the acceptability of any final resource management decision.

References

Arrow, K., 1951. Social Choice and Individual Values. Wiley, New York.

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

Boyle, K.J., Johnson, F.R., McCollum, D.W., 1997. Anchoring and adjustment in single-bounded, contingent-valuation questions. Am. J. Agric. Econ. 79 (5), 1495–1500.

Brown, T.C., Daniel, T.C., 1987. Context effects in perceived environmental quality assessment: scene selection and landscape quality ratings. J. Environ. Psychol. 7, 233–250.

Brown, T.C., Champ, P.A., Bishop, R.C., McCollum, D.W., 1996. Which response format reveals the truth about donations to a public good. Land Econ. 72 (2), 152–166.

Chuenpagdee, R., Knetsch, J.L., Brown, T.C., 2001. Environmental damage schedules: community judgments of importance and assessments of losses. Land Econ. 77 (1), 1–11.

Cummings, R.G., Elliott, S., Harrison, G.W., Murphy, J., 1997. Are hypothetical referenda incentive compatible. J. Polit. Econ. 105 (3), 609 621.

David, H.A., 1988. The Method of Paired Comparisons. Griffin's Statistical Monographs and Courses, vol. 41. Oxford University Press, New York, NY.

Diamond, P.A., Hausman, J.A., 1994. Contingent valuation: is some number better than no number. J. Econ. Perspect. 8 (4), 45–64.

Dunn-Rankin, P., 1983. Scaling Methods. Lawrence Erlbaum Associates, Hillsdale, NJ.

Gorter, R.B., 1997. Scaling the Importance of Environmental Losses: Social Values, Damage Assessment and the Method of Paired Comparisons. Simon Fraser University, Barnaby, BC.

Green, D., Jacowitz, K.E., Kahneman, D., McFadden, D., 1998. Referendum contingent valuation, anchoring, and willingness to pay for public goods. Resour. Energy Econ. 20, 85–116.

Hanemann, M., Kanninen, B., 1999. The statistical analysis of discrete-response CV data. In: Bateman, I.J., Willis, K.G. (Eds.), Valuing Environmental Preferences. Oxford University Press, New York, pp. 302–441.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. Wiley, New York.

Holmes, T.P., Kramer, R.A., 1995. An independent sample test of yea-saying and starting point bias in dichotomous-choice contingent valuation. J. Environ. Econ. Manage. 29 (1), 121–132.

Kahneman, D., Ritov, I., Jacowitz, K.E., Grant, P., 1993. Stated willingness to pay for public goods: a psychological perspective. Psychol. Sci. 4 (3), 310–315.

Kahneman, D., Schkade, D., Sunstein, C.R., 1998. Shared outrage and erratic awards: the psychology of punitive damages. J. Risk Uncertainty 16 (1), 49–86.

Luce, R.D., 1959. Individual Choice Behavior. Wiley, New York.

McFadden, D., 2001. Economic choices. Am. Econ. Rev. 91 (3), 351–378.

Nunnally, J.C., 1976. Psychometric Theory. McGraw Hill, New York.

Parducci, A., 1968. The relativism of absolute judgments. Sci. Am. 219, 84 90.

Peterson, G.L., Brown, T.C., 1998. Economic valuation by the method of paired comparison, with emphasis on evaluation of the transitivity axiom. Land Econ. 74 (2), 240 261.

Rutherford, M.B., Knetsch, J.L., Brown, T.C., 1998. Assessing environmental losses: judgments of importance and damage schedules. Harvard Environ. Law Rev. 22, 51–101.

Slovic, P., 1975. Choice between equally valued alternatives. J. Exp. Psych.: Hum. Percep. Perform. 1, 280–287.

Smith, V.K., Osborne, L.L., 1996. Do contingent valuation estimates pass a 'scope' test? A meta-analysis. J. Environ. Econ. Manage. 31, 287–301.

Thurstone, L.L., 1927. A law of comparative judgment. Psychol. Rev. 34, 273–286.

Torgerson, W.S., 1958. Theory and Methods of Scaling. Wiley, New York, NY.

Tversky, A., 1969. Intransitivity of preferences. Psychol. Rev. 76 (1), 31–48.

Walker, M.E., Morera, O.F., Vining, J., Orland, B., 1999. Disparate WTA-WTP disparities: the influence of human versus natural causes. J. Behav. Decision Making 12 (3), 219–232.